

# Discourse Relation-Enhanced Neural Coherence Modeling

Wei Liu and Michael Strube

Heidelberg Institute for Theoretical Studies gGmbH  
{wei.liu, michael.strube}@h-its.org

## Abstract

Discourse coherence theories posit relations between text spans as a key feature of coherent texts. However, existing work on coherence modeling has paid little attention to discourse relations. In this paper, we provide empirical evidence to demonstrate that relation features are correlated with text coherence. Then, we investigate a novel fusion model that uses position-aware attention and a visible matrix to combine text- and relation-based features for coherence assessment. Experimental results on two benchmarks show that our approaches can significantly improve baselines, demonstrating the importance of relation features for coherence modeling.

## 1 Introduction

Coherence is a property of well-written texts that makes them easier to read and understand than a sequence of randomly strung sentences (Lapata and Barzilay, 2005). Its modeling has been applied to many downstream tasks, such as text summarization (Wu and Hu, 2018), machine translation (Tan et al., 2019), and document-level text generation (Wang et al., 2021).

Discourse relations, such as *Cause* and *Contrast*, describe the logical relation between two text spans. In discourse coherence theory (Rohde et al., 2018; Jurafsky and Martin, 2021), discourse relations between text spans play a key role in establishing the coherence of texts. Referring to the example in Figure 1, which contains four sentences. This text is considered highly coherent because it is organized with specific discourse relations. Specifically, a *Contrast* relation is used to connect the first two sentences, followed by an *Instantiation* to provide more details for the strike, and finally, a *Cause* relation is applied to introduce the last sentence. Despite the potential usefulness of discourse relations, existing works on coherence modeling primarily focus on integrating entity-based

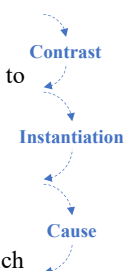
- Tom was late for the meeting this morning.
  - However, it was not his fault but rather due to the citywide strike.
  - All the roads were blocked, and the buses were canceled.
  - Therefore, he had to walk to the office, which took a lot of time.
- 

Figure 1: A coherent text with discourse relations.

features (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013; Tien Nguyen and Joty, 2017; Jeon and Strube, 2022) or applying powerful pre-trained models (Shen et al., 2021; Laban et al., 2021; Abhishek et al., 2021; Liu et al., 2023), with little attention paid to whether discourse relations can contribute to coherence assessment.

To fill this gap, this paper empirically shows that text coherence is correlated with the sequence of discourse relations inferred from documents. Specifically, we first use a PDTB parser to obtain discourse relations in documents. Then, we conduct a statistical analysis to demonstrate that text coherence is correlated with discourse relation transition patterns. Finally, we construct a BiLSTM classifier based on the discourse relation sequence in documents for coherence assessment and show that its performance is comparable to a counterpart built on the textual input of documents.

Based on these observations, we further investigate whether discourse relations can be used to enhance the performance of existing neural coherence models. To this end, we propose a novel fusion model to combine text- and relation-based features for coherence assessment. Specifically, we convert sentences and relations of a document into a flat structure so that the Transformer can simultaneously handle these two different inputs. To promote the correct interaction between sentences

and relations, we further investigate position-aware attention and a visible matrix to guide the attention between these two types of inputs.

We evaluate our methods on two benchmark tasks: assessing discourse coherence and automatic essay scoring. Experiments<sup>1</sup> show that our model outperforms strong baselines, indicating the importance of relation features for coherence modeling. Furthermore, we conduct a detailed analysis showing that relation features contribute to learning coherence patterns, which achieve better performance in long documents and cross-domain settings.

## 2 Related Work

There is a continued research interest in assessing text coherence. Early work is dominated by entity-based approaches. Motivated by the Centering Theory (Grosz et al., 1995), Barzilay and Lapata (2005) propose the well-known entity-grid, which represents entities in a two-dimensional array to track their transitions between sentences for coherence modeling. The entity-grid was extended by many subsequent studies, such as considering semantically-related entities (Filippova and Strube, 2007), adding entity-specific features (Elsner and Charniak, 2011), and modeling entity transition beyond adjacent sentences (Guinaudeau and Strube, 2013). With the advent of deep learning, subsequent studies have employed neural networks for coherence modeling. For example, Tien Nguyen and Joty (2017); Joty et al. (2018) extend the entity-grid with convolutional neural networks while Mesgar and Strube (2018) use LSTM to learn text representations. More recently, pre-trained model-based methods have soon been popular for coherence modeling due to their powerful representation learning. Abhishek et al. (2021) design different transformer-based models for coherence assessment while Shen et al. (2021) examine the performance of a broad range of pre-trained Language Models (LM) for this task. Our work also uses neural networks and pre-trained LM for coherence assessment. However, we focus on investigating the benefits of discourse relations to the task.

Lin et al. (2011) is one of the few studies that use discourse relations for coherence assessment. Specifically, they follow an idea similar to the entity-grid to build a two-dimensional matrix for texts, where rows and columns are sentences and entities, respectively. Each cell  $(s_i, e_j)$  corresponds

to the set of discourse relations entity  $e_j$  is involved within sentence  $s_i$ . Feng et al. (2014) propose a similar approach but replace the PDTB relations with RST ones. However, Mesgar and Strube (2015) point out that these methods are intuitively implausible because they treat discourse relations as features of entities, which contradicts the definition of discourse relations working between sentences (or elementary discourse units). Additionally, these works are constrained by the performance of the discourse parser at that time. For example, the PDTB parser used in Lin et al. (2011) obtains an F1-score of only 25.46 in the top-level implicit relations recognition. By contrast, our study investigates the benefits of discourse relations at the sentence level and is based on a more advanced discourse parser.

Our work is also related to studies using discourse relations for downstream NLP tasks. Hewett et al. (2019) use discourse relations as extra features for predicting argumentation structure. Mi-haylov and Frank (2019) propose a discourse relation guided attention mechanism to enhance the transformer for reading comprehension. Liu et al. (2021b) design a method to fuse discourse information with contextualized features derived from pre-trained models for argument impact classification. Lei and Huang (2023) utilize discourse structures to guide propaganda identification via knowledge distillation. Our work differs from the above studies in both motivation and approach. For example, the key inspiration of this work is derived from the linguistic definition of text coherence.

## 3 Discourse Relations and Coherence

In this section, we first briefly introduce discourse relations and how to extract relations from documents. Then, we provide empirical evidence to demonstrate the correlation between relation features and coherence levels. Finally, we show that a BiLSTM classifier using relation sequences as input is able to achieve performance comparable to the counterpart based on textual input.

### 3.1 Discourse Relations

Discourse relations are a means of logically connecting two segments of discourse. Over the past few decades, various frameworks have been introduced to annotate discourse relations. Among them, the most widely used are Rhetorical Structure Theory (RST, Mann and Thompson, 1988) and

<sup>1</sup><https://github.com/liuwei1206/RelCoh>

	GCDC Enron		TOEFL P1			
		coef	p-value	coef	p-value	
2-gram	Synchronous → Conjunction	0.3924	<0.01	Disjunction → Cause	0.5242	<0.01
	Asynchronous → Asynchronous	0.3675	<0.01	Synchronous → Conjunction	0.4733	<0.01
	Level-of-detail → Asynchronous	0.3040	0.042	Instantiation → Level-of-detail	0.3483	<0.01
	Cause → NoRel	-0.2300	0.015	Conjunction → Synchronous	0.3477	<0.01
3-gram	Cause → NoRel → Conjunction	-0.4835	<0.01	Level-of-detail → Conjunction → Instantiation	0.5239	<0.01
	NoRel → Conjunction → Cause	-0.4359	<0.01	Conjunction → Contrast → Conjunction	0.5234	<0.01
	Cause → Level-of-detail → Conjunction	0.4160	0.012	Conjunction → Conjunction → Contrast	0.5227	<0.01
	Conjunction → Cause → Asynchronous	0.3133	0.056	Level-of-detail → Concession → Cause	0.4882	<0.01

Table 1: Correlation between discourse relation N-gram patterns and coherence levels. Only the top four patterns with the highest absolute correlation coefficients are shown.

the Penn Discourse Treebank (PDTB, Prasad et al., 2008). In the RST framework, a text is represented as a hierarchical discourse tree, where relations are used to link different text spans. By contrast, PDTB does not postulate any structural constraints on discourse relations and focuses on labeling local discourse relations between sentences and clauses. In this work, we follow previous work (Lin et al., 2011) adopting PDTB relations and leave the RST relations for future work.

