Entity-level Interaction via Heterogeneous Graph for Multimodal Named Entity Recognition

Gang Zhao, Guanting Dong, Yidong Shi, Haolong Yan, Weiran Xu and **Si Li***
School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
{zhaogang, dongguanting, yidongshi, haolongy, xuweiran, lisi}@bupt.edu.cn

Abstract

Multimodal Named Entity Recognition (MNER) faces two specific challenges: 1) How to capture useful entity-related visual information; 2) How to alleviate the interference of visual noise. Previous works have gained progress by improving interacting mechanisms or seeking for better visual features. However, existing methods neglect the integrity of entity semantics and conduct cross-modal interaction at token-level, which cuts apart the semantics of entities and makes non-entity tokens easily interfered with by irrelevant visual noise. Thus in this paper, we propose an end-to-end heterogeneous Graph-based Entity-level Interacting model (GEI) for MNER. GEI first utilizes a span detection subtask to obtain entity representations, which serve as the bridge between two modalities. Then, the heterogeneous graph interacting network interacts entity with object nodes to capture entity-related visual information, and fuses it into only entity-associated tokens to rid non-entity tokens of the visual noise. Experiments on two widely used datasets demonstrate the effectiveness of our method. Our code will be available at https://github.com/GangZhao98/GEI.

1 Introduction

Multimodal Named Entity Recognition (MNER) aims at combining both textual and visual contents to detect and classify named entities from multimodal social media posts (e.g., tweets). Different from traditional NER (Torisawa et al., 2007; Lample et al., 2016; Ma and Hovy, 2016) that focuses on formal single-modal texts, MNER confronts two specific challenges: **1)** How to capture useful entity-related visual information; **2)** How to alleviate the interference of visual noise. As shown in Figure 1, the *MISC* entity "*Oscars*" appearing in the tweet may be wrongly recognized as *PER*, since it can refer to both a person name and the movie award. But the trophies in image can help figure out that the "*Oscars*" actually indicates the latter. Effectively capturing entity-related information from the image is essential and challenging. Though helpful, incorporating images may also interfere the non-entity tokens that have no corresponding visual information, and makes them easily misidentified as entities. Effectively alleviating the interference brought by images is also a critical challenge.

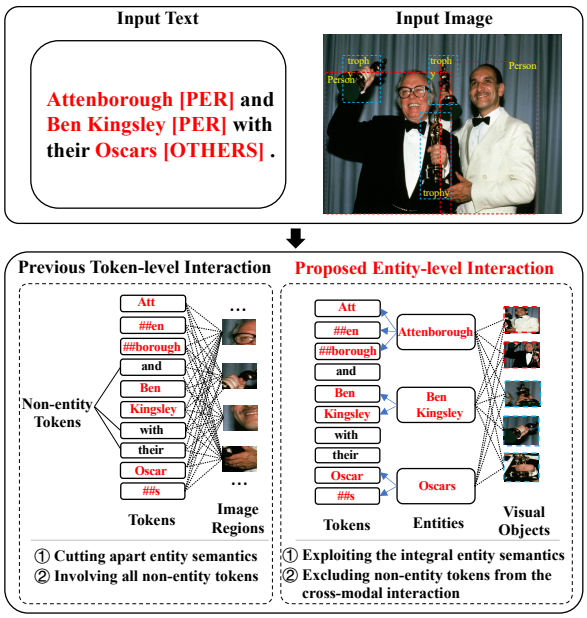

Figure 1: An example from the public social media MNER dataset Twitter-2015, and the difference between previous and proposed cross-modal interacting methods.

Recent works on MNER have gained progress by either improving cross-modal interacting mechanisms (Zhang et al., 2018; Lu et al., 2018; Yu et al., 2020), or seeking for better visual features (Wu et al., 2020; Chen et al., 2020; Zhang et al., 2021). However, existing methods neglect the integrity of entity semantics and directly interact all textual tokens with visual features, which we regard as token-level interaction. Though straightforward,

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6374 - 6379

token-level interaction fails to use integral entity semantics to capture related visual information, and makes non-entity tokens easily interfered with by visual noise.

