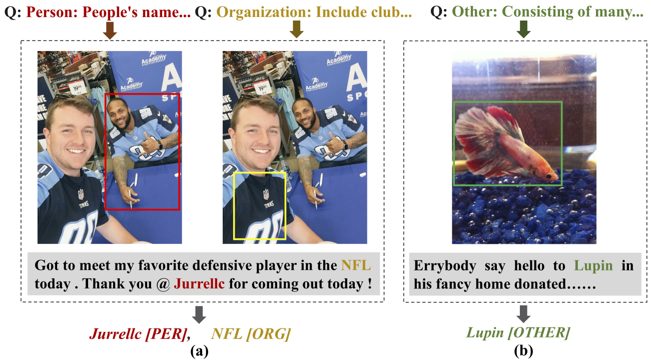
MNER-QG: An End-to-End MRC Framework for Multimodal Named Entity Recognition with Query Grounding

Meihuizi Jia[1,2], **Lei Shen**[2], **Xin Shen**[3], **Lejian Liao**[1], **Meng Chen**[2*], **Xiaodong He**[2], **Zhendong Chen**[1], **Jiaqi Li**[1]

[1]School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China
[2]JD AI Research, Beijing, China
[3]Australian National University, Canberra, Australia
{jmhuizi24, liaolj, chenzd, 3120195492}@bit.edu.cn, {shenlei20, chenmeng20, xiaodong.he}@jd.com, u6498962@anu.edu.au

Abstract

Multimodal named entity recognition (MNER) is a critical step in information extraction, which aims to detect entity spans and classify them to corresponding entity types given a sentence-image pair. Existing methods either (1) obtain named entities with coarse-grained visual clues from attention mechanisms, or (2) first detect fine-grained visual regions with toolkits and then recognize named entities. However, they suffer from improper alignment between entity types and visual regions or error propagation in the two-stage manner, which finally imports irrelevant visual information into texts. In this paper, we propose a novel end-to-end framework named MNER-QG that can simultaneously perform MRC-based multimodal named entity recognition and query grounding. Specifically, with the assistance of queries, MNER-QG can provide prior knowledge of entity types and visual regions, and further enhance representations of both text and image. To conduct the query grounding task, we provide manual annotations and weak supervisions that are obtained via training a highly flexible visual grounding model with transfer learning. We conduct extensive experiments on two public MNER datasets, Twitter2015 and Twitter2017. Experimental results show that MNER-QG outperforms the current state-of-the-art models on the MNER task, and also improves the query grounding performance.


Q: Person: People's name...
Q: Organization: Include club...
Q: Other: Consisting of many...

Got to meet my favorite defensive player in the NFL today . Thank you @ Jurrellc for coming out today !

Errybody say hello to Lupin in his fancy home donated……

Jurrellc [PER],
NFL [ORG]
(a)
Lupin [OTHER]
(b)

Figure 1: Two examples of MNER-QG with entity type “ORG”, “PER”, and “OTHER”.

Introduction

Multimodal named entity recognition (MNER) is a vision-language task that extends the traditional text-based NER and alleviates ambiguity in natural languages by taking images as additional inputs. The essence of MNER is to effectively capture visual features corresponding to entity spans and incorporate certain visual regions into textual representations.

Existing MNER datasets contain few fine-grained annotations in each sentence-image pair, i.e., the relevant image is given as a whole without regional 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 (Moon, Neves, and Carvalho 2018; Lu et al. 2018; Zhang et al. 2018; Arshad et al. 2019; Yu et al. 2020; Chen et al. 2021a; Xu et al. 2022). However, it is hard to interpret and evaluate the effectiveness of implicit alignments. Recently, visual grounding toolkits (Yang et al. 2019) are exploited to explicitly extract visual regions related to different entity types (Zhang et al. 2021). The detected regions are then bound with the input sentence and fed into the recognition model together (Jia et al. 2022c). Because of the two-stage manner, incorporating inaccurate visual regions from the first stage will hurt the final results (error propagation).

With respect to the problem formalization, early methods regard MNER as a sequence labeling task that integrates image embeddings into a sequence labeling model and assigns type labels to named entities. Recently, the machine reading comprehension (MRC) framework is employed in many natural language processing tasks due to its solid language understanding capability (Li et al. 2019a,b; Chen et al. 2021b). To take advantage of the prior knowledge encoded in MRC queries (Li et al. 2019a), we consider MNER as a MRC task, which extracts entity spans by answering queries about entity types. In addition, to capture the fine-grained alignment between entity types and visual regions, we ground the MRC queries on image regions and output their positions as bounding boxes. For example, as shown in Figure 1

[*]Corresponding author

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

(a), recognizing entities with type PER and ORG in sentence "Got to meet my favorite defensive player in the NFL today. Thank you @ Jurrellc for coming out today!" is formalized as extracting answer spans from the input sentence given the query "Person: People's name..." and "Organization: Include club...". Then, answer spans "Jurrellc" and "NFL" are obtained along with their visual regions marked by red and yellow boxes.

To this end, we propose an end-to-end MRC framework for **M**ultimodal **N**amed **E**ntity **R**ecognition with **Q**uery **G**rounding (**MNER-QG**). This joint-training approach forces the model to explicitly align entity spans with the corresponding visual regions, and further improves the performance of both named entity recognition and query grounding. Specifically, we design unified queries with prior information as navigators to pilot our joint-training model. Meanwhile, we extract multi-scale visual features and design two interaction mechanisms, multi-scale cross-modality interaction and existence-aware uni-modality interaction, to enrich both textual and visual information. Since there are few fine-grained annotations for visual regions in existing MNER datasets, we provide two types of bounding box annotations, weak supervisions and manual annotations. The former is obtained by training a visual grounding model with transfer learning, while the latter aims to provide oracle results.