Specifically, we choose the *discopy* (Knaebel, 2021) as discourse parser to extract relations from documents, but make some adjustments. First, we use the relations in PDTB 3.0 (Webber et al., 2019) instead of PDTB 2.0 (Prasad et al., 2008) since the former defines more relations and is an improved version of the latter. Second, we adopt the connective-enhanced approach from Liu and Strube (2023) for implicit relation recognition as it achieves state-of-the-art performance. We train the parser on the PDTB 3.0 corpus using the data split introduced by Ji and Eisenstein (2015), and evaluate its performance on second-level relations. It can achieve an accuracy of 89.61% on explicit relations and 67.80% on implicit ones. This suggests that the parser performs quite well, laying a solid foundation for further analysis. See Appendix A for more information about the parser.

### 3.2 Correlation Analysis

In coherence theories, discourse relations between text spans play a key role in achieving text coherence (Jurafsky and Martin, 2021). Furthermore, Lin et al. (2011) observed that coherent text exhibits preferences for specific discourse relation ordering. This is somehow verified by Biran and McKeown (2015), which shows that relation N-gram planning (transitions between discourse relations) helps generate coherent text. Inspired by

these works, we aim to provide evidence to demonstrate the correlation between relation N-gram patterns and text coherence.

**Dataset.** We conduct analyses on two widely used corpora in coherence modeling: the Grammarly Corpus of Discourse Coherence (GCDC) dataset (Lai and Tetreault, 2018) and the TOEFL dataset (Blanchard et al., 2013). GCDC is a corpus constructed for assessing discourse coherence (ADC), containing texts from four domains, **Yahoo** online forum posts, emails from **Enron**, emails from Hillary **Clinton**’s office, and **Yelp** online business reviews. Each text in this corpus is annotated by expert raters with scores of {1, 2, 3}, representing low, medium, and high levels of coherence, respectively. The TOEFL dataset was originally used for automated essay scoring (AES) but has been used to evaluate coherence models (Burstein et al., 2010; Jeon and Strube, 2020b). It contains essays from **eight prompts** (P1 to P8) along with score levels (low/medium/high) for each essay. See Appendix B for statistics on these two corpora.

For each document  $d$  in the two corpora, we use Stanza (Qi et al., 2020) to segment it into sentences  $\{s_1, s_2, \dots, s_L\}$  and employ the enhanced *discopy* to recognize the relations between adjacent sentences, obtaining a relation sequence  $\{r_1, r_2, \dots, r_{L-1}\}$ , where  $r_i$  denotes the parsed relation between  $s_i$  and  $s_{i+1}$ . Then, we extract transition patterns, i.e., relation N-grams, from the relation sequence, and count their frequency. Finally, we calculate Spearman’s rank correlation<sup>2</sup> between each n-gram pattern and the ground-truth coherence label.

**Results.** Table 1 shows the results on the GCDC Enron and TOEFL P1 datasets. In general, relation N-gram features are empirically correlated

<sup>2</sup>We choose Spearman’s rank correlation because coherence level and relation N-gram frequency are ranked variables.

Input Type	GCDC Enron		TOEFL P1	
	Acc	F1	Acc	F1
Raw Text	46.20 <sub>0.77</sub>	42.86 <sub>0.97</sub>	57.55 <sub>1.24</sub>	50.39 <sub>0.78</sub>
Rel Sequence	44.15 <sub>0.92</sub>	39.43 <sub>1.24</sub>	59.17 <sub>0.87</sub>	53.51 <sub>0.99</sub>
Rel Sequence (shuffled)	37.40 <sub>1.05</sub>	31.62 <sub>1.07</sub>	50.54 <sub>0.92</sub>	43.03 <sub>1.53</sub>

Table 2: The performance (with std) of BiLSTM classifier when using text, relation sequence, and shuffled relation sequence as input, respectively.

with coherence levels. For example, in GCDC Enron, the relation 3-grams containing *NoRel*, e.g., Cause → NoRel → Conjunction, are negatively correlated with coherence level. In the TOEFL P1 dataset, essays containing *Cause* and *Level-of-detail* relations, e.g., Disjunction → Cause, tend to be more coherent. This aligns with existing theories where discourse relations play a key role in achieving text coherence (Rohde et al., 2018). The relation 3-gram patterns seem to be more correlated with text coherence than the relation 2-gram ones. Taking results on TOEFL P1 as an example, the correlation coefficients for 3-gram patterns are generally above 0.5, higher than that of around 0.4 for 2-gram ones. We also observe that the two corpora have different relation N-gram patterns which are correlated with text coherence. This may be due to the genre distinctions (Webber, 2009) for discourse in different texts. In the TOEFL corpus, essays are viewpoint-oriented, using evidence (*Cause* relation) and examples (*Instantiation* relation) to support opinions. In contrast, the documents in the GCDC Enron dataset are narrative texts, typically employing *Conjunction* relations. See Appendix C.1 for the distribution of discourse relations parsed from the two corpora.

### 3.3 Text vs. Relations

We devise another experiment to demonstrate the importance of discourse relations for coherence modeling. Specifically, we train two BiLSTM classifiers for coherence assessment, in which the first uses the raw text of the document as input while the other inputs the discourse relation sequence parsed from the document.

Table 2 shows the accuracy and macro-f1 results on GCDC Enron and TOEFL P1 datasets. Surprisingly, the classifier built on the discourse relation sequence (Rel Sequence) can achieve comparable performance to that built on raw text. On the GCDC Enron dataset, the classifier based on the relation sequence only lags behind that on raw text by 2 to 3 points, despite the relation sequence

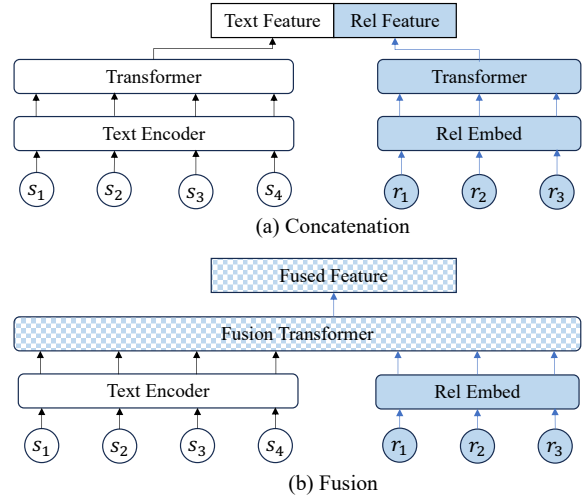


Figure 2: Two ways to combine text- and relation-based features: concatenation vs. fusion.

being much shorter than the word sequence of the text. The results on the TOEFL P1 dataset are more encouraging, with the BiLSTM classifier using relations as input outperforming the counterpart based on raw text. These results indicate that discourse relations parsed from the document are useful for coherence modeling. We further investigate the importance of the relation order by training another classifier on the shuffled relation sequence, and show the result in Table 2. The results of the classifier trained on the shuffled relations lag behind the counterpart trained on the original relation sequence by more than 7 points, suggesting that transition patterns between relations are crucial for coherence modeling. Finally, we compare the correct predictions between classifiers trained on raw text and the relation sequence, and find that only about 60% of them are overlapping. This suggests that raw text and the relation sequence provide different information for coherence assessment.

