Dual-Gated Fusion with Prefix-Tuning for Multi-Modal Relation Extraction

Qian Li[1,2], **Shu Guo**[3], **Cheng Ji**[1,2], **Xutan Peng**[4], **Shiyao Cui**[5], **Jianxin Li**[1,2*], **Lihong Wang**[3]

[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing, China
[3]National Computer Network Emergency Response Technical Team/Coordination Center of China
[4]The University of Sheffield, South Yorkshire, UK
[5]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{liqian, jicheng, lijx}@act.buaa.edu.cn, guoshu@cert.org.cn,
x.peng@shef.ac.uk, cuishiyao@iie.ac.cn, wlh@isc.org.cn

Abstract

Multi-Modal Relation Extraction (MMRE) aims at identifying the relation between two entities in texts that contain visual clues. Rich visual content is valuable for the MMRE task, but existing works cannot well model finer associations among different modalities, failing to capture the truly helpful visual information and thus limiting relation extraction performance. In this paper, we propose a novel MMRE framework to better capture the deeper correlations of text, entity pair, and image/objects, so as to mine more helpful information for the task, termed as DGF-PT. We first propose a prompt-based autoregressive encoder, which builds the associations of intra-modal and inter-modal features related to the task, respectively by entity-oriented and object-oriented prefixes. To better integrate helpful visual information, we design a dual-gated fusion module to distinguish the importance of image/objects and further enrich text representations. In addition, a generative decoder is introduced with entity type restriction on relations, better filtering out candidates. Extensive experiments conducted on the benchmark dataset show that our approach achieves excellent performance compared to strong competitors, even in the few-shot situation.

1 Introduction

As a fundamental subtask of information extraction, relation extraction (RE) aims to identify the relation between two entities (Cong et al., 2022; Xue et al., 2022). Recently, there is a growing trend in multi-modal relation extraction (MMRE), aiming to classify textual relations of two entities as well as introduce the visual contents. It provides additional visual knowledge that incorporates multi-media information to support various cross-modal tasks such as the multi-modal knowledge graph construction (Zhu et al., 2022; Wang et al., 2019) and visual question answering systems (Wang et al., 2022; Shih et al., 2016).

*Corresponding author.

Figure 1: An example of the MMRE task. The task is to predict the relation of given entity pairs for the specific text and image which contains multiple objects.

Existing methods achieved considerable success by leveraging visual information (Zheng et al., 2021a; He et al., 2022; Chen et al., 2022) since the visual contents provide valuable pieces of evidence to supplement the missing semantics for MMRE. Previous work (Zheng et al., 2021a) introduced the visual relations of objects in the image to enrich text embedding via an attention-based mechanism. Next, HVPNet (Chen et al., 2022) used an object-level prefix and a multi-scale visual fusion mechanism to guide the text representation learning. Nevertheless, these methods primarily focus on the relations between objects and text and ignore the finer associations (entity pair, text, and image/objects). Furthermore, they usually suffered from the failure of identifying truly helpful parts/objects of the image to the corresponding entity pair on account of introducing all the objects. This may cause severe performance degradation of downstream tasks.

For multi-modal relation extraction, not all images or their objects are helpful for prediction. As illustrated in Figure 1, given three different inputs with the same relation *Member_of* and entity pair, each of the inputs contains a text, an image, and an entity pair. There are two situations: (a) The image is helpful for relation extraction. For entity pair *LeBron James* and *Lakers*, in the image *LeBron James*

wears the Lakers jersey revealing the implied relationship between the two entities. Therefore, we can improve relation extraction by considering the entity-pair relationships in visual information. (b) The image is unhelpful for the entity pair *LeBron James* and *Lakers* since it only contains *Lakers* object, rather than the association information for entity pairs. Furthermore, the image can provide an incorrect extraction signal, for example, the third image in Figure 1 shows that the relation between *LeBron James* and *Lakers* is more likely to be misjudged as *Coach_of* or *Owner_of*. Unhelpful visual content is prone to providing misleading information when predicting the relation. In general, it is necessary to identify the truly helpful visual information to filter the useless and misleading ones, but it is still under-explored.

To overcome the above challenges, we propose a novel MMRE framework DGF-PT to better incorporate finer granularity in the relations and avoiding unhelpful images misleading the model[1]. Specifically, we propose a prompt-based autoregressive encoder containing two types of prefix-tuning to integrate deeper associations. It makes the model focus on associations of intra-modal (between entity pair and text) by entity-oriented prefix and inter-modal (between objects and text) by the object-oriented prefix. In order to distinguish the importance of image/objects, we design a dual-gated fusion module to address the unhelpful visual data by utilizing interaction information via local and global visual gates. Later, we design a generative decoder to leverage the implicit associations and restrict candidate relations by introducing entity type. We further design joint objective to allow the distribution of representations pre and post-fusion to be consistent while enhancing the model to identify each sample in the latent space. Experimental results show that our approach achieves excellent performance in the benchmark dataset. Our contributions can be summarized as follows.

- We technically design a novel MMRE Framework to build deeper correlations among entity pair, text, and image/objects and distinguish helpful visual information.

- We propose a prompt-based autoregressive encoder with two types of prefixes to enforce the intra-modal and inter-modal association. We

design dual-gated fusion with a local object-importance gate and a global image-relevance gate to integrate helpful visual information.

- Experimental results indicate that the framework achieves state-of-the-art performance on the public multi-modal relation extraction dataset, even in the few-shot situation.