In summary, the contribution of this paper is three-fold:

- We propose a novel end-to-end MRC framework, MNER-QG. Our model simultaneously performs MRC-based multimodal named entity recognition and query grounding in a joint-training manner. To the best of our knowledge, this is the first attempt on MNER.
- To fulfill the end-to-end training, we contribute weak supervisions via training a visual grounding model with transfer learning. Meanwhile, we offer manual annotations of bounding boxes as oracle results.
- We conduct extensive experiments on two public MNER datasets, Twitter2015 and Twitter2017, to evaluate the performance of our framework. Experimental results show that MNER-QG outperforms the current state-of-the-art models on both datasets for MNER, and also improves the QG performance.

Related Work

Multimodal Named Entity Recognition

With the increasing popularity of multimodal data on social media platforms, multimodal named entity recognition (MNER) has become an important research direction, which assists the NER models (Li et al. 2021b,a, 2022) in better identifying entities by taking images as the auxiliary input. The critical challenge of MNER is how to align and fuse textual and visual information. Yu et al. (2020) proposed a multimodal transformer architecture for MNER, which captures expressive text-image representations by incorporating the auxiliary entity span detection. Zhang et al. (2021) created the graph connection between textual words and visual objects acquired by a visual grounding toolkit (Yang et al. 2019), and proposed a graph fusion approach to conduct graph encoding. Xu et al. (2022) proposed a matching and alignment framework for MNER to improve the consistency of representations in different modalities.

Lacking prior information of entity types and accurate annotations of visual regions corresponding to certain entity types, the above methods feed visual information (an entire image, image patches, or retrieved visual regions from toolkits) with the entire sentence into an entity recognition model, which inevitably makes it difficult to obtain the explicit alignment between images and texts.

Machine Reading Comprehension

Machine Reading Comprehension (MRC) aims to answer natural language queries given a set of contexts where the answers to these queries can be inferred. In various forms of MRC, span extraction MRC (Peng et al. 2021; Jia et al. 2022a) is challenging, which extracts a span as the answer from context. The span extraction can be regarded as two multi-class classification or two binary classification tasks. For the former, the model needs to predict the start and end positions of an answer. For the latter, the model needs to decide whether each token is the start/end position. Recurrent Neural Network (RNN) was used to encode textual information, then a linear projection layer was adopted to predict answer spans (Yang et al. 2018; Nishida et al. 2019). The performance was boosted with the development of large-scale pre-trained models (Qiu et al. 2019; Tu et al. 2020), such as ELMo (Peters et al. 2018), BERT (Devlin et al. 2019), and RoBERTa (Liu et al. 2019).

Recently, there is a trend of converting NLP tasks to the MRC form, including named entity recognition (Li et al. 2019a), entity relation extraction (Li et al. 2019b), and sentiment analysis (Chen et al. 2021b). Due to the powerful understanding ability contained in MRC, the model performance of these tasks is improved.

Visual Grounding

Visual grounding aims to localize textual entities or referring expressions in an image. This task is divided into two paradigms: two-stage and one-stage. For the former, the first stage is exploited to extract region proposals as candidates via some region proposal methods (e.g., Edgebox (Zitnick and Dollár 2014), selective search (Uijlings et al. 2013), and Region Proposal Networks (Ren et al. 2015)), and then the second stage is designed to rank region-text candidate pairs. For the latter, researchers utilize one-stage model (e.g., YOLO (Redmon et al. 2016; Redmon and Farhadi 2018; Bochkovskiy, Wang, and Liao 2020)) combined with extra features to directly output the final region(s). Compared with the two-stage manner, the one-stage framework is simplified and accelerates the inference by conducting detection and matching simultaneously.

To connect visual grounding and MRC-based named entity recognition effectively, we use queries from MRC as input texts and force model to perform query grounding. Since queries contain the prior knowledge of entity types, our work can achieve the explicit alignment between entity types and visual regions.

