An Effective Span-based Multimodal Named Entity Recognition with Consistent Cross-Modal Alignment

Yongxiu Xu[1,2], **Hao Xu**[1,2], **Heyan Huang**[3], **Shiyao Cui**[1]
Minghao Tang[1,2], **Longzheng Wang**[1,2], **Hongbo Xu**[1*]
[1] Institute of Information Engineering Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
hhy63@bit.edu.cn, {xuyongxiu, xuhao, tangminghao, wanglongzheng,hbxu}@iie.ac.cn

Abstract

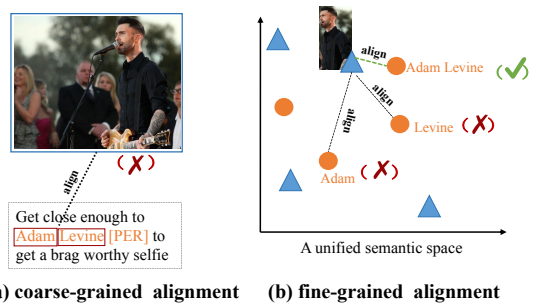
With the increasing availability of multimodal content on social media, consisting primarily of text and images, multimodal named entity recognition (MNER) has gained a wide-spread attention. A fundamental challenge of MNER lies in effectively aligning different modalities. However, the majority of current approaches rely on word-based sequence labeling framework and align the image and text at inconsistent semantic levels (whole image-words or regions-words). This misalignment may lead to inferior entity recognition performance. To address this issue, we propose an effective span-based method, named SMNER, which achieves a more consistent multimodal alignment from the perspectives of information-theoretic and cross-modal interaction, respectively. Specifically, we first introduce a cross-modal information bottleneck module for the global-level multimodal alignment (whole image-whole text). This module aims to encourage the semantic distribution of the image to be closer to the semantic distribution of the text, which can enable the filtering out of visual noise. Next, we introduce a cross-modal attention module for the local-level multimodal alignment (regions-spans), which captures the correlations between regions in the image and spans in the text, enabling a more precise alignment of the two modalities. Extensive experiments conducted on two benchmark datasets demonstrate that SMNER outperforms the state-of-the-art baselines.

Keywords: Multimodal named entity recogniton, Multimodal alignment, Multimodal fusion

1. Introduction

Named Entity Recognition (NER) involves identifying named entities within a given sentence and categorizing them into the pre-defined types. (Li et al., 2020). NER is a critical natural language processing task and plays a key component in information retrieval (Dietz, 2019), question answering (Min et al., 2021), knowledge graph (Zhao et al., 2022), etc. However, in practical scenarios such as social media platforms, the text is often limited, informal, and accompanied by images, which presents a significant challenge for the traditional text-based NER. Multimodal named entity recognition (MNER) has become a new direction and attracts widespread attention attributed to its excellent performance in entity recognition for social media posts.

MNER extends the traditional text-based NER by incorporating images as additional input (Zhang et al., 2018), which can offer complementary benefits to alleviate ambiguity in natural languages. However, MNER poses a fundamental challenge of effectively aligning information across two modalities: text and image. The existing MNER methods primarily utilize various attention networks (such as self-attention or cross-attention) to solve this challenge, which can be categorized into two strategies: coarse-grained alignment and fine-grained alignment, as shown in Figure 1.



(a) coarse-grained alignment (b) fine-grained alignment

Figure 1: An example of coarse-grained alignment and fine-grained alignment. Both of these two strategies align text with image at inconsistent semantic levels, leading to misalignment noises.

In the early stages, some efforts (Zhang et al., 2018; Moon et al., 2018) directly consider the entire image as global-level visual cues, which guides the words in the text to learn a vision-aware representation of a whole image, as shown in Figure 1(a). However, this coarse-grained alignment inevitably introduces image noise (e.g., background) and simultaneously results in the loss of some representative information. Subsequently, increasing studies (Zheng et al., 2020; Yu et al., 2020; Zhang et al., 2021; Xu et al., 2022; Jia et al., 2023) started focusing on fine-grained semantic alignment between text and images. These methods typically involve capturing the interactions between words in the

* Corresponding author

LREC-COLING 2024, pages 1063–1072

text and regions in the image in a unified semantic space, as shown in Figure 1(b). Actually, the regions of the objects in the image should align with the corresponding entity spans in the text rather than the individual words, as individual words may not adequately capture the overall semantics of an entity span. As shown in Figure 1, for the semantic representations of the two modalities, the regions of the person object *Adam Levine* in the image should have a higher similarity to the span “Adam Levine " in the text than the word “Adam" or word “Levine". Given that neither of the aforementioned alignment strategies successfully achieves consistent semantic alignment between text and images, resulting in the introduction of noise and subsequently inferior performance, we collectively refer to these issues as “misalignment noise".

Taking the considerations above, we propose an effective **S**pan-based **M**ultimodal **N**amed **E**ntity Recognition method, named **SMNER**, which regards MNER as a span-based classification task rather than a word-based sequence labeling task. SMNER is intensively designed for learning informative multimodal span representations by effectively aligning and fusing the information contained in text and image. SMNER consists of two key modules: a cross-modal information bottleneck (CMIB) module for global multimodal alignment and denoising, and a cross-modal attention (CMA) module for local multimodal alignment and interaction.

