PromptMNER: Prompt-Based Entity-Related Visual Clue Extraction and Integration for Multimodal Named Entity Recognition

Xuwu Wang[1], Junfeng Tian[2], Min Gui[3], Zhixu Li[1(✉)], Jiabo Ye[4], Ming Yan[2], and Yanghua Xiao[1,5(✉)]

[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China
{xwwang18,zhixuli,shawyh}@fudan.edu.cn
[2] Alibaba DAMO Academy, Hangzhou, China
{tjf141457,ym119608}@alibaba-inc.com
[3] Shopee, Singapore, Singapore
min.gui@shopee.com
[4] East China Normal University, Shanghai, China
jiabo.ye@stu.ecnu.edu.cn
[5] Fudan-Aishu Cognitive Intelligence Joint Research Center, Shanghai, China

Abstract. Multimodal named entity recognition (MNER) is an emerging task that incorporates visual and textual inputs to detect named entities and predicts their corresponding entity types. However, existing MNER methods often fail to capture certain entity-related but text-loosely-related visual clues from the image, which may introduce task-irrelevant noises or even errors. To address this problem, we propose to utilize entity-related prompts for extracting proper visual clues with a pre-trained vision-language model. To better integrate different modalities and address the popular semantic gap problem, we further propose a modality-aware attention mechanism for better cross-modal fusion. Experimental results on two benchmarks show that our MNER approach outperforms the state-of-the-art MNER approaches with a large margin.

Keywords: Named entity recognition · Multimodal learning · Knowledge graph

1 Introduction

Named entity recognition (NER) is an indispensable task for information extraction, knowledge graph construction, and question answering, etc. Recently, posts on social media platforms are becoming increasingly multimodal. It becomes more common that the text information can only be understood with correlated

This work was conducted when Min Gui worked at Alibaba.

A. Bhattacharya et al. (Eds.): DASFAA 2022, LNCS 13247, pp. 297–305, 2022.
https://doi.org/10.1007/978-3-031-00129-1_24

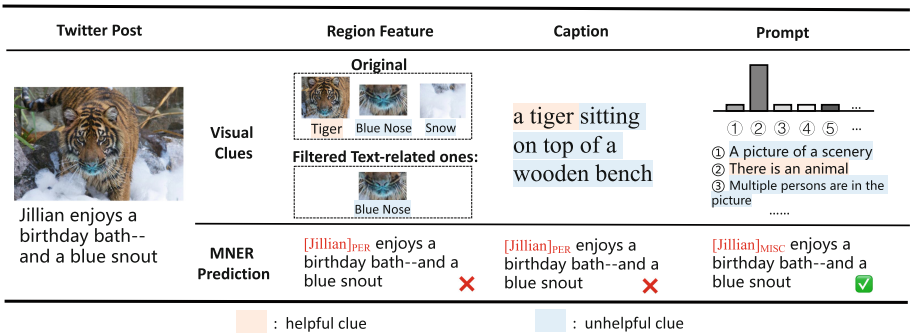

Fig. 1. Comparison of methods used to get the visual clues. The region feature is obtained from an object detector. The caption is generated from an image captioning model. (Color figure online)

images. Therefore, the research of multimodal named entity recognition (MNER) is flourishing, which aims at detecting named entities and classifying them based on a text and the related image.

Previous researches of MNER mainly focus on two issues: extracting visual clues [12] from images and integrating the information of the two modalities. For extracting visual clues, existing works tend to: 1) encode the entire image as a global feature in the form of a visual feature vector or a caption [2,9,11,18], or 2) extract local fine-grained region features from the image [17]. For the integration of multiple modalities, existing efforts focus on: 1) using attention mechanism to fuse the two modalities based on the relevance between them [9,11,18], or 2) transforming the image into a caption or a label set and then directly integrating them with the original sentence [2,14].

However, there are two critical problems that existing works may often neglect. **Problem 1.** It is entity-related visual clues, instead of text-related visual clues or the global image features, that really helps to improve MNER. The potential entities in the sentence are often presented with specific entity names. It is difficult to align these names with the image through image-text relevance (e.g., it is hard to align the entity 'Jillian' to the 'tiger' in the image as shown in Fig. 1). And some text-related visual features may only correspond to non-entity words (e.g., the 'blue nose' in the image corresponds to the word 'blue snout'), which has few benefits to the MNER task. **Problem 2.** Besides, there are semantic gap and structure gap between the two modalities: the visual features of images and the textual features of texts belong to different semantic spaces and have different structures, which hinders the fusion between the two modalities.

