Grounded Multimodal Procedural Entity Recognition for Procedural Documents: A New Dataset and Baseline

Haopeng Ren[1,2,†], **Yushi Zeng**[1,2,†], **Yi Cai**[1,2,‡], **Zhenqi Ye**[1,2], **Li Yuan**[1,2] **Pinli Zhu**[1,2]
[1]1School of Software Engineering, South China University of Technology
[2]Key Laboratory of Big Data and Intelligent Robot
(South China University of Technology) Ministry of Education
ycai@scut.edu.cn

Abstract

Much of commonsense knowledge in real world is in the form of procudures or sequences of steps to achieve particular goals. In recent years, knowledge extraction on procedural documents has attracted considerable attention. However, they often focus on procedural text but ignore a common multimodal scenario in the real world. Images and text can complement each other semantically, alleviating the semantic ambiguity suffered in text-only modality. Motivated by these, in this paper, we explore a problem of grounded multimodal procedural entity recognition (GMPER), aiming to detect the procedural entity and the corresponding bounding box groundings in images (i.e., visual entities). A new dataset (Wiki-GMPER) is built and extensive experiments are conducted to evaluate the effectiveness of our proposed model.

Keywords: Multimodal Procedure Knowledge, Procedural Entity Recognition, Procedural Entity Grounded

1. Introduction

In our daily life, much of commonsense knowledge is in the form of sequences of actions to achieve particular goals (e.g., cooking recipes, crafting and maintenance manuals), which is called *Procedural Knowledge* (Georgeff and Lansky, 1986; Ren et al., 2023). For the large and growing amount of unstructured or semi-structured procedural documents on media platforms such as *WikiHow*[1], *EHow*[2] and *Instructables*[3], it is a pressing need to automatically extract procedural knowledge (e.g., entities or relations) for knowledge graph constructions and downstream procedures understanding applications (e.g., sequence ordering (Wu et al., 2022), question answering system (Zhang et al., 2022) and operation diagnosis (Luo et al., 2021)).

Generally, procedural documents often appear in a multimodal manner. As shown in Figure 1, each step in a procedural document contains an image and the corresponding text description. Nevertheless, current existing procedural entity recognition (PER) methods (Jermsurawong and Habash, 2015; Leopold et al., 2018; Mysore et al., 2019; Jiang et al., 2020; Luo et al., 2021) mainly focus on the text-only settings, which is insufficient for entity disambiguation (Yu et al., 2023). For example shown in Figure 1, without the red bounding box, it is difficult to refer to what state the procedural entity "*Tomato*" is in each step depending only on text description (e.g., a whole tomato in Step 1, while tomato slices in Step 3). Capturing the visual entities (e.g., the red bounding box in Figure 1) in images are beneficial for the procedural document understanding and reasoning (Wu et al., 2022; Zhang et al., 2022). Motivated by this, our work in this paper considers a multimodal setting where the multimodal procedural knowledge extraction system not only detects the procedural entities from the procedural text description but also links the procedural entities to their corresponding bounding boxes in images, as shown in Figure 1. The research on this subject can be called as *Grounded Multimodal Procedural Entity Recognition (GMPER)*.

To tackle the GMPER task, two kinds of related solutions, i.e., Multimodal Named Entity Recognition *MNER* (Zhang et al., 2018) and Grounded Multimodal Named Entity Recognition *GMNER* (Yu et al., 2023) are proposed to extract entities from social media posts. Specifically, existing MNER methods (Moon et al., 2018; Lu et al., 2018; Yu et al., 2020; Zhang et al., 2021a; Chen et al., 2022; Wang et al., 2022a; Jia et al., 2022, 2023) are designed to extract the textual entities with the help of visual features from images, but fail to build the link or correspondence between textual entities and visual entities. To solve this problem, Yu et al. 2023 propose a new task *GMNER*, aiming to simultaneously recognize the textual entities and the corresponding visual regions in images.

Though recent GMNER methods (Yu et al., 2023) achieve remarkable performance, but still face several main challenges when directly adapted to the GMPER task. Firstly, different from the GMNER task which mainly focuses on short multimodal posts, the GMPER task is based on long multimodal procedural documents with multiple steps

[†]The authors contribute equally.
[‡]Corresponding author
[1]https://www.wikihow.com/Main-Page
[2]https://www.ehow.com/
[3]https://www.instructables.com/

LREC-COLING 2024, pages 7971–7981
20-25 May, 2024.

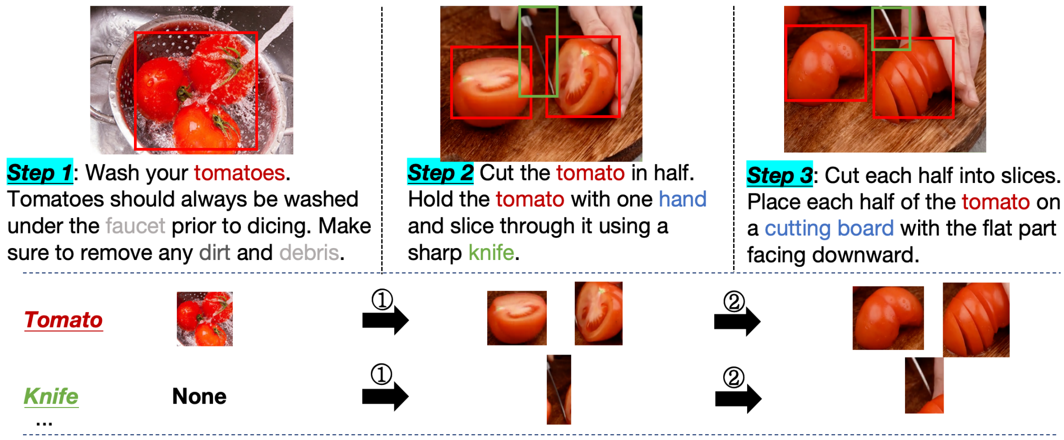


Figure 1: A Case of Grounded Multimodal Procedural Entity Recognition for the Multimodal Procedure document "How to Dice Tomatoes".