More specifically, motivated by the multi-view information bottleneck principle (Federici et al., 2020), we consider the text and image as two different views of the same posts. Firstly, we formulate the cross-modal global semantic alignment from an information-theoretic perspective by maximizing the mutual information and minimizing the distributional divergence between the two modalities. This module can bring the visual semantic distribution closer to the textual semantic distribution and filter out irrelevant information from visual representations. Secondly, for fine-grained multimodal alignment, we feed the contextual unimodal representations into a cross-modal attention module that captures the correlations between spans in the text and regions in the image. This module can enable a more precise alignment between two modalities and acquire informative cross-modal features. Finally, the obtained cross-modal features are aggregated effectively to enhance the representation of spans, thereby improving the performance of entity classification.

In summary, the main contributions of this paper are as follows:

- We propose SMNER[1], a span-based classification method for MNER, aiming to reduce the impact of misalignment and achieve more consistent multimodal alignment at two levels (image-text and regions-words, respectively). To the best of our knowledge, we are the first to explore the span-based MNER model for the issue of misalignment.
- We introduce two modules (CMIB and CMA) from the perspective of information-theoretic principle and cross-modal interaction, respectively. These modules work in synergy to generate more expressive cross-modal representations, enhancing the final entity classification performance.
- We conduct extensive experiments on two widely used MNER datasets to prove the effectiveness of our method. Experimental results show that SMNER outperforms the state-of-the-art models on both datasets.

[1]The code of our model will be released for future research.

2. Related Work

In this section, we review the related works of our method from: mulitmodal named entity recognition and information bottleneck.

2.1. Multimodal Named Entity Recognition

As multimodal data become increasingly popular on social media paltforms, starting with Moon et al. (2018); Lu et al. (2018); Zhang et al. (2018), MNER has attracted broad concerns in named entity recognition.

From the perspective of multimodal alignment and fusion. some studies (Moon et al., 2018; Zhang et al., 2018) tried to encode the entire image, which implicitly interacts the information of two modalities using attention mechanism. For example, Moon et al. (2018) proposed to utilize LSTM-CNN architecture that combines text with image information via a general modality attention, and Zhang et al. (2018) proposed an adaptive co-attention network to dynamically control the fusion of two modalities. Different from above works of using the whole image, subsequent works (Lu et al., 2018; Yu et al., 2020; Wu et al., 2020; Zheng et al., 2020; Zhang et al., 2021) primarily focused on combining the fine-grained regions visual information with the words information in text to boost the MNER performance. Lu et al. (2018) extracted the image regions that are most related to the text and utilized the attention-based model to implicitly interact the information of two modalities. Yu et al. (2020) introduced a multimodal interaction module designed to capture both image-aware word representation and word-aware visual representation. Zhang et al. (2021)

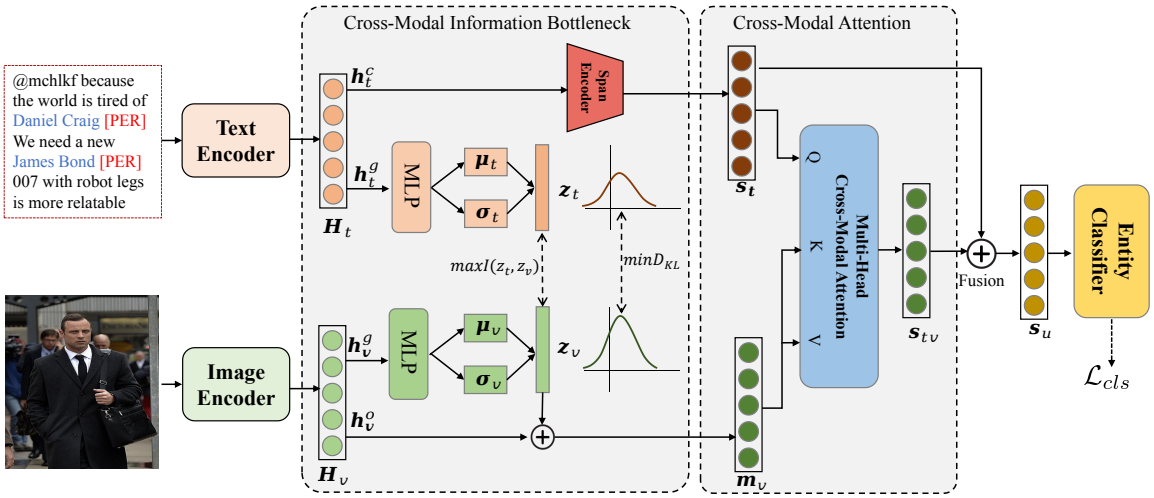


Figure 2: Model architecture overview of SMNER. The cross-modal information bottleneck module for global multimodal alignment and denoising, and the cross-modal attention module for local multimodal alignment and interaction.

exploited a unified multimodal graph to capture the interactions between words in the text and regions in the image.

Despite the studies above have achieved promising results, most of these methods ignore the problem of the visual noise caused by irrelevant images. More recently, Xu et al. (2022), Chen et al. (2022) and Zhang et al. (2023) alleviate this problem by text-image matching, hierarchical visual prefix and contrastive learning, respectively.

Different from the aforementioned methods, we focus on the noise caused by misalignment. Additionally, the above studies are under the word-based sequence labeling framework, whereas we utilize the span-based classification framework, ensuring the alignment and interaction between text and images at consistent semantic levels. It is worth noting that while Zhou et al. (2022a) also employed a span-based framework for MNER, it takes more concerns on mulimodal representations, overlooking the mulitmodal alignment and interaction.

