SCANNER: Knowledge-Enhanced Approach for Robust Multi-modal Named Entity Recognition of Unseen Entities

Anonymous ACL submission

Abstract

Recent advances in named entity recognition (NER) have pushed the boundary of the task to incorporate visual signals, leading to many variants, including multi-modal NER (MNER) or grounded MNER (GMNER). A key challenge to these tasks is that the model should be able to generalize to the entities unseen during the training, and should be able to handle the training samples with noisy annotations. To address this obstacle, we propose SCANNER (Span CANdidate detection and recognition for NER), a model capable of effectively handling all three NER variants. SCANNER is a two-stage structure; we extract entity candidates in the first stage and use it as a query to get knowledge, effectively pulling knowledge from various sources. We can boost our performance by utilizing this entity-centric extracted knowledge to address unseen entities. Furthermore, to tackle the challenges arising from noisy annotations in NER datasets, we introduce a novel self-distillation method, enhancing the robustness and accuracy of our model in processing training data with inherent uncertainties. Our approach demonstrates competitive performance on the NER benchmark and surpasses existing methods on both MNER and GMNER benchmarks. Further analysis shows that the proposed distillation and knowledge utilization methods improve the performance of our model on various benchmarks.

1 Introduction

Named entity recognition (NER) is a fundamental task in natural language processing to identify textual spans that correspond to named entities in the given text, and classify them into pre-defined categories, such as persons, locations, and organizations (Li et al., 2020). The extracted information can be utilized for various downstream tasks, including entity linking and relation extraction.

The rapid growth of the amount of multi-modal contents on social media platforms has given rise to the multi-modal variants of NER. The most prominent example is multi-modal NER (MNER; Zhang et al. (2018)) , which extends traditional NER to identifying named entities in the text based on additional image input paired with the text (Fig. 1b). Another recent example is the grounded MNER (GMNER; Yu et al. (2023)); here, one additionally aims to predict the bounding boxes of named entities appearing in the given image (Fig. 1c).

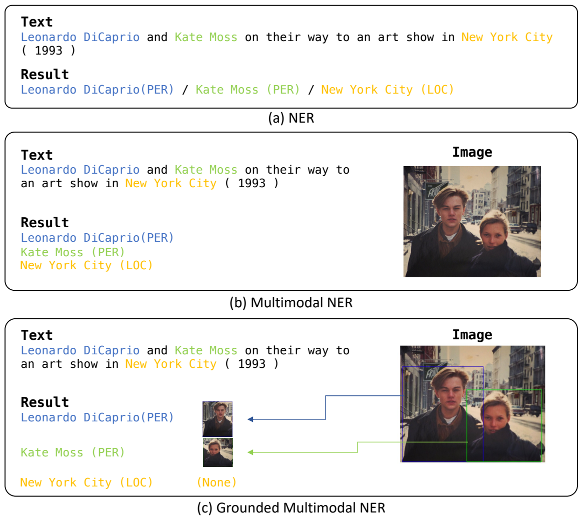

Figure 1: Illustrations of NER, MNER, and GMNER tasks. The NER task aims to identify named entities from the given text. MNER extends this task to utilize additional image informations. GMNER additionally requires the model to predict entity bounding boxes in the given image, if they are present.

A major challenge in NER, MNER, and GMNER tasks is the presence of unseen entities in the test datasets, which are not found in the training datasets. Traditional models often struggle with low performance on these unseen entities (see Table 1). To tackle this problem effectively, it is important to use knowledge about unseen entities in a way that boosts ability of the model to generalize and perform well across different types of data. In this paper, we introduce SCANNER, which stands

Datasets	Methods	Seen entities	Unseen entities
CoNLL2003	BERT-base	93.78	80.90
	Ours (w/o Knowledge)	96.29	89.68
Twitter-2015	BERT-base	79.81	57.81
	Ours (w/o Knowledge)	87.18	73.84
Twitter-2017	BERT-base	93.81	67.76
	Ours (w/o Knowledge)	95.68	82.96

Table 1: A comparison of test F1 scores for the named entities that have appeared at least once in the training dataset, versus the entities that have not appeared.

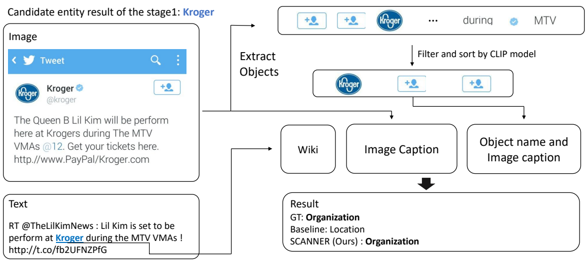

Figure 2: ‘Kroger’ is an unseen entity that is hard to recognize as an Organization or Location. By our knowledge base model, it brings to successful prediction.

for Span CANdidate detection and recognition for Named Entity Recognition. Our approach is designed to effectively use knowledge about unseen entities, addressing NER, MNER, and GMNER tasks with improved robustness. SCANNER adopts a two-stage structure, comprising a span candidate detection module and entity recognition module. The span candidate detection module identifies named entity candidates within sentences. Following this, the entity recognition module uses these candidates as queries to extract relevant knowledge from various sources, effectively recognizing the class of the entity candidate. To our knowledge, this entity-centric knowledge extract method represents the first attempt in this field. As illustrated in Fig. 2, we were able to accurately identify ‘Kroger’ as an ‘organization’ by utilizing object knowledge. SCANNER effectively gathers and uses knowledge from various sources, boosting its performance in the challenging NER, MNER, and GMNER benchmarks. Notably, the GMNER challenge involves the intricate process of identifying entities and determining their bounding boxes within images. The architecture of SCANNER, leveraging its comprehensive knowledge, is effective in addressing the GMNER task. The effectiveness of SCANNER in the GMNER task is highlighted by establishing a new baseline that is over 12 points higher than the previous standard, as measured by the F1 score. Additionally, we introduce the novel self-distillation