Figure 2: Two Cases of Object Detection by Grounded Language-Image Pretrained Model (GLIP) (Li et al., 2022) with the Prompt Text "Tomato or Persimmon"; The highest confidence score for cting object *"tomato"* in left image is **0.76**, while that in the right image is **0.57**.

and complex interactions between procedural entities. As shown in Figure 1, the same procedural entity has multiple visual regions with different states and meanwhile, there will be mutual occlusion between visual procedural entities. Secondly, the state of the same visual entity, such as shape, color and forms (e.g., solid, liquid and gaseous) will dynamically change as the procedure progresses. Existing *MNER* or *GMNER* methods only consider one descriptive text and the corresponding image. It is a challenge for them to track the state changes of visual entities between steps on multimodal procedural documents. For example shown in Figure 2, when the target object *"tomato"* is in a complete state (left picture), GLIP (which is a well-known language-image model pretrained with large-scale multimodal data) can correctly detect the visual region with a high confidence score. However, as the state of target object *"tomato"* changes (i.e., *"tomato"* is cut into slices in the right picture), GLIP detects the visual region of *"tomato"* with a low confidence score and even is prone to wrongly recognize it as another entity type (e.g., *persimmon*).

In our paper, we propose a sequence-aware grounded multimodal procedural entity recognition (SeqGMPER) method to detect both the textual procedural entities and the corresponding visual regions in images from multimodal procedural documents. Specifically, to capture the state changes of procedural entities as the procedure progresses, a Textual or Visual Sequential Feature Fusion (TSFF or VSFF) module is designed. The state features of textual or visual entities in current step take into account to that in the previous steps. Furthermore, to conduct the evaluation on GMPER task, we construct a new dataset, called *Wiki-GMPER* based on the WikiHow resource (Anthonio et al., 2020), in which we manually annotate the textural procedural entities and the corresponding bounding boxes in images.

To summarize, the main contributions of this paper are listed as follows:

- We explore a new problem named Grounded Multimodal Procedural Entity Recognition (GMPER), aiming to automatically recognize textual procedural entities and link the corresponding visual regions in images from multimodal procedural documents.

- We design a textual and visual sequential feature fusion method to capture the state changes of entities as the sequence or procedure progresses, which effectively assist the detection of both textual and visual entities from multimodal procedural documents.

- We create a new grounded multimodal procedural entity recognition dataset *Wiki-GMPER* based on the multimodal procedural documents. Extensive experiments are conducted on the Wiki-GMPER dataset to evaluate the effectiveness of our model in automatically detecting procedural textual and visual entities.

2. Related Work

One kind of important commonsense knowledge in our daily life is the instructions or procedures which are the form of a sequence of actions to complete the particular goals. Current well-known knowledge bases such as *Wiki-Data* (Vrandečić and Krötzsch, 2014), *Wikipedia* (Lehmann et al., 2015), *Freebase* (Bollacker et al., 2007) and *ConceptNet* (Speer et al., 2017) mainly focus on modeling *descriptive knowledge* (i.e., the attributions or features of things (Yang and Nyberg, 2015; Yuan et al., 2023)), but neglect another commonsense knowledge—*Procedural Knowledge* (i.e., the Knowledge of procedures or sequence of actions to achieve the specific goals). To automatically extract the procedural knowledge, existing work (Jermsurawong and Habash, 2015; Feng et al., 2018; Mysore et al., 2019; Qian et al., 2020; Yamakata et al., 2020; Anthonio et al., 2020; Jiang et al., 2020; Pal et al., 2021; Fang et al., 2022; Ren et al., 2023) are designed to identify the procedural entities or their relations from the textual procedure documents (e.g., food recipes and crafting). However, procedural documents often generally appear in a multimodal manner. Therefore, another kind of related works (Pan et al., 2020; Xu et al., 2020) to construct the step-level or entity-level workflow from the multimodal procedural documents. Nevertheless, they only treat the visual features as additional clues but fail to identify the fine-grained entity groundings in images, which suffer from the entity ambiguity (Yu et al., 2023).

Currently, there are two kinds of related works: Multimodal Named Entity Recognition (MNER) and Grounded Multimodal Named Entity Recognition (GMNER). Specifically, MNER has recently attracted considerable attention on social media, aiming to recognize the named entity in text posts with the help of visual features as additional clues. Most of MNER methods (Moon et al., 2018; Lu et al., 2018; Zhang et al., 2018; Arshad et al., 2019; Yu et al., 2020; Zheng et al., 2020; Arshad et al., 2019; Chen et al., 2021, 2022; Wu et al., 2020; Zhang et al., 2021a; Wang et al., 2022b; Jia et al., 2023) mainly focus on the multimodal features alignment and fusion to recognize the textual entities. However, they only regard the visual features as significant clues for textual entity detection but neglect the correspondence between the entity groundings in images. To solve this problem, Yu et al. 2023 propose a grounded multimodal named entity recognition (GMNER) method to extract entity-type-region triples from multimodal media posts. Different from current MNER and GMNER methods, GMPER task focuses on the multimodal procedural documents with multiple steps and complex interactions (e.g., state changes) between procedural entities as the procedure progresses. Motivated by these, we explore a problem of grounded multimodal procedural entity recognition (GMPER), aiming to extract the procedural entity and the corresponding bounding box groundings in images. In our paper, a new dataset *Wiki-GMPER* is built based on the WikiHow dataset bases and we propose a Sequence-aware GMPER method to capture the interaction among steps. Extensive experiments are conducted to evaluate the effectiveness of our proposed model.