2 Related Work

Multi-modal relation extraction (MMRE) task, a subtask of multi-modal information extraction in NLP (Sun et al., 2021; Cong et al., 2020; Sui et al., 2021; Lu et al., 2022), aims to identify textual relations between two entities in a sentence by introducing visual content (Zheng et al., 2021a,b; Chen et al., 2022), which compensates for insufficient semantics and helps to extract the relations.

Recently, there are several works (Zheng et al., 2021a,b; Chen et al., 2022) beginning to focus on the multi-modal relation extraction technology. As the first work, MNRE (Zheng et al., 2021b) developed a multi-modal relation extraction baseline model. It demonstrates that introducing multi-modal information supplements the missing semantics and improves relation extraction performance in social media texts. Later, Zheng et al. (2021a) proposed a multi-modal neural network containing a scene graph modeling the visual relations of objects and aligning the relations between objects and text via similarity and attention mechanism. HVPNet (Chen et al., 2022) designs a visual prefix-guided fusion for introducing object-level visual information and further utilizes hierarchical multi-scaled visual features. Moreover, they introduce the information of all objects and thus cannot distinguish the truly helpful visual information, making it impractical to personalize the use of the image and further damaging their performance.

In the multi-modal relation extraction task, we note that images are naturally helpful information in the problem of multi-modal relation extraction. However, the potential for a differentiated use of image information in this task is under-explored. In this paper, we focus on the finer (intra-modal and inter-modal) association and manage to integrate truly useful visual information, promoting the exploitation of limited images. This enables bridging the gap through the transfer multi-modal relation extraction task into MLM pre-train mechanism (Devlin et al., 2019; Liu et al., 2021).
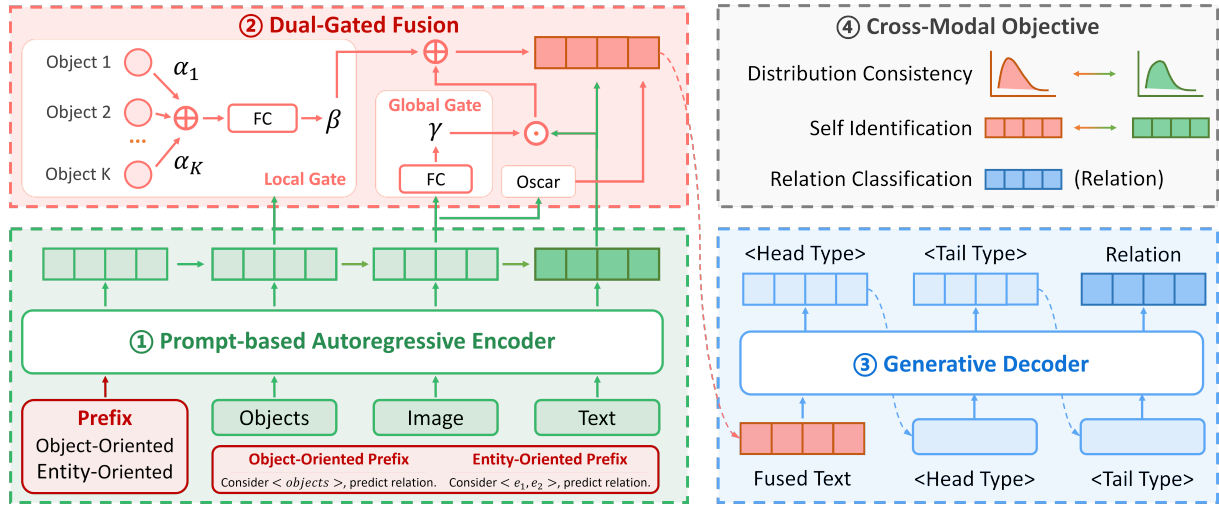

Figure 2: The DGF-PT framework for multi-modal relation extraction.

3 Problem Formulation

We provide the definition of MMRE. For a given sentence text $T = \{w_1, w_2, \ldots, w_L\}$ with L words as well as the image I related to the sentence, and an entity pair (e_1, e_2), an MMRE model takes (e_1, e_2, T, I) as input and calculates a confidence score $p(r_i|e_1, e_2, T, I)$ for each relation $r_i \in R$ to estimate whether T and I can reflect the relation r_i between e_1 and e_2. $R = \{r_1, \ldots, r_C, \text{None}\}$ is a pre-defined relation set, where "None" means that no relation is maintained for the mentions.

4 Framework

This section introduces our proposed DGF-PT framework, as shown in Figure 2. We first design the prompt-based autoregressive encoder to acquire the fine-grained representations (entity pair, object, and text), which contains two types of prefixes for integrating helpful information and characterizing intra-modal and inter-modal interactions. To avoid unhelpful visual information misleading the model, we design the dual-gated fusion module to distinguish the importance of image/objects via local and global gates. It also integrates the semantics of image transferred by Oscar (Li et al., 2020). The Fusion module outputs an enhanced representation. Later, the generative decoder is proposed for relation prediction, leveraging the implicit associations and restricting candidate relations by introducing entity types. Finally, we design the joint objective including distribution-consistency constraint, self-identification enhancement, and relation classification for model optimization.

4.1 Prompt-based Autoregressive Encoder

In order to acquire the finer granularity in the associations (entity pair, objects, and text), we propose a prompt-based autoregressive encoder. After initialization, two specific prefix-tuning strategies are implemented to guide the encoder to attend to task-relevant inter-/intra-modal associations. Subsequently, prefixes, objects, image, and text are progressively fed into an autoregressive encoder stage by stage to obtain fine-grained representations for use in subsequent fusion module (Section 4.2).

