TEPrompt: Task Enlightenment Prompt Learning for Implicit Discourse Relation Recognition

Wei Xiang and **Chao Liang** and **Bang Wang** [*]
School of Electronic Information and Communications,
Huazhong University of Science and Technology, Wuhan, China
{xiangwei, liangchao111, wangbang}@hust.edu.cn

Abstract

Implicit Discourse Relation Recognition (IDRR) aims at classifying the relation sense between two arguments without an explicit connective. Recently, the ConnPrompt (Xiang et al., 2022b) has leveraged the powerful prompt learning for IDRR based on the fusion of multi-prompt decisions from three different yet much similar connective prediction templates. Instead of multi-prompt ensembling, we propose to design auxiliary tasks with enlightened prompt learning for the IDRR task. Although an auxiliary task is not used to directly output final prediction, we argue that during the joint training some of its learned features can be useful to boost the main task. In light of such motivations, we propose a task enlightenment prompt learning model, called TEPrompt, to fuse learned features from three related tasks for IDRR. In particular, the TEPrompt contains three tasks, viz., Discourse Relation Recognition (DRR), Sense Semantics Classification (SSC) and Annotated Connective Prediction (ACP), each with a unique prompt template and an answer space. In the training phase, we jointly train three prompt learning tasks with shared argument representation. In the testing phase, we only take the DRR output with fused features as the final IDRR decision. Experiments with the same conditions have shown that the proposed TEPrompt outperforms the ConnPrompt. This can be attributed to the promoted decision features and language models benefited from joint-training of auxiliary tasks.

1 Introduction

Implicit Discourse Relation Recognition (IDRR) is to detect and classify some latent relation in between a pair of text segments (called arguments) without an explicit connective (Xiang and Wang, 2023). Fig. 1 illustrates an argument pair example with a Contingency relation in the Penn Discourse TreeBank (PDTB) corpus, and the implicit connective 'so' is inserted by annotators. IDRR is of great importance for many downstream Natural Language Processing (NLP) applications, such as question answering (Liakata et al., 2013), machine translation (Guzmán et al., 2014), summarization (Huang and Kurohashi, 2021), and etc. However, due to the absence of an explicit connective, inferring discourse relations from the contextual semantics of arguments is still a challenging task.

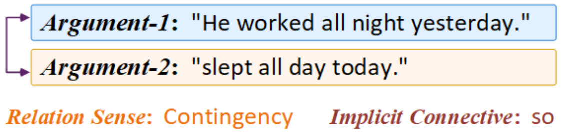

Argument-1: "He worked all night yesterday."
Argument-2: "slept all day today."
Relation Sense: Contingency
Implicit Connective: so

Figure 1: An example of implicit discourse relation annotation with manually inserted connective.

Conventional *pre-train and fine-tuning* paradigm (Liu et al., 2021) designs sophisticated neural networks to encode the representation of argument pairs upon a Pre-trained Language Model (PLM) for relation classification (Chen et al., 2016b; Liu and Li, 2016; Ruan et al., 2020; Li et al., 2020; Liu et al., 2020). On the one hand, these task-specific neural networks introduce some additional parameters that need to be trained by a large amount of labelled data. On the other hand, the task objective function is often not in accordance with that of the PLM, so that the PLM needs to be fine-tuned for solving downstream tasks, resulting in poor utilization of the encyclopedic linguistic knowledge embedded in the pre-training process.

The recent ConnPrompt model (Xiang et al., 2022b) has successfully applied the *pre-train, prompt, and predict* paradigm, i.e. the so-called *prompt learning*, in the IDRR task by transforming the IDRR as a connective-cloze task to predict an answer word and map it to a relation sense. The ConnPrompt has achieved the new state-of-

[*] Corresponding author: Bang Wang

Findings of the Association for Computational Linguistics: ACL 2023, pages 12403–12414
July 9-14, 2023 ©2023 Association for Computational Linguistics

the-art performance on the commonly used PDTB corpus (Webber et al., 2019), however it designs three different yet much similar connective prediction templates which inserts the [MASK] token in between two arguments or at the beginning of one argument for answer prediction. Moreover, to fuse different prompt predictions, the ConnPrompt employs a simple majority voting decision fusing as for final relation sense prediction.

Instead of simple multi-prompt ensemble, we argue that some auxiliary prompt tasks can be designed to enlighten the main prompt task with promoted decision features. For example, as the top relation labels in the PDTB corpus are those plain vocabulary words, we can design an auxiliary task to directly predict such label words from the PLM vocabulary. Furthermore, as the PDTB corpus also contains manually annotated implicit connectives, we can design another auxiliary task to directly predict an annotated connective. Although such auxiliary tasks are not necessarily used to output the final IDRR prediction, they can be jointly trained with the main task on a shared PLM, by which some features learned from the auxiliary tasks can be fused into the main task to promote its decision features for the final prediction.