Text	Dataset
[The [Oval]$_{\text{ORG}}$]$_{\text{ORG}}$	CoNLL2003
[The [World Cup]$_{\text{MISC}}$]$_{\text{MISC}}$	Twitter2015
[Taste of [Toronto]$_{\text{LOC}}$]$_{\text{MISC}}$	Twitter2015
[Mrs. [Brozik]$_{\text{PER}}$]$_{\text{PER}}$	Twitter2017
[[Robert Downey]$_{\text{PER}}$ Jr]$_{\text{PER}}$	Twitter2017

Table 2: Examples of gold annotation and potential alternatives. The gold annotations are marked in blue [*], whereas the alternative annotations are in red [*].

method, called as Trust Your Teacher. The NER task faces challenges with noisy annotations (Wang et al., 2019; Zhu and Li, 2022), particularly at entity boundaries where exact span matching is crucial and ambiguity often leads to increased noise (see Table 2). Our distillation method, which softly utilizes both the prediction of the teacher model and ground truth (GT) logit, addresses the challenges of noisy annotations.

Our approach demonstrates competitive performance on NER and surpasses existing methods on both MNER and GMNER. Further analysis shows that the proposed distillation and knowledge utilization methods improve the performance of our model on various benchmarks.

The contributions of SCANNER are summarized in three key aspects:

- We propose a new distillation method that softly blends the predictions of the teacher model with ground truth annotations to enhance data quality and model training.
- We develop SCANNER, a two-stage structured model that effectively utilizes knowledge to improve performance, particularly in recognizing unseen entities.
- The SCANNER model shows competitive performance in NER benchmarks and demonstrates higher performance than existing methods in MNER and GMNER benchmarks.

2 Related work

Prior works on MNER typically operates by first extracting the NER-related features from the image, and then combining these features with text features to recognize name entities. Roughly, existing works fall into two categories according to how they extract image features.

Textual features. Several works extract the textual metadata from the given image and utilize them as features for the subsequent NER task (Wang et al., 2022b,a; Li et al., 2023b). For instance, ITA (Wang et al., 2022b) extracts object tags, image captions, and OCR results from the given image. Similarly, Li et al. (2023b); Chen and Feng (2023) also extracts image captions, but additionally utilizes large language model as an implicit knowledge source to further refine the features. MoRe (Wang et al., 2022a) takes a slightly different approach, using an image-based retrieval system to retrieve textual descriptions of the closest images in the database.

Visual encoders. Another line of work attempts to extract the image features using a visual encoder, such as pre-trained ResNets, ViTs, or CLIP vision encoder (Wang et al., 2022d; Zhang et al., 2023; Chen et al., 2023). The extracted features are then combined with the text features extracted from a separate text encoder, which often involves additional alignment via cross-modal attention (Chen et al., 2022; Lu et al., 2022; Wang et al., 2022d; Zhang et al., 2023; Chen et al., 2023). Notably, PromptMNER (Wang et al., 2022c) calculates the similarity between visual features and various text prompts to extract visual cues that are loosely related to the input text.

In this paper, we take a different path and extract the image features *conditioned* on the information extracted from the given text. Up to our knowledge, it is the first such attempt in the context of MNER.

In addition, a new task has been introduced, which not only incorporates image inputs but also actively addresses the task of grounding entity locations within images (Yu et al., 2023).

3 Method

In this section, we first introduce the architecture of the proposed method, which comprises the span candidate detection module and the named entity recognition module (Sec. 3.1). Then, we describe the named entity recognition module, which performs entity recognition and visual grounding in the image for each entity candidate (Sec. 3.2). Finally, we explain a novel distillation method, named Trust Your Teacher, which is designed to robustly train our model even in the presence of noisy dataset annotations (Sec. 3.3).

3.1 SCANNER Architecture

The primary focus of this paper is to perform MNER using both knowledge extracted from within images and external knowledge, even for entities not encountered during training. To achieve this, as illustrated in Fig. 3, we propose a two-stage architecture, known for its efficiency in extracting and searching for knowledge from various sources. In the first stage, we extract named entity candidates, and in the second stage, we efficiently search and extract only knowledge relevant to these candidates. This acquired knowledge is then utilized for entity recognition.

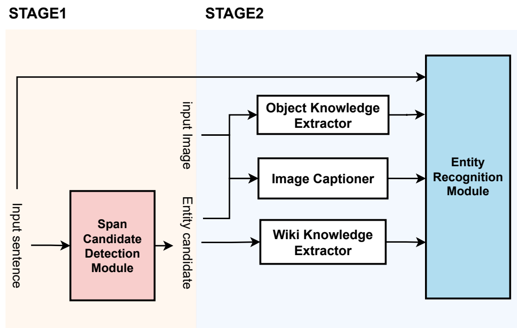

Figure 3: The overall architecture of the proposed SCANNER method. The two-stage structure allows for efficient extraction and utilization of knowledge, as knowledge is extracted only for those entity candidates that were filtered through in stage 1.