Thus in this paper, we propose an end-to-end heterogeneous Graph-based Entity-level Interacting model (GEI) for MNER. As shown in Figure 1, the key insight of entity-level interaction is obtaining entity representations to query related visual information from object features, which has several benefits: **1)** Entity representations carry integral entity semantics, which can capture entity-related information effectively. **2)** Interacting visual features with only entity representations instead of all token representations can protect non-entity tokens from the interference brought by images. In detail, GEI first introduces a span detection subtask to obtain entity representations, which serve as the bridge between two modalities. Then, a multimodal heterogeneous graph is constructed with token, entity and object nodes, whose semantic relationships are modeled by four kinds of edges. After that, GEI interacts entity nodes with object nodes to capture related visual information, and fuses it to token nodes that are connected with entity nodes. Finally, a CRF layer is employed to decode named entities from object-aware token representations.

Overall, our contributions are as follows:

1) We propose a novel end-to-end model GEI for MNER. GEI interacts entity representations with visual objects to capture useful entity-related visual information, and excludes non-entity tokens from the interaction to rid them of the visual noise.

2) We conduct experiments on two widely used datasets *Twitter-2015* (Zhang et al., 2018) and *Twitter-2017* (Lu et al., 2018). The results show that GEI tackles MNER challenges effectively and demonstrate the effectiveness of our GEI.

2 Methodology

Figure 2 shows the architecture of GEI, which contains the following components: Entity Representation Extractor (ERE), Object Feature Encoder (OFE), Heterogeneous Graph Interacting Network (HGIN), CRF Decoding modules.

2.1 Entity Representation Extractor

Given an input sentence $\boldsymbol{X} = \{x_i\}_{i=1}^{|X|}$, where x_i is the i^{th} token and $|X|$ is the max sequence length, we employ BERT pre-trained by Devlin et al. (2018) as our text encoder, and obtain contextualized token embeddings $\boldsymbol{C} = \{c_i\}_{i=1}^{|X|}$. Then, we use a Transformer (Vaswani et al., 2017) to gain hidden representation of each token $\boldsymbol{T} = \{t_i\}_{i=1}^{|X|}$:

$$\boldsymbol{T} = Transformer\text{-}1(\boldsymbol{C}) \in \mathbb{R}^{|X|\times 768} \quad (1)$$

After that, we project token representations to the multimodal space via a linear transformation: $\widetilde{\boldsymbol{T}} = W_t\boldsymbol{T}^\top + b_t$, where $\widetilde{\boldsymbol{T}} = \{\widetilde{t}_i\}_{i=1}^{|X|} \in \mathbb{R}^{|X|\times d_m}$ and d_m is dimension of the multimodal space.

We introduce a span detection subtask to construct entity representations, which are used to capture entity-related visual information and serve as the bridge between two modalities in HGIN. Firstly, we feed C to another Transformer layer to obtain specific hidden representations of the subtask:

$$\boldsymbol{T}' = Transformer\text{-}2(\boldsymbol{C}) \in \mathbb{R}^{N\times 768} \quad (2)$$

Then, we use a CRF (Lafferty et al., 2001) layer to recognize possible entity spans $\{(s_i, e_i)\}_{i=1}^{|E|}$, where $|E|$ is the entity number, s_i and e_i are start and end indexes of the i^{th} entity. The span detection loss $\mathcal{L}_{sd}$ is as follows:

$$\mathcal{L}_{sd} = -\sum_i \log P(z|\boldsymbol{X}) \quad (3)$$

$$P(z|\mathbf{X}) = \frac{\prod_{i=1}^{|X|}\varphi_i(z_{i-1}, z_i, \mathbf{X}))}{\sum_{z'\in\mathbf{Z}}\prod_{i=1}^{|X|}\varphi_i(z'_{i-1}, z'_i, \mathbf{X})} \quad (4)$$

where $\varphi_i(z_{i-1}, z_i, \boldsymbol{X})$ and $\varphi_i(z'_{i-1}, z'_i, \boldsymbol{X})$ are potential functions. Finally, we obtain representation of each entity through maxpooling its constituent token representations: $E_n = Max(\{\widetilde{t}_i\}_{i=s_n}^{e_n})$.