Motivated from such considerations, we propose a *Task Enlightenment Prompt Learning* (TEPrompt) model, where the main IDRR task can be enlightened from some auxiliary prompt tasks in terms of its promoted decision features via fusing auxiliary task features. Specifically, the TEPrompt contains a main prompt task: *Discourse Relation Recognition* (DRR), and two auxiliary prompt tasks: *Sense Semantics Classification* (SSC) and *Annotated Connective Prediction* (ACP). We design each prompt task with a unique template and an answer space. We concatenate three prompt templates as an entire word sequence with two newly added special tokens [$\texttt{Arg}_1$] and [$\texttt{Arg}_2$] for shared argument representation, as the input of a PLM. In the training phase, we jointly train three prompt tasks upon one PLM model but with three different answer predictions as objective functions. In the testing phase, we only take the main prompt decision features yet promoted by fusing the features from the two auxiliary prompts to output the final IDRR decision.

Experiment results have shown that our proposed TEPrompt outperforms the ConnPrompt with the same conditions and achieves the new state-of-the-art performance on the latest PDTB 3.0 corpus.

2 Related Work

2.1 pre-train and fine-tuning paradigm

Conventional pre-train and fine-tuning paradigm usually approaches the IDRR task as a classification problem, and the key is to design a sophisticated downstream neural network for argument representation learning (Zhang et al., 2015; Rutherford et al., 2017). For example, the SCNN model (Zhang et al., 2015) obtains each argument representation via a single convolution layer and concatenates two arguments' representations for relation classification. Some hybrid models have attempted to combine CNN, LSTM, graph convolutional networks and etc., for argument representation learning (Zhang et al., 2021; Jiang et al., 2021b).

Attention mechanisms have been widely used in neural model to unequally encode each word according to its importance for argument representation (Zhou et al., 2016; Guo et al., 2020; Ruan et al., 2020; Li et al., 2020). For example, Zhou et al. (2016) apply self-attention to weight a word according to its similarity to its belonging argument. Ruan et al. (2020) propose a pipeline workflow to apply interactive attention after self-attention. Li et al. (2020) use a penalty-based loss re-estimation method to regulate the attention learning.

Word pair features have been exploited to capture interactions between arguments for representation learning (Chen et al., 2016a,b; Xiang et al., 2022a). For example, Chen et al. (2016b) construct a relevance score word-pair interaction matrix based on a bilinear model (Jenatton et al., 2012) and a single layer neural model (Collobert and Weston, 2008). Xiang et al. (2022a) propose an offset matrix network to encode word-pairs' offsets as linguistic evidence for argument representation.

2.2 pre-train, prompt, and predict paradigm

Recently, some large-scale PLMs have been proposed, such as the BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), and etc. The prompt learning has become a new paradigm for many NLP tasks, which uses the probability of text in PLMs to perform a prediction task, and has achieved promising results (Seoh et al., 2021; Wang et al., 2021; Ding et al., 2021). For example, Seoh et al. (2021) propose a cloze question prompt and a natural language inference prompt for

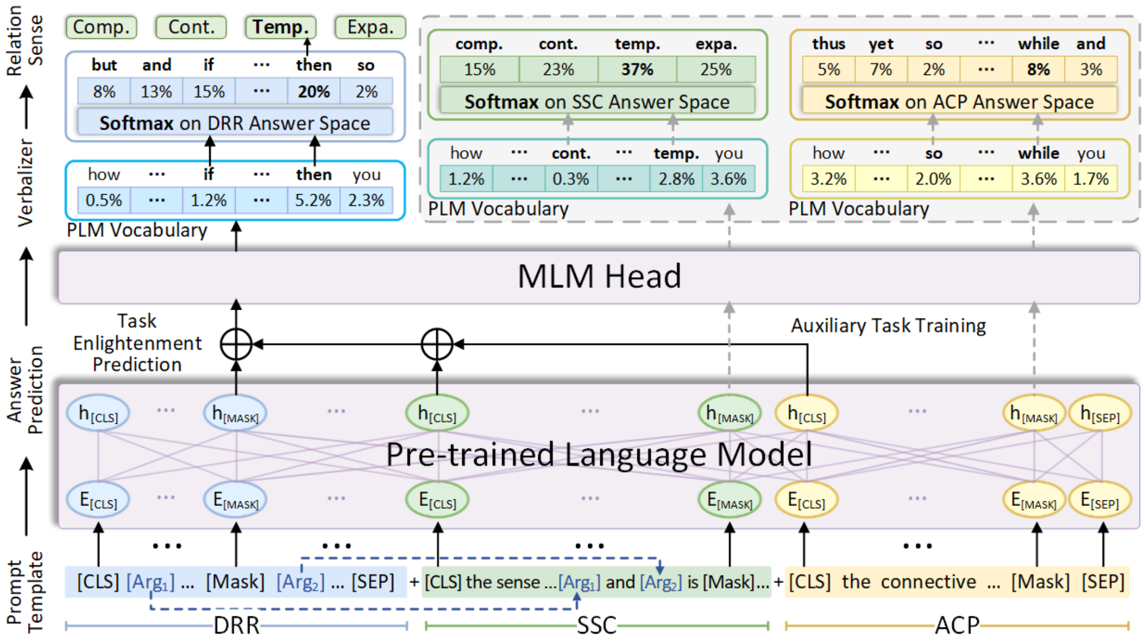


Figure 2: Illustration of our TEPrompt framework. It contains three modules of the prompt templatize, answer prediction and verbalizer for the main prompt task (DRR) and two auxiliary prompt tasks (SSC and ACP).

aspect-based sentiment analysis. Wang et al. (2021) propose a transferable prompting framework to capture cross-task knowledge for few-shot text classification. Ding et al. (2021) apply a cloze-style prompt learning on fine-grained entity typing in fully supervised, few-shot and zero-shot scenarios.