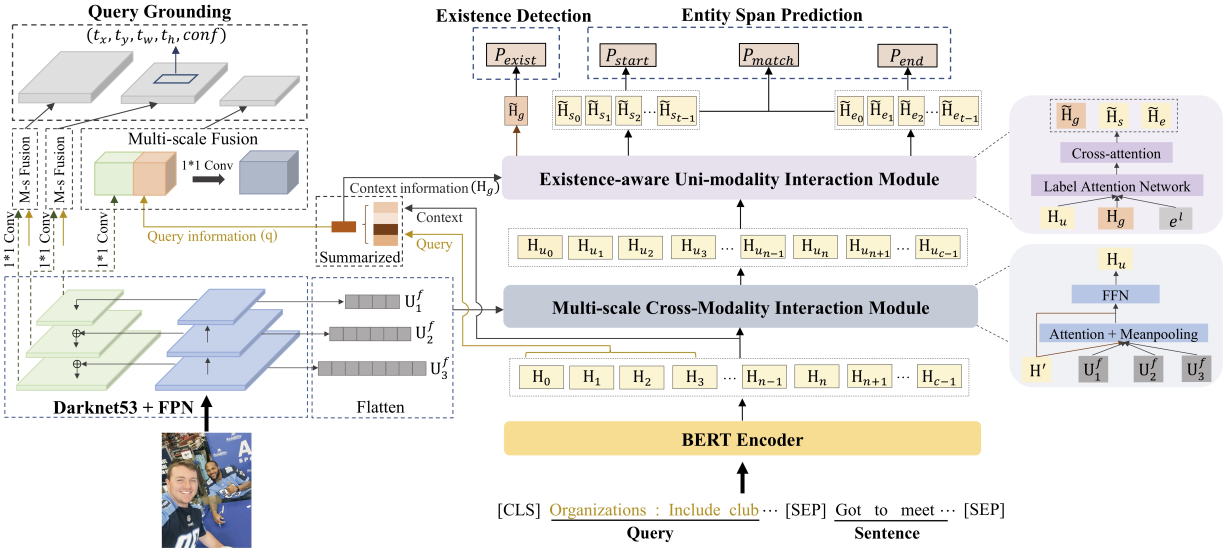

Query Grounding
$(t_x, t_y, t_w, t_h, conf)$
Multi-scale Fusion
1*1 Conv
M-s Fusion
M-s Fusion
1*1 Conv
1*1 Conv
1*1 Conv
Context information (H_g)
Query information (q)
Context
Query
Summarized
U_1^f
U_2^f
U_3^f
Darknet53 + FPN
Flatten
Existence Detection
Entity Span Prediction
P_{exist}
P_{start}
P_{match}
P_{end}
$\tilde{H}_g$
$\tilde{H}_{s_0}$ $\tilde{H}_{s_1}$ $\tilde{H}_{s_2}$ ··· $\tilde{H}_{s_{t-1}}$
$\tilde{H}_{e_0}$ $\tilde{H}_{e_1}$ $\tilde{H}_{e_2}$ ··· $\tilde{H}_{e_{t-1}}$
Existence-aware Uni-modality Interaction Module
H_{u_0} H_{u_1} H_{u_2} H_{u_3} ··· $H_{u_{n-1}}$ H_{u_n} $H_{u_{n+1}}$ ··· $H_{u_{c-1}}$
Multi-scale Cross-Modality Interaction Module
H_0 H_1 H_2 H_3 ··· H_{n-1} H_n H_{n+1} ··· H_{c-1}
BERT Encoder
[CLS] Organizations : Include club ··· [SEP] Got to meet ··· [SEP]
Query
Sentence
$\tilde{H}_g$ $\tilde{H}_s$ $\tilde{H}_e$
Cross-attention
Label Attention Network
H_u H_g e^l
H_u
FFN
Attention + Meanpooling
H' U_1^f U_2^f U_3^f

Figure 2: Overview of our MNER-QG framework (M-s Fusion denotes Multi-scale Fusion).

Method

Overview

Figure 2 illustrates the overall architecture of MNER-QG. Given a sentence $S = \{s_0, s_1, ..., s_{n-1}\}$ and its associated image V, where n denotes the sentence length, we first design a natural language query $Q = \{q_0, q_1, ..., q_{m-1}\}$ with prior awareness about entity types. Then, our model performs multi-scale cross-modality interaction and existence-aware uni-modal interaction to simultaneously detect entity spans $s_{\text{start,end}}$ and the corresponding visual regions via answering the query Q.

Query Construction

Query plays a significant role as the navigator in our MNER-QG, and it should be expressed as generic, precise, and effective as possible. Table 1 shows examples of queries designed by us. We hope that the queries are moderate in difficulty and can provide informative knowledge of MNER and QG tasks. Therefore, the model can stimulate the solid understanding capability of MRC without limiting the performance of QG.

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 club, company, government party, school government, and news organization.

Table 1: Examples of transforming entity types to queries.

Model Architecture

Input Representation. For text information, we concatenate a query and sentence pair $\{[\text{CLS}], Q, [\text{SEP}], S, [\text{SEP}]\}$, and encode the result to a $768D$ real-valued vector with the pre-trained BERT model (Devlin et al. 2019), where [CLS] and [SEP] are special tokens. Then BERT outputs a contextual representation $\mathbf{H} \in \mathbb{R}^{c \times d_c}$, where $c = m + n + 3$ is the length of BERT input, $d_c = 768$. For visual information, inspired by Yang et al. (2019), we use Darknet-53 (Zhu et al. 2016) with feature pyramid networks (Lin et al. 2017) to extract visual features. The images are resized to 256 × 256, and the feature maps are $\frac{1}{32}$, $\frac{1}{16}$, and $\frac{1}{8}$, respectively. Therefore, the three spatial resolutions are $8 \times 8 \times d_1$, $16 \times 16 \times d_2$, and $32 \times 32 \times d_3$, where $d_1 = 1024$, $d_2 = 512$, and $d_3 = 256$ are feature channels.

We unify the dimensions of three visual features and textual feature to facilitate the model computation. Specifically, we add a 1×1 convolution layer with batch normalization and RELU under the feature pyramid networks to map the feature channels d_1, d_2, and d_3 to the same dimension $d = 512$. The new visual features are denoted as $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$. At the same time, we flatten 8×8, 16×16, and 32×32 to 64, 256, and 1024, which are used to generate the visual representations, $\mathbf{U}_1^f$, $\mathbf{U}_2^f$ and $\mathbf{U}_3^f$. For textual information, we use a linear projection to map d_c to $d = 512$, and the mapped representation is $\mathbf{H}^{'}$.