3. Model

In this section, the problem definition of GMPER task is firstly given and then we describe our proposed Sequence-aware Grounded Multimodal Procedural Entity Recognition model (SeqGMPER) in detail.

3.1. Problem Definition and Notations

Given a multimodal procedural document with a sequence of steps $D = \{s_1, s_2, \ldots, s_{L_d}\}$ and a corresponding sequence of images $V = \{v_1, v_2, \ldots, v_{L_d}\}$, the goal of the Grounded Multimodal Procedural Entity Recognition (GMPER) task is to extract a set of entity tuples:

$$Y = \{(e_1, r_1), \ldots, (e_t, r_t)\} \quad (1)$$

where L_d denotes the number of steps, s_i denotes the i-th step containing a sequence of words $s_i = \{w_{i,1}, w_{i,2}, \ldots w_{i,L_s}\}$ in a procedure document; The (e_i, r_i) refers to the i-th entity tuple, where e_i is the i-th procedural entity and r_i is the corresponding bounding box groundings in an image. Note that when the procedural entity e_i does not contain any visual region in an image, the visual region r_i is set as *None*. Meanwhile, the visual region r_i can be defined as a *4-D* vector $(r_i^{x_1}, r_i^{y_1}, r_i^{x_2}, r_i^{y_2})$ which refers the top-left and bottom-right positions of the grounded bounding box in the image, respectively.

3.2. Multimodal Feature Representation

3.2.1. Text & Image Representation

Inspired by the success of grounded language-image pretrained model GLIP (Li et al., 2022) (which pretrained with a large-scale multimodal data) in object detection and phrase grounding tasks, we employ the pretrained multimodal encoder in GLIP (Li et al., 2022) to extract features for both the text and image in each step. Given a procedural document D, Specifically, given one step in a procedural document D, which contains a sequence of words $s_i = \{w_{i1}, w_{i2}, \ldots, w_{iL_s}\}$ and the corresponding image v_i as the input fo the text encoder and visual encoder in GLIP respectively, the

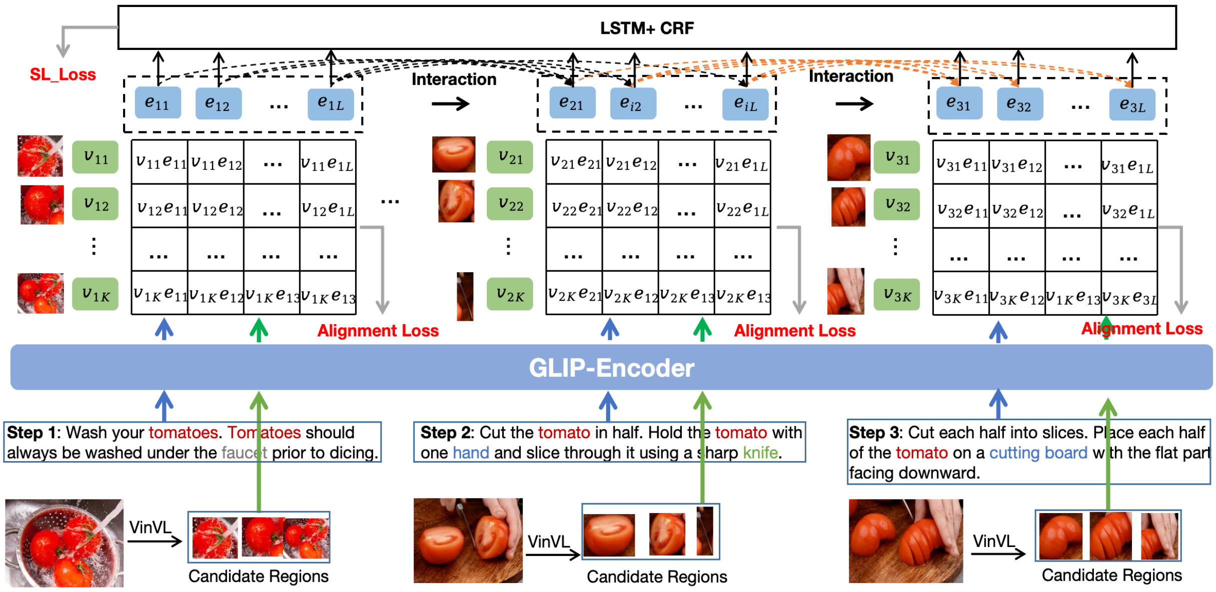


Figure 3: The Framework of Our Proposed Model SeqGMPER

word embedding matrix $\boldsymbol{S}_i = \{\boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, \ldots, \boldsymbol{w}_{iL_s}\}$ and the image feature map $\boldsymbol{M}_i$ can be obtained, where $\boldsymbol{S}_i \in \mathbb{R}^{L_s \times d_t}$, $\boldsymbol{w}_{ij} \in \mathbb{R}^{d_t}$, $\boldsymbol{M}_i \in \mathbb{R}^{d_v \times d_w \times d_h}$ and d_v denotes the number of the convolution kernel (i.e., the number of feature maps). Then, the mean-pooling operation for each feature map $\boldsymbol{M}_i$ is conducted and we can finally obtain the feature representation $\boldsymbol{v}_i$ for the i-th image, where $\boldsymbol{v}_i \in \mathbb{R}^{d_v}$.

