Grounded Multimodal Named Entity Recognition on Social Media

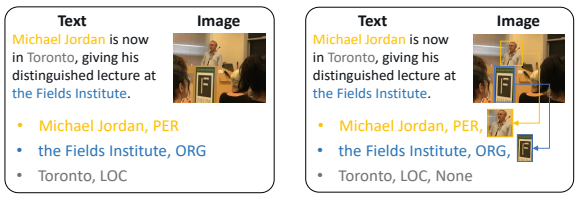
Jianfei Yu*, **Ziyan Li***, **Jieming Wang and Rui Xia**†
School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{jfyu, zyanli, wjm, rxia}@njust.edu.cn

Abstract

In recent years, Multimodal Named Entity Recognition (MNER) on social media has attracted considerable attention. However, existing MNER studies only extract entity-type pairs in text, which is useless for multimodal knowledge graph construction and insufficient for entity disambiguation. To solve these issues, in this work, we introduce a Grounded Multimodal Named Entity Recognition (GMNER) task. Given a text-image social post, GMNER aims to identify the named entities in text, their entity types, and their bounding box groundings in image (i.e., visual regions). To tackle the GMNER task, we construct a Twitter dataset based on two existing MNER datasets. Moreover, we extend four well-known MNER methods to establish a number of baseline systems and further propose a Hierarchical Index generation framework named H-Index, which generates the entity-type-region triples in a hierarchical manner with a sequence-to-sequence model. Experiment results on our annotated dataset demonstrate the superiority of our H-Index framework over baseline systems on the GMNER task. Our dataset annotation and source code are publicly released at https://github.com/NUSTM/GMNER.

1 Introduction

Fueled by the rise of phones and tablets with camera functions, user posts on social media platforms such as Twitter are increasingly multimodal, e.g., containing images in addition to text. The explosive growth of multimodal posts is far beyond humans' capability to digest them. Hence, it presents a pressing need for automatically extracting important information such as entities and relations from the large amount of multimodal posts, which is crucial for structured knowledge graph construction to help people efficiently understand massive content. As an emerging subtask for multimodal knowledge graph construction (Liu et al., 2019), Multimodal Named Entity Recognition (MNER) on social media has recently attracted increasing attention (Zhang et al., 2018; Moon et al., 2018). Given a text-image social post, MNER aims to recognize named entities in text and classify them into pre-defined types such as person (PER), location (LOC), and organization (ORG).



(a) Mulitmodal NER (b) Grounded Mulitmodal NER

Figure 1: Comparison between Multimodal NER (MNER) and our Grounded Multimodal NER (GMNER) task. GMNER is a task to identify the named entities in text, their entity types, and their bounding box groundings in image. *None* denotes there is no grounded bounding box for the entity *Toronto*.

Most previous studies formulate the MNER task as a sequence labeling problem, which focus on (1) designing effective attention mechanisms to model the vision-language interaction to obtain vision-aware word representations (Lu et al., 2018; Yu et al., 2020; Zhang et al., 2021a; Chen et al., 2022b) or (2) converting the images into the textual space by generating image captions and object tags (Chen et al., 2021b; Wang et al., 2022a). Inspired by the success of applying the machine reading comprehension (MRC) framework in NER (Li et al., 2020b), several recent works formalize the MNER task as a MRC problem, which extract entities by answering queries about entity types (Jia et al., 2022, 2023).

However, existing MNER studies mainly regard the visual features as additional clues to help enhance the performance of the text-only NER task, which suffer from several limitations. First, as shown in Fig. 1, previous MNER works only ex-

*Equal contribution.
†Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9141–9154

tract entity-type pairs in text, but failing to link the entities to their corresponding bounding boxes in image. The extracted entity-type pairs are solely useful for constructing text-only knowledge graph rather than multimodal knowledge graph. Moreover, only identifying entity-type pairs in text is often insufficient for entity disambiguation. For example, in Fig. 1, without the grounded yellow bounding box, it is hard to infer the (*Michael Jordan*, *PER*) pair refers to the professor in UC Berkeley rather than the famous basketball player.

To address these issues, in this paper, we propose a new task named Grounded Multimodal Named Entity Recognition (GMNER), aiming to extract the named entities in text, their entity types, and their bounding box groundings in image. Given the example in Fig. 1, the goal is to extract three entity-type-region multimodal triples, i.e., (*Michael Jordan*, *PER*, *yellow box*), (*the Fields Institute*, *ORG*, *blue box*) and (*Toronto*, *LOC*, *None*). GMNER presents the following challenges: (1) apart from extracting entity-type pairs, it requires predicting whether each entity has a grounded region in image; (2) for entities with visually grounded regions, it needs to locate its corresponding bounding box groundings (i.e., visual regions).

To tackle the GMNER task, we first construct a Twitter dataset based on two benchmark MNER datasets, in which we manually annotate the bounding box groundings for each entity-type pair labeled by the two datasets. With the new dataset, we benchmark the GMNER task by establishing a number of baseline systems based on four well-known MNER methods. Furthermore, inspired by the success of the index generation framework in the NER task (Yan et al., 2021), we formulate the GMNER task as a multimodal index generation problem by linearizing all the entity-type-region triples into a position index sequence. We then propose a Hierarchical Index generation framework named H-Index, aiming to address the aforementioned two challenges of GMNER in a hierarchical manner. Specifically, a pre-trained sequence-to-sequence model BART (Lewis et al., 2020) is first employed to encode the textual and visual inputs to generate a set of triples, which contain the indexes of entity positions, entity types, and groundable or ungroundable indicators. Moreover, for groundable entities, we further stack a visual output layer to predict the distribution over candidate visual regions for entity grounding.

