A Parameters

We use the learning rate lr $= 5 \times 10^{-5}$, weight_decay $= 0.0$, smoothing value $\epsilon = 1 \times 10^{-8}$. The maximum number of training epochs was set to 20.

B Classifier Experiments

Input and Output The input consists of a pair of EDUs, one being Nucleus and the other Satellite, with the output being a relation label.

Dataset The RST-DT dataset (Carlson et al., 2001) comprises annotated news articles from which EDU pairs, including Nucleus and Satellite, are extracted for our dataset.

Experimental Setups Table 3 shows the experimental setup. We use BART (Lewis et al., 2020) as the language model, and for comparison, we also conduct experiments in the same setting with BERT (Devlin et al., 2019).

Pre-trained model	facebook/bart-base
Training epochs	20
Optimizer	AdamW
Batch size	Train:10,Valid:5,Test:4
Loss function	cross entropy loss
Learning rate	5×10^{-5}

Table 3: Experimental setups.

Model	Accuracy	F1
BERT	55.17	37.59
BART	54.53	37.70

Table 4: Experimental results.

Results Table 4 shows that the BART-based classifier outperforms BERT in the F1 score, although it is inferior to BERT in the accuracy.

C Implementation Details of PPLM

In an efficient implementation of the Transformer (Wolf et al., 2020), the language model's internal states H_t are utilized as inputs when outputting the token x_{t+1} at time-step $t+1$ conditioned on the output token sequence $\boldsymbol{x}_{:t}$ up to time-step t.

$$o_{t+1}, H_{t+1} = \mathrm{LM}(x_t, H_t) \quad (3)$$

$$x_{t+1} \sim p_{t+1} = \mathrm{Softmax}(W o_{t+1}) \quad (4)$$

Here, the internal states is a matrix that retains Key-Value information used in the attention calculation of the Transformer model. PPLM utilizes the gradient from an attribute model $p(a|X)$ to update the internal states, reflecting attribute a.

$$\Delta H_t \leftarrow \Delta H_t + \alpha \frac{\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)}{||\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)||^{\gamma}} \quad (5)$$

Using the updated internal states $\tilde{H}t = H_t + \Delta H_t$, the language model generates $\tilde{x}t + 1$ based on the token sequence $\boldsymbol{x}$: t up to time-step t.

$$\tilde{o}_{t+1}, H_{t+1} = \mathrm{LM}(x_t, \tilde{H}_t) \quad (6)$$

$$\tilde{x}_{t+1} \sim \tilde{p}_{t+1} = \mathrm{Softmax}(W \tilde{o}_{t+1}) \quad (7)$$

D Results on Recalls

Figure 4 demonstrates that our model significantly improved accuracy for discourse markers like 'since' and 'before', while showing only a slight improvement for 'and' and 'for'. While the former words are closely tied to specific relation labels, the latter are commonly used in text and have weaker associations with relation labels. Consequently, the control based on relation labels proposed in this paper yields a smaller improvement for the latter words.

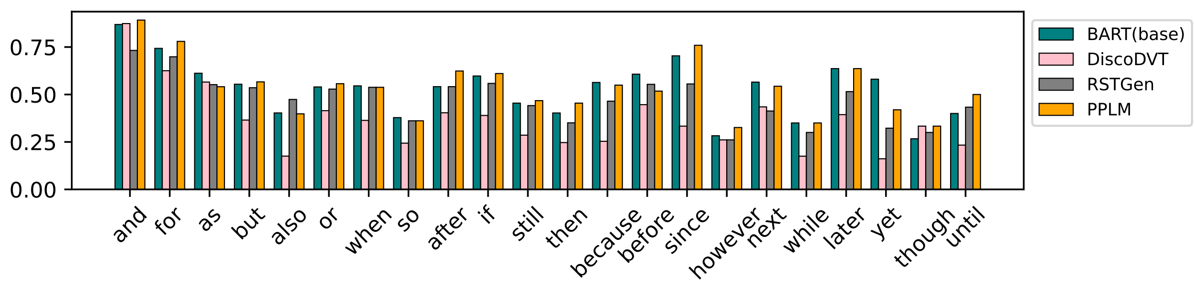

Figure 4: Experimental results for the recall of each discourse marker. We utilize discourse markers listed in Appendix A of the PDTB Annotation Manual (Prasad et al., 2007) We use only those discourse markers from the list that appear more than 30 times in the references.