Some studies design appropriate prompts to reformulate an IDRR task for predicting discourse relations (Jiang et al., 2021a,b; Xiang et al., 2022b). Jiang et al. (2021a) use a masked PLM to generate a pseudo-connective for relation classification. Jiang et al. (2021b) utilize the PLM T5 (Raffel et al., 2020) to generate the target sentence which contains the meaning of discourse relations. Xiang et al. (2022b) propose the ConnPrompt model with the new state-of-the-art performance, which reformulates the IDRR task as a connective-cloze task. They further use a majority voting decision fusion of the same task but with three much similar cloze templates for final relation sense prediction.

The proposed TEPrompt model fuses the learned features of two auxiliary prompt task to boost the main prompt tasks for relation prediction.

3 The Proposed TEPrompt Model

Fig. 2 presents our TEPrompt model, including three modules of prompt templatize, answer prediction and verbalizer for the main prompt task (DRR) and two auxiliary prompt tasks (SSC and ACP). The main DRR prompt task uses a kind of connective-cloze prompt to predict a manually selected answer words between two arguments, and map it to a relation sense; The SSC auxiliary prompt task describes and classifies the sense semantic between two arguments; While the ACP describes and predicts the implicit connective words.

3.1 Prompt Templatize

We first reformulate an input argument pair $x = (Arg_1; Arg_2)$ into a prompt template $T(x)$ by concatenating the main DRR prompt template with two auxiliary prompt templates: SSC and ACP, as the input of a PLM. Some PLM-specific tokens such as [MASK], [CLS] and [SEP] are inserted in the prompt template; While the [MASK] tokens are added for the PLM to predict an answer word v, and the [CLS] and [SEP] tokens are used to indicate the beginning and ending of each prompt template, respectively.

Fig. 3 illustrates the three templates for our DRR, SSC and ACP task. We first use a kind of connective-cloze prompt template as the main DRR prompt template $T_D(x)$, in which argument-1 and argument-2 are concatenated as an entire word sequence, and the [MASK] token is inserted between two arguments. Besides, two newly added specific tokens [Arg$_1$] and [Arg$_2$] are inserted at the front of argument-1 and argument-2 to represent their se-

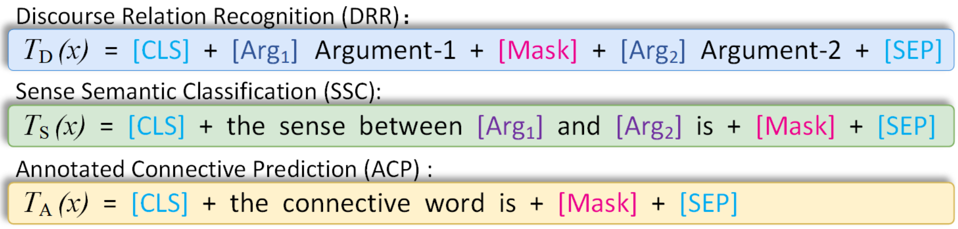

Discourse Relation Recognition (DRR):
$T_D(x)$ = [CLS] + [Arg$_1$] Argument-1 + [Mask] + [Arg$_2$] Argument-2 + [SEP]
Sense Semantic Classification (SSC):
$T_S(x)$ = [CLS] + the sense between [Arg$_1$] and [Arg$_2$] is + [Mask] + [SEP]
Annotated Connective Prediction (ACP) :
$T_A(x)$ = [CLS] + the connective word is + [Mask] + [SEP]

Figure 3: Illustration of our TEPrompt template, which is a concatenation of the three task templates.

mantics which are also shared in the SSC template.

We also design two discrete prompt templates $T_S(x)$ and $T_A(x)$ for the auxiliary task SSC and ACP, respectively. The text of SSC template describes the sense semantics between argument-1 and argument-2; While the text of ACP template describes the implicit connective words. The [MASK] tokens are inserted at the end of SSC and ACP template for prediction. Note that in the SSC template, the specific tokens [Arg$_1$] and [Arg$_2$] are used to represent the semantics of argument-1 and argument-2, which are shared and trained with the main prompt task.

3.2 Answer Prediction

After the PLM, we obtain a hidden state $\mathbf{h}$ for each input token in the prompt templates, where $\mathbf{h} \in \mathbb{R}^{d_h}$ and d_h is the dimension of the hidden state. We use $\mathbf{h}_m^{DRR}$, $\mathbf{h}_m^{SSC}$ and $\mathbf{h}_m^{ACP}$ to denote the hidden state of [MASK] tokens in the DRR, SSC and ACP template, respectively, which are used for the joint training of task enlightenment prompt learning; While the $\mathbf{h}_c^{SSC}$ and $\mathbf{h}_c^{ACP}$ are used to denote the hidden state of the [CLS] token in the SSC and ACP template, respectively, which are used for the feature fusion of auxiliary prompt tasks.