The main contributions of our work can be summarized as follows:

- We introduce a new task named Grounded Multimodal Named Entity Recognition (GMNER), which aims to extract all the entity-type-region triples from a text-image pair. Moreover, we construct a Twitter dataset for the task based on two existing MNER datasets.
- We extend four well-known MNER methods to benchmark the GMNER task and further propose a Hierarchical Index generation framework named H-Index, which generates the entity-type-region triples in a hierarchical manner.
- Experimental results on our annotated dataset show that the proposed H-Index framework performs significantly better than a number of unimodal and multimodal baseline systems on the GMNER task, and outperforms the second best system by 3.96% absolute percentage points on F1 score.

2 Task Formulation

Given a multimodal input containing a piece of text with n words $\boldsymbol{s} = (s_1, \ldots, s_n)$ and an accompanying image $\boldsymbol{v}$, the goal of our Grounded Multimodal Named Entity Recognition (GMNER) task is to extract a set of multimodal entity triples:

$$Y = \{(e_1, t_1, r_1), \ldots, (e_m, t_m, r_m)\}, \quad (1)$$

where (e_i, t_i, r_i) denotes the i-th triple, e_i is one of the entities in text, t_i refers to the type of e_i which belongs to four pre-defined entity types including *PER*, *LOC*, *ORG*, and *MISC*, and r_i denotes the visually grounded region of entity e_i. It is worth noting that if there is no grounded region of entity e_i, r_i is *None*; otherwise, r_i consists of a 4-D spatial feature containing the top-left and bottom-right positions of the grounded bounding box, i.e., $(r_i^{x_1}, r_i^{y_1}, r_i^{x_2}, r_i^{y_2})$.

3 Dataset

Since there was no available corpus for the GMNER task, we construct a Twitter dataset as follows.

Data Collection. Our dataset is built on two benchmark MNER datasets, i.e., *Twitter-15* (Zhang et al., 2018) and *Twitter-17* (Yu et al., 2020), which have already annotated all the entities and their types for each multimodal tweet. To alleviate the annotation difficulty, we filter samples with missing images or with more than 3 entities belonging to

Split	#Tweet	#Entity	#Groundable Entity	#Box
Train	7,000	11,782	4,694	5,680
Dev	1,500	2,453	986	1,166
Test	1,500	2,543	1,036	1,244
Total	10,000	16,778	6,716	8,090

Table 1: Statistics of our Twitter-GMNER dataset.

the same type, and then merge the remaining 12K+ samples as our raw dataset for annotation.

Bounding Box Annotation. We employ three graduate students to independently annotate the grounded regions (i.e., bounding boxes) for each labeled entity based on a widely-used image annotation tool named LabelImg[1]. Fleiss Kappa (Fleiss, 1971) is adopted to measure the annotation agreement. Note that if the Intersection over Union (IoU) score between two annotations is larger than 0.5, we regard them as consistent annotations. The Fleiss score between three annotators is $\mathcal{K} = 0.84$, indicating a substantial annotation agreement. To ensure the quality of our dataset, we remove samples in which the IoU score between annotations is less than 0.5. Finally, we obtain 10,159 samples and randomly select 10K samples as our Twitter-GMNER dataset, followed by averaging the three annotations as the ground-truth bounding box annotation for each sample.

Dataset Analysis. Following Moon et al. (2018), we divide our dataset into train (70%), validation (15%), and test sets (15%). As shown in Table 1, our dataset contains 16,778 entities and around 60% entities do not have a grounded bounding box. For the remaining 6,716 groundable entities, we manually annotate a total of 8,090 bounding boxes, which indicates that each entity may correspond to more than one bounding box.

In Table 2, we compare our dataset with six NER datasets for social media. WNUT16 (Strauss et al., 2016) and WNUT17 (Derczynski et al., 2017) are two text-only NER datasets released at the 2nd and 3rd Workshop on Noisy User-generated Text. Twitter-Snap, Twitter-15, and Twitter-17 are three benchmark MNER datasets released by Lu et al. (2018), Zhang et al. (2018), and Yu et al. (2020), respectively. WikiDiverse is a new dataset introduced by Wang et al. (2022b). Compared with existing datasets, our dataset contains more annotated samples (i.e., 10K) and is the first dataset containing both textual and visual annotations.

Fig. 2 (left) shows the distribution of the number

[1] https://github.com/tzutalin/labelImg

Dataset	Modality	Source	Size
WNUT16	$T_i \rightarrow T_o$	Twitter	5.6K
WNUT17	$T_i \rightarrow T_o$	Reddit et al.	5.7K
Twitter-SNAP	$T_i, V_i \rightarrow T_o$	Twitter	7.2K
Twitter-15	$T_i, V_i \rightarrow T_o$	Twitter	8.3K
Twitter-17	$T_i, V_i \rightarrow T_o$	Twitter	4.8K
WikiDiverse	$T_i, V_i \rightarrow T_o$	News	7.8K
Twitter-GMNER	$T_i, V_i \rightarrow T_o, V_o$	Twitter	10K

Table 2: Comparison with other Named Entity Recognition datasets on social media. T_i and V_i represent textual and visual inputs, and T_o and V_o represent textual and visual outputs.

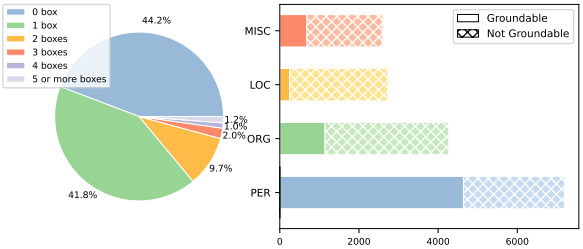


Figure 2: The distribution of the number of bounding boxes in each sample (left) and the distribution of groundable and ungroundable entities for each entity type (right).

of bounding boxes in each sample. We can observe that the image in 44.2% samples is unrelated to any entity mentioned in text, whereas 41.8% samples only contain one bounding box and around 14.0% samples contain two or more bounding boxes. This indicates the necessity and challenge of achieving text-image and entity-image alignments for GMNER. In Fig. 2 (right), we show that most entities with the *PER* type are grounded in the image, whereas entities with the other three types (especially *LOC*) are usually not grounded in the image.