## 4 Discourse Relation-Enhanced Coherence Modeling

Inspired by the above analysis, in this section, we explore approaches to combine text- and relation-based features for coherence modeling. A straightforward way to use both types of information is to extract text- and relation-based features separately, concatenate them, and feed them into a classifier (as shown in Figure 2a). However, simple concatenation does not consider any potential interaction between two types of features. Prior studies (Ji et al., 2016; Yu et al., 2022) have demonstrated that incorporating discourse relations into

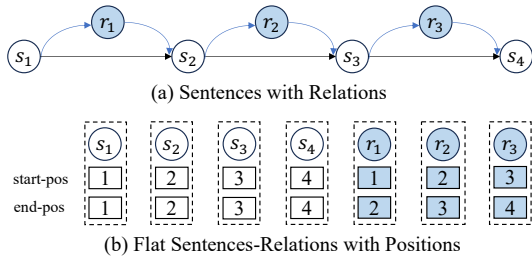


Figure 3: Converting original sentences and parsed relations (a) into a flat sentence-relation structure (b), where start\_pos and end\_pos denote the start and end positions of the node in the original sentence sequence.

language models can lead to better text representations. Therefore, we investigate a fusion model to facilitate the interaction between text and relation information.

Figure 2b shows the overall architecture of the proposed model. First, we use a text encoder and a relation embedding layer to generate sentence and relational representations, respectively. Specifically, given a text  $d = \{s_1, s_2, \dots, s_L\}$  with  $L$  sentences, we input the entire text to a text encoder to obtain representations of tokens  $\{e_1^t, e_2^t, \dots, e_N^t\}$ , where  $N$  is the number of tokens in the text. The text encoder can be a pre-trained language model (PLM), such as RoBERTa (Liu et al., 2019b), or a large-scale language model (LLM), such as Llama (Touvron et al., 2023). Following previous work (Jeon and Strube, 2022), we derive sentence representation by averaging representations of tokens<sup>3</sup> contained in the sentence, i.e.,  $e_j^s = \frac{1}{M} \sum_{t_i \in s_j} e_i^t$ , where  $M$  is the number of tokens in sentence  $s_j$ . Regarding discourse relations  $\{r_1, \dots, r_{L-1}\}$  parsed from the text, we embed each relation  $r_j$  into a vector  $e_j^r = \text{Embed}(r_j)$ , where Embed denotes a relation embedding lookup table. Then, we input sentences and relations into a fusion transformer. The challenge here is how to promote the interaction between sentence and relation representations while ensuring that sentences attend to the right relations (and vice versa). We address it through three components: (1) a flat structure of sentences and relations with positional information; (2) a position-aware attention; and (3) a visibility matrix between sentences and relations.

#### 4.1 Flat structure with positions

After applying the discourse parser, we obtain sentences of the text (the lower part of Figure 3a) and

<sup>3</sup>We also experimented with [CLS] pooling but found average pooling is consistently better, see Appendix D.2.

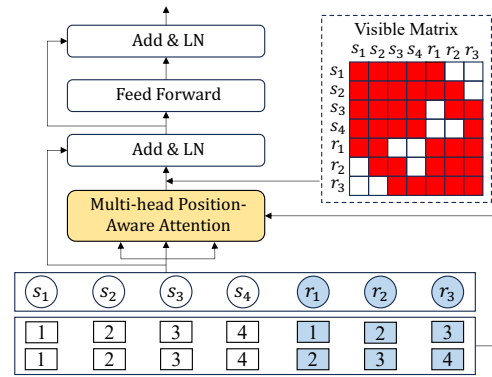


Figure 4: Fusion Transformer.

discourse relations between adjacent sentences (the upper part of Figure 3a), forming a graph structure. However, since the Transformer is designed for sequence modeling (Vaswani et al., 2017), it is not straightforward for the Transformer to process a graph-structured input. One possible solution is to insert relations into the sentence sequence, i.e.,  $[s_1, r_2, s_2, \dots]$ , but the resulting new sequence is no longer natural text.

To address these issues, we introduce a flat structure to organize sentences and relations, in which these two types of sequences are concatenated and equipped with positional information (as shown in Figure 3b). Specifically, sentences and relations are represented as a sequence of triples, where each triple contains three elements: (1) a node, which can be either a sentence or a relation; (2) start\_pos and (3) end\_pos, denoting the start and end position of the node in the original sentence sequence, respectively. If the node is a sentence, the start and end positions are the same. If the node is a relation, the start and end positions are different, indicating which two sentences the relation works on. For example,  $(s_1, 1, 1)$  denotes that this is the first sentence in the text, while  $(r_1, 1, 2)$  means that this is a discourse relation connecting the first and second sentences of the text. With this flat structure, we can maintain the original order information of sentences while facilitating the Transformer’s processing of these two features (see next section).

#### 4.2 Position-aware attention

The vanilla Transformer encodes the sequence using absolute positions, which does not apply to our flat structure input. Taking  $s_1$  and  $r_1$  in Figure 3b as an example, they are related, but their absolute positions are far apart. Inspired by the self-attention in Dai et al. (2019) and Li et al. (2020), we investigate position-aware attention to facilitate the

interaction between relevant sentence and relation nodes. The position-aware attention between the  $i$ -th and the  $j$ -th nodes is defined as:

$$\mathbf{A}_{ij} = \mathbf{q}_i \mathbf{k}_j^T + \mathbf{q}_i \mathbf{r}_{i-j}^T + \mathbf{u} \mathbf{k}_j^T + \mathbf{v} \mathbf{r}_{i-j}^T \quad (1)$$

where  $\mathbf{q}_i, \mathbf{k}_j, \mathbf{r}_{i-j} = \mathbf{e}_i \mathbf{W}_q, \mathbf{e}_j \mathbf{W}_k, \mathbf{p} \mathbf{e}_{i-j} \mathbf{W}_r$ ,  $\mathbf{e}_i$  means the representation of the  $i$ -th node,  $\mathbf{p} \mathbf{e}_{i-j}$  denotes the relative position embedding between the  $i$ -th and the  $j$ -th nodes, and  $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_r, \mathbf{u}, \mathbf{v}$  are trainable parameters. The first and third terms in Eq. 1 are content-based addressing, where the former calculates weight between query and key, and the latter governs a global content bias (Dai et al., 2019). The second and last terms compute weight with relative positional information, which can be used to guide the attention between relevant sentences and relations. Specifically, since each triple in the flat structured input contains two positional information (i.e., `start_pos` and `end_pos`), we can calculate four types of relative distances between the  $i$ -th and the  $j$ -th nodes: (i) `starti - startj`; (ii) `starti - endj`; (iii) `endi - startj`; (iv) `endi - endj`. Under the guidance of relative positional information, a sentence will not only attend neighboring sentences but also the relation that acts upon it. Taking  $s_1$  and  $r_1$  in Figure 3b as an example, the distance between the start positions (`start_pos`) of the two nodes is 0, indicating they are very related. The final relative position embedding between the  $i$ -th and the  $j$ -th nodes, i.e.,  $\mathbf{p} \mathbf{e}_{i-j}$ , is defined as a non-linear transformation over the four relative distances:

$$\mathbf{p} \mathbf{e}_{i-j} = (\mathbf{p}_{s_i-s_j} \otimes \mathbf{p}_{s_i-e_j} \otimes \mathbf{p}_{e_i-e_j} \otimes \mathbf{p}_{e_i-e_j}) \mathbf{W}_p \quad (2)$$

The position embedding  $\mathbf{p}$  is initialized as in Transformer, where  $\mathbf{p}_{pos}^{2k} = \sin(pos/10000^{2k/d_{model}})$  and  $\mathbf{p}_{pos}^{2k+1} = \cos(pos/10000^{2k/d_{model}})$  (Vaswani et al., 2017).