Multi-scale Cross-modality Interaction. This module is shown in Figure 2 with the grey box. We first truncate the token-level representations of query $\mathbf{Q}$ from $\mathbf{H}^{'}$. The query encoding now contains messages from the original sentence S, which can be passed to the QG task. Then, we use an attention-based approach (Rei and Søgaard 2019) to acquire the summarized query representation $\mathbf{q} \in \mathbb{R}^{1 \times d}$ that will be fed into QG.

$$\alpha = \text{softmax}\left(\text{MLP}\left(\mathbf{Q}\right)\right), \quad \mathbf{q} = \sum_{k=0}^{m-1} \alpha_k \mathbf{Q}\left[k, :\right] \quad (1)$$

To fully exploit image information, we employ multi-scale visual representations to update textual representation through a cross-modality attention mechanism, where $\mathbf{H}^{'}$ works as the query matrix, while each of $\mathbf{U}_1^f$, $\mathbf{U}_2^f$, and $\mathbf{U}_3^f$ works as the key and value matrix. The visual-enhanced attention outputs are denoted as $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3 \in \mathbb{R}^{c\times d}$. Then, we merge these matrices to a unified textual representation $\mathbf{H}_a$ using $\mathrm{MeanPooling}$. Finally, we concatenate $\mathbf{H}_a$ and $\mathbf{H}^{'}$, and feed the result into a feed-forward neural network to get the final textual representation $\mathbf{H}_u$.

Existence-aware Uni-modality Interaction. Since the sentence does not always contain the entity asked by the current query, we design a global existence signal to enhance the model's awareness of entity existence. Similar to Equation (1), we summarize contextual representation $\mathbf{H}^{'}$ to acquire the existence representation $\mathbf{H}_g \in \mathbb{R}^{1\times d}$. Inspired by Qin et al. (2021) and Li et al. (2021b), we then employ a label attention network to update both the textual representation with the encoding of start/end label and the existence representation with the encoding of existence label. (Note that the $\mathbf{e}^l$ in Figure 2 denotes a label embedding lookup table). Details of the label attention network are provided in the Appendix[1]. Then, we get start/end label-enhanced textual representation, $\mathbf{H}_s$/$\mathbf{H}_e$, which can be regarded as the start/end representation of entity span, and also label-enhanced existence representation $\widehat{\mathbf{H}}_g$.

We calculate attention scores between $\mathbf{H}_s$ and $\widehat{\mathbf{H}}_g$, where $\mathbf{H}_s$ works as the query matrix, while $\widehat{\mathbf{H}}_g$ works as the key and value matrix, and define the existence-aware start representation $\widetilde{\mathbf{H}}_s$ as follows:

$$\mathbf{Z}_s = \mathrm{softmax}\left(\frac{\mathbf{Q}_s\mathbf{K}_g^\top}{\sqrt{d_k}}\right)\mathbf{V}_g, \widetilde{\mathbf{H}}_s = \mathrm{LN}\left(\mathbf{H}_s + \mathbf{Z}_s\right), \tag{2}$$

where LN denotes the layer normalization function (Ba, Kiros, and Hinton 2016), $\mathbf{Q}_s \in \mathbb{R}^{c\times d}$, $\mathbf{K}_g, \mathbf{V}_g \in \mathbb{R}^{1\times d}$, and $\widetilde{\mathbf{H}}_s \in \mathbb{R}^{c\times d}$. Similarly, we can obtain the updated end representation $\widetilde{\mathbf{H}}_e \in \mathbb{R}^{c\times d}$, and the updated existence representation $\widetilde{\mathbf{H}}_g \in \mathbb{R}^{1\times d}$.

Query Grounding

Following Yang et al. (2019), we first broadcast the query representation $\mathbf{q}$ to each spatial location, denoted as (i, j), and then concatenate the query feature and visual feature $\mathbf{U}_i$, where $i = 1, 2, 3$. The feature dimension after concatenation is $512+512 = 1024$. Another 1×1 convolution layer is appended to better fuse above features at each location independently and map them to the dimension $d = 512$.

Next, we perform the grounding operation. There are $8\times 8+16\times 16+32\times 32 = 1344$ locations in three spatial resolutions, and each location is related to a $512D$ feature vector from the fusion layer. YOLOv3 network centers around each of the location's three anchor boxes, hence it predicts bounding boxes at three scales. The output of YOLOv3 network is $[3 \times (4 + 1)] \times r_i \times r_i$ at each scale for shifting the center, width, and height (t_x, t_y, t_w, t_h) of the anchor box, along with the confidence score $conf$ on this shifted box, where $r_i \times r_i$ denotes the shape size of each spatial resolution. Ultimately, only one region is desired as the output for query grounding. More details can be found in Yang et al. (2019).

The objective function $\mathcal{L}_{QG}$ of QG task consists of regression loss on bounding box $\mathcal{L}_{bbox}$ and object score loss $\mathcal{L}_{object}$. $\mathcal{L}_{bbox}$ is expected to assign bounding box regions to ground truth objects via mean squared error (MSE). $\mathcal{L}_{object}$ is used to classify the bounding box regions as object or non-object via binary cross-entropy (BCE).

Multimodal Named Entity Recognition

The core of multimodal named entity recognition is to predict the entity span in sentence. In this section, we design an auxiliary task named existence detection (ED) after receiving the existence representation $\widetilde{\mathbf{H}}_g$ to predict whether a sentence contains entities with specific type and cooperate with the entity span prediction task to extract entity span.

Existence Detection. This task and the entity span prediction task can share the corresponding mutual information with the co-interactive attention mechanism. The existence of entity is detected as follows:

$$\mathrm{P}_{exist} = \mathrm{softmax}\left(\widetilde{\mathbf{H}}_g\mathbf{W}_{exist}\right) \tag{3}$$

where $\mathbf{W}_{exist} \in \mathbb{R}^{d\times 2}$ and $\mathrm{P}_{exist} \in \mathbb{R}^{1\times 2}$. We formulate the ED sub-task as a text classification task. The loss function is denoted by $\mathcal{L}_{ED}$ and the binary cross-entropy (BCE) loss is taken as the training objective.