To fuse the features of auxiliary prompt SSC and ACP into the main DRR task, we use the fusion gate mechanism to integrate their [CLS] representations into the [MASK] representation of the main DRR task, which is next used for the final answer word prediction. Specifically, we first use a fusion gate mechanism to integrate the [CLS] representations of SSC and ACP, the transition functions are computed as follows:

$$\mathbf{g}_c = sigmoid(\mathbf{W}_c\mathbf{h}_c^{SSP} + \mathbf{U}_c\mathbf{h}_c^{CEP}), \quad (1)$$

$$\tilde{\mathbf{h}}_c = \mathbf{g}_c \odot \mathbf{h}_c^{SSP} + (1 - \mathbf{g}_c) \odot \mathbf{h}_c^{CEP}, \quad (2)$$

where $\mathbf{W}_c \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{U}_c \in \mathbb{R}^{d_h \times d_h}$ are learnable parameters and $\odot$ donates the element-wise product of vectors.

With the fusion gate, we adaptively assign different importance to the features of SSC and ACP prompt task, and outputs $\tilde{\mathbf{h}}_c \in \mathbb{R}^{d_h}$ as the auxiliary prompt vector. We next use another fusion gate to integrate the auxiliary prompt vector $\tilde{\mathbf{h}}_c$ into the [MASK] hidden state of the main DRR prompt $\mathbf{h}_m^{DRP}$ for the final answer prediction. The transition functions are:

$$\mathbf{g}_m = sigmoid(\mathbf{W}_m\mathbf{h}_m^{DRP} + \mathbf{U}_m\tilde{\mathbf{h}}_c), \quad (3)$$

$$\tilde{\mathbf{h}}_m = \mathbf{g}_m \odot \mathbf{h}_m^{DRP} + (1 - \mathbf{g}_m) \odot \tilde{\mathbf{h}}_c, \quad (4)$$

where $\mathbf{W}_m \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{U}_m \in \mathbb{R}^{d_h \times d_h}$ are learnable parameters.

Finally, the Masked Language Model (MLM) classifier of the PLM uses the fused hidden state $\tilde{\mathbf{h}}_m$ to estimates the probability of each word in its vocabulary V for the [MASK] token of the DRR task as follows:

$$P_D(\texttt{[MASK]}_{DRP} = v_d \in V \mid T(x)). \quad (5)$$

Note that, the MLM classifier also estimates an answer word probability P_S and P_A for the [MASK] token of the auxiliary prompt task SSC and ACP without feature fusion in the joint training.

3.3 Verbalizer

We define a discrete answer space for the DRR, SSC and ACP prompt task, respectively, which are all subsets of the PLM vocabulary. Specifically, we use sixteen manually selected answer words as the answer space V_d of the DRR, the same as that of ConnPrompt (Xiang et al., 2022b). Besides, we use four top-level sense labels in the PDTB corpus as the SSC answer space, V_s = {Comparison, Contingency, Expansion, Temporal}, and we use the 174 manually annotated implicit connectives in the PDTB corpus as the ACP answer space V_c of ACP. We note that the answer space of DRR is next mapped to a relation sense in verbalizer process, while the answer space of SSC and ACP are only used in the auxiliary task training.

Relation Sense	Answer words
Comparison	*similarly*, *but*, *however*, *although*
Contingency	*for*, *if*, *because*, *so*
Expansion	*instead*, *by*, *thereby*, *specifically*, *and*
Temporal	*simultaneously*, *previously*, *then*

Table 1: Answer space of the DRR prompt and the connection to the top-level class discourse relation sense labels in the PDTB corpus.

After answer prediction, a softmax layer is applied on the prediction scores of our pre-defined answer space to normalize them into probabilities:

$$P(v \in V \mid T(x)) = \frac{e^{p_{v_i}}}{\sum_{j=1}^{n} e^{p_{v_j}}}. \tag{6}$$

Then, the predicted answer word of DRR is projected into a unique discourse relation sense based on the pre-defined connection regulation. Table 1 presents the verbalizer connection from the answer word to the PDTB discourse relation sense labels.

3.4 Training and Prediction

In the training phase, we tune the PLM parameters based on the DRR, SSC and ACP prompt task jointly to fuse their learned features. We compute a cross entropy loss for the DRR loss L_d, SSC loss L_s and ACP loss L_c, respectively.

$$J(\theta) = -\frac{1}{K}\sum_{k=1}^{K} \mathbf{y}^{(k)} \log(\hat{\mathbf{y}}^{(k)}) + \lambda\|\theta\|^2, \tag{7}$$

where $\mathbf{y}^{(k)}$ and $\hat{\mathbf{y}}^{(k)}$ are the answer label and predicted answer of the k-th training instance respectively. λ and θ are the regularization hyper-parameters. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with $L2$ regularization for model training. The cost function of our TEPrompt is optimized as follows:

$$L = L_d + \beta L_s + \gamma L_c, \tag{8}$$

where β and γ are weight coefficients to balance the importance of the SSC loss and ACP loss.

4 Experiment Setting

In this section, we present our experiment settings, including the dataset, PLMs, competitors, and parameter settings.

The PDTB 3.0 Dataset: Our experiments are conducted on the Penn Discourse TreeBank (PDTB) 3.0 corpus [1] (Webber et al., 2019), which contains more than one million words of English texts from the Wall Street Journal. Following the conventional data splitting, we use sections 2-20 as the full training set, sections 21-22 as the testing set and 0-1 as the development set (Ji and Eisenstein, 2015). Our experiments are conducted on the four top-level classes of relation sense, including Comparison, Contingency, Expansion, Temporal. Table 2 presents the dataset statistics.