3.2.2. Candidate Region Representation

Following Yu et al. (2023), a widely-adopted object detection model VinVL (Zhang et al., 2021b) is utilized to extract the candidate semantic visual region (i.e., the candidate bounding box groundings). Then, we rank candidate visual regions based on their detection probabilities. Specifically, for the i-th step in a multimodal procedure document, we identify the top-K candidate visual regions $C_i = \{c_{i1}, c_{i2}, \ldots, c_{iK}\}$, where the region c_{ij} can be denoted as a *4-D* vector $(c_{ij}^{x1}, c_{ij}^{y1}, c_{ij}^{x2}, c_{ij}^{y2})$ which respectively refers to the top-left and bottom-right positions of the candidate bounding box. Then, the feature map of each candidate visual region can be obtained by extracting the scaling feature area of the corresponding original image feature map $\boldsymbol{M}_i$. In the same way, the mean-pooling operation is used to obtain the final feature representation of the candidate visual regions $\boldsymbol{R}_i = \{\boldsymbol{r}_{i1}, \boldsymbol{r}_{i2}, \ldots, \boldsymbol{r}_{iK}\}$, where $\boldsymbol{r}_{ij} \in \mathbb{R}^{d_v}$.

3.3. Multimodal Sequential Feature Fusion

Since the steps of the multimodal procedural documents are interdependent and interrelated, the entities and regions discovered in the previous step can provide the important clues for the identification of entities and regions in the following steps. For example shown in Figure 1, we can observe that the token "*tomato*" would be regarded as a procedural entity with a high probability since the procedural entity "*tomato*" appears in previous step. Likewise, the identified visual region in current step would be also beneficial for the following steps' visual region detection. Motivated by this observation, we design a Multimodal Sequential Feature Fusion Module to capture the interactions between procedural entities and visual regions between steps. For the sequence data of different modalities (i.e., textual and visual modalities), we respectively conduct the sequential interaction feature fusion.

Sequential Element Attention Mechanism: For both textual sequence (i.e., word sequence in each step) and visual sequences (i.e., candidate visual region sequence in each step), we respectively adopt the sequential element attention mechanism to capture the interaction features among steps. For the convenience of description, we uniformly use $\boldsymbol{X} = \{\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_n\}$ to represent the sequences in both textual and visual modalities, where $\boldsymbol{T}_i = \{\boldsymbol{t}_{i,1}, \boldsymbol{t}_{i,2}, \ldots, \boldsymbol{t}_{i,m}\}$ denotes the sequence of element feature representation and m denotes the length of sequence. Thus, given the previous step T_{i-1} and current step T_i, each fused element representation $\boldsymbol{t}_{i,j}^{fuse}$ can be calculated as follows:

$$\boldsymbol{t}_{i,j}^{fuse} = [\sum_{j=0}^{m} \alpha_{i-1,j} \boldsymbol{t}_{i-1,j}; \boldsymbol{t}_{i,j}] \quad (2)$$

where $\boldsymbol{t}_{i-1,j}$ denotes the feature representation of the j-th element in previous step T_{i-1}. The important degree $\alpha_{i-1,j}$ for the each sequential element t_{ij} in step T_i can be calculated as follows:

$$\alpha_{i-1,j} = \frac{e^{\boldsymbol{t}_{i-1,j}\boldsymbol{t}_{i,j}}}{\sum_{k=0}^{m} e^{\boldsymbol{t}_{i-1,k}\boldsymbol{t}_{i,j}}} \tag{3}$$

Textual Sequential Feature Fusion: For the textual modality, given a word sequence $\boldsymbol{W_i} = \{\boldsymbol{w}_{i,1}, \boldsymbol{w}_{i,2}, \ldots, \boldsymbol{w}_{i,L_s}\}$ in each step from the multimodal procedural document, the Sequential Element Attention can transfer them into another representation sequence $\boldsymbol{W_i^f} = \{\boldsymbol{w}^f_{i,1}, \boldsymbol{w}^f_{i,2}, \ldots, \boldsymbol{w}^f_{i,L_s}\}$. Thus, based on the sequential element attention module, given Each token representation in current step would fuse semantic features from previous step.

Visual Sequential Feature Fusion: Similar to the Textual Sequential Interaction Feature Fusion, given the the Top-K visual regions embeddings $\boldsymbol{R}_i = \{\boldsymbol{r}_{i1}, \boldsymbol{r}_{i2}, \ldots, \boldsymbol{r}_{iK}\}$ in the i-th step, the Sequential Element Attention is utilized to transfer them into another visual region representation sequence $\boldsymbol{R}_i^f = \{\boldsymbol{r}^f_{i,1}, \boldsymbol{r}^f_{i,2}, \ldots, \boldsymbol{r}^f_{i,K}\}$. Thus, based on the sequential element attention module, the visual region representation in current step would try to capture the similar visual regions in the previous step, which builds the feature interaction between steps.

3.4. Grounded Multimodal Procedural Entity Recognition (GMPER)

Based on Section *Multimodal Sequential Feature Fusion*, we can obtain the representation of token sequence and visual region sequence in each step. To conduct the GMPER task, three tasks i.e., Procedural Entity Recognition (PER), Binary Groundable Classification (BGC) and Grounded Procedural Entity (GPE) are conducted.

