Employing Discourse Coherence Enhancement to Improve Cross-Document Event and Entity Coreference Resolution

Xinyu Chen, Peifeng Li*, Qiaoming Zhu,
School of Computer Science and Technology, Soochow University, China
xychennlper@stu.suda.edu.cn, {pfli, qmzhu}@suda.edu.cn
Correspondence: pfli@suda.edu.cn

Abstract

Cross-Document Coreference Resolution (CDCR) aims to identify and group together mentions of a specific event or entity that occur across multiple documents. In contrast to the within-document tasks, in which event and entity mentions are linked by rich and coherent contexts, cross-document mentions lack such critical contexts, which presents a significant challenge in establishing connections among them. To address this issue, we introduce a novel task Cross-Document Discourse Coherence Enhancement (CD-DCE) to enhance the discourse coherence between two cross-document event or entity mentions. Specifically, CD-DCE first selects coherent texts and then adds them between two cross-document mentions to form a new coherent document. Subsequently, the coherent text is employed to represent the event or entity mentions and to resolve any coreferent mentions. Experimental results on the three popular datasets demonstrate that our proposed method[1] outperforms several state-of-the-art baselines.

1 Introduction

Coreference resolution (CR) aims to recognize the same event and entity mentions from various textual spans and then gather them into the same cluster. This task can benefit many downstream tasks in natural language processing (NLP), such as information extraction (Yan et al., 2023), topic detection (Vahidnia), and question answering (Ramesh et al., 2023). CR can be further divided into within-document (WDCR) and cross-document (CDCR) coreference resolution depending on whether the event and entity mentions are in the same document. While most previous work focused on the single within-document event or entity coreference resolution task, this paper focuses on cross-document event and entity coreference resolution, a more challenging task.

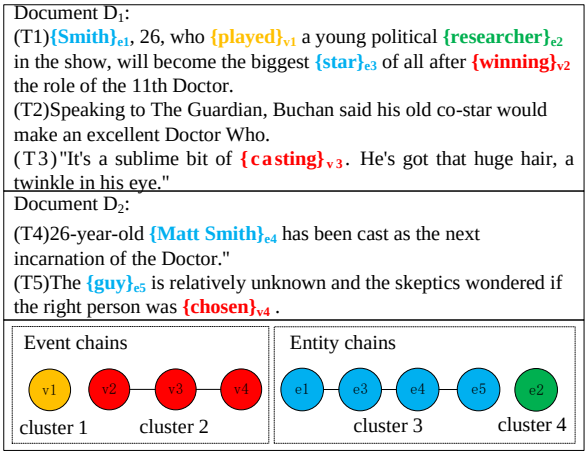

Figure 1: Examples of the event ("v") and entity ("e") mentions, and their coreference resolution results.

Figure 1 illustrates an example of CDCR, given two documents D_1 and D_2 with the mentions marked in different colors (event mentions are marked by their triggers, such as "played" and "winning"), where the same color mentions are coreferent. CDCR first predicts whether there is a coreference relation between the same type (entity or event) mentions, and then clusters the coreferent mentions into the same chains as shown at the bottom of Figure 1.

Coreference resolution is typically modeled as a prediction task based on pairwise similarity (Cattan et al., 2021; Yu et al., 2022; Chen et al., 2023), where the initial step involves using a language model to encode mention spans. In this manner, sufficient contextual information is available to bridge the semantic coherence between two mention spans, as it renders the mention text more comprehensible to benefit the CR model. The utilisation of this pivotal element can facilitate the straightforward implementation of WDCR; however, CDCR is markedly disparate from it. As shown in Figure 1, the cross-document pair (v_2, v_4)

*Corresponding author
[1] https://github.com/chenxinyu-nlp/CDCR

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23272–23286
July 27 - August 1, 2025

does not have a sentence that maintains their semantic coherence, unlike the within-document pair (v_2, v_3), where there is a sentence T2 that maintains their coherence. Hence, the cross-document mentions like (v_2, v_4) cannot benefit from coherent discourse due to the nonexistence of bridging text between them, which is ignored in previous work.