Relation	Train	Dev.	Test
Expansion	8645	748	643
Comparison	1937	190	154
Contingency	5916	579	529
Temporal	1447	136	148
Total	17945	1653	1474

Table 2: Statistics of implicit discourse relation instances in PDTB 3.0 with four top-level relation senses.

Pre-trained Language Models: We use two of the most representative masked pre-trained language models (PLM) for comparison: **BERT** (Devlin et al., 2019) is the first Transformer-based large-scale pre-trained PLM proposed by Google [2], which is pre-trained using a *cloze task* and a *next sentence prediction* task; **RoBERTa** (Liu et al., 2019) is a BERT-enhanced PLM proposed by Facebook [3], which removes the next sentence prediction objective and is pre-trained on a much larger dataset with some modified key hyper-parameters.

Competitors: We compare our TEPrompt with the following advanced models:

- • DAGRN (Chen et al., 2016b) encodes word-pair interactions by a neural tensor network.
- • NNMA (Liu and Li, 2016) combines two arguments' representations for stacked interactive attentions.
- • IPAL (Ruan et al., 2020) propagates self-attention into interactive attention by a cross-coupled network.
- • PLR (Li et al., 2020) uses a penalty-based loss re-estimation to regulate the attention learning.
- • BMGF (Liu et al., 2020) combines bilateral multi-perspective matching and global information fusion to learn a contextualized representation.
- • MANF (Xiang et al., 2022a) encodes two kinds of attentive representation for arguments and fuses

[1] We have purchased the PDTB 3.0 liscence for experiments.

[2] https://github.com/google-research/bert

[3] https://github.com/pytorch/fairseq/

them with the word-pairs features.

• ConnPrompt (Xiang et al., 2022b) applies the prompt learning for IDRR based on the fusion of multi-prompt decisions.

Parameter Setting: We implement the PLM models with 768-dimension provided by HuggingFace transformers [4] (Wolf et al., 2020), and run PyTorch [5] framework with CUDA on NVIDIA GTX 3090 Ti GPUs. The maximum length of our TEPrompt template is set to 150 tokens, in which the maximum length of arguments are 70 tokens. We set the mini-batch size to 32, the learning rate to 1e-5, the weight coefficients β and γ to 0.3 and 0.4 respectively, and all trainable parameters are randomly initialized from normal distributions. We release the code at: https://github.com/HustMinsLab/TEPrompt.

5 Result and Analysis

5.1 Overall Result

Table 3 compares the overall performance between our TEPrompt and the competitors. We implement a four-way classification on the top-level relation sense of the PDTB dataset and adopt the commonly used macro $F1$ score and accuracy (Acc) as performance metrics.

We note that the competitors in the first group all use the pre-train and fine-tuning paradigm; While our TEPrompt and the ConnPrompt use the pre-train, prompt, and predict paradigm, i.e. the prompt learning. Besides, the first two competitors both use a kind of distributed and static word embeddings: Word2vec and Glove; while the others use Transformer-based PLM models: BERT and RoBERTa.

The first observation is that the DAGRN and NNMA cannot outperform the other competitors. This is not unexpected, as the others employ the more advanced dynamic PLMs pre-trained with deeper neural networks and larger scale of parameters, which have been proven more effective for many downstream NLP tasks (Devlin et al., 2019; Liu et al., 2019). The gaps between large PLM fine-tuning and static embedding for representation learning also have a certain impact on the performance of the IDRR task.

The second observation is that our TEPrompt and the ConnPrompt adopting the prompt learning paradigm can significantly outperform the other competitors in terms of much higher macro F1 score (8%+) and Acc(5%+). The outstanding performance can be attributed to the task transformation of connective-cloze prediction into the training of PLMs, other than designing a task-specific model upon PLM, by which the model can better enjoy the encyclopedic linguistic knowledge embedded in a PLM during the model training.

Model	PLM	Acc (%)	F1 (%)
DAGRN (ACL, 2016)	Word2vec	57.33	45.11
NNMA (EMNLP, 2016)	Glove	57.67	46.13
IPAL (COLING, 2020)	BERT	57.33	51.69
PLR (COLING, 2020)	BERT	63.84	55.74
BMGF (IJCAI, 2020)	RoBERTa	69.95	62.31
MANF (ACL-Findings, 2022)	BERT	64.04	56.63
ConnPrompt (COLING, 2022)	BERT	69.67	64.00
Our TEPrompt	BERT	70.08	65.12
ConnPrompt (COLING, 2022)	RoBERTa	75.17	70.88
Our TEPrompt	RoBERTa	**75.51**	**72.26**

Table 3: Comparison of overall results on the PDTB.

Finally, our TEPrompt achieves better performance than the ConnPrompt with the same PLM and outperforms all the other models in both higher macro F1 score and accuracy. Similar results can also be observed in the binary classification (i.e. one-versus-others) of implicit discourse relation recognition, in Table 4. We attribute the outstanding performance of our TEPrompt to the use of auxiliary tasks for enlightenment prompt learning, by which the jointly trained features of auxiliary SSC and ACP prompt task can be well fused into the main DRR task to improve the final answer prediction. This will be further analyzed in our ablation study.