Stage 1: Span Candidate Detection Module. In the first stage of SCANNER, the transformer encoder (Liu et al., 2019) is employed to detect entity candidates from the input text. During this phase, we utilizes BIO (Beginning, Inside, Outside) tagging to classify each token in the input text, determining whether it corresponds to the beginning, inside, or outside of an entity span. The classification process is guided by cross-entropy loss.

Stage 2: Entity Recognition Module. In Stage 2, SCANNER performs named entity recognition and visual grounding for each entity candidate detected in Stage 1. It utilizes each entity candidate as a query to extract and leverage the necessary knowledge for the tasks. During this process, SCANNER efficiently searches and extracts knowledge by focusing on the initially detected entity candidates rather than the entire input text. SCANNER utilizes both internal (image-based) and external (e.g., Wikipedia) knowledge sources to perform MNER on unseen entities, not encountered in training. Detailed information about these modules will be provided in Section 3.2.

3.2 Entity Recognition Module

For each entity candidate identified by the span candidate detection module, the entity recogni-

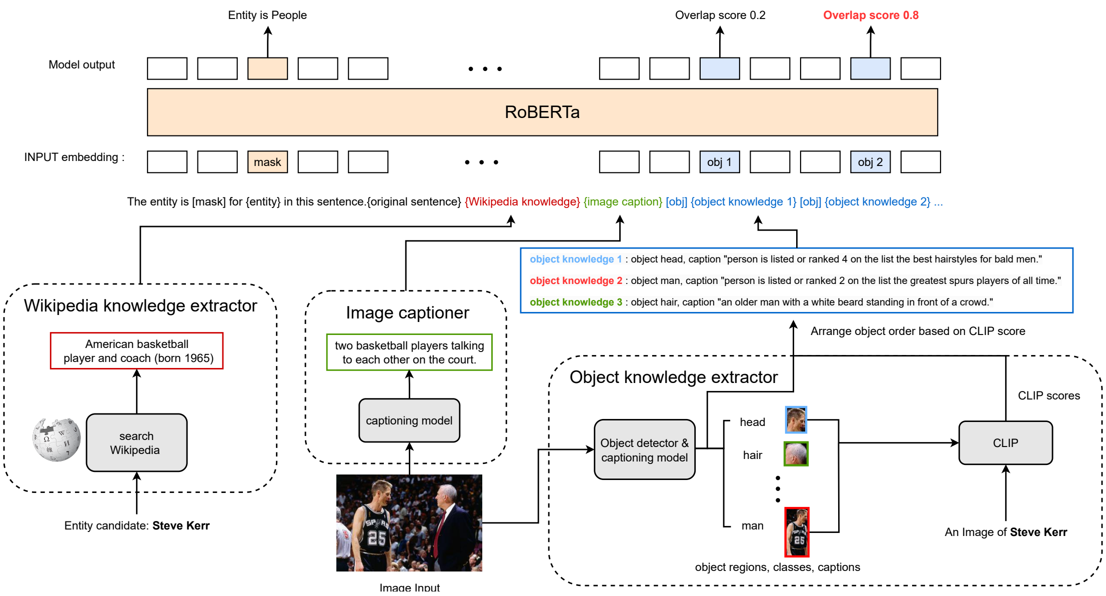

Figure 4: An illustration of the entity recognition module (stage 2). Based on the entity candidates (extracted in stage 1), SCANNER utilizes various knowledge sources such as Wikipedia, image captioner, and object knowledge extractor. The knowledge collected from these sources are then processed by RoBERTa to give the final prediction.

tion module processes a text prompt that includes both the entity candidate and associated knowledge. This knowledge, extracted from images and external knowledge sources, allows for performing MNER on unseen entities that were not encountered during training. Our methodology involves extracting this knowledge from a variety of sources, utilizing the identified entity candidates as the basis for the extraction process. Then, this module classifies the class of each entity candidate and performs grounding to determine which object in the image corresponds to the entity. A detailed illustration is shown in Fig. 4.

3.2.1 Prompt construction with knowledge

The entity recognition module extracts and utilizes useful knowledge from various sources when constructing the text prompt corresponding to the input. The knowledge applied for constructing text prompts in our method includes the following.

Wikipedia knowledge. Initially, information is searched using the entity candidate as a query in external knowledge source, which is Wikipedia. This information can be valuable for classifying the type of entity for each candidate and, moreover, enables the model to classify unseen entities that were not encountered during training. As illustrated in Fig. 4, for entity candidates like 'Steve Kerr', it enhances entity recognition performance by providing valuable information for classification as an American basketball player and coach.

Image caption. To effectively utilize visual information, image captioning results are also used. We use the BLIP-2 (Li et al., 2023a) to extract synthetic captions for the whole image.

Object knowledge. In addition to global information about the image, object-level information is also beneficial for entity recognition. To achieve this, results obtained from the object detector are employed as knowledge. Initially, object classes are converted into text format and used as knowledge. Then, synthetic captions for each object region are also utilized in conjunction with class names. This information is structured as details corresponding to each object, along with a special token denoted as [obj], as shown in Fig. 4. Additionally, during this process, the visual-language similarity between each object and entity candidate is calculated, and objects are arranged in order of high similarity, which is then included in the text prompt. One of the problems with existing methods for the MNER task is that the model sometimes references objects in the image that are irrelevant to the entity, leading to incorrect recognition. By arranging the object details in the text prompt according to the visual-language similarity order with the entity, our model can focus more on the object regions that are highly related to the entity. In this paper, CLIP (Radford et al., 2021) is employed for visual-language similarity, specifically calculating the similarity between the text representation of the entity candidate and the visual representation of each Region of Interest (RoI).