2.2. Information Bottleneck

Information Bottleneck (IB) (Tishby and Zaslavsky, 2015) principle provides a theoretical framework for analyzing deep neural networks, which formulates the goal of representation learning as an information trade-off between predictive power and representation compression. Later, variational information bottleneck (VIB) (Alemi et al., 2016) bridges the gap between IB and deep learning with variational inference. More recently, Federici et al. (2020) provides a variant of IB which extends the ability of IB to the multi-view unsupervised setting, enabling the identification of superfluous information that is not shared by both views. Nowadays, owing to its capacity for learning minimal informative representations, IB has been extensively applied in computer vision (Peng et al., 2018), sentiment analysis (Mai et al., 2022), and natural language processing (Zhou et al., 2022b). Motivated by this, instead of directly applying IB principle to MNER task, we adopt the multi-view IB principle for enhancing the distribution consistency between the two modalities and filtering out irrelevant information from the images.

3. Method

3.1. Overview

Task Definition. Given the input pair containing a text sentence T and its associated image V, the goal of MNER is to detect entity spans from T, and classify them to corresponding entity types. Unlike the existing MNER models that regard MNER as a sequence labeling task, we regard MNER as a span classification task. Let $T = \{w_1, w_2, ..., w_N\}$ denote the input sentence with N words and the label for the text T is formulated as a set $Y = \{(s_k, e_k, y_k)\}_{k=1}^{N_e}$, where N_e is the number of the named entities. (s_k, e_k) is the span of an entity that corresponds to the phrase $T_{(s_k,e_k)} = \{w_{s_k}, w_{s_k+1}, ..., w_{e_k}\}$ and y_k represents the corresponding entity type that belongs to a predefined entity type set.

Model Architecture. The overall architecture of the SMNER is illustrated in Figure 2. Given image-text pairs, we first obtain the unimodal representations by the modal-specific encoder. Then, the representations of both text and image primarily

flow into two modules: 1) the cross-modal information bottleneck module for global multimodal alignment and denoising, 2) the cross-modal attention module for local multimodal alignment and interaction. Finally, we fuse the representations of the two modalities to obtain the multimodal span representation, and feed it into a entity classification layer to get the final predictions. These modules are trained simultaneously using an end-to-end framework.

3.2. Modal-specific Encoder

Given a multimodal dataset with $|\mathcal{D}|$ samples, is formulated as $\{\mathcal{X}, \mathcal{Y}\}_i^{|\mathcal{D}|}$. Each example $(x, y) \in \mathcal{D}$ contains the multimodal post $x = \{x^t, x^v\}$ and the task defined label y, where x^t and x^v are text and image respectively. For each post $\{x^t, x^v\}$, we first utilize the pre-trained models to obtain its unimodal representations $\boldsymbol{H}_t$ and $\boldsymbol{H}_v$, respectively.

Text Encoder. To precisely capture both the global and contextual representations, we adopt a pretrained BERT-base-uncased model (Kenton and Toutanova, 2019) as our textual encoder. Given a text $x^t = \{w_1, w_2, ..., w_n\}$ with n words, it needs to add a [CLS] token at the beginning and a [SEP] token at the end. We denote the text input as $x' = (w_0, w_1, w_2, ..., w_n, w_{n+1})$, where w_0 is the [CLS] token and w_{n+1} is the [SEP] token. We feed the input x' into BERT to obtain the factual output $\boldsymbol{H}_t = \{\boldsymbol{h}_0, \boldsymbol{h}_1, ..., \boldsymbol{h}_n\}$, where $\boldsymbol{h}_t^g = \boldsymbol{h}_0 \in \mathbb{R}^{1*d}$ represents the global text representation, $\boldsymbol{h}_t^c = \{\boldsymbol{h}_1, ..., \boldsymbol{h}_n\} \in \mathbb{R}^{n*d}$ is the contextual word representations for x^t, and d is the dimension of textual representations.

Image Encoder. To extract meaningful feature representations from images, we leverage a pretrained 152-layer ResNet (He et al., 2016) as the image encoder, which essentially splits each input image into m visual blocks. Specifically, we fist rescale the whole image into 224*224 piexls, and then feed them into ResNet to obtain visual representation $\boldsymbol{V} = \{\boldsymbol{v}_0, \boldsymbol{v}_1, ..., \boldsymbol{v}_m\} \in \mathbb{R}^{(m+1)*2048}$. To project the visual representations into the same dimension as the textual representations, we further convert $\boldsymbol{V}$ with a linear transformation: $\boldsymbol{H}_v = \boldsymbol{W}_v^T \boldsymbol{V}$, where $\boldsymbol{W}_v \in \mathbb{R}^{2048*d}$ is the weight matrix. Finally, we obtain $\boldsymbol{H}_v = \{\boldsymbol{h}_{v0}, \boldsymbol{h}_{v1}, ..., \boldsymbol{h}_{vm}\} \in \mathbb{R}^{(m+1)*d}$, where $\boldsymbol{h}_v^g = \boldsymbol{h}_{v0} \in \mathbb{R}^{1*d}$ is the representation of the whole image, and $\boldsymbol{h}_v^o = \{\boldsymbol{h}_{v1}, ..., \boldsymbol{h}_{vm}\} \in \mathbb{R}^{m*d}$ is the representations of the regional objects.

3.3. Cross-Modal Information Bottleneck

One challenge of multimodal alignment is how to establish a unified semantic representation space to bridge the semantic gap between two different modalities. Additionally, we should consider that text representations play a predominant role in MNER task, as all the entities to be recognized originate from text. To achieve these objectives, we present a Cross-Modal Information Bottleneck (CMIB) module from an information-theoretic perspective, aims to bring visual semantic distribution closer to the textual semantic distribution while filtering noise from the images.