4 Methodology

In this section, we present the details of our proposed Hierarchical Index generation (H-Index) framework.

Overview. As illustrated in Fig. 3, H-Index formulates the GMNER task as an index generation problem, which resorts to a pre-trained sequence-to-sequence model BART (Lewis et al., 2020) to encode the textual and visual inputs, followed by decoding the indexes of entities, types, and groundable or ungroundable indicators. For entities with groundable indicators, another output layer is added to predict the distribution over visual regions for entity grounding.

4.1 Feature Extraction

Text Representation. Given the input text $\boldsymbol{s} = (s_1, \ldots, s_n)$, we feed it to the embedding matrix

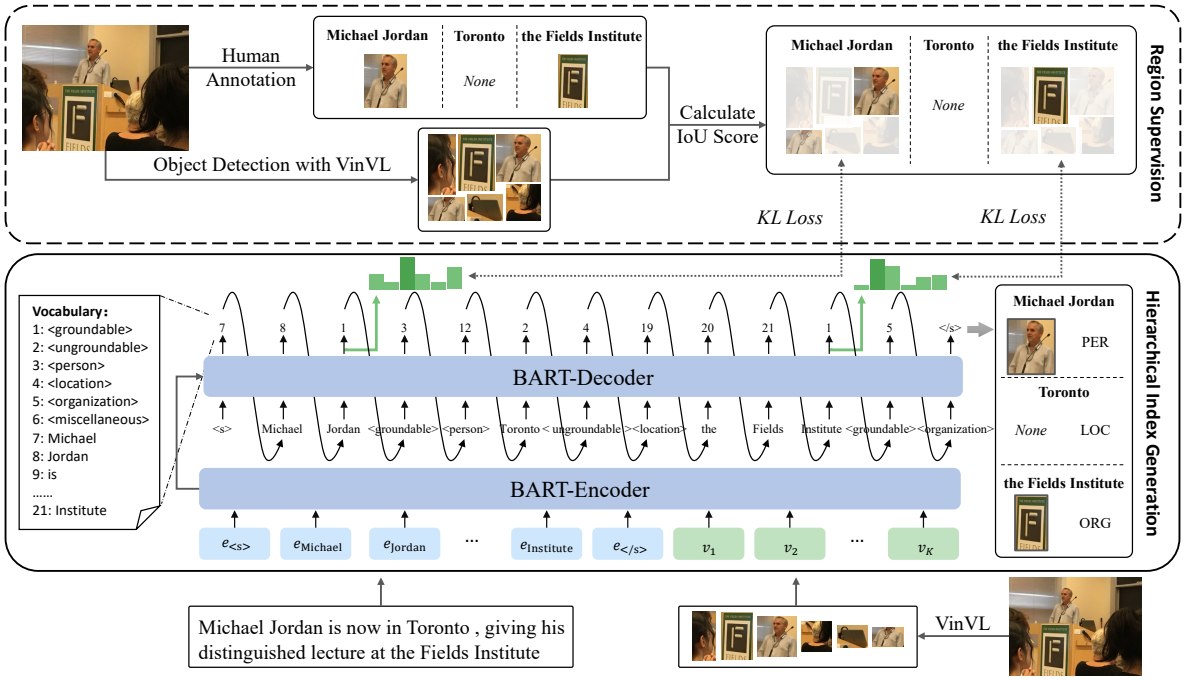


Figure 3: The overview of our proposed Hierarchical Index generation framework (H-Index).

of BART to obtain the text representation as $\mathbf{T} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, where $\mathbf{e}_i \in \mathbb{R}^d$.

Visual Representation. Given the input image $\boldsymbol{v}$, we employ a widely-adopted object detection method VinVL (Zhang et al., 2021b) to identify all the candidate objects (i.e., visual regions). After ranking these objects based on their detection probabilities, we keep the top-K objects and extract the mean-pooled convolutional features from VinVL to obtain fixed-size embeddings for visual regions, denoted by $\mathbf{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_K\}$, where $\mathbf{r}_i \in \mathbb{R}^{2048}$ is the representation of the i-th region. We then use a linear layer to transform $\mathbf{R}$ into the same dimension of text, and thus the regional representation is denoted as $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$, where $\mathbf{v}_i \in \mathbb{R}^d$.

4.2 Design of Multimodal Indexes

As mentioned before, the GMNER task requires extracting three kinds of information, including entity mentions in text, entity types, and visual regions in image. To project these different information into the same output space, we draw inspiration from the NER task (Yan et al., 2021) and use unified position indexes to pointing to these information.

Specifically, we can infer from Table 1 that around 60% entities do not have grounded visual regions in image, which indicates the correct prediction of the relevance between entities and images is essential to entity grounding. Thus, we first transform the entity-image relation into two indexes (i.e., *1* and *2* in the left of Fig. 3) to indicate whether each entity is groundable or ungroundable. Next, we use four indexes (i.e., *3* to *6*) to refer to four entity types. Because the input text $\boldsymbol{s}$ is a sequence with n words, we directly use n position indexes starting from *7* to refer to each word in $\boldsymbol{s}$.

For example, given the textual and visual inputs in Fig. 3, its output index sequence contains three entity-relation-type triples. The first triple *[7,8,1,3]* refers to {*Michael Jordan*, *groundable*, *PER*}, the second triple *[12,2,4]* denotes {*Toronto*, *ungroundable*, *LOC*}, and the third triple *[19,20,21,1,5]* refers to {*the Fields Institute*, *groundable*, *ORG*}. Formally, let us use $\boldsymbol{y}$ to denote the output index sequence.