All such knowledge mentioned above is converted into a textual format and integrated with the text prompt for entity recognition and visual grounding.

The text prompt, structured to include entity candidates, the entire input text sentence, and extracted knowledge, is presented as *"The entity is* [*mask*] *for {entity} in this sentence. {original sentence} {Wikipedia} {image caption}* [*obj*] *{object 1}* [*obj*] *{object 2} ..."*.

3.2.2 Encoder and Objective

The prompts constructed for each entity candidate are input into a transformer encoder model (Liu et al., 2019). For entity recognition, the output token representation of the [mask] token in the text prompt x_i for the i-th entity candidate is fed into a linear layer to predict the probability distribution $\hat{y}_i$. Given the ground truth y, the objective function is to minimize the cross-entropy loss between the predicted entity class distribution and the ground truth logit:

$$\mathcal{L}_c = -\sum_{i=1}^{N} y_i \log \hat{y}_i, \tag{1}$$

where N is the total number of the entity candidates.

Additionally, the visual grounding is performed by feeding the output token representation of the j-th [obj] token from the text prompt x_i into a linear layer. This is followed by a sigmoid function, which aids in predicting the overlap score $\hat{o}_{ij}$ between the ground truth image region grounding entity candidate i and object j. The objective function of visual grounding is calculated based on the binary cross-entropy loss between the overlap score and the ground truth Intersection over Union (IoU):

$$\begin{aligned}\mathcal{L}_g = -\sum_{i=1}^{N} \sum_{j} & o_{ij} \log \hat{o}_{ij} \\ & + (1 - o_{ij}) \log(1 - \hat{o}_{ij}),\end{aligned} \tag{2}$$

where o_{ij} is the ground truth IoU between the ground truth image region of the entity i and object region j.

In training stage, we combine two losses as the final loss of our model:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_g, \tag{3}$$

where λ is the weighting coefficient, we set λ to 1 for the GMNER task and to 0 for the NER and MNER tasks in this paper.

3.3 Trust Your Teacher

We introduce the novel self-distillation method, called as Trust Your Teacher (TYT). Our distillation method, which softly utilizes both the prediction of the teacher model and ground truth (GT) logit, addresses the challenges of noisy annotations. First, we train the teacher model using equation 1, and then train the final student model using both the predictions of the teacher model and the ground truth labels. The most significant feature of our proposed method is that it assesses the reliability of each sample by utilizing the prediction of the teacher model to determine if it is trustworthy or noisy. Based on this assessment, the method sets the weights between the model prediction and the gt label, which are then reflected in the loss calculation. The objective of the our proposed distillation method composes a cross-entropy loss with ground truth and Kullback-Leibler Divergence (KLD) loss with teacher predictions:

$$\begin{aligned}\mathcal{L}_{TYT} = & \sum_{i} a_i \, \mathcal{L}_{CE}\big(y_i, S(x_i, \theta_S)\big) \\ & + (1 - a_i)\mathcal{L}_{KLD}\big(S(x_i, \theta_S), T(x_i, \theta_T)\big),\end{aligned} \tag{4}$$

where x_i is the input sample, θ_S and θ_T are the model parameters of the student and teacher, S and T are the prediction distributions of the student and teacher and a_i is a balancing factor proposed in this paper. In detail, a_i determines whether to trust the teacher model prediction or the ground truth, and it represents the prediction score of the teacher model for the ground truth class index, which is $T(x_i, \theta_T)[y_i]$. This implies that since the teacher model is well-trained, if the score for the ground truth class is high, then the sample is considered reliable and more weight is given to the cross-entropy with the ground truth label. Conversely, if the score is low, the sample is assumed to be an unreliable, noisy sample, and more weight is placed on the KLD loss with the prediction of the teacher model, rather than the ground truth label.

To demonstrate the significant impact of our TYT approach, we have carried out some experiments. Fig. 5 illustrates our experiments on a text classification task in MNLI dataset. We extracted about 30% of the train set for experimental efficiency and intentionally added label noise at rates

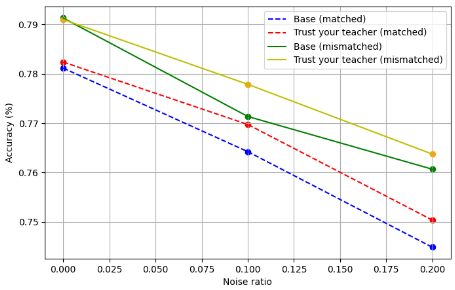

Figure 5: Experiments of text classification task in MNLI datasets. ‘matched’ is in-domain, and ‘mismatched‘ is out-domain.

Methods	Twitter-2015			Twitter-2017		
	Pre.	Rec.	F1.	Pre.	Rec.	F1
Base	83.28	87.68	85.43	**90.96**	93.23	92.08
Half	83.36	87.69	85.47	90.31	92.92	91.60
Full	**83.63**	87.72	85.63	90.53	92.95	91.72
TYT	83.59	**87.98**	**85.73**	90.94	**93.28**	**92.09**

Table 3: Ablation study on the MNER dataset in first stage. ‘Half’ is when a_i is 0.5 and ‘Full’ is 0. In ‘TYT’, a_i is adjusted through the trust your teacher method.

of 10% and 20% to this subset. We then compared the performance of the model trained with our TYT method on the train set with added label noise, against the baseline that does not use distillation. Fig. 5 indicates that using TYT demonstrates relatively robust performance under moderate noise conditions. Additionally, we compare our method with the conventional soft distillation methods that do not dynamically vary the a_i parameter in the entity detection task, stage 1 of MNER. Table 3 shows that our method has better performance on MNER benchmarks, and adaptively varying the a_i is more effective than keeping it fixed.