### 4.3 Visible matrix

While relative position embedding can effectively guide attention calculation, sentence nodes may still attend to irrelevant relation nodes, such as  $s_1$  attending to  $r_3$  (see Figure 3a), leading to a poor text representation. Thus, we further introduce a visible matrix  $\mathbf{M}$  to prevent this. The  $\mathbf{M}$  is defined as:

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{if } \text{cond}_1 \mid \text{cond}_2 \mid \text{cond}_3 \\ -\infty, & \text{otherwise} \end{cases} \quad (3)$$

where  $\text{cond}_1$  and  $\text{cond}_2$  are defined as nodes  $i$  and  $j$  are both sentences or relations, and  $\text{cond}_3$  is defined as nodes  $i$  and  $j$  is one sentence and one relation, and the relation works on the sentence. We apply the visible matrix to the attention calculation:

$$\mathbf{A}^* = \text{Softmax}(\mathbf{A} + \mathbf{M}) \quad (4)$$

Then layer normalizations and a feed-forward network (as shown in Figure 4) are applied to produce the text representation  $\mathbf{v}$ . Finally, we input  $\mathbf{v}$  into a softmax classifier, and use the cross-entropy loss for training.

## 5 Experiments

We conduct experiments on the GCDC (Lai and Tetreault, 2018) and TOEFL (Blanchard et al., 2013) datasets to show the effectiveness of relation features for coherence modeling.

**Implementation Details.** We implement our model using the Pytorch library, experiment with two different text encoders, a pre-trained language model RoBERTa<sub>base</sub> (Liu et al., 2019b), and a large language model LLama2<sub>7B</sub> (Touvron et al., 2023), and initialize the relation embedding with Glove (Pennington et al., 2014). We use the AdamW optimizer with an initial learning rate of 1e-3, a batch size of 32, and a maximum epoch number of 20 for training. Considering the training variability in GCDC, we follow the setting in Lai and Tetreault (2018) to perform 10-fold cross-validation over the training dataset. Regarding the TOEFL dataset, we conduct 5-fold cross-validation on the dataset of each prompt, which is a common setting for the AES task (Taghipour and Ng, 2016). Like previous work (Frag and Yanakoudakis, 2019; Jeon and Strube, 2022), we use standard accuracy (Acc, %) as evaluation metrics.<sup>4</sup> We show more detailed settings in Appendix D.1.

**Baselines.** To investigate the usefulness of discourse relations, we compare with a baseline using only textual input without any relation information:

- **TextOnly.** This model consists of a text encoder to obtain sentence representation, a sentence-level transformer to extract coherence patterns, and a softmax classifier for prediction.

To show the effectiveness of our fusion strategy, we compare it with the concatenate baseline:

<sup>4</sup>We also report the results of Macro-F1 in Appendix D.3.

Model		Clinton	Enron	Yahoo	Yelp	Avg
Jeon and Strube (2022)		64.20 <sub>0.40</sub>	55.30 <sub>0.30</sub>	58.40 <sub>0.20</sub>	57.30 <sub>0.20</sub>	58.90
Liu et al. (2023)		66.20 <sub>0.81</sub>	57.00 <sub>0.81</sub>	<b>63.65</b> <sub>0.74</sub>	58.05 <sub>1.21</sub>	61.23
LLama-Prompt		62.60 <sub>1.59</sub>	57.35 <sub>1.42</sub>	60.05 <sub>1.41</sub>	57.50 <sub>1.02</sub>	59.36
GPT4-Prompt		53.00	53.00	50.00	49.00	51.25
RoBERTa	TextOnly	64.55 <sub>0.69</sub>	57.50 <sub>0.89</sub>	60.05 <sub>0.35</sub>	58.20 <sub>0.75</sub>	60.10
	Concat	65.45 <sub>0.79</sub>	58.30 <sub>0.56</sub>	61.35 <sub>0.67</sub>	59.05 <sub>0.57</sub>	61.04
	Our Method	<b>66.25</b> <sub>0.64</sub>	<b>59.60</b> <sub>1.26</sub>	63.05 <sub>0.42</sub>	<b>60.20</b> <sub>0.95</sub>	<b>62.28</b>
LLama	TextOnly	63.90 <sub>0.49</sub>	57.05 <sub>0.79</sub>	59.60 <sub>0.49</sub>	57.35 <sub>0.74</sub>	59.47
	Concat	64.10 <sub>0.66</sub>	57.15 <sub>0.50</sub>	61.15 <sub>0.81</sub>	58.35 <sub>0.71</sub>	60.19
	Our Method	65.75 <sub>0.46</sub>	59.30 <sub>0.98</sub>	61.70 <sub>0.78</sub>	59.45 <sub>0.99</sub>	61.55

Table 3: Mean accuracy results (with std) on the GCDC dataset.

- **Concat.** This baseline simply concatenates text- and relation-based features without considering interactions between them (see Figure 2a).

Recently, prompt-based methods using LLM have significantly impacted various NLP tasks. Therefore, we also compare to baselines of this trend:

- **LLama-Prompt.** Using LoRA (Hu et al., 2022) to tune LLama2-7B, and predict the coherence of an input document with the designed template.<sup>5</sup>
- **GPT4-Prompt.** Calling GPT 4 API and applying in-context learning<sup>5</sup> (Min et al., 2022) for coherence assessment.

Further, we compare our method against previous state-of-the-art models on each corpus.

## 5.1 Overall Results

**GCDC.** Table 3 presents the results on the GCDC dataset, where the last two blocks show the results based on RoBERTa and LLama. When using RoBERTa as the text encoder, both Concat and Our Method outperform the TextOnly baseline, indicating that relation features are helpful for coherence assessment. The improvement of Concat over the TextOnly baseline is limited, with an increase in accuracy of less than one point (60.10 → 61.04). We argue simply concatenating text- and relation-based features can not fully utilize relation information since the two features are processed separately, without considering the interaction between them. Compared to Concat, Our Method shows a greater improvement, increasing by 2.18% in accuracy, suggesting that our approach is more efficient in utilizing relation information.

<sup>5</sup>We show the prompts used for LLama and GPT 4 baselines in figures 6 and 7.

When using LLama as the text encoder, similar results are observed, showing that relation features are useful across different encoders. Surprisingly, our method implemented with RoBERTa performs better than the counterpart with LLama (62.28 vs. 61.55) despite the latter having more parameters and pre-trained on more corpus than the former. We suspect this is because RoBERTa learns bidirectional context-aware representation while LLama is limited by its uni-directional context (Yang et al., 2019). Recent work also observed similar results of RoBERTa and LLama on other text classification tasks (Rodriguez-Garcia et al., 2024).