In this paper, we propose a novel framework named PROMPTMNER to address the above problems. Firstly, we leverage Prompt Learning [7] to extract entity-related visual clues from images. Prompt learning is an emerging way of releasing the immense power of Pre-trained Language Models (PLMs) to tackle downstream tasks. With the pre-trained vision language models' ability to predict the matching degree between an image and a text, we design multiple

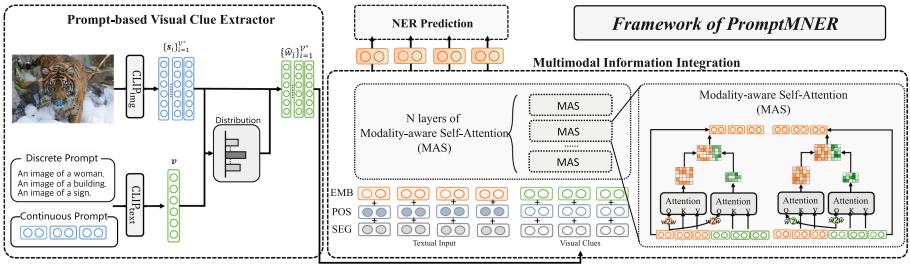

Fig. 2. Framework of the PROMPTMNER

entity-related prompts (see Fig. 1 for example) and use the matching degrees between the image and the prompts to select the entity-related visual clues. Secondly, to bridge the semantic gap in integrating the entity-related visual clues and the textual features, a modality-aware transformer encoder is adopted to achieve better intra-modal homogeneous attention and inter-modal heterogeneous attention. The experiments demonstrate that PROMPTMNER can achieve significant improvement compared with the state-of-the-art (SOTA) methods.

2 Methodology

2.1 Task Definition

Let $T = (w_1, w_2, ..., w_n)$ denote a sentence consisting of multiple words and I denote an image. Given a sentence T and its corresponding image I as input, MNER aims at detecting a set of entities from T and classifying each entity into one of the pre-defined types including person (`PER`), location (`LOC`), organization (`ORG`) and miscellaneous (`MISC`).

2.2 Overview

As shown in Fig. 2, the proposed PROMPTMNER mainly consists of the following components: Firstly, a Prompt-based Visual Clue Extractor (Sect. 2.3) is used to extract entity-related visual clues with a pre-trained vision-language model (VLM) from the input image. Secondly, a Multimodal Information Integration Module (Sect. 2.4) is designed to fuse the extracted visual clues and the input sentence. Finally, to train the model for NER (Sect. 2.5), we adopt the span-based method [3] that enumerates all candidate spans of the sentence and predicts the corresponding entity types.

2.3 Prompt-Based Visual Clue Extractor

Let $\mathbf{v}$ represent the presentation of the input image $\mathbf{I}$. As discussed in Sect. 1, as a form of expression with huge amount of information, $\mathbf{v}$ contains lots of task-independent noises. By designing a set of entity-related prompts, a VLM is able

to predict the relevance between the image and every prompt due to the valuable priors obtained from pre-training. It helps us to extract the entity-related visual clues with corresponding weights, as well as fade out task-irrelevant noises.

Prompt Design. We define each entity-related prompt as a sentence in the form of $P_i = \textit{an image of } [w_i]$, $w_i \in \mathcal{V}^*$, where w_i represents a word or phrase from a entity-related vocabulary $\mathcal{V}^*$. Following recent work of prompt learning [1,4], we design both discrete and continuous entity-related prompts.

Discrete Prompts. For discrete prompts, the words $w_i \in \mathcal{V}^*$ come from a human-readable vocabulary. We propose three methods to find the $\mathcal{V}^*$:

- **Task Labels.** The labels of MNER: `person, location, organization` can be used to build $\mathcal{V}^*$.
- **KB-Retrieved Labels.** However, the above task labels have multiple shortcomings. 1) The limited number of task labels cannot guarantee the high coverage of visual clues. 2) The task labels may be quite different from the frequently used expressions of the VLM's pre-training corpus. It would severely impair VLM's ability to predict the relevance between the image and the prompt. Thus, we incorporate off-the-shelf Related Words[1] to retrieve related words of the task labels to enrich the label set. Related Words is a KB of related words that incorporates multiple KBs including WordNet [10], Concept Net [6], etc. For example, `person` can be expanded with `people, someone, individual, worker, child`, etc.
- **Expert Provided Labels.** Some related words that rely on the complicated relationships such as whole-part and collective-individual are difficult to be retrieved through a KB, because this retrieving procedure often involves complex knowledge such as common sense and cognition. So we also obtain entity-related labels from experts who are familiar with the task as well as VLMs. For example, `person` can be expanded with `player, pants, hat, suit, group of people, team`, etc.