4.3 Index Generation Framework

Given a multimodal input, we employ a sequence-to-sequence model BART (Lewis et al., 2020) to generate the output index sequence $\boldsymbol{y}$.

Encoder. We first feed the concatenation of text and visual representations to the BART encoder to obtain the hidden representation as follows:

$$\mathbf{H}^e = [\mathbf{H}^e_{\mathrm{T}}; \mathbf{H}^e_{\mathrm{V}}] = \mathrm{Encoder}([\mathbf{T}; \mathbf{V}]), \quad (2)$$

where $\mathbf{H}^e_{\mathrm{T}} \in \mathbb{R}^{n \times d}$ and $\mathbf{H}^e_{\mathrm{V}} \in \mathbb{R}^{K \times d}$ are textual and visual parts of $\mathbf{H}^e \in \mathbb{R}^{(n+K) \times d}$, respectively.

Decoder. At the i-th time step, the decoder takes $\mathbf{H}^e$ and the previous decoder output $\boldsymbol{y}_{<i}$ as inputs

to predict the output probability distribution $p(\boldsymbol{y}_i)$:

$$\mathbf{h}_i = \text{Decoder}(\mathbf{H}^e; \boldsymbol{y}_{<i}), \quad (3)$$

$$\bar{\mathbf{H}}^e_{\text{T}} = \big(\mathbf{T} + \text{MLP}(\mathbf{H}^e_{\text{T}})\big)/2, \quad (4)$$

$$p(\boldsymbol{y}_i) = \text{Softmax}([\mathbf{C}; \bar{\mathbf{H}}^e_{\text{T}}] \cdot \mathbf{h}_i), \quad (5)$$

where MLP refers to a multi-layer perceptron, $\mathbf{C}$ = TokenEmbed(c) refers to the embeddings of two indicator indexes, four entity type indexes, and special tokens such as the "end of sentence" token </*s*>, and $\cdot$ denotes the dot product.

The cross entropy loss is used to optimize the parameters of the generative model as follows:

$$\mathcal{L}^T = -\frac{1}{NM}\sum_{j=1}^{N}\sum_{i=1}^{M}\log p(\boldsymbol{y}_i^j), \quad (6)$$

where N and M refer to the number of samples and the length of output index sequence, respectively.

4.4 Entity Grounding

Lastly, for groundable entities, we further stack another output layer to perform entity grounding.

Specifically, let us use $\mathbf{h}_k$ to refer to the time step whose predicted index is the groundable indicator (i.e., index *1*). We then obtain the probability distribution over all the visual regions from VinVL, denoted by $p(\boldsymbol{z}_k)$ as follows:

$$\bar{\mathbf{H}}^e_{\text{V}} = \big(\mathbf{V} + \text{MLP}(\mathbf{H}^e_{\text{V}})\big)/2, \quad (7)$$

$$p(\boldsymbol{z}_k) = \text{Softmax}(\bar{\mathbf{H}}^e_{\text{V}} \cdot \mathbf{h}_k). \quad (8)$$

Region Supervision. As shown in the top of Fig. 3, since the visual regions from VinVL are different from the ground-truth (GT) bounding boxes, we first compute the overlap between visual regions and GT bounding boxes based on their Intersection over Union (IoU) scores. Note that for each visual region, we compute its IoU scores with respect to all GT bounding boxes of a given entity and take the maximum value as its IoU score. Moreover, for visual regions whose IoU score is less than 0.5, we follow the practice in visual grounding (Yu et al., 2018b) by setting its IoU score as 0. Next, we renormalize the IoU score distribution as the region supervision for a given entity, denoted by $g(\boldsymbol{z}_k)$.

The objective function of entity grounding is to minimize the Kullback-Leibler Divergence (KLD) loss between the predicted region distribution $p(\boldsymbol{z}_k)$ and the region supervision $g(\boldsymbol{z}_k)$:

$$\mathcal{L}^V = \frac{1}{NE}\sum_{j=1}^{N}\sum_{k=1}^{E} g(\boldsymbol{z}_k^j)\log\frac{g(\boldsymbol{z}_k^j)}{p(\boldsymbol{z}_k^j)}, \quad (9)$$

Algorithm 1 Our Entity-Groundable/Ungrounable-Type Triple Recovery Algorithm

Input: Predicted sequence $\hat{\boldsymbol{y}} = [\hat{y}_1, ..., \hat{y}_l]$ and $\hat{y}_l \in [1, n + |c|]$, where c is the list of two indicator indexes, four entity type indexes, and special tokens.

Output: Triples E

```
E = {}, e = [], i = 1
while i <= l do
    y_i = Y[i]
    if y_i < |c| then
        if len(e) > 0 then
            if indexes in e is ascending then
                if y_i = 1 or y_i = 2 then
                    E.add(e, c_{y_i}, c_{y_{i+1}})
                end if
            end if
        end if
        e = []
        i = i + 2
    else
        e.append(y_i)
    end if
    i = i + 1
end while
return E
```

where E is the number of groundable entities.

In the training stage, we combine $\mathcal{L}^T$ and $\mathcal{L}^V$ as the final loss of our H-Index model:

$$\mathcal{L} = \mathcal{L}^T + \mathcal{L}^V. \quad (10)$$

4.5 Entity-Type-Region Triple Recovery

In the inference stage, given a multimodal input, we use the trained H-Index model to generate the index sequence $\hat{\boldsymbol{y}}$ in an autoregressive manner based on greedy search, and predict the region distribution $p(\hat{\boldsymbol{z}}_k)$ for the k-th groundable entity.

With the output index sequence, we can first convert each index to its original meaning and then recover (*entity*, *groundable/ungroundable*, *type*) triples based on the index span of each element. The full algorithm is shown in Algorithm 1.