The second block in Table 3 shows the results of prompt-based methods, including LLama-Prompt and GPT4-Prompt. Similar to using LLama as a text encoder, the performance of LLama-prompt also underperforms our method using RoBERTa, with an accuracy gap of 2.92%. The performance of GPT 4 on this task is even worse, lagging behind our method (RoBERTa) by 11% in accuracy. This is consistent with previous findings that GPT 4 achieves a certain level of accuracy in scoring essays but still underperforms trained models (Mizumoto and Eguchi, 2023). To further show the importance of relation features and the efficiency of our method, we compare against two state-of-the-art models (Jeon and Strube, 2022; Liu et al., 2023) on this corpus. The two models are entity-based, and their results are shown in the first block of Table 3. Our method, using relation features, outperforms the two entity-based models for coherence assessment, indicating its superiority for this task. **TOEFL.** Results on the TOEFL dataset are shown in Table 4. Similar to the observations on the GCDC dataset, relation features contribute to coherence modeling. When using RoBERTa as the text encoder, Concat and Our Method outperform the

Model		Prompt								Avg
		1	2	3	4	5	6	7	8	
Jeon and Strube (2022)		78.38 <sub>0.00</sub>	75.70 <sub>0.30</sub>	76.58 <sub>0.00</sub>	76.56 <sub>0.00</sub>	<b>79.10</b> <sub>0.00</sub>	76.41 <sub>0.00</sub>	75.03 <sub>0.00</sub>	74.54 <sub>0.00</sub>	76.54
Liu et al. (2023)		75.79 <sub>1.14</sub>	76.25 <sub>1.07</sub>	74.14 <sub>1.18</sub>	75.81 <sub>0.71</sub>	77.01 <sub>0.94</sub>	77.08 <sub>1.14</sub>	73.55 <sub>0.80</sub>	72.91 <sub>0.66</sub>	75.34
LLama-Prompt		76.81 <sub>1.36</sub>	76.12 <sub>1.12</sub>	76.57 <sub>1.23</sub>	75.55 <sub>1.06</sub>	76.93 <sub>1.16</sub>	76.33 <sub>1.04</sub>	76.10 <sub>0.96</sub>	74.73 <sub>1.37</sub>	76.14
GPT4-Prompt		59.21	58.65	64.28	58.27	58.48	65.10	60.23	59.34	57.25
RoBERTa	TextOnly	76.36 <sub>0.90</sub>	75.10 <sub>1.03</sub>	75.29 <sub>0.51</sub>	75.33 <sub>1.47</sub>	75.90 <sub>1.01</sub>	75.61 <sub>1.88</sub>	73.76 <sub>0.91</sub>	73.34 <sub>1.06</sub>	75.08
	Concat	77.63 <sub>1.31</sub>	75.87 <sub>0.36</sub>	76.72 <sub>0.93</sub>	76.66 <sub>1.87</sub>	78.20 <sub>1.14</sub>	77.08 <sub>1.31</sub>	75.48 <sub>0.69</sub>	74.92 <sub>1.15</sub>	76.57
	Our Method	<b>78.97</b> <sub>0.75</sub>	<b>77.21</b> <sub>0.99</sub>	<b>77.59</b> <sub>0.92</sub>	<b>77.19</b> <sub>0.90</sub>	78.45 <sub>1.14</sub>	<b>78.22</b> <sub>1.57</sub>	<b>76.78</b> <sub>0.96</sub>	<b>75.85</b> <sub>1.06</sub>	<b>77.49</b>
LLama	TextOnly	74.96 <sub>1.17</sub>	74.45 <sub>1.55</sub>	74.71 <sub>0.43</sub>	73.81 <sub>1.45</sub>	75.65 <sub>1.55</sub>	75.62 <sub>0.96</sub>	74.64 <sub>0.93</sub>	73.34 <sub>1.02</sub>	74.65
	Concat	75.94 <sub>0.75</sub>	75.85 <sub>1.21</sub>	75.31 <sub>0.56</sub>	74.47 <sub>1.47</sub>	76.50 <sub>1.19</sub>	76.35 <sub>0.98</sub>	75.12 <sub>0.74</sub>	73.58 <sub>1.23</sub>	75.39
	Our Method	77.16 <sub>1.12</sub>	76.89 <sub>1.33</sub>	76.29 <sub>0.71</sub>	76.19 <sub>1.04</sub>	77.41 <sub>1.12</sub>	77.29 <sub>1.06</sub>	76.31 <sub>0.82</sub>	75.19 <sub>0.94</sub>	76.59

Table 4: Mean accuracy results (with std) on the TOEFL dataset.

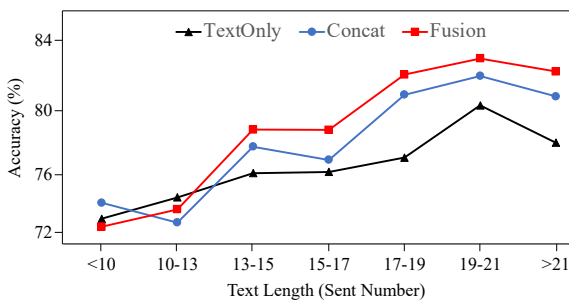


Figure 5: Accuracy against text length.

TextOnly baseline by 1.49% and 2.41% in accuracy, respectively. The same results are observed when using LLama as a text encoder, where the improvement of Concat and Our Method over the TextOnly baseline is 0.72% and 2.05%, respectively. In both settings, Our Method outperforms Concat, demonstrating the effectiveness of fusion for text- and relation-based features. Despite being quite popular in recent research, prompt-based methods, including LLama-Prompt and GPT4-Prompt, are comparable to other baselines and slightly inferior to our method using RoBERTa or LLama as the text encoder. We further compare our method with previous entity-based approaches. Results in Table 4 show that our approach performs better than the two models, highlighting the usefulness of relation features for this task.

## 5.2 Performance Analysis

We conducted two analyses to understand why relation features perform well in coherence assessment. First, we compare the performance of Our Method and Concat with the TextOnly baseline across different document lengths (in terms of sentence number). Figure 5 shows the accuracy trends of these three models (using RoBERTa) on the TOEFL P1 dataset as the number of sentences increases. Our

Model	Enron → Others		TOEFL P1 → Others	
	RoBERTa	LLama	RoBERTa	LLama
TextOnly	51.83	47.50	71.88	67.70
Concat	53.50(+1.67)	49.83(+2.33)	74.86(+2.98)	70.93(+3.23)
Our Method	56.33(+4.50)	52.33(+4.83)	75.52(+3.64)	72.49(+4.79)

Table 5: Cross-domain accuracy of models.

Method and Concat show comparable performance to the TextOnly baseline at the beginning but gradually outperform the baseline when the sentence number increases, demonstrating that relation information contributes to learning better coherence patterns for long documents. Our Method consistently outperforms Concat, indicating that it is more efficient in exploiting relation features.

To probe whether our model has truly learned better coherence patterns, we further examine its transferability in cross-domain settings. Specifically, we train TextOnly, Concat, and Our Method on Enron of GCDC (or Prompt 1 of TOEFL), and evaluate their performance on other parts of GCDC (or other prompts of TOEFL) datasets. Table 5 shows the results of these three models. With relation information, Concat and Our Method consistently show better performance than the TextOnly baseline in the cross-domain setting, indicating the relation sequence of texts can serve as domain-agnostic features for coherence assessment. Our Method outperforms the Concat baseline in all cross-domain experiments, showing the superiority of fusion toward simple concatenation.

## 5.3 Ablation Study

We conduct ablation studies to evaluate the effectiveness of position-aware attention (PAA) and visible matrix (VM). Specifically, we first remove the visible matrix, and then replace the position-aware attention with a vanilla one. Table 6 shows the



Model	RoBERTa		LLama	
	Enron	TOEFL P1	Enron	TOEFL P1
Our Method	59.60	78.97	59.30	77.16
- VM	59.20	78.12	58.55	76.26
- VM, PAA	58.15	77.24	57.40	75.68

Table 6: Ablation study for visible matrix (VM) and position-aware attention (PAA) in our method.

results on the GCDC Enron and TOEFL P1 dataset using RoBERTa. We can observe that each component contributes to the performance, showing their essential to achieve good performance. Furthermore, the performance drop from removing the position-aware attention mechanism is greater than that from eliminating the visible matrix, indicating that relative position information is more important in guiding fusion.