4.1.1 Initialization

Given text T, the word embeddings $\mathbf{w} \in \mathbb{R}^{1 \times N}$ are obtained through the GPT-2 model (Radford et al., 2019) and then fed into a fully-connection layer, where N is the word dimension. The initial text representation is $\mathbf{T} = [\mathbf{w}_1; \mathbf{w}_2; \ldots; \mathbf{w}_L] \in \mathbb{R}^{L \times N}$.

Given an image I, the global image feature $\mathbf{I} \in \mathbb{R}^{M \times N}$ is obtained by VGG16 (Simonyan and Zisserman, 2015) with a fully-connection layer, transferring the feature into M-block N-dimensional vectors. We then extract object features using Faster R-CNN (Ren et al., 2015) and select top-K objects using the ROI classification score. Each object feature is obtained by the average grouping of ROI regions. The initial object representation is $\mathbf{O} = [\mathbf{o}_1; \mathbf{o}_2; \ldots; \mathbf{o}_K] \in \mathbb{R}^{K \times N}$.

4.1.2 Object & Entity Oriented Prefixes

Utilizing the advantages of prefix-tuning, the pre-trained encoder (e.g., GPT-2 as our encoder) can be guided to learn task-specific features for fast adaptation to the MMRE task (Liu et al., 2021). However, the design of appropriate prefixes for finer associa-

tions learning in the MMRE task remains an open research question, and the direct use of prefixes from other tasks is not reasonable. Therefore, we construct two types of prefixes, an object-oriented prefix for inter-modal relevance (objects & text) and an entity-oriented prefix for intra-modal correlations (entity pair & text), encouraging the encoder to leverage text as a medium to strengthen multi-granular associations to acquire enhanced semantic representations.

Object-Oriented Prefix. Given that objects related to entities are indeed useful information for the MMRE task, we propose an object-oriented prefix, termed as $\mathcal{P}_o(\cdot)$, which provides guidance information of inter-modal relevance to the encoder. For the input text T, we define the following pattern "Consider $\langle objects \rangle$, predict relation.", where $\langle objects \rangle$ means the objects relevant to the entity pair of T which is different for each input. It emphasizes specific key textual contents and introduces the visual features of relevant objects.

Entity-Oriented Prefix. Due to the visual information may be incomplete or misleading, we argue that only an object-oriented prefix is insufficient to capture classification information. Thus, we propose an entity-oriented prefix, termed as $\mathcal{P}_e(\cdot)$, to capture intra-modal association to adapt the task. We define the following pattern "Consider $\langle e_1, e_2 \rangle$, predict relation.", where $\langle e_1, e_2 \rangle$ is the entity pair to predict the relation.

4.1.3 Multi-Stage Autoregressive Encoder

Prompt-based learning keeps the parameters of the whole PLM frozen and prepends the prefixes before the task inputs (Liu et al., 2021). The bidirectional encoder (e.g., BERT) cannot effectively integrate the proposed dual-gated fusion module (Section 4.2) in model testing. Therefore, we deploy a unidirectional encoder (e.g., GPT and GPT-2) and design multiple stages to integrate multi-granular textual and visual knowledge, where the prefixes, objects, image, and text are fed stage by stage.

First stage (S_1). The input of the first stage S_1 contains the prefixes and objects to learn the relevance from local granularity and obtain the representations of objects. To introduce task-related prefix knowledge, the two types of trainable prefixes $\mathcal{P}_o(\cdot)$ and $\mathcal{P}_e(\cdot)$ are prepended before the input sequence as the prefix tokens, obtained through the GPT-2 vocabulary. In S_1, the encoder learns the representations of objects and updates the prefix tokens of

each model layer:

$$
\begin{aligned}
\mathcal{P}_o^*(\mathbf{o}_{e_1}, \mathbf{o}_{e_2}), \mathcal{P}_e^*(\mathbf{T}[e_1], \mathbf{T}[e_2]), \mathbf{h}_o &= \mathcal{S}_1(\cdot) \\
&= \text{Encoder}(\mathcal{P}_o(\mathbf{o}_{e_1}, \mathbf{o}_{e_2}), \mathcal{P}_e(\mathbf{T}[e_1], \mathbf{T}[e_2]), \mathbf{O}),
\end{aligned}
\tag{1}
$$

where $\mathcal{P}_o^*(\cdot)$ and $\mathcal{P}_e^*(\cdot)$ are updated prefixes after S_1, $\mathbf{h}_o$ is the representations of objects, and $\mathbf{o}_{e_1}$ and $\mathbf{o}_{e_1}$ are the initial embeddings of entities e_1 and e_2. After the S_1 stage, the object information is introduced into the prefix embedding.

Second stage (S_2). The inputs of the second stage S_2 are the outputs of the first stage $\mathcal{S}_1$ (including the updated prefixes and the representations of objects) and the image feature $\mathbf{I}$ to get the representations of images $\mathbf{h}_i$. We hope the model can learn to capture the inter-modal relevance from global granularity, which is useful information for relation extraction that may improve performance. Thus, we introduce the image information in S_2. The S_2 embedding is therefore updated as:

$$
\mathbf{h}_i = \mathcal{S}_2(\cdot) = \text{Encoder}(\mathcal{S}_1(\cdot), \mathbf{I}).
\tag{2}
$$

Third stage (S_3). To learn text representation $\mathbf{h}_t$, the third stage inputs $\mathcal{S}_2$ and text T using interactive objects and images.

$$
\mathbf{h}_t = \mathcal{S}_3(\cdot) = \text{Encoder}(\mathcal{S}_2(\cdot), \mathbf{T}).
\tag{3}
$$

4.2 Dual-Gated Fusion

Unhelpful/Task-irrelevant information in image is often ignored by simply utilizing all objects for aggregation. To solve this, we propose dual-gated fusion to effectively integrate helpful visual information while filtering out misleading information. This module utilizes local and global gates to distinguish the importance and relevance of image/objects, and filters out task-irrelevant parts. By integrating semantic information of the image, a final fused representation containing associations among image, object, and text is obtained.