3.4.1. Procedural Entity Recognition

Given the multimodal procedural document (including the textual sequence and the visual region sequence), PER task aims to detect the procedural entities from token sequence in each step by understanding the language-visual features. Specifically, given the feature representation of token sequence in each step $\boldsymbol{W_i^f}$ (which obtained by section *Multimodal Sequential Feature Fusion*), we employ LSTM-CRF (Huang et al., 2015) layer to predict the corresponding sequence of labels $y_{per} = \{y_1, y_2, \ldots, y_{L_s}\}$,

$$p(y_{per}|w_{i,j}) = LSTM\text{-}CRF(\boldsymbol{W}_i^f) \tag{4}$$

where the label $y_i \in$ {B-Object, I-Object, O} and $w_{i,j}$ denotes the j-th token of the word sequence in the step s_i. Then, the CRF loss L_{entity} is adopted to optimize the model's parameters.

3.4.2. Binary Groundable Classification

Based on the procedural entities identified in PER, a binary classification task is employed to determine each predicted procedural entity is groundable or ungroundable. Specifically, considering multi-token procedural entities, the entity embeddings $\boldsymbol{E}_i^f = \{\boldsymbol{e}^f_{i,1}, \boldsymbol{e}^f_{i,2}, \ldots, \boldsymbol{e}^f_{i,L_e}\}$ are obtained by mean-pooling the feature representations of multiple token belonging to the same entity, where $\boldsymbol{e}_{ij} \in \mathbb{R}^{d_t}$ and L_e is the number of identified procedural entities. Finally, the probability of the groudable and ungroundable procedural entity can be calculated as follows:

$$p(y_{bgc}|e_{i,j}) = Softmax(\boldsymbol{W}\boldsymbol{e^f_{i,j}} + b) \tag{5}$$

where $y_{bgc} \in \{0, 1\}$ and $\boldsymbol{W} \in \mathbb{R}^{d_t \times 2}$ and b are the parameter fo the BGC classifer. Then, the cross-entropy loss $L_{groudable}$ is utilized to optimize the model's parameters.

3.4.3. Grounded Procedural Entity

For the i-th step, we employ a crossmodal attention module (Tsai et al., 2019) to fuse entity-level embeddings $\boldsymbol{E}_i^f$ and visual region embeddings $\boldsymbol{R}_i^f$. We then obtain the probability distribution over all the visual regions for each procedural entity denoted by z_i, as follows:

$$\boldsymbol{H}_i = Crossmodal\text{-}Attention(\boldsymbol{E}_i, \boldsymbol{R}_i) \tag{6}$$

$$\boldsymbol{Z}_i = Sigmoid(\boldsymbol{W}_G\boldsymbol{H}_i + \boldsymbol{B}) \tag{7}$$

where $\boldsymbol{H}_i \in \mathbb{R}^{L_g \times d_c}, \boldsymbol{W}_G \in \mathbb{R}^{d_c \times K}, \boldsymbol{B} \in \mathbb{R}^{1 \times K}, \boldsymbol{Z}_i \in \mathbb{R}^{L_g \times K}$ and L_g denotes the number of groundable entities. Specifically, we set a threshold for the GPE task, the visual region $\boldsymbol{r}_{i,j}$ belongs to the entity if the predcition probability greater than 0.5. Then, the BCE loss is used to optimize the parameters of the entity grounding recovery module as follows:

$$L_{grounding} = BCE(\boldsymbol{Z}_i, \boldsymbol{Y}_{gpe}) \tag{8}$$

where $\boldsymbol{Y}_{gpe} \in \mathbb{R}^{L_g \times K}$ denotes the matrix of the ground true labels.

Finally, in the training stage, the three losses (i.e. L_{entity} and $L_{groundable}$ and $L_{grounding}$) are simultaneously used to conduct the parameter optimization, as follows:

$$L = L_{entity} + L_{groundable} + L_{grounding} \tag{9}$$

Dataset Statistics	Train	Test	Validation
# Doc.	809	299	122
# Step	4794	1341	701
Avg Step of Doc.	5.90	5.00	5.75
# Entity	19869	5836	2738
# Groundable	11029	2854	1252
# Ungroundable	8840	2982	1486

Table 1: The statistics of our annotated dataset Wiki-GMPER

4. Experiment

We firstly introduce the construction of the new dataset *Wiki-GMPER* and then analyze the experimental results in detail.

4.1. Dataset Collection & Annotation

We collect the corpus from the benchmark sequence ordering dataset i.e. WikiHow (Anthonio et al., 2020; Wu et al., 2022) which provides a collection of human-created *how-to* articles about various topics (e.g., Crafts, Computers and Recipes). Two topics i.e., *Crafts* and *Recipes* are selected to build *Wiki-GMPER*, a dataset of multimodal procedural documents with the procedural entity taggings and the corresponding bounding box annotations in images, as shown in Figure 1. For the procedural entity taggings, three well-educated annotators are employed to make annotations by averaging the candidate procedural corpus with the BRAT tool[4]. Then, the bounding box annotation is conducted with the graphical image annotation tool *LabelImg* tool[5]. To ensure the quality of human-annotation, each annotator is required to give the confidence score for each annotated label. We weigh the confidence score of each annotator for the same label and the label with the highest score will be preserved.

Statistically, the final dataset contains 1230 multimodal procedural documents with 6836 steps. Each step consists of a text description and a corresponding image. We split the final annotated dataset into train, test and validation sets with 7:2:1 ratio. Table 1 depicts the detailed statistics of the annotated dataset *Wiki-GMPER*.