Given $\boldsymbol{x}^t$ and $\boldsymbol{x}^v$ that are derived from the same post, they share the same predictive task for a target y. Therefore, in this paper, we consider $\boldsymbol{x}^t$ and $\boldsymbol{x}^v$ to be two views for the same object and suppose the $\boldsymbol{x}^t$ is sufficient for y. Motivated by the multi-view IB (Federici et al., 2020), we can subdivide $I(\boldsymbol{x}^v, \boldsymbol{z}^v)$ into two components by using the chain of mutual information (MI):

$$I(\boldsymbol{x}^v; \boldsymbol{z}^v) = \underbrace{I(\boldsymbol{z}^v; \boldsymbol{x}^t)}_{\text{consistent}} + \underbrace{I(\boldsymbol{x}^v; \boldsymbol{z}^v | \boldsymbol{x}^t)}_{\text{irrelevant}} \tag{1}$$

where $\boldsymbol{z}^t$ and $\boldsymbol{z}^v$ are the representations of the entire text $\boldsymbol{x}^t$ and image $\boldsymbol{x}^v$, respectively. $I(\boldsymbol{z}^v; \boldsymbol{x}^t)$ denotes the information that is consistent between two modalities, and $I(\boldsymbol{x}^v; \boldsymbol{z}^v | \boldsymbol{x}^t)$ denotes the information in $\boldsymbol{z}^v$ which is unique to $\boldsymbol{x}^v$ but is not predictable by observing $\boldsymbol{x}^t$, i.e., irrelevant information in the image.

We would like to define an objective function for the representation $\boldsymbol{z}^v$ of $\boldsymbol{x}^v$ that discards as much information as possible without losing any entity information. For this purpose, we should ensure that the representation $\boldsymbol{z}^v$ is sufficient for $\boldsymbol{x}^t$ (maximizing $I(\boldsymbol{z}^v; \boldsymbol{x}^t)$), and that the irrelevant information is discarded (minimizing $I(\boldsymbol{x}^v; \boldsymbol{z}^v | \boldsymbol{x}^t)$). So the loss function of the cross-modal information bottleneck in our model is defined as:

$$\mathcal{L}_{cmib} = -I(\boldsymbol{z}^v; \boldsymbol{x}^t) + \beta I(\boldsymbol{x}^v; \boldsymbol{z}^v | \boldsymbol{x}^t) \tag{2}$$

where β represents the Lagrangian multiplier introduced by the constrained optimization. With the gradients from back-propagation, semantic regularization can automatically enforce semantic agreement among heterogeneous representations.

It is challenging to compute the mutual information $I(\boldsymbol{z}^v; \boldsymbol{x}^t)$ and $I(\boldsymbol{x}^v; \boldsymbol{z}^v | \boldsymbol{x}^t)$ directly. The same as Federici et al. (2020), we use variational inference to compute a variational upper bound for $I(\boldsymbol{x}^v; \boldsymbol{z}^v | \boldsymbol{x}^t)$ as follow:

$$\begin{aligned} & I(\boldsymbol{x}^v; \boldsymbol{z}^v | \boldsymbol{x}^t) \\ & = D_{KL}(p(\boldsymbol{z}^v | \boldsymbol{x}^v) || p(\boldsymbol{z}^t | \boldsymbol{x}^t)) \\ & - D_{KL}(p(\boldsymbol{z}^v | \boldsymbol{x}^t) || p(\boldsymbol{z}^t | \boldsymbol{x}^t)) \\ & \leqslant D_{KL}(p(\boldsymbol{z}^v | \boldsymbol{x}^v) || p(\boldsymbol{z}^t | \boldsymbol{x}^t)) \end{aligned} \tag{3}$$

Therefore, we replace it in (2) with this upper bounder, which can be optimized via evaluating

the Kullback-Leibler (KL) divergence between the unimodal distributions approximated by two modal-specific Variational AutoEncoders (VAEs). Mathematically, the posterior distribution $p(\boldsymbol{z}^u|\boldsymbol{x}^u), u \in \{t, v\}$ of each unimodal representation is estimated as follows:

$$\begin{aligned} p(\boldsymbol{z}^t|\boldsymbol{x}^t) &= \mathcal{N}(\boldsymbol{z}^t|\mu(\boldsymbol{h}_t^g), \sigma(\boldsymbol{h}_t^g)) \\ p(\boldsymbol{z}^v|\boldsymbol{x}^v) &= \mathcal{N}(\boldsymbol{z}^v|\mu(\boldsymbol{h}_v^g), \sigma(\boldsymbol{h}_v^g)) \end{aligned} \tag{4}$$

where the mean μ and variance σ of Gaussian distribution can be obtained from the modal-specific multilayer perceptron (MLP) layers. Then, we use reparameterization trick to sample $\boldsymbol{z}^t$ and $\boldsymbol{z}^v$:

$$\begin{aligned} \boldsymbol{z}^t &= \mu(\boldsymbol{h}_t^g) + \sigma(\boldsymbol{h}_t^g) \times \boldsymbol{\epsilon} \\ \boldsymbol{z}^v &= \mu(\boldsymbol{h}_v^g) + \sigma(\boldsymbol{h}_v^g) \times \boldsymbol{\epsilon} \end{aligned} \tag{5}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is a standard normal Gaussian distribution.