Specifically, the local object-importance gate vector $\boldsymbol{\beta}$ by the local object features and the global image-relevance gate vector $\boldsymbol{\gamma}$ by the global image features are calculated as:

$$
\alpha_k = \frac{\cos(\mathbf{h}_t, \mathbf{h}_o[k])}{\sum_{j=1}^{K} \cos(\mathbf{h}_t, \mathbf{h}_o[j])},
\tag{4}
$$

$$
\boldsymbol{\beta} = \text{FC}_\beta(\overline{\mathbf{h}}_o) = \text{FC}_\beta\left(\sum_{k=1}^{K} \alpha_k \mathbf{h}_o[k]\right),
\tag{5}
$$

$$
\boldsymbol{\gamma} = \tanh(\text{FC}_\gamma(\mathbf{h}_i)),
\tag{6}
$$

where FC is fully-connected layer and $\mathbf{h}_o[k]$ is the k-th object in the object set O. $\overline{\mathbf{h}}_o$ calculates attention between the selected top-K objects of divergent modalities. Subsequently, the textual characteristic of fusion $\tilde{\mathbf{h}}_t$ is calculated by

$$\tilde{\mathbf{h}}_t = \text{MLP}\left(\mathbf{h}_t \odot \boldsymbol{\gamma} + \boldsymbol{\beta}\right) + \mathbf{h}_t, \qquad (7)$$

where MLP is multilayer perceptron and $\odot$ means hadamard product.

In order to further integrate the semantics of visual information, we use Oscar (Li et al., 2020) to transfer $\mathbf{h}_i$ into a text description $\tilde{\mathbf{h}}_{i2t}$ for each image, using objects as anchor points to align visual and textual features in a common space. It learns multi-modal alignment information of entities from a semantic perspective. The detail is given in Appendix A.

While local representations can capture valuable clues, global features provide condensed contextual and high-level semantic information. Given this insight, we leverage the global information from one modality to regulate the local fragment of another modality, enabling the entity to contain semantic information and filter out irrelevant visual information. The final fused representation is:

$$\tilde{\mathbf{h}}_t = \tilde{\mathbf{h}}_t + \tilde{\mathbf{h}}_{i2t} \odot \delta, \qquad (8)$$

where δ is trade-off factor between text embedding $\tilde{\mathbf{h}}_t$ and the inter-modal text representation $\tilde{\mathbf{h}}_{i2t}$.

4.3 Generative Decoder

To leverage the implicit associations and restrict candidate relations by introducing entity type, we design a generative decoder.

The type of entity pair is helpful for relation classification. For example, the relation of entity type *Person* and *Organization* must not be *born* and *friend*, but maybe *CEO* and *staff*. Thus, we introduce head type $\mathbf{T}_{e_1}^t$ and tail type $\mathbf{T}_{e_2}^t$ one by one to leverage the implicit associations and restrict candidate relations.

To maintain the consistency of the relation extraction task with the MLM pre-trained model, we use the generative decoder to predict the relation. The prediction of the generative decoder is:

$$\mathbf{h}_{e_1}^t, \mathbf{h}_{e_2}^t, \mathbf{r} = \text{Decoder}\left(\mathcal{S}_3(\cdot), \mathbf{T}_{e_1}^t, \mathbf{T}_{e_2}^t\right), \quad (9)$$

where $\mathbf{h}_{e_1}^t, \mathbf{h}_{e_2}^t$ are the representation of types, and $\mathbf{r}$ is the representation of relation.

4.4 Joint Objective

In order to address distribution consistency within the dual-gated fusion module, we introduce the distribution-consistency constraint loss, which is applied on a single-sample basis. Additionally, to meet the need for inter-sample identification, we propose self-identification enhancement loss. The overall joint objective is then formed by combining the relation classification loss with the aforementioned constraints.

Distribution-Consistency Constraint. In order to ensure the dual-gated fusion module effectively integrates helpful visual features while avoiding the introduction of task-irrelevant information, we introduce distribution-consistency constraint to measure and optimize the change in representation distribution pre and post-fusion. Thus, we propose to use KL divergence to measure the distance between the probability distribution of $\tilde{\mathbf{h}}_t$ and $\mathbf{h}_t$, which is equal to calculating the cross-entropy loss over the two distributions:

$$\mathcal{L}_d(\theta) = \text{KL}\left(p_\theta(\boldsymbol{r}|\tilde{\mathbf{h}}_t)\|p_\theta(\boldsymbol{r}|\mathbf{h}_t)\right) \qquad (10)$$

$$= \sum_{\boldsymbol{r} \in R} p_\theta(\boldsymbol{r}|\tilde{\mathbf{h}}_t) \log p_\theta(\boldsymbol{r}|\mathbf{h}_t). \qquad (11)$$

Self-Identification Enhancement. The MMRE task requires the model to have the ability to correctly classify relations from individual samples. However, relation labels are unevenly distributed or lacking in the real world. Therefore, further enhancement is needed. We design a negative-sampling-based self-supervised loss function to enhance the model. Moreover, the dual-gated fusion module is treated as the augmentation function leveraging the modality information. Specifically, textual representation $\mathbf{h}_t$ and fused representation $\tilde{\mathbf{h}}_t$ are the mutually positive samples:

$$\mathcal{L}_s = [s(x, \tilde{x}) - s(x_n, \tilde{x})]_+ + [s(x, \tilde{x}) - s(x, \tilde{x}_n)]_+, \qquad (12)$$

where $\{x, \tilde{x}\}$ are $\{\mathbf{h}_t, \tilde{\mathbf{h}}_t\}$, $[a]_+ = \max(a, 0)$, and $s(\cdot, \cdot)$ is the cosine similarity. x_n and $\tilde{x}_n$ are the hardest negatives of $\mathbf{h}_t$ and $\tilde{\mathbf{h}}_t$ in a mini-batch based on a similarity-based measurement.