Model	Expa.	Comp.	Cont.	Temp.
DAGRN (ACL, 2016)	64.71	27.34	62.56	38.91
NNMA (EMNLP, 2016)	65.10	29.15	63.33	41.03
DERM (COLING, 2018)	64.96	41.71	67.73	46.73
IPAL (COLING, 2020)	66.86	37.31	66.40	41.25
PLR (COLING, 2020)	69.33	35.16	66.97	43.40
BMGF (IJCAI, 2020)	72.61	50.85	72.42	45.23
MANF (ACL, 2022)	70.00	35.83	66.77	40.22
Our TEPrompt	**77.34**	**53.42**	**77.98**	**53.55**

Table 4: Comparison of binary classification results on the PDTB (F1 score %). We have reproduced some of the competitors on PDTB 3.0 for fair comparison.

5.2 Ablation Study

To examine the effectiveness of different prompt tasks, we design the following ablation studies.

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

[5] pytorch.org

- Prompt-SSC is only the SSC prompt concatenating argument-1 and argument-2 in front, without the DRR and ACP task.
- TEPrompt-SSC combines the SCC prompt with DRR and ACP, and only uses the predicted answer of SSC for relation sense mapping.
- Prompt-ACP is only the ACP prompt concatenating argument-1 and argument-2 in front, without the DRR and SSC.
- TEPrompt-ACP combines the ACP prompt with the DRR and SSC, and uses the predicted answer of ACP for relation sense mapping [6].
- Prompt-DRR is only the DRR prompt without the auxiliary prompt SSC and ACP.
- TEPrompt w/o Gate is our task enlightenment prompt model without fusion mechanisms.

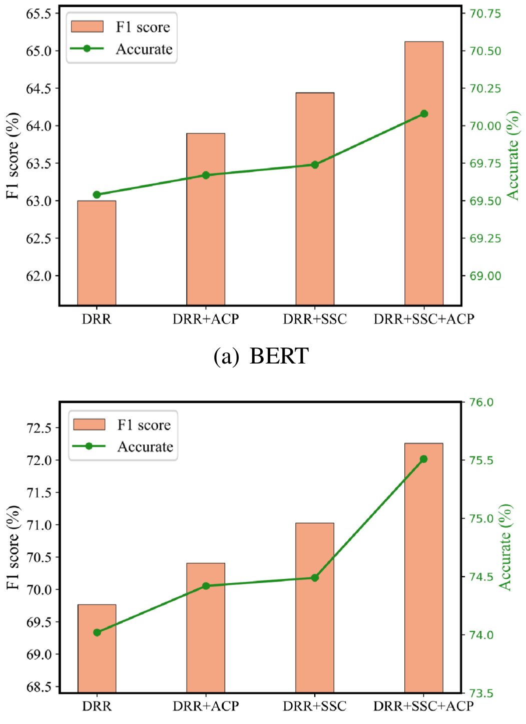
Table 5 compares the results of our ablation study models with both single-prompt and multi-prompt ConnPrompt.

PLM	BERT		RoBERTa	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)
ConnPrompt-1	69.74	63.95	74.36	69.91
ConnPrompt-2	69.34	63.69	73.61	69.63
ConnPrompt-3	67.64	62.65	73.54	69.00
ConnPrompt-Multi	69.67	64.00	75.17	70.88
Prompt-SSC	67.37	60.64	70.62	66.09
TEPrompt-SSC	67.64	62.73	74.22	69.93
Prompt-ACP	66.08	59.08	72.73	67.89
TEPrompt-ACP	67.23	61.44	73.13	68.83
Prompt-DRR	69.54	63.00	74.02	69.77
TEPrompt w/o Gate	68.32	63.48	75.03	70.58
TEPrompt	**70.08**	**65.12**	**75.51**	**72.26**

Table 5: Results of ablation study on the PDTB corpus.

Task enlightenment prompt: We can observe that the Prompt-DRR has comparable performance to each single-ConnPrompt, viz. ConnPrompt-1/2/3. This is not unexpected. All the three single-ConnPrompts are with the same connective-cloze prompt model, and the only difference is the location of the cloze-mask in each template; While the Prompt-DRR is with the same connective-cloze prompt model and answer space as a single-ConnPrompt. The ConnPrompt-Multi uses multi-prompt majority voting and outperforms any of the single-ConnPrompt; While the TEPrompt designs two auxiliary tasks to augment the main task and outperforms both Prompt-DRR and ConnPrompt-Multi, which validates the effectiveness of our task enlightenment prompt learning via fusing features from both main and auxiliary prompt tasks by joint training.

Prompt ablation study: Among the second group of prompt ablation models, it can be observed that the Prompt-SSC and Prompt-ACP cannot outperform the Prompt-DRR; While the TEPrompt-SSC and TEPrompt-ACP also cannot outperform the TEPrompt. Although both the SSC and ACP prompt model can each output the final prediction by mapping its predicted answer to a relation sense, however, their objectives are not completely in accordance with the IDRR task. The SCC prompt is designed to classify sense semantics; While the ACP prompt aims at predicting manually annotated connectives. Furthermore, we can also observe that the TEPrompt-SSC and TEPrompt-ACP have achieved better performance than the Prompt-SSC and Prompt-ACP, respectively. This again validates our argument that fusing features from jointly trained auxiliary prompt tasks can be useful to boost the main prompt task prediction.