For the j-th ungroundable entity, the predicted triple is $(e_j, t_j, \textit{None})$. For the k-th groundable entity, we regard the visual region with the highest probability in $p(\hat{\boldsymbol{z}}_k)$ as the predicted bounding box, and take its 4-D coordinates $r_k = (x_k^1, y_k^1, x_k^2, y_k^2)$ as the visual output. Thus, the predicted triple of the k-th groundable entity is (e_k, t_k, r_k).

5 Experiments

5.1 Experimental Settings

For our proposed framework H-Index, we employ the pre-trained VinVL model released by (Zhang et al., 2021b) to detect top-K visual regions, and use the pre-trained $\text{BART}_{\text{base}}$ model from (Lewis et al., 2020) to initialize the parameters in the index generation framework in Section 4.3. Hence, the hidden dimension d is set to the default setting 768. The batch size and training epoch are set to 32 and 30, respectively. During training, we use the AdamW optimizer for parameter tuning. For the learning rate and the number of candidate visual regions K, we set their values to 3e-5 and 18 after a grid search over the combinations of [1e-5, 1e-4] and [2, 20] on the development set.

Evaluation Metrics. The GMNER task involves three elements, i.e., entity, type, and visual region. For entity and type, we follow previous MNER works to use the exact match for evaluation (Zhang et al., 2018). For visual region, if it is ungroundable, the prediction is considered as correct only when it is *None*; otherwise, the prediction is considered as correct only when the IoU score between the predicted visual region and one of the ground-truth (GT) bounding boxes is large than 0.5 (Mao et al., 2016). The correctness of each element is computed as follows:

$$C_e/C_t = \begin{cases} 1, & p_e/p_t = g_e/g_t; \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

$$C_r = \begin{cases} 1, & p_r = g_r = None; \\ 1, & \max(\text{IoU}_1, ..., \text{IoU}_j) > 0.5; \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

where C_e, C_t, and C_r denote the correctness of entity, type, and region predictions, p_e, p_t, and p_r denote the predicted entity, type, and region, g_e, g_t, and g_r denote the gold entity, type, and region, and IoU_j denotes the IoU score between the predicted region p_r with the j-th GT bounding box $g_{r,j}$. We then calculate precision (Pre.), recall (Rec.), and F1 score to measure the performance of GMNER:

$$correct = \begin{cases} 1, & C_e \text{ and } C_t \text{ and } C_r; \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

$$Pre = \frac{\#correct}{\#predict}, \quad Rec = \frac{\#correct}{\#gold}, \quad (14)$$

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec}, \quad (15)$$

where *#correct* denotes the number of predicted triples that match the gold triples, and *#predict* and *#gold* are the number of predicted and gold triples.

5.2 Baseline Systems

Since GMNER is a new task and there is no existing method for comparison, we first consider several text-only methods as follows:

- *HBiLSTM-CRF-None*, which uses the hierarchical BiLSTM-CRF model (Lu et al., 2018) to extract entity-type pairs, followed by setting the region prediction to the majority class, i.e., *None*;
- *BERT-None*, *BERT-CRF-None*, and *BARTNER-None*, which replace the hierarchical BiLSTM-CRF model in *HBiLSTM-CRF-None* with BERT (Devlin et al., 2019), BERT-CRF, and BARTNER (Yan et al., 2021), respectively.

Moreover, we develop a pipeline approach as a strong baseline, which first uses any previous MNER method to extract entity-type pairs and then predicts the bounding box for each pair with an Entity-aware Visual Grounding (EVG) model.

Specifically, given the i-th extracted entity-type pair (e_i, t_i) from existing MNER methods as well as the textual input $\boldsymbol{s}$, we construct the textual input as follows: $[[\text{CLS}], \boldsymbol{s}, [\text{SEP}], e_i, [\text{SEP}], t_i, [\text{SEP}]]$, which is fed to a pre-trained $\text{BERT}_{\text{base}}$ model (Devlin et al., 2019) to obtain the text representation $\mathbf{T}$. We then use the feature extraction method in Section 4.1 to obtain the visual representation $\mathbf{V}$. Next, a Cross-Model Transformer layer (Tsai et al., 2019) is utilized to model the interaction between the text and visual representations as follows: $\mathbf{H} = \text{CMT}(\mathbf{V}, \mathbf{T}, \mathbf{T})$, where $\mathbf{V}$ and $\mathbf{T}$ are regarded as queries and keys/values, and $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$ is the generated hidden representation. For each visual region $\mathbf{h}_j \in \mathbb{R}^d$, we add an output layer to predict whether it is the grounded region of (e_i, t_i): $p(y_j) = \text{sigmoid}(\mathbf{w}^\top \mathbf{h}_j)$, where $\mathbf{w} \in \mathbb{R}^d$ is the weight matrix. During inference, we choose the visual region with the highest probability. If the probability is higher than a tuned threshold, it implies the input entity-type pair is groundable and the predicted region is the top visual region; otherwise, the prediction region is *None*.