Relation Classification. The loss for relation classification by the negative log-likelihood function is as follows:

$$\mathcal{L}_c = -\log p(\mathbf{r}|\mathbf{h}_t, \mathbf{h}_i, \mathbf{h}_t[e_1], \mathbf{h}_t[e_2]), \qquad (13)$$

Model Type	Model Name	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
Text-based RE Models	**Glove+CNN (Zeng et al., 2014)**	70.32	57.81	46.25	51.39
	PCNN (Zeng et al., 2015)	72.67	62.85	49.69	55.49
	MTB (Soares et al., 2019)	72.73	64.46	57.81	60.96
MMRE Models	**BERT+SG (Devlin et al., 2019)**	74.09	62.95	62.65	62.80
	BERT+SG+Att. (Devlin et al., 2019)	74.59	60.97	66.56	63.64
	VisualBERT (Li et al., 2019)	-	57.15	59.45	58.30
	MEGA (Zheng et al., 2021a)	<u>76.15</u>	64.51	68.44	66.41
	HVPNet (Chen et al., 2022)	-	<u>83.64</u>	<u>80.78</u>	<u>81.85</u>
Ours	**DGF-PT (BERT Encoder)**	79.82 (↑ 3.67)	79.72 (↑ -3.92)	78.63 (↑ -2.15)	79.24 (↑ -2.61)
	DGF-PT (GPT Encoder)	82.03 (↑ 5.88)	81.23 (↑ -2.41)	82.48 (↑ 1.70)	82.09 (↑ 0.24)
	DGF-PT (GPT-2 Encoder)	**84.25 (↑ 8.10)**	**84.35 (↑ 0.71)**	**83.83 (↑ 3.05)**	**84.47 (↑ 2.62)**

Table 1: Main experiments. The best results are highlighted in bold, "–" means results are not available, and the underlined values are the second-best result. "↑" means the increase compared to the underlined values.

where $\mathbf{r}$ is the relation between the head entity e_1 and the tail entity e_2 for $\mathbf{h}_x$, and $\mathbf{h}_t[e_1]$, $\mathbf{h}_t[e_2]$ are the representations of the two entities. Finally, the overall loss function of our model is as follows.

$$\mathcal{L} = \lambda_d \mathcal{L}_d + \lambda_s \mathcal{L}_s + \lambda_c \mathcal{L}_c, \qquad (14)$$

where λ_d, λ_s, and λ_c are trade-off parameters. We optimize all training inputs in a mini-batch strategy.

5 Experiment

5.1 Dataset and Evaluation Metric

We conduct experiments in a multi-modal relation extraction dataset MNRE (Zheng et al., 2021b), crawling data from Twitter[2]. The MNRE includes $15,484$ samples and $9,201$ images. It contains 23 relation categories. As previous work (Zheng et al., 2021a) recommended, the MNRE dataset is divided into $12,247$ training samples, $1,624$ development samples, and $1,614$ testing samples. We report the official Accuracy (Acc.), Precision (Prec.), Recall, and F1 metrics for relation evaluation.

5.2 Comparision Methods

We compare our method with three text-based RE models and five MMRE models.

Text-based RE Models. We first consider a group of representative text-based RE models, which do not introduce image information, for modeling the connection of words in the sentence: (1) Glove+CNN (Zeng et al., 2014) is a CNN-based model with additional position embeddings to utilize the position association. (2) PCNN (Zeng et al., 2015) is a RE method utilizing external knowledge graph with a distant supervision manner to build

connection by the graph. (3) Matching the Blanks (MTB) (Soares et al., 2019) is a BERT-based RE model to learn context correlation.

MMRE Models. We further consider another group of previous approaches for MMRE to integrate visual information: (4) **BERT+SG** (Devlin et al., 2019) concatenates BERT representations with visual content, which is obtained by the pre-trained scene graph tool (Tang et al., 2020) to learn the connection between text and the object of the image. (5) **BERT+SG+Att** adopts an attention mechanism to compute the relevance between the textual and visual features. (6) **Visual-BERT** (Li et al., 2019) is a single-stream encoder, learning cross-modal correlation in a model. (7) **MEGA** (Zheng et al., 2021a) considers the relevance from the structure of objects in the image and semantics of text perspectives with graph alignment. (8) **HVPNet** (Chen et al., 2022) introduces an object-level prefix with a dynamic gated aggregation strategy to enhance the correlation between all objects and text.

In contrast to these methods, our approach incorporates the correlation between entity pairs, text, and visual information, and effectively identifies useful visual information.

5.3 Implementation Details

For all baselines, we adopt the best hyperparameters and copy results reported in the literature (Zheng et al., 2021a,b; Chen et al., 2022).