(a) BERT

(b) RoBERTa

Figure 4: Comparison of auxiliary prompt effections.

Gate Fusion Mechanism: We also observe that the TEPrompt w/o Gate without gate fusion mechanism cannot outperform the full TEPrompt

[6] Note that some implicit connectives correspond to multiple relation senses, we choose the one with the highest frequency in the training data as the prediction relation sense.

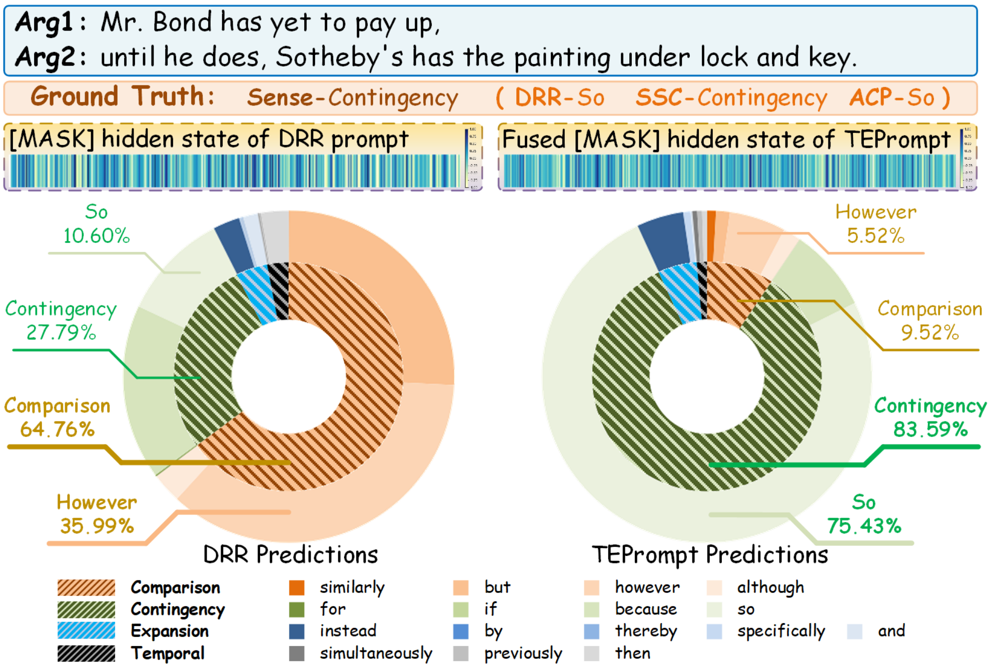


Figure 5: Visualization of the predicted answer words and relation sense for the DRR Prompt and TEPrompt.

model, even it jointly trains a PLM as well as the MLM head with two auxiliary tasks. This indicates that the features learned from auxiliary tasks can indeed augment the main task prediction.

Auxiliary prompt effections: To further investigate the task enlightenment effections, we design several combinations of individual prompt models: the DRR with the only main task, the DRR+SSC and DRR+ACP are the main task enlightened by only one auxiliary task, and DRR+SSC+ACP (viz., TEPrompt) is the main task enlightened by two auxiliary tasks.

Fig. 4 compares the performance of different auxiliary prompt ablation models. We can observe that both the SSC and ACP auxiliary task can help improving the performance of the main DRR task. This suggests that fusing either the sense semantics feature in training SSC or the annotated connective feature in training ACP (viz., the two [CLS] tokens) can help promoting the decision feature of the main DRR task (viz., the [MASK] token) to improve the IDRR prediction. Finally, our TEPrompt joint training with both SSC and ACP auxiliary prompts yields substantial improvements over all ablation models, again approving our arguments and design objectives.

5.3 Case Study

We use a case study to compare the TEPrompt and the DRR prompt. Note that the DRR prompt can be regarded as the ConnPrompt using only one template yet without multi-prompt ensemble. Fig. 5 visualizes the representation of the [MASK] token, as well as its prediction probability and classified relation sense by a pie chart. The [MASK] token representation of the TEPrompt is quite different from that of the DRR prompt, as the former also fuses two auxiliary prompt task features. Such feature fusion from auxiliary tasks may enlighten the main task to make correct predictions.

It can be observed that the DRR prompt itself tends to predict a Comparison relation (64.76%) corresponding to the answer word *'however'* with the highest probability 35.99%. After feature fusion, the TEPrompt can correctly recognize the Contingency relation (83.59%) between the two arguments by predicting the answer word *'so'* with a much higher probability 75.43% than that of the DRR prompt prediction (10.60%). We argue that such benefits from the adjustments of prediction probabilities can be attributed to the feature fusion of the two auxiliary prompt tasks.

6 Concluding Remarks

In this paper, we have argued a main prompt task can be enlightened by some auxiliary prompt tasks for performance improvements. For the IDRR task, we have proposed a TEPrompt, a task enlightenment prompt model that fuses learned features from our designed auxiliary SSC and ACP task into the decision features of the main DRR task. Since the three prompt tasks are trained jointly, the learned auxiliary task features in the training phase can help promoting the main task decision feature and improving the final relation prediction in the testing phase. Experiment results and ablation studies have validated the effectiveness of our arguments and design objectives in terms of improved state-of-the-art IDRR performance.