As shown in Table 3, we stack the EVG model over four well-known MNER methods as follows:

- *GVATT-RCNN-EVG*, which uses GVATT (Lu et al., 2018), a visual attention method based on

	Methods	GMNER Pre.	GMNER Rec.	GMNER F1	MNER Pre.	MNER Rec.	MNER F1	EEG Pre.	EEG Rec.	EEG F1
Text	HBiLSTM-CRF-None (Lu et al., 2018)	43.56	40.69	42.07	78.80	72.61	75.58	49.17	45.92	47.49
	BERT-None (Devlin et al., 2019)	42.18	43.76	42.96	77.26	77.41	77.30	46.76	48.52	47.63
	BERT-CRF-None	42.73	44.88	43.78	77.23	78.64	77.93	46.92	49.28	48.07
	BARTNER-None (Yan et al., 2021)	44.61	45.04	44.82	79.67	79.98	79.83	48.77	49.23	48.99
Text+Image	GVATT-RCNN-EVG (Lu et al., 2018)	49.36	47.80	48.57	78.21	74.39	76.26	54.19	52.48	53.32
	UMT-RCNN-EVG (Yu et al., 2020)	49.16	51.48	50.29	77.89	79.28	78.58	53.55	56.08	54.78
	UMT-VinVL-EVG (Yu et al., 2020)	50.15	52.52	51.31	77.89	79.28	78.58	54.35	56.91	55.60
	UMGF-VinVL-EVG (Zhang et al., 2021a)	51.62	51.72	51.67	79.02	78.64	78.83	55.68	55.80	55.74
	ITA-VinVL-EVG (Wang et al., 2022a)	52.37	50.77	51.56	80.40	78.37	79.37	56.57	54.84	55.69
	BARTMNER-VinVL-EVG	52.47	52.43	52.45	**80.65**	**80.14**	**80.39**	55.68	55.63	55.66
	H-Index (Ours)	**56.16**	**56.67**	**56.41**	79.37	80.10	79.73	**60.90**	**61.46**	**61.18**

Table 3: Performance comparison of different methods on tasks of GMNER, MNER, and Entity Extraction & Grounding (EEG).

the BiLSTM-CRF model (Lample et al., 2016), to extract entity-type pairs, followed by applying the EVG model based on the objects detected by Faster R-CNN (Anderson et al., 2018);

- *UMT-RCNN-EVG*, which replaces the MNER method in *GVATT-RCNN-EVG* with UMT (Yu et al., 2020), a Multimodal Transformer approach with an auxiliary entity span detection task. *UMT-VinVL-EVG* is a variant of *UMT-RCNN-EVG*, which replaces Faster R-CNN with VinVL;
- *UMGF-VinVL-EVG* is a variant of *UMT-VinVL-EVG* using UMGF (Zhang et al., 2021a) for MNER, which models text-image interactions with a multimodal graph fusion network;
- *ITA-VinVL-EVG* is another variant of *UMT-VinVL-EVG* using ITA (Wang et al., 2022a) for MNER, which translates images to captions and object tags, followed by sequence labeling.
- *BARTMNER-VinVL-EVG* is a variant of our H-Index approach, which first uses the index generation framework in Section 4.3 to identify entity-type pairs, and then uses the EVG model to predict the grounded bounding box for each pair.

5.3 Main Results

In Table 3, we show the results of different methods on the GMNER task. To better compare these methods, we also report the F1 score of two subtasks of GMNER, including MNER and Entity Extraction & Grounding (EEG). Note that MNER aims to identify the entity-type pairs whereas EEG aims to extract the entity-region pairs.

Results on GMNER. First, for text-based methods, it is clear that *BARTNER-None* significantly outperforms the other methods, which shows the effectiveness of the index generation framework and agrees with the observation in existing NER works (Yan et al., 2021). Second, all the multimodal approaches consistently perform much better than text-based methods. This indicates the usefulness of our proposed Entity-aware Visual Grounding (EVG) baseline. Third, comparing all the multimodal baseline systems, *BARTMNER-VinVL-EVG* obtains the best result, primarily due to its outstanding performance on the MNER subtask. Finally, we can clearly observe that our proposed *H-Index* framework outperforms the best baseline *BARTMNER-VinVL-EVG* by 3.96 absolute percentage points based on F1 score. The main reason for the performance gain is that all the baseline systems extract entity-types pairs followed by visual grounding, which suffer from the common error propagation issue of pipeline methods. In contrast, *H-Index* uses the unified index generation framework to directly generate the entity-type-region triples with a sequence-to-sequence model.

Results on MNER and EEG. First, for the MNER subtask, we can see that our *H-Index* framework performs generally better than most baselines but worse than *BARTMNER-VinVL-EVG*. We conjecture the reason is that all the baselines are pipeline methods and should achieve the best performance on each stage, whereas our *H-Index* model is an end-to-end approach for entity-type-region triple extraction, which may only obtain the sub-optimal model on the MNER task. In addition, for the EEG subtask, *H-Index* significantly outperforms the best baseline by 5.44 absolute percentage points. These observations verify the effectiveness of our *H-Index* framework.

5.4 In-Depth Analysis

Ablation Study. In Table 5, we conduct ablation study of our *H-Index* framework. First, we replace

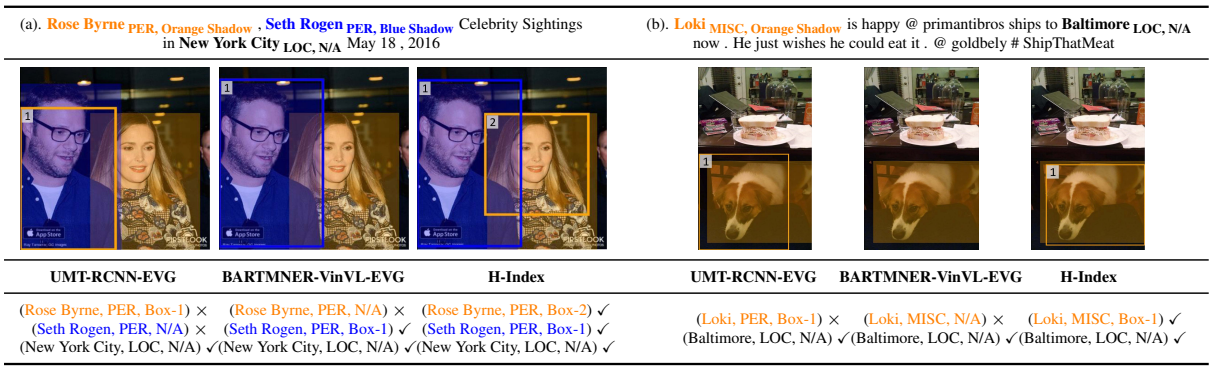


Table 4: Prediction comparison on two test samples. ✓ and × denote correct and incorrect predictions. N/A refers to *None*.