Entity Span Prediction. To tag the entity span from a sentence using MRC framework, it is necessary to find the start and end positions of the entity. We utilize two binary classifiers to predict whether each token in the sentence is the start/end index or not, respectively. The probability that each token is predicted to be a start position is as follows:

$$\mathrm{P}_{start} = \mathrm{softmax}_{\text{eachrow}}\left(\widetilde{\mathbf{H}}_s\mathbf{W}_s\right) \tag{4}$$

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

Since there could be multiple entities of the same type in the sentence, we add a binary classification model to predict the matching probability of start and end positions inspired by Li et al. (2019a).

$$\mathrm{P}_{match} = \mathrm{sigmoid}\left(\mathbf{W}_m\left[\widetilde{\mathbf{H}}_s; \widetilde{\mathbf{H}}_e\right]\right) \tag{5}$$

where $\mathbf{W}_m \in \mathbb{R}^{1\times 2d}$, $\mathrm{P}_{match} \in \mathbb{R}^{1\times 2}$. $[;]$ is denoted the concatenation in columns.

During training, the objective function $\mathcal{L}_{ESP}$ of entity span prediction (ESP) sub-task consists of start position loss $\mathcal{L}_{start}$, end position loss $\mathcal{L}_{end}$ and matching loss $\mathcal{L}_{match}$, where binary cross-entropy (BCE) is used for calculation.

Finally, combining the two tasks QG and MNER, the overall objective function is as follows:

$$\mathcal{L} = \omega_f\mathcal{L}_{QG} + \lambda_1\mathcal{L}_{ED} + \lambda_2\mathcal{L}_{ESP} \tag{6}$$

where ω_f, λ_2 and λ_3 are hyper-parameters to control the contributions of each sub-task.

[1] The appendix will be released at: https://github.com/jmhz24/MNER-QG.

Experiments

Dataset Construction

There are two widely-used MNER datasets, Twitter2015 (Zhang et al. 2018) and Twitter2017 (Lu et al. 2018), used to evaluate the effectiveness of our framework. Both datasets are separated into training, validation, and test sets with the same type distribution. Statistics are listed in Appendix. And then, we contribute two types of labels: weak supervisions and manual annotations for public research.

For weak supervisions, we apply the pre-trained fast and accurate one-stage visual grounding model (Yang et al. 2019) (denoted as FA-VG) as the base model. In the setting of Phrase Localization task, FA-VG was trained and evaluated on the Flickr30K Entities dataset (Plummer et al. 2015) that augments the original Flickr30K (Young et al. 2014) with region-phrase correspondence annotations. However, there are two obstacles: (1) These phrases/queries are from image captions, and not specially constructed for the named entity recognition task. (2) The MNER datasets (i.e. Twitter2015/2017) have different data domains compared with the Flickr30K Entities dataset. Thus, we utilize transfer learning to overcome above issues. In addition, we contribute manual annotations for public research. We hire three crowd-sourced workers who are familiar with the tasks of MNER and object detection to help us annotate the bounding box in the image. The annotators are requested to tag the visual regions in the image corresponding to the entity span in the sentence. After the data annotation, we merge the instances of strong inter-annotator agreement from three crowd-sourced workers to acquire high-quality and explicit text-image alignment data (Chen et al. 2019; Jia et al. 2022b). Details of the annotation with two types of labels are provided in the Appendix.

Experiment Settings

Evaluation Metrics. For the MNER task, we use precision ($Pre.$), recall ($Rec.$), and F1 score ($F1$) to evaluate the performance of overall entity types, and use $F1$ only for each type. For the QG, we follow prior works (Rohrbach et al. 2016) and utilize Accu@0.5 as evaluation protocol. Given a query, an output image region is considered correct if its IoU is at least 0.5 with the ground truth bounding box. In addition, we add Accu@0.75 (IoU is at least 0.75) and Miou (mean of IoU) as additional evaluation metrics.

Implementation Details. 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 datasets after conducting a grid search over the interval [1e-5, 1e-4] and [0.1, 0.6]. We train the model with AdamW optimization. To further evaluate our joint-training model, we take out the images from Twitter2015/2017 to train the QG model separately. For a fair comparison, we use the same configurations such as batch size, learning rate, and optimizer in both the QG model and our joint-training model. For the joint-training loss, we set the hyper-parameters $\lambda_1 = 1$ and $\lambda_2 = 2$ by tuning on the validation set. We specially set a balance factor ω_f to dynamically scale the loss of MNER and QG. Please refer to Appendix for more details.

Baseline Models. Two groups of baselines are compared with our approach. The first group consists of some text-based MNER models that formalize MNER as a sequence labeling task: (1) **BiLSTM-CRF** (Huang, Xu, and Yu 2015); (2) **CNN-BiLSTM-CRF** (Ma and Hovy 2016); (3) **HBiLSTM-CRF** (Lample et al. 2016); (4) **BERT** (Devlin et al. 2019); (5) **BERT-CRF**; (6) **T-NER** (Ritter et al. 2011; Zhang et al. 2018). The second group contains several competitive MNER models: (1) **GVATT-HBiLSTM-CRF** (Lu et al. 2018); (2) **GVATT-BERT-CRF** (Yu et al. 2020); (3) **AdaCAN-CNN-BiLSTM-CRF** (Zhang et al. 2018); (4) **AdaCAN-BERT-CRF** (Yu et al. 2020); (5) **UMT-BERT-CRF** (Yu et al. 2020); (6) **MT-BERT-CRF** (Yu et al. 2020); (7) **ATTR-MMKG-MNER** (Chen et al. 2021a); (8) **UMGF** (Zhang et al. 2021); (9) **MAF** (Xu et al. 2022). The details of these models are illustrated in Appendix.