Continuous Prompts. However, finding as many discrete prompts as possible is a non-trivial task, which requires a large amount of time for word tuning as well as expertise about the task and the VLM. So alternatively, we can use continuous vectors as the prompts and optimize them during the training procedure: $P_i = \mathbf{w}_i$, where $\mathbf{w}_i$ is a learnable embedding. This type of prompt is denoted as the continuous prompt, which is able to automatically learn open-set entity-related visual clues.

Visual Clue Extraction. In this paper, we adopt CLIP as the VLM. Given the input image $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$, its embedding is obtained through: $\mathbf{v} = \text{CLIP}_{\text{img}}(\mathbf{I})$. Here $\text{CLIP}_{\text{img}}(\cdot)$ represents the convolutional neural network-based image encoder of CLIP. For each prompt P_i, we obtain its embedding from the CLIP's textual encoder: $\mathbf{s_i} = \text{CLIP}_{\text{text}}(P_i)$.

[1] http://relatedwords.org.

After that, given the image embedding $\mathbf{v}$ and entity-related prompt embeddings $\{\mathbf{s}_i\}_{i=1}^{|\mathcal{V}^*|}$, we are able to get the relevance between the image $\mathbf{I}$ and every prompt P_i:

$$p\left(P_i|\mathbf{I}\right) = \frac{\exp\left(< \mathbf{s_i}, \mathbf{v} > /\tau\right)}{\sum_{j=1}^{|\mathcal{V}^*|} \exp\left(< \mathbf{s_j}, \mathbf{v} > /\tau\right)} \tag{1}$$

where $< \cdot, \cdot >$ represents the cosine similarity between two vectors and τ is a temperature parameter. Both of them are set as the same paradigm of the pre-training of CLIP.

Finally, the weighted entity-related visual clue $\hat{\mathbf{w}}_i$ is represented as the weighted embedding of the prompt: $\hat{\mathbf{w}}_i = p\left(P_i|\mathbf{I}\right) \times \mathbf{s}_i$.

2.4 Multimodal Information Integration

In this subsection, we introduce how to integrate the visual clues $\{\hat{\mathbf{w}}_i\}_{i=1}^{|\mathcal{V}^*|}$ and the sentence $T = (w_1, ..., w_N)$ to capture the interaction between the two modalities.

Embedding Layer. Initially, for each word w_i of the input sentence $\{w_i\}_{i=1}^{N}$, it is mapped to a distributed vector from the embedding layer of a pre-trained language model (LM): $\mathbf{w}_i = \mathrm{LM}_{\mathrm{WordEmb}}(w_i)$. Given the visual clues $\{\hat{\mathbf{w}}_i\}_{i=1}^{\mathcal{V}^*}$ as well as $\{\mathbf{w}_i\}_{i=1}^{N}$, we then concatenate the corresponding sequences and add the special tokens of `[CLS]` and `[SEP]`:

$$\texttt{[CLS]}, \mathbf{w}_1, ..., \mathbf{w}_N, \texttt{[SEP]}, \hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_{|\mathcal{V}^*|}, \texttt{[SEP]}. \tag{2}$$

For each word $\mathbf{w}_i$ and $\hat{\mathbf{w}}_i$, the position embedding and segmentation embedding are also added to it followed by a LayerNorm layer to capture the positional information and the type information.

Modality-aware Self-Attention (MAS). Then the embeddings sequence of textual and visual tokens are fed into multiple layers of self-attention. Note that the visual clue tokens and textual tokens are heterogeneous, which will influence the information fusion achieved through self-attention. Yamada [15] noticed that it is beneficial to take the target token type into consideration when computing the attention score. Inspired by this discovery, we propose to take the modality type into consideration when computing the attention query. Let $\mathbf{x}_i$ represent the input token that corresponds to $\mathbf{w}_i$ or $\hat{\mathbf{w}}_i$. Its attention weight corresponding to the j^{th} token in the sequence is calculated as:

$$\alpha_{ij} = \mathrm{softmax}\left(\frac{\left(\mathbf{K}\mathbf{x}_j\right)^{\mathrm{T}}\left(\mathbf{Q}_{Q2V}\mathbf{x}_i\right)}{\sqrt{N + |\mathcal{V}^*|}}\right) \tag{3}$$

where $\mathbf{Q}_{Q2V}$, $\mathbf{K}$ represent the matrices used to projecting queries and keys. $\mathbf{Q}_{Q2V}$ is designed to be modality-aware: it may be $\mathbf{Q}_{\mathbf{w2w}}$, $\mathbf{Q}_{\mathbf{w2\hat{w}}}$, $\mathbf{Q}_{\mathbf{\hat{w}2w}}$, or