Similarly, for $I(\boldsymbol{z}^v; \boldsymbol{x}^t)$, we calculate its lower bounder as follows:

$$\begin{aligned} I(\boldsymbol{z}^v; \boldsymbol{x}^t) &= I(\boldsymbol{z}^v; \boldsymbol{z}^t) + I(\boldsymbol{z}^v; \boldsymbol{x}^t|\boldsymbol{z}^t) \\ &\geqslant I(\boldsymbol{z}^v; \boldsymbol{z}^t) \end{aligned} \tag{6}$$

We maximize the mutual information (MI) $I(\boldsymbol{z}^v; \boldsymbol{z}^t)$ for modality pairs by infoNCE (Oord et al., 2018) which is used as a lower bounder on MI. Subsequently, it can be optimized by :

$$I(\boldsymbol{z}^v; \boldsymbol{x}^t) \geqslant -\mathbb{E}_p \left[f_d(\boldsymbol{z}^t, \boldsymbol{z}^v) - \mathbb{E}_{p'} log \sum_{\boldsymbol{z}^{v'}} f_d(\boldsymbol{z}^t, \boldsymbol{z}^{v'}) \right] \tag{7}$$

where $f_d(\cdot)$ is a discriminator function that measures the degree of consistence between text-image representations, $(\boldsymbol{z}^t, \boldsymbol{z}^v)$ refers to the positive sample which sampled text-image representations from the same input pair, and $(\boldsymbol{z}^t, \boldsymbol{z}^{v'})$ denotes in-batch negative example.

3.4. Cross-Modal Attention

Cross-Modal Attention module can further capture the fine-grained semantic interactions between two modalities after semantic regularization, enabling a more precise alignment of the two modalities.

Specifically, given the contextual unimodal representations $\boldsymbol{h}_t^c \in \mathbb{R}^{n*d}$ and $\boldsymbol{h}_v^o \in \mathbb{R}^{m*d}$, we first obtain the representations of span (i, j) by a span encoder, denoted as $\{\boldsymbol{s}_{(i,j)}|1 \leqslant i \leqslant j \leqslant n\}$ where $\boldsymbol{s}_{(i,j)} = [\boldsymbol{h}_i; \boldsymbol{h}_j; \boldsymbol{l}_{(i,j)}; \boldsymbol{p}_{(i,j)}]$, $\boldsymbol{l}_{(i,j)}$ is the span length embedding and $\boldsymbol{p}_{(i,j)}$ is the morph embedding for span (i, j). We could obtain the representations for all spans, and correspondingly acquire the contextual span representations $\boldsymbol{s}_t' \in \mathbb{R}^{n_s*d_s}$ as follows:

$$\boldsymbol{s}_t' = (\boldsymbol{s}_{(1,1)}, ..., \boldsymbol{s}_{(i,j)}, ..., \boldsymbol{s}_{(n,n)}) \tag{8}$$

where n_s is the number of spans and d_s is the dimension of span representation. Notice that, in this paper, the max span length is limited to 4, which can cover almost all entities. Similarly, we further convert $\boldsymbol{s}_t'$ into $\boldsymbol{s}_t$ via a linear transformation matrix $\boldsymbol{W}_s \in \mathbb{R}^{d_s*d}$, so $\boldsymbol{s}_t$ has the same dimension as the visual representations.

Additional, we fuse the global and local representation of images via concatenating, i.e., $\boldsymbol{m}_v = [\boldsymbol{z}_v; \boldsymbol{h}_v^o]$. Subsequently, we use the multi-head cross-modal attention to obtain the text-aware visual representations as follows:

$$\boldsymbol{\alpha}_i = softmax(\frac{[\boldsymbol{W}_{qi}\boldsymbol{s}_t]^T[\boldsymbol{W}_{ki}\boldsymbol{m}_v]}{\sqrt{d/h}}) \tag{9}$$

$$\boldsymbol{s}_{t\to v}^i = \boldsymbol{\alpha}_i[\boldsymbol{W}_{vi}\boldsymbol{m}_v]^T \tag{10}$$

$$\boldsymbol{s}_{t\to v} = \boldsymbol{W}[\boldsymbol{s}_{t\to v}^1; \boldsymbol{s}_{t\to v}^2; ...; \boldsymbol{s}_{t\to v}^h]^T \tag{11}$$

where h is the number of heads, $\{\boldsymbol{W}_{qi}, \boldsymbol{W}_{ki}, \boldsymbol{W}_{vi}\} \in \mathbb{R}^{d/h\times d}$ are the weight matrices for each query, key and value, $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ is the weight matrix for h-head attention. $\boldsymbol{\alpha}_i$ denotes the alignment score between each span in the text and visual block in the image. Finally, we obtain the final text-aware image representation $\boldsymbol{s}_{tv}$ by stacking a fully-connected feed-forward network (FFN) and two residual layers with layer-normalization (LN) as follows:

$$\begin{aligned} \boldsymbol{s}_{tv}' &= LN(\boldsymbol{s}_t + \boldsymbol{s}_{t\to v}) \\ \boldsymbol{s}_{tv} &= LN(\boldsymbol{s}_{tv}' + FFN(\boldsymbol{s}_{tv}')) \end{aligned} \tag{12}$$

3.5. Entity Classification

The input of the entity classifier is the multimodal span representation $\boldsymbol{s}_u$ obtained by concatenating the embeddings $\boldsymbol{s}_t$ and the embeddings $\boldsymbol{s}_{tv}$, as follows:

$$\boldsymbol{s}_u = [\boldsymbol{s}_t; \boldsymbol{s}_{tv}] \tag{13}$$

Then, we feed the final span representation $\boldsymbol{s}_u$ into a fully-connected network to predict the probabilities of entity types:

$$\hat{y} = softmax(MLP(\boldsymbol{s}_u)) \tag{14}$$

Since we regard MNER as a span-based multi-classification task, we apply the span-level cross entropy loss at training phase, as follows:

$$\mathcal{L}_{cls} = \sum_{(i,j,y)\in\{\mathcal{X},\mathcal{Y}\}^{|\mathcal{D}|}} -y_{(i,j)} log\hat{y}_{(i,j)} \tag{15}$$

where $\hat{y}_{(i,j)}$ is the prediction for the span $\boldsymbol{s}_{(i,j)}$, and $y_{(i,j)}$ is the ground truth.