According to different derivations of bounding box labels in the images, we provide two versions of our model **MNER-QG** and **MNER-QG (Oracle)** for evaluation. In addition, we provide a variant of the model, **MNER-QG-Text**, which uses text input only.

Main Results

Table 2 shows the results of our model and baselines. The upper results are from text-based models and the lower results are from multimodal models. Firstly, we compare the multimodal models with their corresponding uni-modal baselines in MNER, such as AdaCAN-CNN-BiLSTM-CRF vs. CNN-BiLSTM-CRF, and MNER-QG vs. MNER-QG-Text. We notice almost all multimodal models can significantly outperform their corresponding uni-modal competitors, indicating the effectiveness of images. And then, we compare our MNER-QG with other multimodal baselines. The result shows MNER-QG outperforms all baselines on Twitter2017 and yields competitive results on Twitter2015. MNER-QG (Oracle) with more accurate manual annotations yields further results in both datasets.

Ablation Study

Table 3 shows the ablation results. We observe that all subtasks are necessary. First, after removing the QG loss, the performance noticeably drops on all metrics. In particular, $F1$ scores on two datasets degrade by 0.71% and 0.62%, respectively. The result shows the QG training promotes explicit alignment between text and image. Besides, removing the ED loss also damages the performance on all metrics. $F1$ scores on the two datasets decrease by 0.47% and 0.41%, respectively. We conjecture that 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 QG and ED loss, the performance degrades significantly, indicating that both the QG and ED tasks are essential in our framework.

Case Study

Here we conduct further qualitative analysis with two specific examples. We compare the results from MNER-QG, MNER-QG (Oracle), and the competitive model UMGF.

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	-	-	-	-	-	-	-
MNER-QG-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
MNER-QG (Ours)	**85.31**	81.65	**63.41**	41.32	**77.43**	72.15	74.70	**92.92**	**86.19**	84.52	**71.67**	**88.26**	85.65	**86.94**
MNER-QG (Oracle) (Ours)	**85.68**	81.42	**63.62**	41.53	**77.76**	72.31	**74.94**	**93.17**	**86.02**	84.64	**71.83**	**88.57**	85.96	**87.25**

Table 2: Results on two MNER datasets. We refer to the results of UMGF from Zhang et al. (2021) and other results from Xu et al. (2022). Our model achieves a statistically significant improvement with p-value<0.05 under a paired two-sided t-test.

Methods	Twitter2015			Twitter2017		
	Pre.	*Rec.*	*F1*	*Pre.*	*Rec.*	*F1*
MNER-QG	77.43	72.15	74.70	88.26	85.65	86.94
- w/o QG loss	77.50	70.79	73.99	88.01	84.69	86.32
- w/o ED loss	77.53	71.20	74.23	87.81	85.28	86.53
- w/o QG+ED loss	77.17	70.29	73.57	87.63	84.47	86.02

Table 3: Ablation study of MNER-QG on test set.

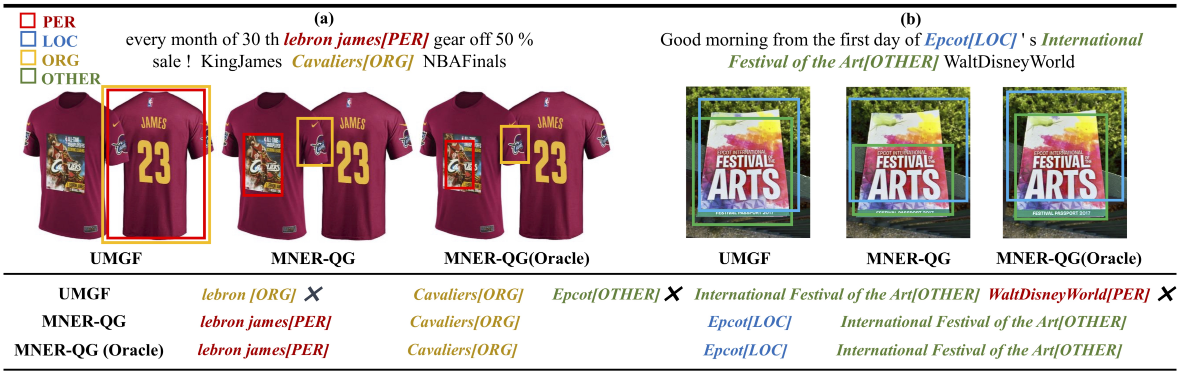
In Figure 3 (a), the sentence contains two entities "*lebron james*", and "*Cavaliers*" with PER, and ORG types respectively. However, UMGF locates the entity "*lebron james*" inaccurately and misjudges its type. We guess it is because UMGF cannot detect the region of person on the red T-shirt. Instead, both MNER-QG and MNER-QG (Oracle) extract region about "*lebron james*" (red box) for PER, and the logo about "*Cavaliers*" (yellow box) for ORG on clothing, and the regions extracted by MNER-QG (Oracle) are more accurate due to the more elaborate manual annotations. Compared with UMGF, our model can locate more relevant visual regions, which can assist the model on accurately recognizing entities. Figure 3 (b) shows a more challenging case, where the image cannot provide useful regions about LOC. It can be seen that UMGF, MNER-QG and MNER-QG (Oracle) cannot locate the relevant visual regions for this entity. However, both MNER-QG and MNER-QG (Oracle) can recognize "*Epcot*" and its type. We conjecture that the solid understanding capability of MRC and the guidance of query prior information contribute to the final correct prediction.