$\mathbf{Q}_{\hat{\mathbf{w}}2\hat{\mathbf{w}}}$ depending on the modality of the query and the value. The output is then calculated as:

$$\mathbf{x}_i := \sum_{j=1}^{N+|\mathcal{V}^*|} \alpha_{ij}\mathbf{x}_j \tag{4}$$

We apply K layers of MAS to embed the input sequence and take the hidden states of the textual words $\{\mathbf{h}_i\}_{i=1}^{N}$ from the last layer for the prediction.

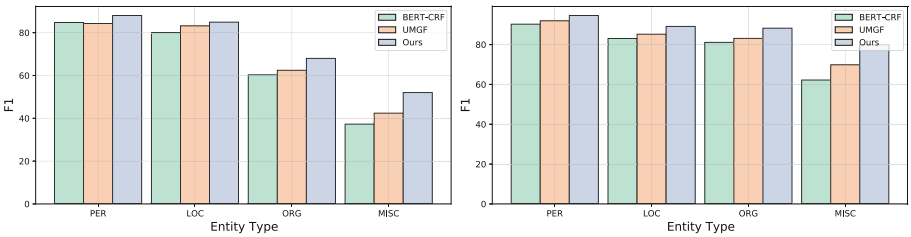

Fig. 3. Model performance in terms of different entity types. Our results come from PROMPTMNER Best.

2.5 Prediction

Given the textual word representations: $\{\mathbf{h}_i\}_{i=1}^{N}$, which have integrated the information of the visual clues, we then reformulate NER as the task of identifying start and end indices of an entity span as well as assigning a category label to the span [5]. So we enumerate all possible spans in the sentence $\{w_1, ..., w_N\}$ and classify whether it is an entity and predict the entity type. For each text span $\{w_i, ..., w_j\}$, we concatenate the embeddings of the first and the last tokens and then feed them into a MLP to predict its entity type:

$$l_c = \text{MLP}\left([\mathbf{h}_i||\mathbf{h}_j]\right) \tag{5}$$

where l_c represents the entity type from {`person, location, organization, misc, not_entity`}.

Our objective is assigning a correct entity type to each enumerated span. So the loss function of MNER is formulated as the softmax cross-entropy loss:

$$p\left(l_c\right) = \frac{\exp(l_c)}{\sum_{\hat{c}=1}^{C}\exp(l_{\hat{c}})}, \quad \mathcal{L}_{\text{MNER}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i_c}\log p(l_c) \tag{6}$$

where C represents the number of the entity types. y_c is the correct span label.

3 Experiment

3.1 Experimental Setups

Datasets. We test on two benchmarks: Twitter-2015 [18] and Twitter-2017 [9].

Baselines. We compare with a wide range of baselines including both textual and multimodal baselines: CNN-BiLSTM-CRF, BERT-CRF, BERT-BiLSTM-CRF, ACoA [18], IAIK [17], UMT [16], RIVA [13], RpBERT [12], UMGF [17].

Implementation Details. We conduct all the experiments on 8 NVIDIA V100 GPUs using Pytorch 1.7. We use RoBERTa [8] as the LM and CLIP as the VLM. We set the number of self-attention layers K as 12. We set AdamW optimizer with the learning rate of 1e−4 and a warmup linear scheduler to control the learning rate. The batch size is set as 32 and the dropout is set as 0.5.

Table 1. Performance of different methods of the MNER task. 'T' and 'T+V' represent textual methods and multimodal methods respectively. 'CPrompt' represents continuous prompts. PROMPTMNER Best is achieved with 100 CPrompts

	Methods	Twitter-2015			Twitter-2017		
		Prec.	Recall	F1	Prec.	Recall	F1
T	CNN-BiLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44
	BERT-BiLSTM-CRF	–	–	71.60	–	–	–
T+V	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
	RIVA	–	–	73.80	–	–	–
	IAIK	74.78	71.82	73.27	–	–	–
	RpBERT	–	–	74.80	–	–	85.51
	UMGF	74.49	75.21	74.85	86.54	84.50	85.51
	PROMPTMNER Best	78.03	**79.17**	**78.60**	**89.93**	90.60	**90.27**
	PROMPTMNER w/ Task Labels	77.89	77.44	77.66	88.27	90.82	89.53
	PROMPTMNER w/ KB Prompts	78.15	78.23	78.19	89.78	89.71	89.74
	PROMPTMNER w/ Expert Prompts	77.96	78.65	78.30	89.59	90.45	90.02
	PROMPTMNER w/ 30 CPrompts	77.84	78.90	78.37	89.32	**90.97**	90.14
	PROMPTMNER w/ 50 CPrompts	78.33	78.75	78.54	89.67	90.60	90.13
	PROMPTMNER w/o MAS	**78.47**	77.62	78.04	89.12	90.07	89.59