4.2. Experimental Settings

We conduct extensive experiments[6] on our annotated dataset *Wiki-GMPER*. Following Yu et al. 2023, the VinVL model (Zhang et al., 2021b) is used to obtain the top-K candidate visual regions. We utilize the grounded language-image pretrained model (GLIP) (Li et al., 2022) to extract the features representation of both text and images. Thus, the dimension of word representation is set as 768. In each optimization step during training, one multimodal procedural document (containing multiple steps) is used (i.e., the hyper-parameter batch size is set as 1). We use the AdamW optimizer for parameter tuning with the learning rate 2e-5. In our experimental evaluation, the precision, recall and F1 metrics are utilized to evaluate the models' performance, following Yu et al. 2023.

In our experiments, we conduct the comparative experiments with two groups of related works, including the *text-only* based methods (BiLSTM-CRF-None (Yu et al., 2023), BERT-None (Kenton and Toutanova, 2019), BERT-CRF-None (Yu et al., 2023) and BARTNER-VinVL-NONE (Yan et al., 2021)) and the *multimodal* based methods (i.e., UMT-RCNN-EVG (Yu et al., 2020), UMT-VinVL-EVG (Yu et al., 2020), UMGF-VinVL-EVG (Zhang et al., 2021a), ITA-VinVL-EV (Wang et al., 2022a), BARTNER-VinVL-EVG (Yu et al., 2023) and H-Index (Yu et al., 2023)).

4.3. Result Analysis

4.3.1. Comparison with Related Models

To demonstrate the effectiveness of our proposed model, we conduct the comparative experiments with current related works on our annotated dataset *Wiki-GMPER*, as shown in Table 2. As we can observe, our proposed model obtains the better performance respectively on Precision, Recall and F1 scores and achieves the state-of-the-art performance. Specifically, comparing with existing text-only NER methods (e.g., BiLSTM-CRF (Huang et al., 2015), BERT-None (Kenton and Toutanova, 2019) and BARTNER (Yan et al., 2021)), our proposed model (i.e., *SeqGMPER*) obtains the higher F1 score with a large margin. Comparing the experimental results between existing text-only methods and our proposed model SeqGMPER-None, we analyze that our proposed model can effectively perform alignment and fusion of text and visual modality data. The comparative experimental results can demonstrate that the visual features from images significantly improve the performance of procedural entity recognition.

Moreover, existing MNER and GMNER methods (i.e., UMT-RCNN-EVG, UMT-VinVL-EVG (Yu et al., 2020), UMGF-VinVL-EVG (Zhang et al., 2021a), ITA-VinVL-EVG (Wang et al., 2022a), BARTNER-VinVL-EVG, H-Index (Yu et al., 2023)) are adapted into GMPER tasks. Compared with them, our proposed model achieves better performances respec-

[4] http://brat.nlplab.org/index.html

[5] https://github.com/HumanSignal/labelImg

[6] The code and datasets are publicly available at https://github.com/betterAndTogether/SeqGMPNER

Model		Pre.	Rec.	F1
Text Only	BiLSTM-CRF-None	16.45	14.08	15.17
	BERT-None (Kenton and Toutanova, 2019)	19.96	20.53	20.24
	BERT-CRF-None	19.86	21.82	20.79
	BARTNER-None (Yan et al., 2021)	20.30	22.92	21.53
Text+Image	UMT-RCNN-EVG (Yu et al., 2020)	32.47	33.91	33.18
	UMT-VinVL-EVG (Yu et al., 2020)	38.14	39.82	38.96
	UMGF-VinVL-EVG (Zhang et al., 2021a)	37.70	39.89	38.76
	ITA-VinVL-EVG (Wang et al., 2022a)	38.85	40.76	39.78
	BARTNER-VinVL-EVG (Yu et al., 2023)	34.08	39.76	36.70
	H-Index (Yu et al., 2023)	41.45	43.37	42.38
	SeqGMPER-None (Ours)	40.20	40.86	40.53
	SeqGMPER (Ours)	**44.86**	**43.74**	**44.28**

Table 2: The Comparative Experimental Results with Current Related Methods. The model "{X}-None" denotes the region predictions default as *None* (i.e., Ungroundable).

Methods	Pre.	Rec.	F1
SeqGMPNER	**44.86**	**43.74**	**44.28**
SeqGMPNER w/o TSFF	44.30	41.06	42.62
SeqGMPNER w/o VSFF	43.82	40.98	42.35

Table 3: Ablation Experiments of Our Model

Tasks	Pre.	Rec.	F1
PER	78.77	83.26	80.96
BGC	69.71	81.97	75.35
GPE	64.00	59.49	61.66

Table 4: The average experimental results on three subtasks: Procedural Entity Recognition (PER), Binary Groundable Classification (BGC) and Grounded Procedural Entity (GPE).