Discussions

Effectiveness of the End-to-End Manner. Table 4 shows the results of our joint-training approach with other single-training approaches on different tasks. MNER-VG is a two-stage MNER model, which uses the VG model trained via transfer learning to acquire visual region in the first stage and integrates it into the second stage to enhance token representation. FA-VG is a one-stage VG model, and we retrain the model using Twitter2015/2017 datasets. As can be seen, compared with models MNER-QG-Text and FA-VG trained on a single data source (e.g., text or image) in different tasks, our joint-training model significantly improves the performance of each task, e.g., $F1$ score and $Accu@0.5$ are improved by 1.96% and 3.1% (max:4.03%), respectively in Twitter2015. Compared with the two-stage model MNER-VG, our end-to-end model still has obvious advantages, e.g., $F1$ scores are increased by 0.76% and 0.91% in Twitter2015/2017, respectively. The above results indicate that the different tasks in our model are complementary with each other under an end-to-end manner and enable the model to yield better performance.

Accuracy of QG. To check the quality of the labels contributed by us for the QG, we present the results of the different models trained with two types of labels. In addition, we provide the result on a high-quality Flickr30K Entities dataset for comparison. The dataset links 31,783 images in Flickr30K with 427K referred entities. Table 5 illustrates that for either MNER-QG or FA-VG, training with manual annotations outperforms that with weak supervisions. But the acquisition of weak supervisions is cost-effective and time-efficient. Regardless of the annotation method, our joint-training MNER-QG significantly improves the performance compared with single-training FA-VG on QG task. Compared with the results of FA-VG on Flickr30K Entities, the results on Twitter2015/2017 are competitive. In particular, $Accu@0.5$ in Twitter2017 with manual annotations is 2.33% higher than the result in Flickr30K Entities[2]. The results indicate two types of labels on Twitter2015/2017 for QG are reliable and leave ample scope for future research.

[2] There are great deviations in the number of images and the distribution of data in Twitter2015/2017 and Flickr30K Entities, and the comparison of the three datasets is shown in the Appendix.



(a) every month of 30 th *lebron james[PER]* gear off 50 % sale ! KingJames *Cavaliers[ORG]* NBAFinals

(b) Good morning from the first day of *Epcot[LOC]* ' s *International Festival of the Art[OTHER]* WaltDisneyWorld

UMGF	*lebron [ORG]* ✗	*Cavaliers[ORG]*	*Epcot[OTHER]* ✗	*International Festival of the Art[OTHER]*	*WaltDisneyWorld[PER]* ✗
MNER-QG	*lebron james[PER]*	*Cavaliers[ORG]*	*Epcot[LOC]*	*International Festival of the Art[OTHER]*	
MNER-QG (Oracle)	*lebron james[PER]*	*Cavaliers[ORG]*	*Epcot[LOC]*	*International Festival of the Art[OTHER]*	

Figure 3: Example comparison among MNER-QG, MNER-QG (Oracle), and UMGF.

Methods	Twitter2015						Twitter2017					
	MNER			QG			MNER			QG		
	Pre.	*Rec.*	*F1*	Accu@0.5	Accu@0.75	Miou	*Pre.*	*Rec.*	*F1*	Accu@0.5	Accu@0.75	Miou
MNER-QG-Text	76.35	69.46	72.74	-	-	-	87.12	84.03	85.55	-	-	-
MNER-VG	77.03	71.08	73.94	-	-	-	87.91	84.22	86.03	-	-	-
FA-VG	-	-	-	50.83	32.69	45.49	-	-	-	56.03	38.92	51.14
MNER-QG (Ours)	77.43	72.15	74.70	53.93 (M:54.86)	40.22 (M:41.13)	49.50 (M:50.41)	88.26	85.65	86.94	57.50 (M:58.49)	43.03 (M:43.67)	54.09 (M:55.3)

Table 4: Performance comparison of joint-training and single-training models on test set. Note that two results were provided for the QG task, one is the QG results when MNER reaches the optimum, and the other is the optimal results in the QG task. (M denotes Max).

Methods	Twitter2015		Twitter2017		Flickr30K
	(W.S)	(M.A)	(W.S)	(M.A)	
	A@0.5	A@0.5	A@0.5	A@0.5	A@0.5
FA-VG	50.83	63.94	56.03	71.02	68.69
MNER-QG (Ours)	54.86	67.41	58.49	73.53	-

Table 5: Results on different bounding box labels on test set (W.S and M.A denote weak supervisions and manual annotations, respectively. A@0.5 is Accu@0.5. The result of FA-VG on Flickr30K derives from Yang et al. (2019).)

Effect of Query Transformations. We explore different ways of query transformations and take entity type ORG as example for illustration. 1) *Keyword:* An entity type keyword. e.g.,"Organization". 2) *Rule-based Template Filling:* Phrases generated by a simple template. e.g.,"Please find Organization". 3) *Keyword's Wikipedia:* The definition of the entity type keyword from Wikipedia. e.g.,"An organization is an entity, such as an institution or an association, that has a collective goal and is linked to an external environment." 4) *Keyword+Annotation:* The concatenation of a keyword and its annotation. e.g.,"Organization: Include club, company, government party, school...". Results are shown in Figure 4. Queries designed by methods 1 and 2 contain insufficient information, which results in friendly QG result but limits the language understanding of MRC. For method 3, definitions from Wikipedia are relatively general, leading to inferior results on both tasks. Compared with other methods, method 4 achieves the highest $F1$ score and Accu@0.5 in both tasks. We conjecture that method 4 contains generic and precise knowledge of certain types, which accords with the requirements for query construction.