By combining the loss functions from the main classification task and the cross-modal information bottleneck module, the overall loss function for SMNER is defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \cdot \mathcal{L}_{cmib} \tag{16}$$

Item	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
#PER	2217	552	1816	2943	626	621
#LOC	2091	522	1697	731	173	178
#ORG	928	247	839	1674	375	395
#MISC	940	225	726	701	150	157
#Tweets	4000	1000	3257	3373	723	723

Table 1: The statistical information of two MNER datasets.

where λ is the hyper-parameter to balance the different loss.

4. Experiments

4.1. Experimental Settings

Datasets. We compare SMNER with the existing methods on the two widely used benchmark datasets collected from social medias, including: Twitter-2015 (Lu et al., 2018) and Twitter-2017 (Zhang et al., 2018). There are four types of named entities including: Person (PER)m Location (LOC), Organization (ORG) and others (MISC) that are annotated in the text. Table 1 shows the detailed statistical information of the two datasets.

Implementation Configurations. We utilize Pytorch framework to conduct experiments with 1 Nvidia 3090 GPU. The dimension size d is set to 768. For span encoder, the dimensions of the span length embedding and morph embedding are set to 50 and 100, respectively. In the cross-modal information bottleneck module, we implement the modal-specific MLPs for obtaining μ and σ using 3 fully-connected layers with ReLU activation function in each layer. We manually tune the hyper-parameter β and λ, and achieve the best results with $\beta = 0.01$ and $\lambda = 0.1$. ReLU is used as the default activation function unless otherwise specified. All optimizations are performed with the AdamW optimizer, where the decay is 0.01, the learning rate is $1e-5$ and batch size is 16.

Baseline Methods. To demonstrate the effectiveness of our model, we mainly compare our model with three groups of state-of-the-art baselines. The first group contains several representative text-based NER approaches: BiLSTM-CRF (Huang et al., 2015), CNN-BiLSTM-CRF (Ma and Hovy, 2016), HBiLSTM-CRF (Lample et al., 2016), BERT-CRF (Kenton and Toutanova, 2019), SpanNER (Fu et al., 2021). The second group contains several competitive word-based MNER approaches with various alignment strategies: VG (Lu et al., 2018), AdaCoAtt (Zhang et al., 2018),UMT (Yu et al., 2020), UMGF (Zhang et al., 2021), MAF (Xu et al., 2022), HVPNeT (Chen et al., 2022), Debias (Zhang et al., 2023). And the third group contains two competitive span-based MNER approaches: SMVAE (Zhou et al., 2022a), MRC-MNER (Jia et al., 2023).

4.2. Main Results

Effectiveness. Table 2 represents the performance comparision between SMNER and all baselines. As shown in the table, SMNER outperforms all the compared methods on two datasets in terms of $F1$ and achieves the second-best results in terms of $Prec.$ and $Rec.$, which verifies the effectiveness of our methods. We also draw the following observations:

(1) Among text-based methods, span-based method performs better than CRF-based methods, by comparing SpanNER with BERT-CRF. This can be explained by the fact that the informal social text usually faces challenges like short length and out-of-vocabulary (OVV), which can be better addressed by the span-based methods.

(2) Visual features are generally helpful for text-based methods on both datasets by comparing the multimodal approaches with their corresponding text-based baselines (such as, VG vs. HBiLSTM-CRF, UMT vs. BERT-CRF and SMNER vs. SpanNER). This indicates the necessary of incorporating visual information for MNER task.

(3) Multimodal methods are not always superior to unimodal methods (e.g., compare VG or AdaCoAtt with BERT-CRF or SpanNER). Since both VG and AdaCoAtt directly incorporate the whole image as global visual clues to enrich word representation, which can inevitably introduce some misalignment noise.

(4) Among multimodal methods, SMNER (ours) not only significantly outperforms the word-based baselines, but also outperforms the current state-of-the-art span-based methods, confirming the advantage of the consistent multimodal alignment. Although De-bias method consider the misalignment issue, it obtained correlation information still by associating words in the text with regions in the image, and fails to align two modalities consistently, causing inferior performance. Both SMVAE and MRC-MNER are also span-based methods, but SMVAE ignores interactions between two modalities, MRC-MNER requires additional tools or external annotation data, which affects model's accuracy and adaptability.

Summarily, our method outperforms all these state-of-the-art methods. We attribute the performance gains of SMNER into its two advantages: 1) a well-aligned global semantic space between the two modalities achieved via the CMIB module, providing semantic regularization for the main task; 2) a more precise fine-grained cross-modal semantic

Modality	Methods	Twitter-2015			Twitter-2017		
		$Pre.$	$Rec.$	$F1$	$Pre.$	$Rec.$	$F1$
Text	BiLSTM-CRF	68.14	61.09	64.42	79.42	73.43	76.31
	CNN-BiLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37
	HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44
	SpanNER*	70.09	74.27	72.54	83.91	84.46	84.18
Text+Image (word-based)	VG	73.96	67.90	70.80	83.41	80.33	81.87
	AdaCoAtt	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
	UMGF	74.49	75.21	74.85	86.54	84.50	85.51
	MAF	71.86	75.10	73.42	86.13	86.38	86.25
	De-bias	74.45	76.13	75.28	87.59	86.11	86.84
	HVPNeT	73.87	**76.82**	75.32	85.84	**87.93**	86.87
Text+Image (span-based)	SMVAE	74.40	75.76	75.07	85.77	86.97	86.37
	MRC-MNER	**77.43**	72.15	74.70	**88.26**	85.65	86.94
	SMNER(ours)	75.34	76.81	**76.06**	88.15	87.47	**87.81**

Table 2: Performance comparison of different competitive baselines on two MNER datasets. The marker $*$ represents the models reproduced by us for MNER. The **bold** numbers indicate the best results and the numbers with underline indicate the second-best results. And all improvement of our model are statistically significant with $p \leq 0.05$ under t-test.