2.2 Object Feature Encoder

We propose OFE module to encode the input image to visual features. Considering that visual objects have similar semantic granularity with entities, we acquire object representations as our image features. Given an input image $\boldsymbol{I}$, we first use the object detection algorithm DETR (Carion et al., 2020) to detect bounding boxes of visual objects $\boldsymbol{O} = \{o_i\}_{i=1}^{|O|}$, where $|O|$ is the number of detected objects. Then, we concatenate $\boldsymbol{I}$ and $\boldsymbol{O}$, feed them to the 152-layer ResNet (He et al., 2016) and take the output from the last pooling layer $\boldsymbol{V} = \{v_i\}_{i=1}^{|O|+1}$ as the visual features:

$$\boldsymbol{V} = ResNet([\boldsymbol{I}; \boldsymbol{O}]) \in \mathbb{R}^{(|O|+1)\times 2048} \quad (5)$$

After that, we project visual features to the multimodal space via a multi-layer perceptron with

ReLU activation function:

$$\widetilde{\boldsymbol{V}} = ReLU(W_v \boldsymbol{V}^\top + b_v) \in \mathbb{R}^{(|O|+1)\times d_m} \quad (6)$$

2.3 Heterogeneous Graph Interacting Network

We design HGIN module to capture entity-related visual information via entity-level interaction, and fuse it to entity-associated token representations.

Graph Construction. As shown in Figure 2, the multimodal heterogeneous graph G contains three kinds of nodes: textual token nodes $N_{T_i} = \widetilde{t}_i$, entity nodes $N_{E_i} = E_i$, and visual object nodes $N_{V_i} = \widetilde{v}_i$. We introduce following kinds of edges for G: 1) Entiy-Object Edge: N_{E_i} and N_{V_j} are fully connected to capture the entity-related visual information. 2) Entity-Token Edge: N_{E_i} is connected with associated token nodes $\{N_{T_j}\}_{j=s_i}^{e_i}$ to enhance them with entity-related visual information. 3) Intra-modal Edge: To capture intra-modal interactions, all token nodes $\{N_{T_i}\}_{i=1}^{|X|}$ are fully connected with each other, so do object nodes.

Cross-modal Interaction. Firstly, we employ multi-head self-attention on the intra-modal edge to exploit contexts of the same modality:

$$\boldsymbol{D}_m^{(l)} = MultiHead(\boldsymbol{H}_m^{(l)}, \boldsymbol{H}_m^{(l)}, \boldsymbol{H}_m^{(l)}) \quad (7)$$

where $m \in \{T, V\}$, $\boldsymbol{H}_m^{(l)} = \{H_{m_i}^{(l)}\}$ and $H_{m_i}^{(l)}$ is the hidden feature of node N_{m_i} at the l^{th} layer. Then, we interact entity nodes with object nodes via a gated cross-attention module:

$$\boldsymbol{R}_E^{(l)} = MultiHead(\boldsymbol{H}_E^{(l)}, \boldsymbol{D}_V^{(l)}, \boldsymbol{D}_V^{(l)}) \quad (8)$$

$$\boldsymbol{\alpha}_E^{(l)} = Sigmoid(W_{e_1}^{(l)} \boldsymbol{R}_E^{(l)} + W_{e_2}^{(l)} \boldsymbol{H}_E^{(l)}) \quad (9)$$

$$\boldsymbol{M}_E^{(l)} = \boldsymbol{\alpha}_E^{(l)} \cdot \boldsymbol{R}_E^{(l)} + (1 - \boldsymbol{\alpha}_E^{(l)}) \cdot \boldsymbol{H}_E^{(l)} \quad (10)$$

where $\boldsymbol{M}_E^{(l)}$ are object-aware entity representations Similarly, we obtain entity-aware object representations $\boldsymbol{M}_V^{(l)}$. After that, we fuse visual information from $\boldsymbol{M}_E^{(l)}$ to its associated token nodes:

$$\alpha_{T_i}^{(l)} = Sigmoid(W_{T_1}^{(l)} M_{E_j}^{(l)} + W_{T_2}^{(l)} D_{T_i}^{(l)}) \quad (11)$$

$$M_{T_i}^{(l)} = \alpha_{T_i}^{(l)} \cdot M_{E_j}^{(l)} + (1 - \alpha_{T_i}^{(l)}) \cdot D_{T_i}^{(l)} \quad (12)$$

where $D_{T_i}^{(l)}$ is the constituent token node of $M_{E_j}^{(l)}$. Finally, we feed $M_m^{(l)}, m \in \{T, V\}$ to feed-forward neural networks to obtain $H_m^{(l+1)}$, and then update $H_E^{(l+1)}$: $H_{E_i}^{(l+1)} = Max(\{H_{T_j}^{(l+1)}\}_{j=s_i}^{e_i})$.

After fusing entity-related visual information into corresponding token representations, we apply a CRF layer to conduct sequence labeling and obtain the entity recognition loss $\mathcal{L}_{mner}$:

$$\mathcal{L}_{mner} = -\sum_i \log P(y|\boldsymbol{X}) \quad (13)$$

$$P(y|\mathbf{X}) = \frac{\prod_{i=1}^{|X|} \varphi_i(y_{i-1}, y_i, \mathbf{X}))}{\sum_{y' \in \mathbf{Y}} \prod_{i=1}^{|X|} \varphi_i(y'_{i-1}, y'_i, \mathbf{X})} \quad (14)$$

where $\varphi_i(y_{i-1}, y_i, \boldsymbol{X})$ and $\varphi_i(y'_{i-1}, y'_i, \boldsymbol{X})$ are potential functions. When training, we sum the loss mentioned above as the final loss: $\mathcal{L} = \lambda_1 \mathcal{L}_{sd} + \lambda_2 \mathcal{L}_{mner}$, where λ_1 and λ_2 are hyperparameters.

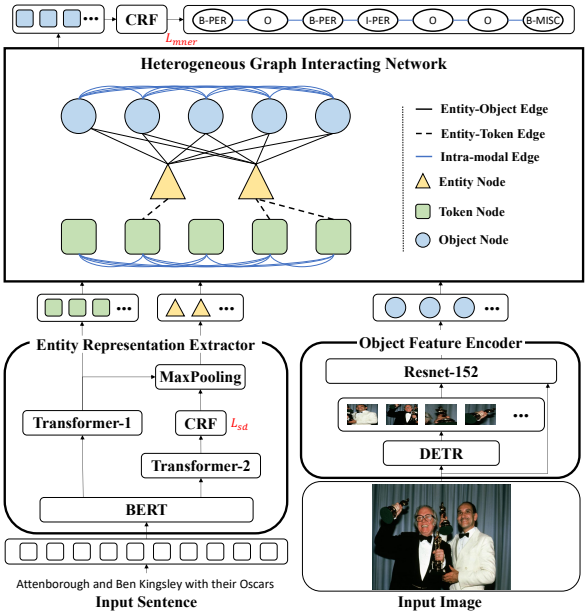

Figure 2: The overall architecture of GEI.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate our method on two public MNER datasets Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018). The datasets contain four different types of entities: Person, Location, Organization, Misc.

Baselines and Metrics. For a thorough comparison, we compare our approach with two groups of baseline models. Firstly, the representative text-based NER approaches: 1) **CNN-BiLSTM-CRF** (Ma and Hovy, 2016), which is a classical text-based neural network for NER with both the word-level and character-level information. 2) **HBiLSTM-CRF** (Lample et al., 2016), which is an improvement of CNN-BiLSTM-CRF, replacing