Methods	Pre.	Rec.	F1
H-Index	**56.16**	**56.67**	**56.41**
- rep. KLD Loss with CE Loss	55.88	53.72	54.78
- w/o Hierarchical Prediction	55.83	52.89	54.32

Table 5: Comparison results of ablated H-Index models.

the KLD loss in Equation (9) with the cross-entropy loss. We find that the performance slightly drops, indicating that the KLD loss can better capture the relations between different visual regions for region detection. Moreover, we remove the hierarchical prediction of the groundable/ungroundable indicator in Section 4.3 and entity grounding in Section 4.4. Specifically, we use a special token to indicate whether the current time step in the decoder is for visual grounding, and then add a binary classification layer for each visual region, which is the same as the EVG baseline. As shown in Table 5, removing the hierarchical prediction leads to a performance drop of 2.09 percentage points.

Sensitivity Analysis of K. We use our *H-Index* model and *BARTMNER-VinVL-EVG* to analyze the impact of the number of object regions from VinVL on GMNER and EEG tasks. In Fig. 4, we can find the two methods gradually perform better as K becomes larger. This is because when K is small, the top-K regions from VinVL may not cover ground-truth visual regions. When K equals to 18, the two methods consistently achieve the best performance.

5.5 Case Study

We further conduct case study to compare the predictions of *UMT-RCNN-EVG*, *BARTMNER-VinVL-EVG*, and *H-Index* on two test samples in our dataset. In Table 4.a, we find that all the methods correctly identify the three entity-type pairs. However, *UMT-RCNN-EVG* is confused with the two *PER* entities and wrongly predicts their grounded regions, while *BARTMNER-VinVL-EVG* fails to

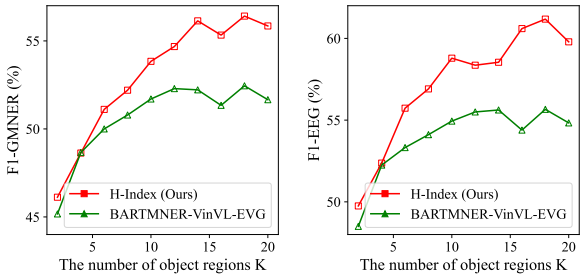


Figure 4: The impact of the value of K on GMNER and Entity Extraction & Grounding (EEG) tasks.

identify the grounded region of *Rose Byrne*. In contrast, our *H-Index* model correctly identifies the bounding box groundings for the two *PER* entities. Similarly, in Table 4.b, the two baselines fail to identify either the correct entity type or the correct bounding box of *Loki*, whereas *H-Index* correctly grounds *Loki* onto the visual region with the dog, and predicts its type as *MISC*.

6 Related Work

NER on Social Media. Many supervised learning methods have achieved satisfactory results on formal text (Li et al., 2020a), including feature engineering methods (Finkel et al., 2005; Ratinov and Roth, 2009) and deep learning methods (Chiu and Nichols, 2016; Ma and Hovy, 2016). However, most of them perform poorly on social media, because the text on social media is often informal and short. To handle this problem, many social text-based features such as hashtags (Gimpel et al., 2010) and freebase dictionary (Ritter et al., 2011) are designed to enhance the performance of both feature-based methods (Baldwin et al., 2015) and deep learning methods (Limsopatham and Collier, 2016; Gerguis et al., 2016; Lin et al., 2017; Suman et al., 2021).

Multimodal NER on Social Media. With the

rapid growth of multimodal posts on social media, MNER has recently attracted much attention. Most existing MNER methods focus on modeling the text-image interactions by designing various kinds of cross-modal attention mechanism (Moon et al., 2018; Lu et al., 2018; Zheng et al., 2020). With the recent advancement of deep learning techniques, many studies focus on designing different neural networks for MNER, including Transformer-based methods (Sun et al., 2021; Xu et al., 2022; Chen et al., 2022a; Jia et al., 2023), Graph Neural Network-based methods (Zhang et al., 2021a; Zhao et al., 2022), Modality Translation-based methods (Chen et al., 2021b; Wang et al., 2022a), and Prompt-based models (Wang et al., 2022c). Despite obtaining promising results, these methods solely utilize the visual clues to better extract entity-type pairs. In contrast, the goal of our work is to extract entity-type-region triples from each multimodal post.

Visual Grounding. Given a natural language query, Visual Grounding (VG) aims to locate the most relevant object or region in an image. Most existing works on VG belong to two categories, i.e., one-stage methods and two-stage methods. The former focuses on utilizing recent end-to-end object detection models such as YOLO (Redmon and Farhadi, 2018) and DETR (Carion et al., 2020) to directly predict the visual region (Yang et al., 2019; Deng et al., 2021; Ye et al., 2022). The latter aims to first leverage object detection models (Ren et al., 2015; Zhang et al., 2021b) to obtain region proposals and then rank them based on their relevance to the text query (Yu et al., 2018a; Yang et al., 2020; Chen et al., 2021a). Our work follows the latter line of methods, which detects candidate visual regions with VinVL, followed by entity grounding.

7 Conclusion

In this paper, we introduced a new task named Grounded Multimodal Named Entity Recognition (GMNER), aiming to identify the named entities, their entity types, and their grounded bounding boxes in a text-image social post. Moreover, we constructed a new Twitter dataset for the task, and then extended four previous MNER methods to benchmark the task. We further proposed a Hierarchical Index generation framework (H-Index), which generates the entity-type-region triples in a hierarchical manner. Experimental results demonstrate the effectiveness of our H-index framework.