We used PyTorch[3] as a deep learning framework to develop the MMRE. The BERT and GPT-2[4] are for text initialization and the dimension is set at

---

[2]https://archive.org/details/twitterstream

[3]https://pytorch.org/

[4]https://github.com/huggingface

768. The VGG version is VGG16[5]. We use Faster R-CNN (Ren et al., 2015) for image initialization and set the dimension of visual objects features at 4096. For hyper-parameters, the best coefficients $\lambda_d, \lambda_s, \lambda_c$ are 2, 2 and 3. The best δ is 0.4. See Appendix B for more details on model training.

5.4 Main Results

To verify the effectiveness of our model, we report the overall average results in Table 1.

From the table, we can observe that: 1) Our model outperforms text-based RE models in terms of four evaluation metrics, indicating the beneficial impact of visual information on relation extraction and the necessity of its integration. 2) Compared to MMRE baselines, our model achieves the best results. Specifically, our model improves at least 2.62% in F1 and 8.10% in Acc., respectively. These results indicate that our method for incorporating and utilizing visual information is superior and effective. 3) Compared to different encoders (e.g., BERT and GPT), the GPT and GPT-2 achieve better results. It demonstrates that the generative encoder can integrate effective visual features more effectively, which is more suitable for the task. For the generative model, the performance is sensitive to the order of input. Thus, we discuss the effect of the order of text, image, and objects in Appendix C.

5.5 Discussion for Model Variants

For a further detailed evaluation of the components of our framework, we performed ablation experiments and reported the results in Table 2. E-P means entity-oriented prefix and O-P means object-oriented prefix. "↓" means the average decrease of all four metrics compared to our model.

Discussions for core module. To investigate the effectiveness of each module, we performed variant experiments, showcasing the results in Table 2. From the table, we can observe that: 1) the impact of the prefixes tends to be more significant. We believe the reason is that the multiple prompts characterize modality interactions, helping for providing more visual clues. 2) By removing each module, respectively, the performance basically decreased. Compared to joint objective modules, the dual-gated fusion is significantly affected. It demonstrates the effectiveness of knowledge fusion introducing useful visual content and addressing

Variants	Acc.	Prec.	Recall	F1	△ Avg
DGF-PT (Ours)	**85.25**	**84.35**	**83.83**	**84.47**	-
w/o All Prefixes	83.32	82.93	83.42	82.35	↓ 1.47
w/o E-P	84.62	83.31	82.83	82.63	↓ 1.13
w/o O-P	84.05	83.20	83.29	83.39	↓ 0.99
w/o dual-gated fusion	84.24	83.56	82.74	83.26	↓ 1.03
w/o joint objective	84.49	83.90	83.26	84.05	↓ 0.55
repl. All Prefixes in S_2	83.29	82.40	83.17	82.24	↓ 1.70
repl. All Prefixes in S_3	84.10	83.24	82.02	82.29	↓ 1.56
repl. E-P in S_2	84.28	83.20	82.41	83.93	↓ 1.02
repl. E-P in S_3	84.37	83.82	83.03	83.41	↓ 0.82
repl. O-P in S_2	84.12	83.71	83.15	83.84	↓ 0.77
repl. O-P in S_3	84.03	83.65	82.20	83.73	↓ 1.07
repl. E-P in S_2 & O-P in S_3	84.09	83.43	82.81	84.15	↓ 0.86
repl. E-P in S_3 & O-P in S_2	84.76	84.24	83.38	84.20	↓ 0.33

Table 2: Variant experiments on different orders to introduce the two prefixes. "w/o" means removing the corresponding module from the complete model. "repl." means replacing the stage of introducing prefix.

noise visual data. All observations demonstrate the effectiveness of each component in our model.

Discussions for the stage of prefix. We explore the effects by introducing the prefixes at a different stage in the encoder, as shown in Table 2. From the table, we can observe that: 1) Compared to feed all Prefixes in the S_2, S_3 stage, the S_1 stage is more effective. It demonstrates that early introducing prefixes may integrate more helpful visual classification information. 2) When the O-P is fixed, and the E-P is fed into the S_3 stage, the performance of our model is the best compared to that introduced in S_2 stages. It demonstrates that the E-P is set nearly to the text features helpful to introduce intra-modal association. 3) When we fix the E-P, and the O-P is introduced in the S_2 stage, which is fed nearly to objects, the performance achieves better than S_3. It demonstrates that the O-P is nearly to object features, which can capture more useful local information for relation classification. 4) When we change the stage of the two prefixes, the performance achieves better as the E-P in S_3 and the O-P in S_2. All observations demonstrate that "E-P in S_1 & O-P in S_1" is the best schema to introduce intra-modal association and inter-modal relevance.

5.6 Discussions for Image Information

To further investigate the impact of images on all the compared methods, we report the results by deleting different proportions of images, as shown in Figure 3. From the figure we can observe that: 1) On all metrics, the more proportion of images
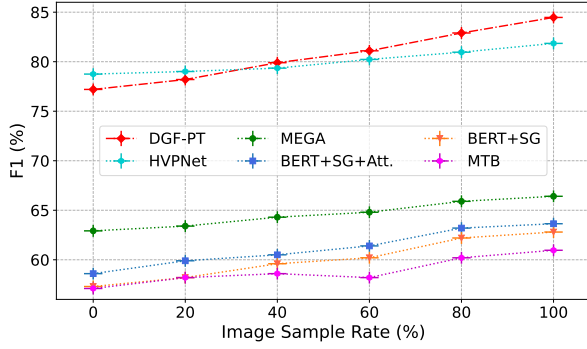
---

[5] https://github.com/machrisaa/tensorflow-vgg

Figure 3: Different proportions of images.