Methods	Twitter-2015		Twitter-2017		Size(M)
	Train(s)	Test(s)	Train(s)	Test(s)	
HVPNet	150.27	54.65	122.87	13.13	177.97
SMNER	**130.60**	**28.94**	**94.54**	**6.52**	**169.89**

Table 3: Comparison of average training and testing time (seconds for one epoch) and the number of parameters (Millions).

Methods	Twitter-2015			Twitter-2017		
	$Pre.$	$Rec.$	$F1$	$Pre.$	$Rec.$	$F1$
SMNER	**75.34**	**76.81**	**76.06**	**88.15**	**87.47**	**87.81**
w/o CMIB	75.21	74.60	74.90	85.50	86.85	86.17
w/o CMA	74.58	75.52	75.05	87.19	86.33	86.76
w/o CMIB+CMA	73.52	74.34	73.93	84.73	86.23	85.48

Table 4: Ablation study of MNER.

alignment and interaction achieved by CMA, further mitigating the noise introduced by misalignment.

Efficiency. We also compare SMNER with the state-of-the-art model (HVPNet) in terms of the runtime and model size. As shown in Table 3, the model size of SMNER is smaller than HVPNet, which indicates our model is simpler than HVPNet. On the other hand, both the training and testing speeds of our model are faster than HVPNet. Notice that, although our model enumerates the spans, social posts text are often short, so it does not affect the efficiency of our model.

4.3. Ablation Study

We do the ablation studies to further investigate the effectiveness of the main components in our model, as shown in Table 4.

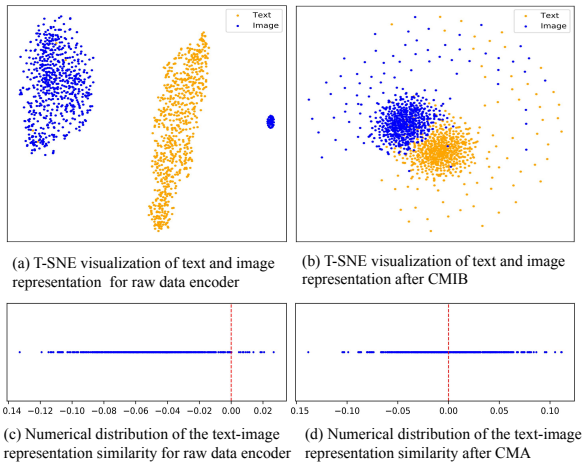
From the information-theoretic view, removing the CMIB module will significantly drop the performance, which justifies that directly use image embeddings produced by image-specific encoder may bring noise and further shows the importance of CMIB for alleviating the visual noise. From the cross-modal interaction view, discarding the CMA module also leads the performance drop, which indicates the usefulness of capturing the fine-grained correlations between two modalities. Furthermore, removing both CMIB and CMA also leads to an obvious performance drop, which indicates that both CMIB module and CMA module make important contributions to the final results.

4.4. Visualization for Modality-consistent

To indicated that our model can produce the consistent text-image representations by applying CMIB and CMA, we perform a text-image representation visualization and a numerical distribution visualization of text-image similarity, respectively, as shown in Figure 3. Specifically, the samples used in this analysis are from the test set of Twitter-2017. We gathered the representations of text and image at three stages: encoding of raw data, after employing the CIMB, and after employing the CMA.

By comparing Figure 3(a) and 3(b), we can observe that the semantic distribution of the text and image, after applying CMIB, exhibits a more consistent distribution shape, and the distribution distance significantly decreases. This phenomenon manifests the effectiveness of the cross-modal semantic alignment regularization by the CMIB component.

By comparing Figure 3(c) and 3(d), we can find that, without applying the CMA, the most values of



(a) T-SNE visualization of text and image representation for raw data encoder

(b) T-SNE visualization of text and image representation after CMIB

(c) Numerical distribution of the text-image representation similarity for raw data encoder

(d) Numerical distribution of the text-image representation similarity after CMA

Figure 3: The visualization for modality-consistent.

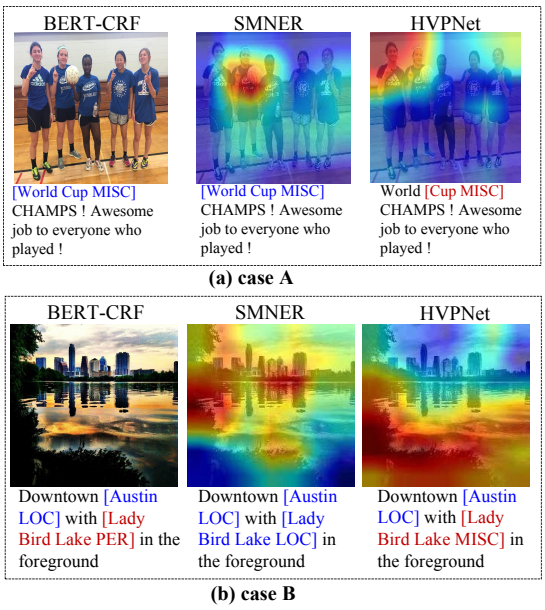


Figure 4: Two cases of the predictions by BERT-CRF, HVPNeT and SMNER. For visualization, both images and attention maps are scaled (from red:high to blue:low). In text, we use different colors mark entities: blue marks correct entities, red marks wrong ones.

text-image representation similarity are less than zero. While the majority of sample pairs have similarity values greater than 0, after applying the CMA. This phenomenon further indicates the effectiveness of the cross-modal alignment by the CMA component. Therefore, we can confirm that adding both the CMIB and CMA components could achieve the semantic agreements between textual and visual representations.