We apply the TYT to both stages 1 and 2. But in NER, we only use it in stage 1. The loss from the TYT is applied only to the classification loss and not to the loss for visual grounding.

4 Experiment

4.1 Dataset

Our methodology’s efficacy was assessed using widely used datasets for each task. We utilize CoNLL2003 (Tjong Kim Sang and De Meulder, 2003) for NER, Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018) for MNER, and Twitter-GMNER (Yu et al., 2023) for GMNER. Details are in appendix B.

4.2 Experimental Setups

Evaluation metrics. To evaluate our method, we use Entity-wise F1, precision, and recall scores for NER and MNER tasks. For the GMNER task, there is an additional evaluation of the visual grounding. For instances, where the visual grounding is ungroundable, a prediction is correct if it is classified as ‘None.’ For others, correctness hinges on the IoU metric. A prediction is considered correct if the IoU score between the predicted visual region and the ground truth bounding boxes exceeds a threshold of 0.5. We use F1, precision, and recall scores, which are calculated based on the aggregate correctness across entity, type, and visual region predictions. Our primary focus is on the F1 score in line with numerous preceding studies.

Implementation details. Following most recent works, we implement our model utilizing RoBERTa-large in NER, XLM-RoBERTa-large (Conneau et al., 2020) for MNER, GMNER both in stage 1 and stage 2. For the object detector, we use VinVL (Zhang et al., 2021b) following the settings with ITA (Wang et al., 2022b). To address the requirements of visual-language similarity and image caption, we use each of them CLIP [1] and BLIP-2 [2] models respectively. Detailed hyper-parameter settings are shown in appendix A. All experiments were done on a single GeForce RTX 4090 GPU or NVIDIA H100 GPU, and we report the average score from 5 runs with different random seeds for each setting.

Also we applied several minor methods to enhance performance. In the second stage, we incorporated a ‘non-entity’ label to account for instances where the model erroneously predicts entity candidates not present in the dataset. That allowed for more accurate handling of such cases. We augmented it with non-entity data by dividing the training set into four folds in stage 1 and validating each fold. Secondly, we employed adversarial weight perturbation (AWP) (Wu et al., 2020) in stage 1 , which enhances the robustness and generalization capabilities of the model. We initiated AWP from an intermediate stage of our training process.

4.3 Experimental results in various NER tasks

Experimental results in NER. To evaluate the effectiveness of our approach in NER, we primarily compared our model against the existing methods

[1] openai/clip-vit-large-patch14
[2] Salesforce/blip2-opt-2.7b

Methods	Twitter-2015							Twitter-2017						
	Single Type(F1)				Overall			Single Type(F1)				Overall		
	PER	LOC	ORG	OTH.	Pre.	Rec.	F1	PER	LOC	ORG	OTH.	Pre.	Rec.	F1
Text														
BERT-CRF†	85.37	81.82	63.26	44.13	75.56	73.88	74.71	90.66	84.89	83.71	66.86	86.10	83.85	84.96
BERT-SPAN† (Yamada et al., 2020)	85.35	81.88	62.06	43.23	75.52	73.83	74.76	90.84	85.55	81.99	69.77	85.68	84.60	85.14
RoBERTa-SPAN† (Yamada et al., 2020)	87.20	83.58	66.33	50.66	77.48	77.43	77.45	94.27	86.23	87.22	74.94	88.71	89.44	89.06
Text+Image														
UMT (Yu et al., 2020)	85.24	81.58	63.03	39.45	71.67	75.23	73.41	91.56	84.73	82.24	70.10	85.28	85.34	85.31
UMGF (Zhang et al., 2021a)	84.26	83.17	62.45	42.42	74.49	75.21	74.85	91.92	85.22	83.13	69.83	86.54	84.50	85.51
MNER-QG (Jia et al., 2023)	85.68	81.42	63.62	41.53	77.76	72.31	74.94	93.17	86.02	84.64	71.83	88.57	85.96	87.25
R-GCN (Zhao et al., 2022)	86.36	82.08	60.78	41.56	73.95	76.18	75.00	92.86	86.10	84.05	72.38	86.72	87.53	87.11
ITA (Wang et al., 2022b)	-	-	-	-	-	-	78.03	-	-	-	-	-	-	89.75
PromptMNER (Wang et al., 2022c)	-	-	-	-	78.03	79.17	78.60	-	-	-	-	89.93	90.60	90.27
CAT-MNER (Wang et al., 2022d)	88.04	84.70	68.04	52.33	78.75	78.69	78.72	94.61	88.40	88.14	80.50	90.27	90.67	90.47
MoRe (Wang et al., 2022a)	-	-	-	-	-	-	79.21	-	-	-	-	-	-	90.67
PGIM ‡ (Li et al., 2023b)	88.34	84.22	70.15	52.34	79.21	**79.45**	79.33	96.46	89.89	89.03	79.62	**90.86**	**92.01**	**91.43**
SCANNER(Ours)	88.24	85.16	69.86	52.23	**79.72**	79.03	**79.38**	95.18	88.52	88.45	79.71	90.40	90.67	90.54
	± 0.27	± 0.22	± 0.31	± 1.39	± 0.56	± 0.64	± 0.14	± 0.23	± 0.26	± 0.66	± 2.98	± 0.19	± 0.53	± 0.32

Table 4: Experiment results on the Twitter-15 and Twitter-17. The results for methods marked with † are from Wang et al. (2022d). The methods marked with ‡ denotes that they utilize LLMs (of ChatGPT scale) as knowledge sources.