3.2 Experimental Results

We report the performance of different methods in Table 1. We can see that: 1) Our method greatly outperforms previous methods. So the prompt-based entity-related visual clues can effectively extract helpful information from the image. 2) Both KB-retrieved prompts and expert-provided prompts achieve competitive

performance. But the continuous prompts help the model to achieve the best results. And more continuous prompts lead to better model performance.

We also present the performance of each entity category in Fig. 3. In all entity categories, our method shows superior performance compared with BERT-CRF and UMGF, especially in the categories of `ORG` and `MISC`. We speculate that the reason may be that existing methods are difficult to extract visual clues helpful to these two types of entities.

4 Conclusion

In this paper, we propose a novel framework PROMPTMNER to extract and integrate entity-related visual clues for multimodal named entity recognition. The model includes a novel prompt-based visual clue extractor to obtain useful task-related image features and a new multimodal integration module to fuse the visual features with the textual features. Experiments show the superiority of our method compared with previous methods on two MNER datasets.

Acknowledgement. This research was supported by the National Key Research and Development Project (No. 2020AAA0109302), National Natural Science Foundation of China (No. 62072323), Shanghai Science and Technology Innovation Action Plan (No. 19511120400), Shanghai Municipal Science an Technology Major Project (No. 2021SHZDZX0103) and Alibaba Research Intern Program.

References

1. Chen, D., Li, Z., Gu, B., Chen, Z.: Multimodal named entity recognition with image attributes and image knowledge. In: Jensen, C.S., et al. (eds.) DASFAA 2021. LNCS, vol. 12682, pp. 186–201. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73197-7_12
2. Chen, S., Aguilar, G., et al.: Can images help recognize entities? A study of the role of images for multimodal NER (2021)
3. Fu, J., Huang, X., Liu, P.: SpanNER: Named entity re-/recognition as span prediction. arXiv preprint arXiv:2106.00641 (2021)
4. Li, X.L., Liang, P.: Prefix-Tuning: optimizing continuous prompts for generation. In: Proceedings of ACL, pp. 4582–4597 (2021)
5. Liu, C., Fan, H., Liu, J.: Span-based nested named entity recognition with pretrained language model. In: Jensen, C.S., et al. (eds.) DASFAA 2021. LNCS, vol. 12682, pp. 620–628. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73197-7_42
6. Liu, H., Singh, P.: ConceptNet-a practical commonsense reasoning tool-kit. BT Technol. J. **22**(4), 211–226 (2004). https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d
7. Liu, P., Yuan, W., et al.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. arXiv:2107.13586 (2021)
8. Liu, Y., Ott, M., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692 (2019)

9. Lu, D., Neves, L., et al.: Visual attention model for name tagging in multimodal social media. In: Proceedings of ACL, pp. 1990–1999 (2018)
10. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
11. Moon, S., Neves, L., et al.: Multimodal named entity recognition for short social media posts. In: Proceedings of NAACL, pp. 852–860 (2018)
12. Sun, L., Wang, J., et al.: RpBERT: a text-image relation propagation-based BERT model for multimodal NER. In: Proceedings of AAAI, vol. 35 (2021)
13. Sun, L., Wang, J., et al.: RIVA: a pre-trained tweet multimodal model based on text-image relation for multimodal NER. In: COLING, pp. 1852–1862 (2020)
14. Wu, Z., Zheng, C., et al.: Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In: MM (2020)
15. Yamada, I., Asai, A., et al.: LUKE: deep contextualized entity representations with entity-aware self-attention. arXiv preprint arXiv:2010.01057 (2020)
16. Yu, J., Jiang, J., et al.: Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In: Proceedings of ACL (2020)
17. Zhang, D., Wei, S., et al.: Multi-modal graph fusion for named entity recognition with targeted visual guidance. In: Proceedings of AAAI, pp. 14347–14355 (2021)
18. Zhang, Q., Fu, J., et al.: Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of AAAI (2018)