In our future work, we shall investigate other types of auxiliary tasks for the IDRR task as well as the applicability of such task enlightenment prompt learning for other NLP tasks.

Limitations

The two auxiliary prompt tasks are closely related to the PDTB corpus, as the top-level relation sense labels are those plain vocabulary words and the PDTB provides manually annotated connectives.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China (Grant No: 62172167). The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

Jifan Chen, Qi Zhang, Pengfei Liu, and Xuanjing Huang. 2016a. Discourse relations detection via a mixed generative-discriminative framework. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2921–2927, Phoenix, Arizona, USA.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016b. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1726–1735, Berlin, Germany.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, Helsinki, Finland.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, MN, USA.

Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. 2021. Prompt-learning for fine-grained entity typing. *arXiv preprint*, arXiv:2108.10604:1–12.

Fengyu Guo, Ruifang He, Jianwu Dang, and Jian Wang. 2020. Working memory-driven neural networks with a novel knowledge enhancement paradigm for implicit discourse relation recognition. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7822–7829, New York, NY, USA.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, Baltimore, MD, USA.

Yin Jou Huang and Sadao Kurohashi. 2021. Extractive summarization considering discourse and coreference relations based on heterogeneous graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, EACL '21, pages 3046–3052, Stroudsburg, PA, USA.

Rodolphe Jenatton, Nicolas L Roux, Antoine Bordes, and Guillaume R Obozinski. 2012. A latent factor model for highly multi-relational data. In *Advances in neural information processing systems*, pages 3167–3175, Lake Tahoe, Nevada, United States.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Congcong Jiang, Tieyun Qian, Zhuang Chen, Kejian Tang, Shaohui Zhan, and Tao Zhan. 2021a. Generating pseudo connectives with mlms for implicit discourse relation recognition. In *The 18th Pacific Rim International Conference on Artificial Intelligence*, pages 113–126, Hanoi, Vietnam.

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2021b. Not just classification: Recognizing implicit discourse relation on joint modeling of classification and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2418–2431, Punta Cana, Dominican Republic.

Xiao Li, Yu Hong, Huibin Ruan, and Zhen Huang. 2020. Using a penalty-based loss re-estimation method to improve implicit discourse relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1513–1518, Online.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 747–757, Seattle, Washington, USA.

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. *arXiv preprint*, arXiv:2107.13586:1–46.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3830–3836, Virtual.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1233, Austin, Texas, USA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint*, arXiv:1907.11692:1–13.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, pages 1–18, New Orleans, LA, USA.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Huibin Ruan, Yu Hong, Yang Xu, Zhen Huang, Guodong Zhou, and Min Zhang. 2020. Interactively-propagative attention learning for implicit discourse relation recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3168–3178, Online.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 281–291, Valencia, Spain.

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough. 2021. Open aspect target sentiment classification with natural language prompts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6311–6322, Punta Cana, Dominican Republic.

Chengyu Wang, Jianing Wang, Minghui Qiu, Jun Huang, and Ming Gao. 2021. Transprompt: Towards an automatic transferable prompting framework for few-shot text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2792–2802, Punta Cana, Dominican Republic.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 1:1–81.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 1:1–34.

Wei Xiang, Bang Wang, Lu Dai, and Yijun Mo. 2022a. Encoding and fusing semantic connection and linguistic evidence for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3247–3257, Dublin, Ireland.

Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022b. ConnPrompt: Connective-cloze prompt learning for implicit discourse relation recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911, Gyeongju, Republic of Korea.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Lisbon, Portugal.

Yingxue Zhang, Fandong Meng, Li Peng, Jian Ping, and Jie Zhou. 2021. Context tracking network: Graph-based context modeling for implicit discourse relation recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association Computational Linguistics*, pages 1592–1599, Online.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, page 207–212, Berlin, Germany.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation section before Reference.

☒ A2. Did you discuss any potential risks of your work?
This paper is a foundational research for discourse understanding, to our knowledge, there should be no potential risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section Abstract and section I Introduction.

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

B ☑ Did you use or create scientific artifacts?

We have purchased the PDTB 3.0 corpus for experiments, and cite the corpus in Section I introduction and Section IV experiment dataset

☑ B1. Did you cite the creators of artifacts you used?
We cite the corpus in Section I introduction and Section IV experiment dataset. We have purchased the PDTB 3.0 corpus with liscence.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In Section IV, we state that we have purchased the PDTB 3.0 liscence for experiments.

☑ 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)?
In Section IV, we state that we have purchased the PDTB 3.0 liscence for experiments.

☒ 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?
The PDTB 3.0 corpus contains documents/articels from public available Wall Street Journal. Our use of PDTB 3.0 does not involve with any privacy information and offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section IV Experiment Setting, we provide brief introduction about PDTB

☑ 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 IV Experiment Settings, we provide details of train/test/dev splits.

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

C ☑ Did you run computational experiments?

Section V Results and Anlysis.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section IV Experiments settings.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section IV Experiments settings.

☑ 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?
Section V Experiment Results and Analysis

☑ 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.)?
Section IV experiments settings.

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.?
No response.

□ 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)?
No response.

□ 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?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.