Figure 4: Impact of differences in sample number.

were introduced, the better the model performed. It demonstrates that more images provide more meaningful information for relation classification and utilize visual information more effectively. 2) Compared to other methods, our model performs best (except for HVPNet in introducing only 0%-20% percentage image data). It demonstrates that without visual information, our model is still more capable to capture intra-modal associations for relation classification. 3) Compared to the HVPNet with an object-level prefix, our model performs poorly with fewer visual data. The main reason is that the prefix is outstanding in the few-shot situation and incorporating deeper correlations of our model needs enough visual information compared to the prefix-based method. Our model performs better than HVPNet with the visual data increase. Observations indicate that our model incorporates visual features more effectively.

5.7 Discussions for Sample Number

We investigate the impact of the sample number of different relations. To do so, we divide the dataset into multiple blocks based on the sample number of each relation and evaluate the performance by varying the sample number of relations in [0, 1000] compared with the outstanding baselines, as shown in Figure 4. From the figure, we can observe that: 1) The increasing of sample number performance improvements to all methods. The main reason is that the smaller the sample number, the more difficult it is to distinguish the relation. 2) Our model could also advance the baseline methods with the decrease in sample number, demonstrating the superiority of our method in tackling the relation with fewer samples. This phenomenon confirms the prefixes are suitable for few-shot situations. All the observations demonstrate that our method reduces the impact of the sample number.
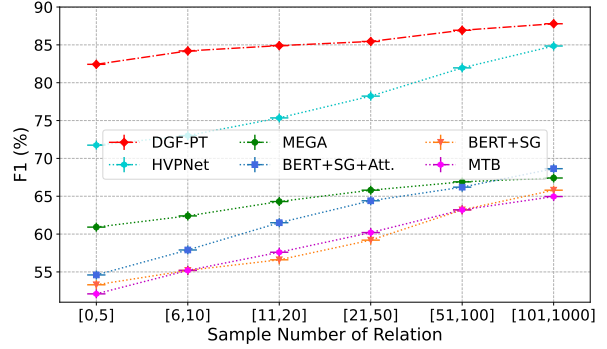
Text Input: *P.Wilson* and *Dagmara* went to the *Boundaries Premiere* to support *Vera*, ...

	Head Entity	Tail Entity	Truth Relation	Prediction Result			
				DGF-PT	HVPNet	MEGA	BERT+SG+Att.
	P.Wilson	Dagmara	couple	√	√	√	√
	P.Wilson	Boundaries Premiere	present_in	√	√	√	×
	Vera	P.Wilson	peer	√	×	×	×

Figure 5: Case study of the same image and text with different entity pairs. √/×: correct/wrong prediction.

5.8 Case Study

To illustrate our model can effectively identify useful visual information, we provide an example involving various entity pairs. As shown in Fig. 5, the helpful information varies depending on the entity pair. From the figure, we can observe that: 1) Our model achieves superior performance across different entity pairs, demonstrating its ability to effectively extract useful visual information while avoiding the negative influence of unhelpful information on prediction. 2) When presented with the entity pair of *Vera* and *P.Wilson* that contains limited useful visual information, our model remains the best, while other baselines make incorrect predictions. These observations further demonstrate the effectiveness of our model in leveraging visual information while avoiding the negative influence of unhelpful information on predictions.

6 Conclusion

We propose DGF-PT, a novel multi-modal relation extraction framework, to capture deeper correlations among entity pair, text, and image/objects and integrate more helpful information for relation extraction. Our framework effectively integrates intra-modal and inter-modal features, distinguishes helpful visual information, and restricts candidate relations. Extensive experiments conducted on the benchmark dataset show that our approach achieves excellent performance.

Limitations

Our work overcomes visual noise data that limit extraction performance, incorporating multi-modal knowledge of different levels. Empirical experiments demonstrate that our method avoids noise data misleading the MMRE model. However, there are still some limitations of our approach can be summarized as follows:

- Due to the limitation of the existing MMRE datasets, we only experiment on two modalities to explore the influence of image features. We will study more modalities in future work.

- Our method neglects the multiple relations for an input, which may not consider the multiple semantics of entities. We leave the multiple relation extraction method for future work.

Ethics Statement

In this work, we propose a new MMRE framework that captures deeper correlations and fuses helpful visual information to benchmark our architecture with baseline architectures on the MNRE dataset.

Data Bias. Our framework is designed for multi-modal relation extraction for Twitter data. However, when applied to data with vastly different distributions or in new domains, the model's performance may be biased. The results reported in the experiment section are based on specific benchmark datasets, which may be affected by these biases. Therefore, caution should be taken when evaluating the generalizability and fairness.

Computing Cost/Emission. Our research, which entails the utilization of large language models, necessitates a significant computational burden. We recognize that this computational burden results in a negative environmental impact in terms of carbon emissions. Specifically, our work required a cumulative 425 GPU hours of computation utilizing Tesla V100 GPUs. The total emissions generated by this computational process are estimated to be 47.18 kg of CO_2 per run, with a total of two runs being performed.

Acknowledgment

References

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 *NAACL*, pages 1607–1618.

Xin Cong, Jiawei Sheng, Shiyao Cui, Bowen Yu, Tingwen Liu, and Bin Wang. 2022. Relation-guided few-shot relational triple extraction. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2206–2213. ACM.