## 6 Conclusions

In this paper, we provide empirical evidence to demonstrate the correlation between discourse relations and text coherence. Then, we introduce a novel fusion model to combine text- and relation-based features for coherence assessment. Experiments on two benchmarks show that our method consistently outperforms various baseline models, demonstrating the importance of relation features and the effectiveness of our approach.

## 7 Limitations

This study focuses on investigating whether discourse relations contribute to coherence assessment. Despite achieving position results, our method has several limitations that can be further explored in future work. Firstly, the performance of the PDTB parser used in this work is far from perfect. Future efforts should focus on building more powerful parsers to facilitate the analysis of discourse relations' role in coherence modeling. Secondly, we only experiment with PDTB relations. Extending our findings to other relations, such as RST relations (Mann and Thompson, 1988), would be very interesting. Finally, this study only considered discourse relations for coherence modeling and did not investigate whether they can be combined with other coherence patterns, such as entity-based patterns (Barzilay and Lapata, 2008). Therefore, exploring the integration of different patterns to improve coherence assessment would be an encouraging direction.

## Acknowledgements

The authors would like to thank the four anonymous reviewers for their comments. We also thank Xiyang Fu for her valuable suggestions on the model design. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany.

## References

- Tushar Abhishek, Daksh Rawat, Manish Gupta, and Vasudeva Varma. 2021. [Transformer models for text coherence assessment](#). *CoRR*, abs/2109.02176.
- Regina Barzilay and Mirella Lapata. 2005. [Modeling local coherence: An entity-based approach](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.
- Regina Barzilay and Mirella Lapata. 2008. [Modeling local coherence: An entity-based approach](#). *Computational Linguistics*, 34(1):1–34.
- Or Biran and Kathleen McKeown. 2015. [Discourse planning with an n-gram model of relations](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1973–1977, Lisbon, Portugal. Association for Computational Linguistics.
- Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. [Toefl11: A corpus of non-native english](#). *ETS Research Report Series*, 2013(2):i–15.
- Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. [Using entity-based features to model coherence in student essays](#). In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 681–684, Los Angeles, California. Association for Computational Linguistics.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. [Transformer-XL: Attentive language models beyond a fixed-length context](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.
- Micha Elsner and Eugene Charniak. 2011. [Extending the entity grid with entity-specific features](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129, Portland, Oregon, USA. Association for Computational Linguistics.
- Youmna Farag and Helen Yannakoudakis. 2019. [Multi-task learning for coherence modeling](#). In *Proceedings of the 57th Annual Meeting of the Association for*

- Computational Linguistics*, pages 629–639, Florence, Italy. Association for Computational Linguistics.
- Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. [The impact of deep hierarchical discourse structures in the evaluation of text coherence](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Katja Filippova and Michael Strube. 2007. [Extending the entity-grid coherence model to semantically related entities](#). In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 139–142, Saarbrücken, Germany. DFKI GmbH.
- Xiyan Fu and Anette Frank. 2023. [SETI: Systematicity evaluation of textual inference](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4101–4114, Toronto, Canada. Association for Computational Linguistics.
- Xiyan Fu and Anette Frank. 2024a. [Exploring continual learning of compositional generalization in NLI](#). *Transactions of the Association for Computational Linguistics*, 12:912–932.
- Xiyan Fu and Anette Frank. 2024b. [The mystery of compositional generalization in graph-based generative commonsense reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8376–8394, Miami, Florida, USA. Association for Computational Linguistics.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. [Centering: A framework for modeling the local coherence of discourse](#). *Computational Linguistics*, 21(2):203–225.
- Camille Guinaudeau and Michael Strube. 2013. [Graph-based local coherence modeling](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria. Association for Computational Linguistics.
- Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. [The utility of discourse parsing features for predicting argumentation structure](#). In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Sungho Jeon and Michael Strube. 2020a. [Centering-based neural coherence modeling with hierarchical discourse segments](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7458–7472, Online. Association for Computational Linguistics.
- Sungho Jeon and Michael Strube. 2020b. [Incremental neural lexical coherence modeling](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6752–6758, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Sungho Jeon and Michael Strube. 2022. [Entity-based neural local coherence modeling](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7787–7805, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. [A latent variable recurrent neural network for discourse-driven language models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342, San Diego, California. Association for Computational Linguistics.
- Shafiq Joty, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen. 2018. [Coherence modeling of asynchronous conversations: A neural entity grid approach](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 558–568, Melbourne, Australia. Association for Computational Linguistics.
- Daniel Jurafsky and James H Martin. 2021. [Speech and language processing \(3rd ed. draft\)](#).
- René Knaebel. 2021. [discopy: A neural system for shallow discourse parsing](#). In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. [Can transformer models measure coherence in text: Re-thinking the shuffle test](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.
- Alice Lai and Joel Tetreault. 2018. [Discourse coherence in the wild: A dataset, evaluation and methods](#). In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.
- Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI’05*,

- page 1085–1090, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Yuan Yuan Lei and Ruihong Huang. 2023. [Discourse structures guided fine-grained propaganda identification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 331–342, Singapore. Association for Computational Linguistics.
- Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. [FLAT: Chinese NER using flat-lattice transformer](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842, Online. Association for Computational Linguistics.
- Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. [Automatically evaluating text coherence using discourse relations](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.
- Wei Liu, Xiyan Fu, and Michael Strube. 2023. [Modeling structural similarities between documents for coherence assessment with graph convolutional networks](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7792–7808, Toronto, Canada. Association for Computational Linguistics.
- Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021a. [Lexicon enhanced Chinese sequence labeling using BERT adapter](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5847–5858, Online. Association for Computational Linguistics.
- Wei Liu and Michael Strube. 2023. [Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.
- Wei Liu, Stephen Wan, and Michael Strube. 2024. [What causes the failure of explicit to implicit discourse relation recognition?](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2738–2753, Mexico City, Mexico. Association for Computational Linguistics.
- Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. 2019a. [An encoding strategy based word-character LSTM for Chinese NER](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2379–2389, Minneapolis, Minnesota. Association for Computational Linguistics.
- Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021b. [Exploring discourse structures for argument impact classification](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3958–3969, Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. [RoBERTa: A robustly optimized BERT Pretraining Approach](#). *CoRR*, abs/1907.11692.
- William C. Mann and Sandra A. Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of text organization](#). *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Mohsen Mesgar and Michael Strube. 2015. [Graph-based coherence modeling for assessing readability](#). In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 309–318, Denver, Colorado. Association for Computational Linguistics.
- Mohsen Mesgar and Michael Strube. 2018. [A neural local coherence model for text quality assessment](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium. Association for Computational Linguistics.
- Todor Mihaylov and Anette Frank. 2019. [Discourse-aware semantic self-attention for narrative reading comprehension](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2541–2552, Hong Kong, China. Association for Computational Linguistics.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Atsushi Mizumoto and Masaki Eguchi. 2023. [Exploring the potential of using an ai language model for automated essay scoring](#). *Research Methods in Applied Linguistics*, 2(2):100050.
- Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. [On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers](#). In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*,

- pages 68–82, Online. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. [The Penn Discourse TreeBank 2.0](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Raquel Rodriguez-Garcia, Julio Reyes Montesinos, Jesus M. Fraile-Hernandez, and Anselmo Peñas. 2024. [HAMiSoN-ensemble at ClimateActivism 2024: Ensemble of RoBERTa, llama 2, and multi-task for stance detection](#). In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 118–124, St. Julians, Malta. Association for Computational Linguistics.
- Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. [Discourse coherence: Concurrent explicit and implicit relations](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia. Association for Computational Linguistics.
- Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. 2021. [Evaluating document coherence modeling](#). *Transactions of the Association for Computational Linguistics*, 9:621–640.
- Kaveh Taghipour and Hwee Tou Ng. 2016. [A neural approach to automated essay scoring](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.
- Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. [Hierarchical modeling of global context for document-level neural machine translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.
- Dat Tien Nguyen and Shafiq Joty. 2017. [A neural local coherence model](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Ziao Wang, Xiaofeng Zhang, and Hongwei Du. 2021. [Building the directed semantic graph for coherent long text generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2563–2572, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Bonnie Webber. 2009. [Genre distinctions for discourse in the Penn TreeBank](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, Suntec, Singapore. Association for Computational Linguistics.
- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse TreeBank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.
- Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32, pages 5602–5609.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Changlong Yu, Hongming Zhang, Yangqiu Song, and Wilfred Ng. 2022. [CoCoLM: Complex commonsense enhanced language model with discourse relations](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1175–1187, Dublin, Ireland. Association for Computational Linguistics.