Limitations

Although we introduce a new GMNER task and propose a number of baseline systems and an H-Index framework, there are still some limitations in this work.

First, our GMNER task only requires identifying the visual regions that are correspondent to named entities mentioned in text. However, for each image, many visual regions may contain real-world entities that are not mentioned in text. Therefore, it would be interesting to further annotate the entities that only occur in the image and explore a more complete MNER task in the future.

Second, our work is a preliminary exploration of the GMNER task, and the proposed approaches are primarily based on previous representative NER or MNER methods. We hope this work can encourage more research to apply the recent advanced techniques from both the NLP and computer vision communities to improve its performance.

Ethics Statement

Our dataset is constructed based on two public MNER datasets, i.e., *Twitter-15* (Zhang et al., 2018) and *Twitter-17* (Yu et al., 2020). Three graduate students are employed as our annotators. The average time to annotate every 1,000 samples for each annotator is around 17 hours. Since the two datasets publicly released the text, images, and named entities, each annotator is asked to independently annotate the bounding box groundings for each entity without accessing to the user account. To ensure that the annotators were fairly compensated, we paid them at an hourly rate of CNY 36 (i.e., USD 5.2 per hour), which is higher than the current average wage in Jiangsu Province, China. We do not share personal information and do not release sensitive content that can be harmful to any individual or community. Because it is easy to retrieve multimodal tweets via image IDs from the two MNER datasets, we will release our annotation based on the textual modality and unique image IDs.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. This work was supported by the Natural Science Foundation of China (62076133 and 62006117), and the Natural Science Foundation of Jiangsu Province for Young Scholars (BK20200463) and Distinguished Young Scholars (BK20200018).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of CVPR*, pages 6077–6086.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Long Chen, Wenbo Ma, Jun Xiao, Hanwang Zhang, and Shih-Fu Chang. 2021a. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1036–1044.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021b. Can images help recognize entities? a study of the role of images for multimodal ner. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 87–96.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022a. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 904–915.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1769–1779.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Michel Naim Gerguis, Cherif Salama, and M Watheq El-Kharashi. 2016. Asu: An experimental study on applying deep learning in twitter named entity recognition. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 188–196.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2010. Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. 2023. Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *Proceedings of AAAI*.

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 ACM MM*, pages 3549–3558.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*, pages 7871–7880.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020a. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020b. A unified mrc framework for named entity recognition. In *Proceedings of ACL*, pages 5849–5859.

Nut Limsopatham and Nigel Collier. 2016. Bidirectional LSTM for named entity recognition in twitter messages. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*.

Bill Yuchen Lin, Frank F Xu, Zhiyi Luo, and Kenny Zhu. 2017. Multi-channel bilstm-crf model for emerging named entity recognition in social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. Mmkg: multi-modal knowledge graphs. In *The Semantic Web: 16th International Conference, ESWC 2019*, pages 459–474.

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 ACL*, pages 1990–1999.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of ACL*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of NAACL*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL*.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of EMNLP*, pages 1524–1534.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

Chanchal Suman, Saichethan Miriyala Reddy, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Why pay more? a simple and efficient named entity recognition system for tweets. *Expert Systems with Applications*, 167:114101.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13860–13868.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of ACL*, pages 6558—-6569.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022a. ITA: Image-text alignments for multi-modal named entity recognition. In *Proceedings of NAACL*, pages 3176–3189.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022b. Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of ACL*, pages 4785–4797.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022c. : Prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In *International Conference on Database Systems for Advanced Applications*, pages 297–305.

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*, pages 1215–1223.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various ner subtasks. In *Proceedings of ACL-IJCNLP*, pages 5808–5822.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2020. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9952–9961.

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*, pages 4683–4693.

Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. 2022. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15502–15512.

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 ACL*, pages 3342–3352.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018a. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315.

Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018b. Rethinking diversified and discriminative proposal generation for visual grounding. In *Proceedings of IJCAI*, pages 1114–1120.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of AAAI*, pages 14347–14355.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of CVPR*, pages 5579–5588.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of AAAI*, pages 5674–5681.

Fei Zhao, Chunhui Li, Zhen Wu, Shangyu Xing, and Xinyu Dai. 2022. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal ner. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3983–3992.

Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2020. Object-aware multimodal named entity recognition in social media posts with adversarial learning. *IEEE Transactions on Multimedia*, 23:2520–2532.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We report our limitations in the last section.

☒ A2. Did you discuss any potential risks of your work?
We do not think there are any potential risks of our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
We summarize our contributions in the introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We introduce our new dataset in section 3.

☑ B1. Did you cite the creators of artifacts you used?
The artifacts we use are referenced and briefly introduced in section 5.1.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We use the publicly available Twitter datasets released by previous MNER works to study our GMNER task with further annotations. We will state the original licenses when releasing the dataset.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We discuss the intended use of our proposed dataset in section 1.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data we use is based on the publicly available datasets, which have been checked and pre-processed by previous works.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We analyze our proposed Twitter-GMNER dataset in section3.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We analyze our proposed Twitter-GMNER dataset in section3.

C ☑ Did you run computational experiments?

We introduce our experiments in section 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We introduce the experimental details in appendix A.3.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We introduce the experiment settings in section 5.1.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Because we have divided our dataset into train, dev, and test sets, we choose a model which obtains the best result on the dev set with a single run, and report its performance on the test set.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We introduce the parameter settings in section 5.1.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

We introduce human annotation details in section3 and appendix A.1.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We introduce the annotation procedure in appendix A.1.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We introduce this information in appendix A.1.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
We build the Twitter-GMNER dataset based on the public datasets and follow their usage requirements.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The dataset we use is based on publicly available datasets, which have been approved by an ethics review board in previous works.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We include this information in section 3.