Visual Feature	Model	Twitter-2015			Twitter-2017		
		Pre.	Rec.	F1	Pre.	Rec.	F1
-	CNN-BiLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37
	HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37
	BERT	68.30	74.61	71.32	82.19	83.72	82.95
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44
Image Region	VG	73.96	67.90	70.80	83.41	80.38	81.87
	ACoA	72.75	68.74	70.69	84.16	80.24	82.15
	UMT	71.67	75.23	73.41	85.28	85.34	85.31
Visual Object	OCSGA	**74.71**	71.21	72.92	-	-	-
	Object-AGBAN	<u>74.13</u>	72.39	73.25	-	-	-
	UMGF*	72.43	74.09	73.25	85.33	84.32	84.82
Image Caption	Captions	68.52	74.61	71.49	86.16	**87.49**	**86.82**
Ours (Visual Object)	**GEI**†	73.39	**75.51**	**74.43**	**87.50**	<u>86.01</u>	<u>86.75</u>
	-Entity-level interaction	72.49	75.11	73.78	<u>87.09</u>	85.42	86.25
	-Span Detection	71.67	74.00	72.87	83.12	85.27	84.18
	+Image Region	72.64	<u>75.40</u>	<u>73.99</u>	86.42	85.27	85.84

Table 1: Performance comparison on the two MNER datasets. Result marked by * is conducted via the released code of Zhang et al. (2021). The boldface and underlined numbers are the best two results in each column. † refers to significant with p-values < 0.05 when comparing unimodal baselines.

the bottom CNN layer with LSTM to build the hierarchical structure. 3) **BERT** (Devlin et al., 2018), which is a competitive baseline for NER with multi-layer bidirectional Transformer encoder and followed by stacking a softmax layer for entity prediction. 4) **BERT-CRF**, a variant of BERT, which replaces the softmax layer with a CRF layer. Secondly, several competitive multimodal approaches for MNER: 5) **VG** (Lu et al., 2018), which utilizes a visual attention and a gate mechanism to expoit implicit information from a whole image to guide word representation learning based on HBiLSTM-CRF. 6) **ACoA** (Zhang et al., 2018), which designs an adaptive co-attention network to learn word-aware visual representations and vision-aware word representations based on CNN-BiLSTM-CRF. 7) **UMT** (Yu et al., 2020), which extends Transformer to multi-modal version and incorporates the auxiliary entity span detection module. 8) **Object-AGBAN** (Zheng et al., 2020), which proposes an adversarial bilinear attention network to capture the correlations of visual objects and textual entities. 9) **OCSGA** (Wu et al., 2020), which combines dense co-attention network (self-attention and guide attention) to model the correlations between visual objects and textual entities. 10) **UMGF** (Zhang et al., 2021) , which proposes a unified multimodal graph fusion approach for MNER and achieves current SOTA on Twitter-2015. 11) **Captions** (Chen et al., 2020), which uses image captions as visual features and achieves current SOTA on Twitter-2017. Following previous works, we take Micro F1-score as the evaluation metric.

Implementation Details. We use the Adam (Kingma and Ba, 2014) optimizer with a learning rate 3e-05. We set the batch size to 16. The number of gnn layer is set to 6, the λ_1 and λ_2 are set to 0.5 and the dropout rate is set to 0.4. The head number of multi-head attention is set to 8. For all experiments, we train and test our model on a Tesla-V100 GPU. We take the average F1 scores of three experiments as our final result. To alleviate the error propagation caused by the gap between training and predicting, we take the scheduled sampling strategy (Bengio et al., 2015). Specifically, when training, GEI gradually switches the span detection results from golden label to the model predictions on its own. From epoch 2 to epoch 6, GEI linearly increases the proportion of predicted span detection results from 0% to 90%.

3.2 Results and Analysis

Table 1 shows the main results of GEI compared with the baseline models on both Twitter-2015 and Twitter-2017. Results show that our proposed framework GEI significantly outperforms UMGF by 1.93% and 1.18% on Twitter-2017 and Twitter-2015, respectively. Further, comparing with Captions, GEI also surpasses 2.94% F1-scores on Twitter-2015 and has a competitive performance on Twitter-2017. Besides, GEI outperforms all baseline models that also use visual objects, which suggests that conducting cross-modal interaction at entity-level can effectively exploit useful visual

information from object features.