Explicit	Distribution	Implicit	Distribution
Asynchronous	8.69%	Asynchronous	4.64%
Cause	7.87%	Cause	24.23%
Concession	19.94%	Cause+Belief	0.82%
Condition	5.99%	Concession	6.72%
Conjunction	36.55%	Condition	0.85%
Contrast	4.58%	Conjunction	20.84%
Disjunction	1.23%	Contrast	3.86%
Instantiation	1.30%	Equivalence	1.21%
Level-of-detail	1.01%	Instantiation	6.84%
Manner	1.23%	Level-of-detail	14.60%
Negative-condition	0.54%	Manner	0.74%
Purpose	1.63%	Purpose	3.31%
Similarity	0.42%	Substitution	1.34%
Substitution	0.96%	Synchronous	2.35%
Synchronous	8.07%	NoRel	8.18%

Table 7: Explicit and Implicit relations used in this study and their distribution in the training corpus.

## A PDTB Parser

We use an updated version of *discopy* (Knaebel, 2021) to parse discourse relations from documents. The first update is to replace PDTB 2.0 (Prasad et al., 2008) relations with PDTB 3.0 (Webber et al., 2019) ones. In detail, we consider explicit and implicit relations between adjacent sentences of a text. For the explicit part, we use 15 discourse relations since they have enough examples for training (Liu et al., 2024). Similarly, we use the 14 most common relations and also one "NoRel" label since no relation is common in low coherent texts. Table 7 shows the relations used in this work and their distributions in PDTB 3.0. The second update is to adopt the model of Liu and Strube (2023) for implicit relation recognition since it achieves state-of-the-art performance. We implement this parser with RoBERTa and train it on the PDTB 3.0 using the data splitting of Ji and Eisenstein (2015). This parser achieves an accuracy of 89.61% and 67.80% on explicit and implicit test sets of PDTB 3.0, respectively.

## B Dataset

The GCDC dataset contains texts from four domains: **Yahoo** online forum posts, emails from **Enron**, emails from Hillary **Clinton**'s office, and **Yelp** online business reviews. Similarly, the TOEFL dataset consists of essays from eight different prompts. The statistics of the GCDC and TOEFL datasets are shown in Table 8.

Dataset	Split	#Doc	Avg #Sent	Avg #Word	
GCDC	Clinton	Train	1000	8.9	182.9
		Test	200	8.8	186.0
	Enron	Train	1000	9.2	185.1
		Test	200	9.3	191.1
	Yahoo	Train	1000	7.8	157.2
		Test	200	7.8	162.7
	Yelp	Train	1000	10.4	178.2
		Test	200	10.1	179.1
TOEFL	Prompt 1	Total	1656	13.7	339.1
	Prompt 2	Total	1562	15.7	357.8
	Prompt 3	Total	1396	14.7	343.5
	Prompt 4	Total	1509	15.1	338.0
	Prompt 5	Total	1648	15.2	358.4
	Prompt 6	Total	960	15.3	358.3
	Prompt 7	Total	1686	14.0	336.6
	Prompt 8	Total	1683	14.7	340.9

Table 8: Statistics of datasets, where #Doc, #Sent, and #Word mean the number of documents, sentences, and words, respectively.

Relation	GCDC Enron	TOEFL P1
Conjunction	33.47%	19.29%
Cause	25.72%	37.40%
Concession	8.92%	7.60%
Level-of-detail	13.48%	11.99%
Asynchronous	4.51%	1.69%
Synchronous	0.80%	0.61%
Contrast	1.09%	4.65%
Instantiation	1.49%	10.58%
NoRel	8.55%	1.12%
Condition	0.32%	0.41%
Purpose	0.32%	0.17%
Substitution	0.57%	1.31%
Manner	0.01%	0.01%
Disjunction	0.01%	0.06%
Equivalence	0.39%	2.56%
Cause+belief	0.26%	0.37%
Negative-condition	0.08%	0.14%
Similarity	0.01%	0.04%

Table 9: The distribution of discourse relations parsed from GCDC Enron and TOEFL P1 corpora.

## C More Analysis Results

### C.1 Distribution of parsed relations

Table 9 shows the distribution of parsed relations from GCDC Enron and TOEFL P1 datasets. Texts in the Enron dataset are emails, using *Conjunction* relations frequently to link pieces of information. By contrast, essays in TOEFL datasets are viewpoint-oriented, using *Cause* and *Instantiation* relations to support their opinions. Therefore, we see the relation distribution of the two corpora are very different.

### C.2 Top-level or second-level relations?

In Sections 3.2 and 3.3, we use the second-level discourse relations for analysis because fine-grained relations can provide more information for coherence assessment. To verify this point, we re-

Model		Clinton	Enron	Yahoo	Yelp	Avg
Liu et al. (2023)		48.49 <sub>1.61</sub>	45.67 <sub>1.57</sub>	<b>51.92</b> <sub>1.06</sub>	44.18 <sub>1.10</sub>	47.66
LLama-Prompt		47.63 <sub>1.91</sub>	51.37 <sub>1.87</sub>	47.50 <sub>1.96</sub>	46.41 <sub>1.87</sub>	48.23
GPT4-Prompt		44.22	45.03	43.00	41.96	43.55
RoBERTa	TextOnly	47.58 <sub>1.09</sub>	48.74 <sub>1.05</sub>	45.71 <sub>0.86</sub>	45.63 <sub>0.78</sub>	46.92
	Concat	50.49 <sub>1.21</sub>	49.42 <sub>1.25</sub>	47.97 <sub>1.29</sub>	46.74 <sub>0.98</sub>	48.66
	Our Method	52.28 <sub>1.02</sub>	52.51 <sub>1.58</sub>	49.61 <sub>1.80</sub>	47.44 <sub>0.95</sub>	50.46
LLama	TextOnly	48.51 <sub>1.10</sub>	48.99 <sub>0.84</sub>	47.31 <sub>1.25</sub>	46.38 <sub>0.98</sub>	47.80
	Concat	51.98 <sub>0.43</sub>	48.69 <sub>1.00</sub>	48.42 <sub>1.34</sub>	46.93 <sub>1.34</sub>	49.00
	Our Method	<b>52.90</b> <sub>1.17</sub>	<b>52.62</b> <sub>0.79</sub>	49.94 <sub>1.02</sub>	<b>47.67</b> <sub>0.52</sub>	<b>50.78</b>

Table 10: Mean macro-F1 results (with std) on the GCDC dataset.