Methods	CoNLL2003		
	Pre.	Rec.	F1.
W^2NER (Li et al., 2022)	92.71	**93.44**	93.07
DiffusionNER (Shen et al., 2023a)	92.99	92.56	92.78
PromptNER (Shen et al., 2023b)	92.96	93.18	93.08
SCANNER (Ours)	**93.07**	**93.44**	**93.26**
	± 0.20	± 0.23	± 0.21

Table 5: Experiment results on the CoNLL2003.

Methods	Twitter-GMNER		
	Pre.	Rec.	F1.
Text			
HBiLSTM-CRF-None (Lu et al., 2018)	43.56	40.69	42.07
BERT-None (Devlin et al., 2019)	42.18	43.76	42.96
BERT-CRF-None	42.73	44.88	43.78
BARTNER-None (Yan et al., 2021a)	44.61	45.04	44.82
Text+Image			
GVATT-RCNN-EVG (Lu et al., 2018)	49.36	47.80	48.57
UMT-RCNN-EVG (Yu et al., 2020)	49.16	51.48	50.29
UMT-VinVL-EVG (Yu et al., 2020)	50.15	52.52	51.31
UMGF-VinVL-EVG (Zhang et al., 2021a)	51.62	51.72	51.67
ITA-VinVL-EVG (Wang et al., 2022b)	52.37	50.77	51.56
BARTMNER-VinVL-EVG (Yu et al., 2023)	52.47	52.43	52.45
H-Index (Yu et al., 2023)	56.16	56.67	56.41
SCANNER (Ours)	**68.09**	**68.96**	**68.52**
	± 0.73	± 0.61	± 0.67

Table 6: Experiment results on the Twitter-GMNER. The reported figures for the baseline models are taken from Yu et al. (2023).

in Table 5. It shows that SCANNER exhibits a competitive performance compared to the existing NER methods.

Experimental results in MNER. In assessing the effectiveness of SCANNER in MNER, we conducted comparative analyses against various leading models in this task. The results, detailed in Table 4, reveal that our model achieves superior performance in Twitter-2015 and exhibits markedly impressive results in Twitter-2017. Notably, while PGIM shows outstanding performance on Twitter-2017, it utilizes large language models (LLM) like ChatGPT, which incurs API costs, a notable drawback. In contrast, our model does not rely on LLM knowledge, freeing it from such disadvantages and demonstrating better performance on Twitter-2015.

Experimental results in GMNER. To show our effectiveness in GMNER, we make broad comparisons with all existing methods. Text-only models made to predict the visual groundings all 'None'. The Table 6 shows that our model achieves significant performance improvements over prior research and establishes a new powerful baseline for future GMNER studies.

4.4 Ablation study

We conduct ablation experiments on the MNER task to evaluate the effectiveness of the proposed method. These results are shown in Table 7. We observe that removing the Trust Your Teacher method led to a decrease in performance. Our proposed distillation method effectively alleviates the dataset noise issue, making our model more robust to learning from noisy dataset. Additionally, to verify the effectiveness of the various types of knowledge used in our study, we compare the results with experiments where each type of knowledge was removed. We confirm that the object knowledge, Wikipedia knowledge, and image caption knowledge used in our paper all contribute to the performance improvement of the MNER task.