Xin Cong, Bowen Yu, Tingwen Liu, Shiyao Cui, Hengzhu Tang, and Bin Wang. 2020. Inductive unsupervised domain adaptation for few-shot classification via clustering. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part II*, volume 12458 of *Lecture Notes in Computer Science*, pages 624–639. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Shwai He, Liang Ding, Daize Dong, Boan Liu, Fuqiang Yu, and Dacheng Tao. 2022. Cherry hypothesis: Identifying the cherry on the cake for dynamic networks. *CoRR*, abs/2211.05528.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

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

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *ACL*, pages 5755–5772.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language

models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99.

Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *CVPR*, pages 4613–4621.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *ACL*, pages 2895–2905.

Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, and Jun Zhao. 2021. A large-scale chinese multimodal NER dataset with speech clues. In *ACL/IJCNLP*, pages 2807–2818.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based BERT model for multimodal NER. In *AAAI*, pages 13860–13868.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *CVPR*, pages 3713–3722.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7716–7729. Association for Computational Linguistics.

Meng Wang, Guilin Qi, HaoFen Wang, and Qiushuo Zheng. 2019. Richpedia: a comprehensive multimodal knowledge graph. In *Joint International Semantic Technology Conference*, pages 130–145.

Fuzhao Xue, Aixin Sun, Hao Zhang, Jinjie Ni, and Eng-Siong Chng. 2022. An embarrassingly simple model for dialogue relation extraction. In *ICASSP*, pages 6707–6711.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021a. Multimodal relation extraction with efficient graph alignment. In *ACM MM*, pages 5298–5306.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021b. MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *ICME*, pages 1–6.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, pages 13041–13049.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *arXiv preprint arXiv:2202.05786*.

A Oscar for Image Caption Generation

To generate the text description of the image for multi-modal knowledge alignment without additional pre-training on multi-modal relation extraction, we directly utilize the image captioning method, generating a natural language description of the content of an image. In this paper, we use Oscar (Object-Semantics Aligned Pre-training) (Li et al., 2020) to transfer the image into a text description for each image, which integrates multi-modal alignment information of entities from a semantic perspective.

Oscar uses object tags detected in images as anchor points to significantly facilitate alignment learning. Input samples are processed into triples involving image region features, captions, and object tags similar to the pre-training. It randomly masks 15% of caption tokens and uses the corresponding output representations to perform a classification to predict tokens. Similarly to VLP (Zhou et al., 2020), the self-attention mask is constrained so that a caption token can only attend to the tokens before its position to simulate a uni-directional generation process. It eases the learning of semantic alignments between images and texts on the public corpus of 6.5 million text-image pairs, creating new state-of-the-art on the image caption task. Thus, we use Oscar to integrate useful images by transferring them into textual descriptions.

B Hyper-parameter Settings

Our implementation is based on PyTorch[6]. All experiments were carried out on a server with one GPU (Tesla V100). For re-implementation, we report our hyper-parameter settings on the dataset in Table 3. Note that the hyper-parameter settings are tuned in the validation data by grid search with 5 trials. The learning rate is $2e-4$, the batch size is 100, and the dropout rate is 0.6. We use AdamW (Loshchilov and Hutter, 2019) to optimize the parameters. The maximum length of the text is 128 and the objects of each image are 10. For the learning rate, we adopt the method of grid search with a step size of 0.0001.

C Discussions for Input Order

Due to utilizing a generative encoder, where the prefix, object, image, and text are input stage by stage, thus the order affects the performance of

Hyper-parameter	MNRE dataset
word embedding dimension	768
image embedding dimension	4,096
dropout rate	0.6
batch size	100
training epoch	20
maximum length of text	128
learning rate	$2e-4$
threshold λ_d	2
threshold λ_s	2
threshold λ_c	3
threshold δ	0.4

Table 3: Hyper-parameter settings of DGF-PT.

Variants	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
DGF-PT ($I_o \rightarrow I_i \rightarrow I_t$)	**85.25**	**84.35**	**83.83**	**84.47**
$I_o \rightarrow I_t \rightarrow I_i$	84.64	83.32	82.09	83.53
$I_i \rightarrow I_o \rightarrow I_t$	84.92	83.17	82.24	84.85
$I_i \rightarrow I_t \rightarrow I_o$	84.85	83.38	83.70	84.26
$I_t \rightarrow I_o \rightarrow I_i$	83.01	82.96	82.61	82.73
$I_t \rightarrow I_i \rightarrow I_o$	82.27	82.04	81.75	82.29

Table 4: Impact of the input order of the image I_i, objects I_o, and text I_t.

the model. As shown in Table 4, we exploit the best input order for multi-modal relation extraction. From the figure, we can observe that: 1) Our model is affected by the input order of text, image, and objects. The reason we think that prompt-based autoregressive encoder is a more efficient way to integrate multi-grained information. 2) The best input order is $I_o \rightarrow I_i \rightarrow I_t$. Furthermore, when the text I_t is input before others, the performance of our model dramatically decreases. It demonstrates that visual information is fed before textual information usually integrating more helpful extraction knowledge.

---

[6]https://pytorch.org/

A For every submission:

☑ A1. Did you describe the limitations of your work?
We discuss this in Section Limitations and Ethics Statement.

☑ A2. Did you discuss any potential risks of your work?
We discuss this in Section Limitations and Ethics Statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
We summarize at the end of Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We use Oscar, VGG, and Faster R-CNN in Section 4.

☑ B1. Did you cite the creators of artifacts you used?
We provide the reference for all artifacts.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In section 5.1.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In section 5.1.

C ☑ Did you run computational experiments?

In section 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In section 5.3 and Appendix B.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 5 and Appendix C.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In Section 5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Section 5.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.