4.5. Case Study via Attention Visualization

To further validate the effectiveness of our proposed model, Figure 4 presents two cases with the cross-attention maps and the predicted results, produce by BERT-CRF (unimodal, text-based), HVPNet (multimodal, word-based) and SMNER (multimodal, span-based).

In case A, it is evident that both BERT-CRF (based on text-only) and SMNER successfully recognize the entity correctly. Additionally, the cross-attention of SMNER demonstrates a positive effect on its prediction (e.g., in the second image, the cross-attention of SMNER attends mostly to the regions of object *football* in the image which is highly related to text “World Cup"). However, the entity predicted by HVPNet is incorrect and incomplete, since lacking effective ways to map the semantic of the objects in the image to the spans in the text.

In case B, Only SMNER successfully predicts all entities, which further manifests the effectiveness of consistent cross-modal alignment for the MNER task. Since our model performs multimodal alignment at two consistent levels, it often captures complementary information from each level, and then combines them to predict the answer effectively. The BERT-CRF model relies solely on the textual information, may not be able to distinguish whether “Miss Bird Lake" refers to a location and instead, it may tend to predict it as a *PER* entity (based on the capitalization features of the words). From the cross-attention heatmaps, we can see that the HVPNet model allocates more interaction attention to the tree regions in the image, which results in the incorrect identification of “Miss Bird Lake" as a *MISC* entity.

All in all, these results further validate the assumption that our model is able to achieve more consistent text-image alignment for multimodal named entity recognition. However, we also found that when a sentence contains multiple types of entities, our model tends to make incorrect predictions. We speculate that the main reason is that the entity spans in the text do not have the corresponding regions in the image (a case of modality missing), and the textual information is insufficient, making it difficult to make accurate predictions. An effective solution would be to leverage external knowledge, which can be further explored as future work.

5. Conclusion

In this paper, we propose SMNER, an effective span-based method for MNER that achieves consistent alignment of text and image modalities at two levels: the global level and local level. Specifically, SMNER comprises two key modules: the CMIB module for the global semantic alignment and denoising, and the CMA module for the local semantic alignment and interaction. Through

ablation studies, we further demonstrate the contributions of both CMIB and CMA to our final performance. Extensive experimental studies illustrate that SMNER outperforms all the baselines on two public benchmarks. In the future, we will explore the application of CMIB in other multimodal tasks for information compression and denoising.

6. Acknowledgement

This work is supported by the National Key Research and Development Program of China (Grant No.2021YFB3100600).

7. Ethics Statement

This work provides a span-based method for multimodal named entity recognition in social media. We are committed to upholding the highest ethical standards throughout this research and ensuring the integrity and welfare of all involved parties. All authors are responsible for the content of this paper.

8. Bibliographical References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. In *International Conference on Learning Representations*.

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.

Laura Dietz. 2019. Ent rank: Retrieving entities for topical information needs through entity-neighbor-text relations. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 215–224.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning robust representations via multi-view information bottleneck. In *8th International Conference on Learning Representations*. OpenReview. net.

Jinlan Fu, Xuan-Jing Huang, and Pengfei Liu. 2021. Spanner: Named entity re-/recognition as span prediction. 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 7183–7195.

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

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

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.

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*, pages 4171–4186.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Changhee Lee and Mihaela van der Schaar. 2021. A variational information bottleneck approach to multi-omics data integration. In *International Conference on Artificial Intelligence and Statistics*, pages 1513–1521. PMLR.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified mrc framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859.

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

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*.

Sewon Min, Kenton Lee, Ming-Wei Chang, Kristina Toutanova, and Hannaneh Hajishirzi. 2021. Joint passage ranking for diverse multi-answer retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6997–7008.

Seungwhan Moon, Leonardo Neves, Vitor Carvalho, 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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. 2018. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. In *International Conference on Learning Representations*.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE.

Peng Wang, Xiaohang Chen, Ziyu Shang, and Wenjun Ke. 2023. Multimodal named entity recognition with bottleneck fusion and contrastive learning. *IEICE TRANSACTIONS on Information and Systems*, 106(4):545–555.

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.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: a general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1215–1223.

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

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

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023. Reducing the bias of visual objects in multimodal named entity recognition. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 958–966.

Yu Zhao, Han Zhou, Anxiang Zhang, Ruobing Xie, Qing Li, and Fuzhen Zhuang. 2022. Connecting embeddings based on multiplex relational graph attention networks for knowledge graph entity typing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4608–4620.

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.

Baohang Zhou, Ying Zhang, Kehui Song, Wenya Guo, Guoqing Zhao, Hongbin Wang, and Xiaojie Yuan. 2022a. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6293–6302.

Jie Zhou, Qi Zhang, Qin Chen, Liang He, and XuanJing Huang. 2022b. A multi-format transfer learning model for event argument extraction via variational information bottleneck. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1990–2000.