As shown in Fig. 6, all three types of knowledge

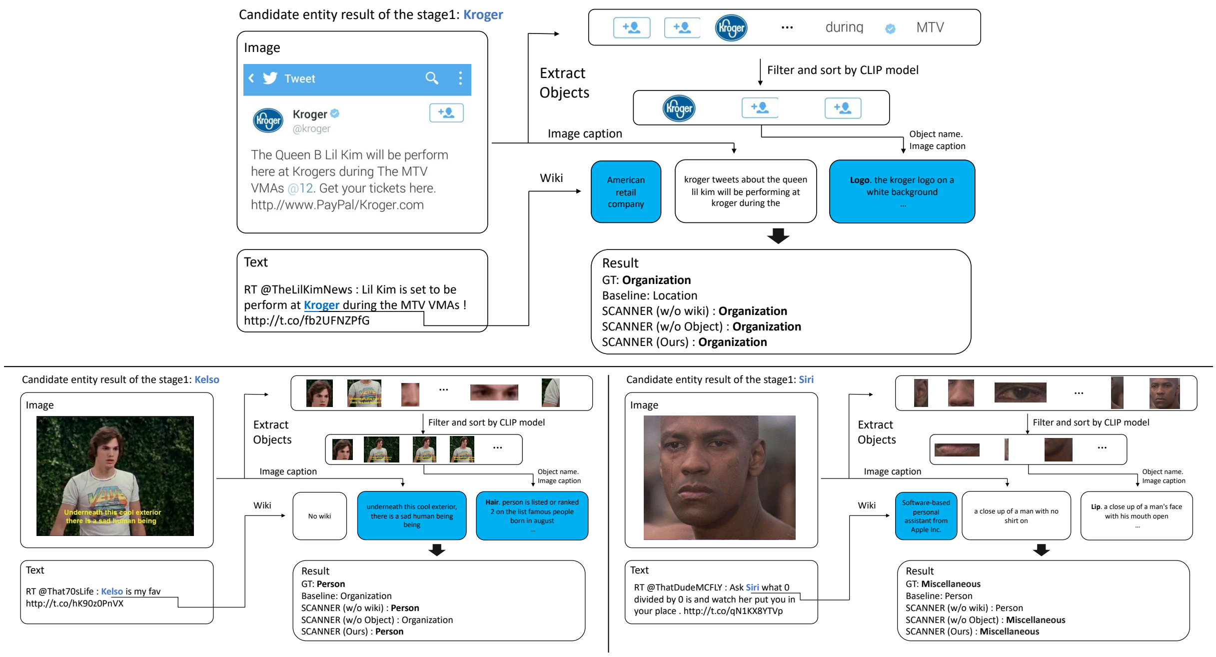

Figure 6: Visualization results showing how various types of knowledge are brought in and utilized differently to perform the MNER task. Knowledge highlighted in blue positively influences correct predictions.

Methods	Twitter-2015			Twitter-2017		
	Pre.	Rec.	F1.	Pre.	Rec.	F1
SCANNER	79.72	79.03	79.38	90.40	90.67	90.54
- TYT	-0.26	-0.17	-0.21	-0.24	-0.11	-0.18
- OBK	+0.11	-0.60	-0.26	-0.23	-0.22	-0.22
- WKK	-1.12	-0.14	-0.64	-0.51	-0.44	-0.48
- ICK	-0.08	-0.54	-0.31	-0.29	-0.31	-0.29

Table 7: Ablation studies on MNER datasets. '-TYT' is without trust your teacher method. '-OBK' is without object knowledge. '-WKK' is without Wikipedia knowledge.'-ICK' is without image caption knowledge.

Datasets	w/o Knowledge		w/ Knowledge	
	Seen	Unseen	Seen	Unseen
CoNLL2003	96.29	89.68	96.35	89.70
Twitter-2015	87.18	73.84	87.50	75.45
Twitter-2017	95.68	82.96	95.90	83.71

Table 8: The result comparing the test F1 scores in unseen entities of knowledge extracted and baseline.

can be utilized as useful information for named entity recognition. In the case of the first image, knowledge from Wikipedia such as "American retail company" and object knowledge containing the logo information of "Kroger" both help in predicting the "Kroger" entity as an organization. For the image on the bottom left, image caption and object knowledge aided in named entity recognition. Moreover, in the image on the bottom right, vision information like image caption and object knowledge led to incorrect entity recognition results, but it was corrected through external knowledge from Wikipedia. Thus, the three types of knowledge proposed in this paper complement each other, enabling accurate MNER performance.

Table 8 shows the effectiveness of knowledge in unseen entities. As SCANNER utilizes various knowledge in MNER, it greatly increases performance in unseen entities. In NER, lack of various knowledge causes there to be no image, which slightly improves the performance.

5 Conclusions

We introduce SCANNER, a novel approach for performing NER tasks by utilizing knowledge from various sources. To efficiently fetch diverse knowledge, SCANNER employs a two-stage structure, which detects entity candidates first, and performs named entity recognition and visual grounding on these candidates. Additionally, we propose the novel distillation method, which robustly trains the model against dataset noise, demonstrating superior performance in various NER benchmarks. We believe that our method can be easily extended to utilize knowledge from multiple sources that were not covered in this paper.

Limitations

In this study, we extract knowledge from various sources and utilize it to perform MNER tasks. By leveraging several vision experts such as CLIP, and also fetching external knowledge, our method takes relatively longer inference time compared to approaches that do not use knowledge. However, the use of vision experts and knowledge is essential for a MNER model that functions well even with unseen entities, and we efficiently extract information through a two-stage structure.

Additionally, the aspect of combining the extracted knowledge with Large Language Models (LLMs) is not been explored in this paper. LLMs themselves are massive models containing a wealth of information. Therefore, similar to the other sources of knowledge used in the paper, the LLM response results for entity candidates can be effectively utilized for entity recognition. We leave the utilization of LLMs as future work and will release our implementation to facilitate future research.

Ethics statement

All experimental results we provide in this paper is based on publicly available datasets and open-source models.

References

F. Chen and Y. Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. *arXiv preprint arXiv:2306.14122*.

F. Chen, J. Liu, K. Ji, W. Ren, J. Wang, and J. Chen. 2023. Learning implicit entity-object relations by bidirectional generative alignment for multimodal ner. In *ACM MM*.

X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, F. Huang, L. Si, and H. Chen. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR*.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*.

M. Jia, L. Shen, X. Shen, L. Liao, M. Chen, X. He, Z. Chen, and J. Li. 2023. Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *AAAI*.

J. Li, H. Fei, J. Liu, S. Wu, M. Zhang, C. Teng, D. Ji, and F. Li. 2022. Unified named entity recognition as word-word relation classification. In *AAAI*.

J. Li, D. Li, S. Savarese, and S. Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

J. Li, H. Li, Z. Pan, and G. Pan. 2023b. Prompt chatgpt in mner: Improved multimodal named entity recognition method based on auxiliary refining knowledge from chatgpt. *arXiv preprint arXiv:2305.12212*.

J. Li, A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34.

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

I. Loshchilov. and . Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji. 2018. Visual attention model for name tagging in multimodal social media. In *ACL*.

J. Lu, D. Zhang, J. Zhang, and P. Zhang. 2022. Flat multi-modal interaction transformer for named entity recognition. In *COLING*.

M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang. 2023a. DiffusionNER: Boundary diffusion for named entity recognition. In *ACL*.