Input Type	GCDC Enron		TOEFL P1	
	Acc	F1	Acc	F1
Raw Text	46.20 <sub>0.77</sub>	42.86 <sub>0.97</sub>	57.55 <sub>1.24</sub>	50.39 <sub>0.78</sub>
Rel Sequence (Top)	41.35 <sub>0.74</sub>	34.75 <sub>0.84</sub>	55.56 <sub>0.71</sub>	49.16 <sub>1.29</sub>
Rel Sequence (Second)	44.15 <sub>0.92</sub>	39.43 <sub>1.24</sub>	59.17 <sub>0.87</sub>	53.51 <sub>0.99</sub>

Table 11: The performance (with std) of BiLSTM classifier when using text, top-level relation sequence, and second-level relation sequence as input.

conduct the analysis in Section 3.3 but replace the second-level relations with the top-level ones<sup>6</sup>. Table 11 shows the performance comparison between different settings. The classifier trained on the second-level relations consistently outperforms the counterpart trained on the top-level relations, with a gap of about 3 to 4 points. This demonstrates that more fine-grained relations are more helpful in coherence assessment.

## D Experiments

### D.1 Detailed experimental setting

For experiments on both GCDC and TOEFL datasets, whether using RoBERTa or LLama as the text encoder, we use an Adam optimizer with an initial learning rate of 0.001, a batch size of 32, and a maximum epoch number of 20 for training. However, different dropout values are used for different text encoders, with RoBERTa and LLama using values of 0.1 and 0.5, respectively. Following the previous work (Lai and Tetreault, 2018), we perform 10-fold cross-validation over the training dataset for the GCDC corpus. Similarly, a 5-fold cross-validation is performed on the dataset of each prompt, which is the common setting for this corpus (Jeon and Strube, 2020a). All the experiments

<sup>6</sup>This is achieved by training the same discourse parser but with top-level relation set, i.e., *Contingency*, *Comparison*, *Expansion*, and *Temporal* relations.

Text: [Tom was late for the meeting this morning. However, it was ....]  
Level of Coherence: High

Figure 6: Template for the LLama-Prompt baseline.

Replace the MASK token by selecting only one of the following coherence labels: [low, medium, high].

Examples:

Text: [text\_1]

Coherence level: low

Text: [text\_2]

Coherence level: medium

Text: [text\_3]

Coherence level: high

<other two examples for each coherence level>

Text: [target\_text]

Coherence level: [MASK]

Figure 7: Template for the GPT4-Prompt baseline.

are run on a single Tesla P40 GPU with 24 GB memory.

For the LLama-Prompt baseline, we follow existing work (Rodriguez-Garcia et al., 2024) formulating the coherence modeling as a generation task and using LoRA (Hu et al., 2022) to efficiently tune the LLama on the training set. For LoRA, we set the rank  $r$  as 64,  $lora\_alpha$  as 16,  $lora\_dropout$  as 0.1, and  $target\_modules$  as ["q\_proj", "o\_proj", "k\_proj", "gate\_proj", "up\_proj", "down\_proj"]. The template we used for the LLama-Prompt baseline is shown in Figure 6. Regarding the GPT4-Prompt baseline, we employ in-context learning with 3 examples for each coherence level. The version of GPT used in this work is "GPT-4o", which is very fast for generation and also affordable for us. We show the template for this baseline in Figure 7.

### D.2 [CLS] pooling or average pooling

We have tried both [CLS] pooling and average pooling in our experiments. We found that using

Model		Prompt								Avg
		1	2	3	4	5	6	7	8	
Liu et al. (2023)		72.56 <sub>1.65</sub>	72.31 <sub>1.68</sub>	73.14 <sub>1.90</sub>	74.39 <sub>1.49</sub>	73.02 <sub>1.70</sub>	73.93 <sub>1.25</sub>	71.66 <sub>1.16</sub>	70.39 <sub>1.40</sub>	72.68
LLama-Prompt		74.10 <sub>2.41</sub>	72.77 <sub>2.18</sub>	74.50 <sub>1.00</sub>	73.37 <sub>1.34</sub>	74.87 <sub>2.19</sub>	74.41 <sub>1.96</sub>	73.27 <sub>2.68</sub>	72.55 <sub>1.61</sub>	73.73
GPT4-Prompt		53.37	50.25	64.12	58.57	50.53	64.92	59.29	56.91	57.25
RoBERTa	TextOnly	73.07 <sub>1.24</sub>	71.36 <sub>0.68</sub>	74.44 <sub>0.59</sub>	72.28 <sub>1.60</sub>	72.91 <sub>1.28</sub>	72.25 <sub>1.79</sub>	71.28 <sub>1.46</sub>	70.30 <sub>1.50</sub>	72.23
	Concat	74.35 <sub>1.56</sub>	72.31 <sub>1.18</sub>	74.31 <sub>0.76</sub>	74.08 <sub>2.06</sub>	74.06 <sub>1.53</sub>	74.05 <sub>1.23</sub>	72.99 <sub>1.35</sub>	71.74 <sub>2.40</sub>	73.49
	Our Method	<b>76.42</b> <sub>1.24</sub>	<b>72.86</b> <sub>1.86</sub>	<b>75.60</b> <sub>0.76</sub>	<b>75.37</b> <sub>1.46</sub>	<b>75.70</b> <sub>0.43</sub>	<b>75.73</b> <sub>1.57</sub>	<b>72.76</b> <sub>2.02</sub>	<b>73.02</b> <sub>1.64</sub>	<b>74.71</b>
LLama	TextOnly	71.00 <sub>0.49</sub>	69.38 <sub>1.30</sub>	71.13 <sub>0.87</sub>	71.29 <sub>1.37</sub>	71.23 <sub>0.87</sub>	71.58 <sub>1.42</sub>	69.71 <sub>0.98</sub>	68.64 <sub>1.70</sub>	70.49
	Concat	71.90 <sub>1.04</sub>	71.63 <sub>0.64</sub>	72.14 <sub>1.03</sub>	71.91 <sub>0.85</sub>	72.78 <sub>1.31</sub>	72.82 <sub>1.58</sub>	70.05 <sub>1.07</sub>	70.57 <sub>1.50</sub>	71.72
	Our Method	73.36 <sub>1.10</sub>	<b>72.89</b> <sub>1.30</sub>	72.88 <sub>1.04</sub>	73.50 <sub>0.92</sub>	73.23 <sub>1.19</sub>	73.38 <sub>1.64</sub>	71.82 <sub>0.80</sub>	70.55 <sub>1.09</sub>	72.70

Table 12: Mean macro-F1 results (with std) on the TOEFL dataset.

the average pooling (to get sentence representations) consistently outperforms the [CLS] pooling. For instance, on the TOEFL P1 dataset (using a RoBERTa encoder), the accuracy of the TextOnly baseline and our method with average pooling are 76.36% and 78.97%, respectively, compared to 72.58% and 76.32% with [CLS] pooling. Average pooling captures information from all tokens in the sequence, preserving richer linguistic features, whereas [CLS] pooling relies solely on the [CLS] token representation, which may omit important contextual details. Similar findings are reported in Mosbach et al. (2020).

### D.3 Macro-F1 results

Typically, accuracy is commonly used as an evaluation metric for NLP tasks (Liu et al., 2019a, 2021a; Fu and Frank, 2023, 2024a,b), including coherence assessment (Frag and Yannakoudakis, 2019; Lai and Tetreault, 2018; Jeon and Strube, 2020a). Due to the uneven distribution of labels in the GCDC and TOEFL dataset, Liu et al. (2023) suggests to also report the Macro-F1 results for this task. We follow this setup and show the Macro-F1 results on the GCDC and TOEFL datasets in Tables 10 and 12, respectively. We can see from the tables 10 and 12 that the trend of the Macro-F1 results is similar to that of the accuracy results, where relation features contribute to performance and Our Method utilizes relations more efficiently than the Concat baseline.