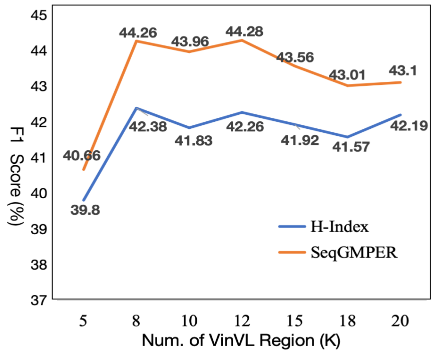
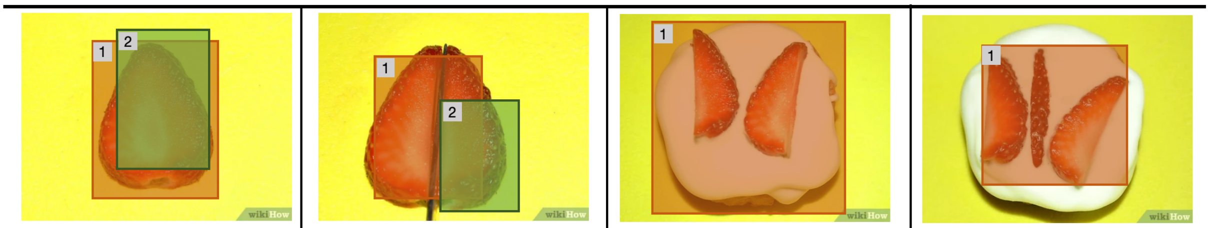


Figure 4: The impact of the value of K (Num. of the VinVL regions) on GMPER task for *H-Index* and our proposed model *SeqGMPER*.

tively on Precision, Recall and F1 scores, as shown in Table 2. We analyze that existing MNER or GMNER methods can only recognize the procedural entity and identify bounding box groundings for each step individually (i.e., including a text description and a corresponding image). Thus, they cannot capture the state changes of visual entities as the procedure progress, which impacts the detection of bounding box groundings. Instead, our proposed model with the sequential feature fusion module can build the connections between steps respectively for textual and visual feature representation. The comparative experimental results in Table 2 can demonstrate the effectiveness of our proposed model in capturing state changes in both textual and visual entities between steps.

4.3.2. Ablation Experiments

To further evaluate the effectiveness of each module in our proposed model, the ablation experiments are conducted. Specifically, we conduct the ablation experiments for the textual sequential feature fusion (TSFF) and visual sequential feature fusion (VSFF), as shown in Table 3. The performance of our proposed model drops significantly without the TSFF or the VSFF module, which can evaluate the effectiveness of our proposed TSFF and VSFF modules. Specifically, according to our observation, the procedural entities mentioned in current step would often appear in the later steps in a procedural document (e.g., the procedural entity "*tomato*" in Figure 1). Thus, the contextual steps would provide important clues for the procedural entity recognition in current step. The ablation experimental results can demonstrate that our proposed TSFF module can effectively capture the interaction among steps, which is beneficial to the procedural entity detection. In the same way, the state of visual entity would change as the procedure progresses. The ablation experimental results can evaluate that VSFF can effectively capture the state changes of visual entity to detect the bounding box groundings.



H-Index	Ours	H-Index	Ours	H-Index	Ours	H-Index	Ours
(strawberry, Box-2) ✓	(strawberry, Box-1) ✓	(strawberry, Box-2) ✗	(strawberry, Box-1) ✓	(cake, None) ✗	(cake, Box-1) ✓	(strawberry, None) ✗ (wings, None) ✓	(strawberry, Box-1) ✓ (wings, None) ✓

Figure 5: Prediction comparison on a multimodal procedural document "*How to Make Strawberry Butterflies*" between *H-Index* and our proposed model *SeqGMPER*. The symbols ✓ and ÃŮ denote correct and incorrect predictions.

4.3.3. Analysis of Sub-Tasks in GMPER

We also conduct the experiment to evaluate the performance of the three sub-tasks in our proposed model: Procedural Entity Recognition (PER), Binary Groundable Classification (BGC) and Grounded Procedural Entity (GPE). In training stage, all subtasks (i.e., PER, BGC and GPE) will be conducted to optimize the models' parameters. In order to independently evaluate our proposed model in BGC task, the ground-true labels of PER task are given to evaluate the performance in testing stage. In the same way, both the ground-true labels of PER and BGC tasks are given to predict the bounding box groundings in GPE task. As shown in Table 4, our proposed model can effectively recognize the textual procedural entities based on the multimodal semantic understanding. From the experimental results on BGC and GPE tasks, we can analyze that our proposed model can learn the multimodal language-image features and effectively detect the groundable procedural entities in images. To some extent, it can evaluate that our proposed model can effectively capture the interactions between steps.

4.3.4. Impact Analysis for Hyper-Parameter K

As shown in Figure 4, we also conduct the comparative experiments for our proposed model with different number of candidate VinVL regions. Compared with H-Index (Yu et al., 2023), our proposed model SeqGMPER obtains the better performance in all K-value settings. According to our observation, both our proposed model and H-Index obtain the lowest F1 score in GMPER task when the hyper-parameter K is set as 5. We analyze that most steps in procedural documents contain more than 5 visual entities in our annotated dataset. As the value of K increases, the performance of both H-Index and our proposed model improves significantly. They both achieve the highest F1 scores when the hyper-parameter K is set between 8 and 12, which can indicate that most of steps in our annotated dataset have around 8-12 visual regions. When the hyper-parameter K is set higher than 12, the performance gradually decreases.

4.3.5. Case Study

To intuitively explain the effectiveness of our proposed model, we conduct the case studies on GMPER task for H-Index (Yu et al., 2023) and our proposed model SeqGMPER. As shown in Figure 5, we can observe that both SeqGMPER and H-Index can correctly recognize the procedural entity "*strawberry*" in step 1. However, as the shape of "*strawberry*" changes in the following steps (i.e., step 2, 3 and 4), H-Index gradually fails to localize its bounding boxes. We analyze that existing works cannot effectively capture the interaction (e.g, the state changes of procedural entities) between steps. Instead, our proposed model SeqGMPER can correctly recognize both the procedural entities and their corresponding bounding boxes in images. The experimental results in this case study demonstrate that SeqGMPER can effectively capture the state changes of visual entities as the procedure progresses and achieve the better performance than H-Index.