Y. Shen, Z. Tan, S. Wu, W. Zhang, R. Zhang, Y. Xi, W. Lu, and Y. Zhuang. 2023b. PromptNER: Prompt locating and typing for named entity recognition. In *ACL*.

E. F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *NAACL*.

X. Wang, J. Cai, Y. Jiang, P. Xie, K. Tu, and W. Lu. 2022a. Named entity and relation extraction with multi-modal retrieval. In *EMNLP findings*.

X. Wang, M. Gui, Y. Jiang, Z. Jia, N. Bach, T. Wang, Z. Huang, and K. Tu. 2022b. Ita: Image-text alignments for multi-modal named entity recognition. In *NAACL-HLT*.

X. Wang, J. Tian, M. Gui, Z. Li, J. Ye, M. Yan, and Y. Xiao. 2022c. Promptmner: prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In *DASFAA*.

X. Wang, J. Ye, Z. Li, J. Tian, Y. Jiang, M. Yan, J. Zhang, and Y. Xiao. 2022d. Cat-mner: Multimodal named entity recognition with knowledge-refined cross-modal attention. In *ICME*.

Z. Wang, J. Shang, L. Liu, L. Lu, J. Liu, and J. Han. 2019. Crossweigh: Training named entity tagger from imperfect annotations. In *EMNLP-IJCNLP*.

D. Wu, S.-T. Xia, and Y. Wang. 2020. Adversarial weight perturbation helps robust generalization. In *NeurIPS*.

I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *EMNLP*.

H. Yan, T. Gui, J. Dai, Q. Guo, Z. Zhang, and X. Qiu. 2021a. A unified generative framework for various ner subtasks. In *ACL*.

H. Yan, T. Gui, J. Dai, Q. Guo, Z. Zhang, and X. Qiu. 2021b. A unified generative framework for various NER subtasks. In *ACL*.

J. Yu, J. Jiang, L. Yang, and R. Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *ACL*.

J. Yu, Z. Li, J. Wang, and R. Xia. 2023. Grounded multimodal named entity recognition on social media. In *ACL*.

D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, and G. Zhou. 2021a. Multi-modal graph fusion for named entity recognition with targeted visual guidance. *AAAI*.

P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*.

Q. Zhang, J. Fu, X. Liu, and X. Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *AAAI*.

X. Zhang, J. Yuan, L. Li, and J. Liu. 2023. Reducing the bias of visual objects in multimodal named entity recognition. In *WSDM*.

F. Zhao, C. Li, Z. Wu, S. Xing, and X. Dai. 2022. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal ner. In *ACM MM*.

E. Zhu and J. Li. 2022. Boundary smoothing for named entity recognition. In *ACL*.

A Hyper-parameter settings

Datasets	Stage1			
	epochs	batch size	lr	weight decay
CoNLL2003	5	8	5×10^{-6}	1
Twitter-2015	10	4	1×10^{-5}	2
Twitter-2017	10	8	1×10^{-5}	2
Twitter-GMNER	10	8	1×10^{-5}	2

Table 9: Hyper-parameter settings in Stage 1 were used in the experiments for NER, MNER, and GMNER.

Datasets	Stage2				
	epochs	batch size	lr	weight decay	max objects
CoNLL2003	20	8	3×10^{-6}	0.01	-
Twitter-2015	5	8	1×10^{-5}	0.01	15
Twitter-2017	7	8	5×10^{-6}	0.01	15
Twitter-GMNER	5	8	5×10^{-6}	2	15

Table 10: Hyper-parameter settings in Stage2 were used in the experiments for NER, MNER, and GMNER.

We conducted our experiments with hyper-parameter settings as outlined in the following Table 9 and Table 10, and we utilize AdamW (Loshchilov. and Hutter, 2019) optimizer for all tasks. 'max objects' refers to the maximum number of object knowledge inputs. We performed a grid search for the learning rate within the range of $[5 \times 10^{-6}, 1 \times 10^{-5}]$. We tested batch sizes of 4, 8, and 16 to determine the optimal value, and we explored weight decay within a range of $[0.01, 2]$.

B Detailed dataset statistics

To demonstrate the superiority of our method for various NER tasks, we conduct experiments on a range of datasets. The overall dataset statistics are shown in Table 11, and each task description is in below.

NER dataset. CoNLL2003 (Tjong Kim Sang and De Meulder, 2003), a dataset with four named entities: PER, LOC, ORG, and MISC. We follow the standard setting (Peters et al., 2017; Yan et al., 2021b; Shen et al., 2023a): use both the train set and dev set for training and evaluate with the test set

MNER dataset. Twitter2015 (Zhang et al., 2018) and Twitter2017 (Lu et al., 2018); collected from social network service posts. Like CoNLL2003, it consists of the same four named entity types. We operate a train set for training and hyper-parameter tuning using a dev set and evaluate it with the test set.

	Text				Image Total
	#Total	#Train	#Dev	#Test	#Groundable Entity
CoNLL2003	20,744	17,291	-	3,453	-
Twitter-2015	8,257	4,000	1,000	3,257	-
Twitter-2017	4819	3373	723	723	-
Twitter-GMNER	10,000	7,000	1,500	1,500	6,716

Table 11: Dataset Statistics of NER, MNER, and GMNER benchmarks

GMNER dataset. Twitter-GMNER (Yu et al., 2023), a dataset collected by extracting some of the data from Twitter2015 and Twitter2017, and employ bounding box annotation. We operate same validate strategy as MNER.