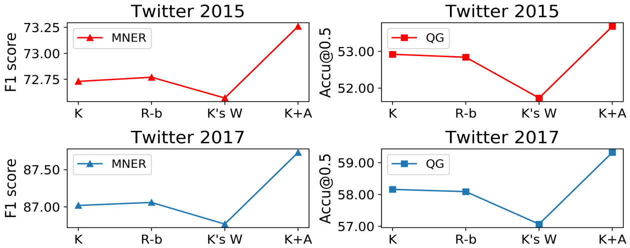


Figure 4: Results with different query transformations in MNER and QG on validation set (K, R-b, K's W, and K+A correspond to methods 1-4 of query transformations).

Conclusion and Future Work

In this work, we propose MNER-QG, an end-to-end MRC framework for MNER with QG. Our model provides prior knowledge of entity types and visual regions with the guidance of queries, then enhances representations of both texts and images after the interactions of multi-scale cross-modality and existence-aware uni-modality, at last, simultaneously extracts entity span and grounds the queries onto visual regions of the image. To perform the query grounding task, we contribute weak supervisions and manual annotations. Experimental results on two datasets show that the joint-training model MNER-QG competes strongly with other baselines in different tasks. MNER-QG leaves ample scope for further research. For future work, we will explore more effective multimodal interaction approaches.

Acknowledgments

This work is supported by the National Key R&D Program of China under Grant No. 2020AAA0106600.

References

Arshad, O.; Gallo, I.; Nawaz, S.; and Calefati, A. 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.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Chen, D.; Li, Z.; Gu, B.; and Chen, Z. 2021a. Multimodal Named Entity Recognition with Image Attributes and Image Knowledge. In *Database Systems for Advanced Applications - 26th International Conference (DASFAA)*, 186–201.

Chen, S.; Wang, Y.; Liu, J.; and Wang, Y. 2021b. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, 12666–12674.

Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; Li, S.; Zhou, X.; and Wang, W. Y. 2019. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations (ICLR)*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jia, M.; Liao, L.; Wang, W.; Li, F.; Chen, Z.; Li, J.; and Huang, H. 2022a. Keywords-aware Dynamic Graph Neural Network for Multi-hop Reading Comprehension. *Neurocomputing*, 501: 25–40.

Jia, M.; Liu, R.; Wang, P.; Song, Y.; Xi, Z.; Li, H.; Shen, X.; Chen, M.; Pang, J.; and He, X. 2022b. E-ConvRec: A Large-Scale Conversational Recommendation Dataset for E-Commerce Customer Service. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, 5787–5796.

Jia, M.; Shen, X.; Shen, L.; Pang, J.; Liao, L.; Song, Y.; Chen, M.; and He, X. 2022c. Query Prior Matters: A MRC Framework for Multimodal Named Entity Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 3549–3558.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 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.

Li, F.; Wang, Z.; Hui, S. C.; Liao, L.; Song, D.; and Xu, J. 2021a. Effective named entity recognition with boundary-aware bidirectional neural networks. In *Proceedings of the Web Conference 2021 (WWW)*, 1695–1703.

Li, F.; Wang, Z.; Hui, S. C.; Liao, L.; Song, D.; Xu, J.; He, G.; and Jia, M. 2021b. 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.

Li, J.; Fei, H.; Liu, J.; Wu, S.; Zhang, M.; Teng, C.; Ji, D.; and Li, F. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 10965–10973.

Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; and Li, J. 2019a. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 5849–5859.

Li, X.; Yin, F.; Sun, Z.; Li, X.; Yuan, A.; Chai, D.; Zhou, M.; and Li, J. 2019b. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, 1340–1350.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 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.

Ma, X.; and Hovy, E. 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)*.

Moon, S.; Neves, L.; and Carvalho, V. 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.

Nishida, K.; Nishida, K.; Nagata, M.; Otsuka, A.; Saito, I.; Asano, H.; and Tomita, J. 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.

Peng, W.; Hu, Y.; Yu, J.; Xing, L.; and Xie, Y. 2021. APER: AdaPtive Evidence-driven Reasoning Network for machine reading comprehension with unanswerable questions. *Knowledge-Based Systems*, 229: 107364.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized

Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational (NAACL)*, 2227–2237.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 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.

Qin, L.; Liu, T.; Che, W.; Kang, B.; Zhao, S.; and Liu, T. 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.

Qiu, L.; Xiao, Y.; Qu, Y.; Zhou, H.; Li, L.; Zhang, W.; and Yu, Y. 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.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 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.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Rei, M.; and Søgaard, A. 2019. Jointly learning to label sentences and tokens. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 6916–6923.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 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.

Ritter, A.; Clark, S.; Etzioni, O.; 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.

Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision (ECCV)*, 817–834.

Tu, M.; Huang, K.; Wang, G.; Huang, J.; He, X.; and Zhou, B. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 9073–9080.

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision*, 104(2): 154–171.

Xu, B.; Huang, S.; Sha, C.; and Wang, H. 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.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 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.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 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.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 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.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 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.

Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; and Zhou, G. 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.

Zhang, Q.; Fu, J.; Liu, X.; and Huang, X. 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.

Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4995–5004.

Zitnick, C. L.; and Dollár, P. 2014. Edge boxes: Locating object proposals from edges. In *European conference on computer vision (ECCV)*, 391–405.