5. Conclusion

In our paper, we explore a problem of automatically recognizing procedural entities in text descriptions and linking their corresponding bounding box groundings in images for multimodal procedural documents, named grounded procedural entity recognition (GMPER). Existing procedural knowledge extraction methods often focus on recognizing procedural entities or relations in text-only modal, but neglect a common multi-modal scenario. Existing related works i.e., MNER and GMNER cannot effectively capture the interaction between steps and suffer from the bounding box grounding prediction errors. To solve these problems, we propose a sequence-aware GMPER method to capture the state changes of procedural entity as the procedure progresses. Extensive experiments are conducted on our constructed dataset to evaluate the effectiveness of our proposed model.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62076100), Fundamental Research Funds for the Central Universities, SCUT (x2rjD2230080), the Science and Technology Planning Project of Guangdong Province (2020B0101100002), Guangdong Provincial Fund for Basic and Applied Basic Research - Regional Joint Fund Project (Key Project) (23201910250000318,308155351064), CAAI-Huawei MindSpore Open Fund, CCF-Zhipu AI Large Model Fund.

6. References

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikihowtoimprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5721–5729.

Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. 2019. Aiding intra-text representations with visual context for multimodal named entity recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 337–342. IEEE.

Kurt Bollacker, Robert Cook, and Patrick Tufts. 2007. Freebase: A shared database of structured general human knowledge. In *AAAI*, volume 7, pages 1962–1963.

Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021. Multimodal named entity recognition with image attributes and image knowledge. In *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part II 26*, pages 186–201. Springer.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. 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*, pages 1607–1618.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the reciperef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3481–3495.

Wenfeng Feng, Hankz Hankui Zhuo, and Subbarao Kambhampati. 2018. Extracting action sequences from texts based on deep reinforcement learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4064–4070.

Michael P Georgeff and Amy L Lansky. 1986. Procedural knowledge. *Proceedings of the IEEE*, 74(10):1383–1398.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jermsak Jermsurawong and Nizar Habash. 2015. Predicting the structure of cooking recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 781–786.

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 the AAAI Conference on Artificial Intelligence*, volume 37, pages 8032–8040.

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 the 30th ACM International Conference on Multimedia*, pages 3549–3558.

Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2020. Recipe instruction semantics corpus (risec): Resolving semantic structure and zero anaphora in recipes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 821–826.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Henrik Leopold, Han van Der Aa, and Hajo A Reijers. 2018. Identifying candidate tasks for robotic process automation in textual process descriptions. In *Enterprise, Business-Process and Information Systems Modeling: 19th International Conference, BPMDS 2018, 23rd International Conference, EMMSAD 2018, Held at CAiSE 2018, Tallinn, Estonia, June 11-12, 2018, Proceedings 19*, pages 67–81. Springer.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

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.

Ruipu Luo, Qi Zhu, Qin Chen, Siyuan Wang, Zhongyu Wei, Weijian Sun, and Shuang Tang. 2021. Operation diagnosis on procedure graph: The task and dataset. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3288–3292.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860.

Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 56–64.

Kuntal Kumar Pal, Kazuaki Kashihara, Pratyay Banerjee, Swaroop Mishra, Ruoyu Wang, and Chitta Baral. 2021. Constructing flow graphs from procedural cybersecurity texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3945–3957.

Liang-Ming Pan, Jingjing Chen, Jianlong Wu, Shaoteng Liu, Chong-Wah Ngo, Min-Yen Kan, Yugang Jiang, and Tat-Seng Chua. 2020. Multimodal cooking workflow construction for food recipes. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1132–1141.

Chen Qian, Lijie Wen, Akhil Kumar, Leilei Lin, Li Lin, Zan Zong, ShuâĂŹang Li, and Jianmin Wang. 2020. An approach for process model extraction by multi-grained text classification. In *Advanced Information Systems Engineering: 32nd International Conference, CAiSE 2020, Grenoble, France, June 8–12, 2020, Proceedings 32*, pages 268–282. Springer.

Haopeng Ren, Yushi Zeng, Yi Cai, Bihan Zhou, and Zetao Lian. 2023. Constructing procedural graphs with multiple dependency relations: A new dataset and baseline. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8474–8486.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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 the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

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 the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022b. Promptmner: 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. Springer.

Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2022. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4525–4542.

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.

Frank F Xu, Lei Ji, Botian Shi, Junyi Du, Graham Neubig, Yonatan Bisk, and Nan Duan. 2020. A benchmark for structured procedural knowledge extraction from cooking videos. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 30–40.

Yoko Yamakata, Shinsuke Mori, and John A Carroll. 2020. English recipe flow graph corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5187–5194.

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 the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822.

Zi Yang and Eric Nyberg. 2015. Leveraging procedural knowledge for task-oriented search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 513–522.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352.

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154.

Li Yuan, Yi Cai, Jin Wang, and Qing Li. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11051–11059.

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 the AAAI conference on artificial intelligence*, volume 35, pages 14347–14355.

Huibin Zhang, Zhengkun Zhang, Yao Zhang, Jun Wang, Yufan Li, Ning Jiang, Xin Wei, and Zhenglu Yang. 2022. Modeling temporal-modal entity graph for procedural multimodal machine comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1179–1189.

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 the IEEE/CVF conference on computer vision and pattern recognition*, 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 the AAAI conference on artificial intelligence*, volume 32.

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.