Ablation Study. To further investigate the effectiveness of entity-level interaction, we conduct ablation experiments on 3 variants: 1) *-Entity-level Interaction*, removing entity nodes from the multimodal graph and directly interacting object nodes with entity-associated token nodes. 2) *-Span Detection*, further removing the span detection subtask and interacting all token nodes with object nodes. 3) *+Image Region*, replacing visual object features with fixed region features following Yu et al. (2020). From Table 1, we can observe that: 1) Employing entity representations that carry integral entity semantics to capture entity-related visual information is important and contributes +0.65% / +0.50% F1-score. 2) Excluding non-entity tokens from the cross-modal interaction to alleviate the visual interference is essential and improves the performance significantly. 3) Compared with fixed image regions, employing visual objects that have similar semantic granularity with entities is preferable and enhances +0.44% / +0.91% F1-score.

Case Study. Figure 3 shows two representative examples which intuitively demonstrate the effectiveness of our method. 1) For the left example, UMT and UMGF misidentify the PER entity "Leonardo" as MISC, while GEI extracts both entities correctly. It shows that our proposed GEI captures the entity-related visual information effectively via entity-level interaction (i.e., "Leonardo" and people appearing in the image). 2) For the right example, both UMT and UMGF suffer from the interference brought by the image and mislabel non-entity token "HURRY" as a PER entity. However, due to excluding non-entity tokens from the cross-modal interaction, our GEI rids the "HURRY" of the visual noise and makes the prediction correctly. This noticeable phenomenon indicates that our framework alleviates the interference brought by images via entity-level interaction.

Visualization of Entity-level Interaction. To gain an insight into the interaction between entities and visual objects, we visualize the cross-modal attention weights between entity nodes and visual object nodes for the example appearing in Figure 1. As shown in Figure 4, it is obvious that two PER entities "Attenborough" and "Ben Kingsley" have greater weights with two person objects than other visual objects during cross-modal interaction. The same phenomenon exists between the MISC entity "Oscars" and two trophy objects. These findings confirm that our framework can effectively capture entity-related visual information through entity-level cross-modal interaction.

Figure 3: Case study of our proposed GEI, previous SOTA methods UMT and UMGF. The bottom three rows are predicted entities of different approaches.

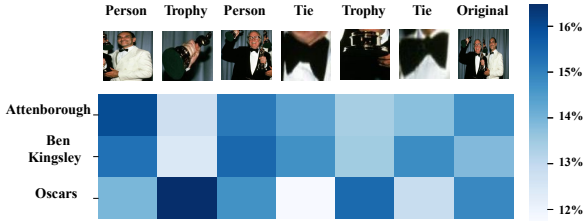

Figure 4: Visualization of cross-modal interaction attention between entity representations and visual objects.

4 Conclusion

In this paper, we propose an heterogeneous Graph-based Entity-level Interacting model (GEI) for MNER. GEI interacts entity representations with visual objects to capture useful entity-related visual information, and excludes non-entity tokens from the interaction to rid them of the visual noise. Experiments on two public MNER datasets demonstrate the effectiveness of our method.

Limitations

MNER methods have gained impressive progress on multimodal social media NER by incorporating complementary visual features. Though helpful in many cases, incorporating images may also bring diverse interference to the task. In this paper, we focus on the interference suffered by non-entity tokens, and alleviate it by excluding non-entity tokens from the cross-modal interaction process. However, existing methods (including our GEI) are still confronted with inevitable interference when the image is irrelevant with the text or contains ironic meaning. How to effectively alleviate such kind of interference remains to be studied in future.

Acknowledgements

We sincerely thank all anonymous reviewers for their valuable comments and suggestions. This work was supported by National Natural Science Foundation of China (61702047).

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28.

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.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2020. Can images help recognize entities? a study of the role of images for multimodal ner. *arXiv preprint arXiv:2010.12712*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

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

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Kentaro Torisawa et al. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 698–707.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1038–1046.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14347–14355.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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.