The discourse coherence theory (Grosz, 1978; Hobbs, 1979) stated that "Coherent discourse lightens the burden of comprehension and enhances the likelihood of being understood. As a result, successive sentences should convey a high degree of overlapping information, including entities and events". Moreover, the cross-document mentions also exhibit gaps in different perspectives of description from various describers.

To address the above issues and inspired by the discourse coherence theory, we introduce a new task called **C**ross-**D**ocument **D**iscourse **C**oherence **E**nhancement (**CD-DCE**) to enhance the discourse coherence among cross-document event/entity mentions, and subsequently extract global information on the coherent text obtained through discourse structure. The contributions of this paper are as follows:

- We propose a new task CD-DCE to enhance CDCR, which select and insert coherent sentences between two cross-document mentions to form a new coherent text.
- Experimental results on the three popular datasets (ECB+, ECB+META, and WEC) indicate that our model outperforms several SOTA baselines.

2 Related Work

2.1 Entity Coreference Resolution

Most previous work on entity coreference resolution focused on the within-document tasks (Joshi et al., 2019; Kirstain et al., 2021; Zhu et al., 2024), including mention-pair classifier methods (Ng and Cardie, 2002; Bengtson and Roth, 2008), latent-tree models(Fernandes et al., 2012; Björkelund and Kuhn, 2014) and mention-ranking models (Wiseman et al., 2016; Clark and Manning, 2016). In the cross-document task, the approach of "encoding first, then clustering" is widely used in early stage (Bagga and Baldwin, 1998; Gooi and Allan, 2004). Singh et al. (2011) focused in improving scalability and jointly learning to entity linking (Dutta and Weikum, 2015); Barhom et al. (2019) represented entity mentions by their lexical span and surrounding context; Caciularu et al. (2021) pretrained a language model via a set of related documents for cross-document task; Yu et al. (2022) augmented pairwise mention representation with structured argument features.

2.2 Event Coreference Resolution

Event coreference resolution is a more challenging task than entity coreference resolution due to the more complex structures of event mentions (Yang et al., 2015). Most previous studies modelled event coreference resolution as a pairwise similarity problem. In the event-level WDCR task, researchers resolved coreferent events via feature engineering (Chen and Ji, 2009; Bejan and Harabagiu, 2010; Krause et al., 2016), multi-task learning (Lu and Ng, 2017, 2021), and event representation enhancing (Tran et al., 2021; Xu et al., 2022, 2023), etc. Early event-level CDCR task contains holistic model on nominal and verbal mentions (Lee et al., 2012), unsupervised method (Bejan and Harabagiu, 2014), and iteratively unfolding inter-dependencies method (Choubey and Huang, 2017).

The recent methodologies of event-level CDCR can be divided into the following three categories.
Feature representation Argument information was widely introduced into event representations (Barhom et al., 2019; Yu et al., 2022). Recently, some researchers leveraged discourse structure information to enhance event representation. Chen et al. (2023) constructed cross-document discourse rhetoric structure and then extract local and global information from this structure to represent event mentions. Gao et al. (2024) used within-document rhetorical structure and cross-document lexical chains to capture long-distance dependencies.
Encoder enhancement Caciularu et al. (2021) pretrained a cross-document language model via sets of related documents. Held et al. (2021) trained a fine-grained classifier to deeply extract event mentions features from the local perspective.
Data augmentation Ahmed et al. (2023) used a lemma heuristic method to balance the coreferent and non-coreferent event mention pairs to improve the quantity of dataset. Ding et al. (2024) developed a rationale-centric counterfactual data augmentation method to enhance this task. Min et al. (2024) summarized cross-document event mentions by LLMs to enhancing the comprehension capabilities of SLM for event mentions.

Among the above studies, Cattan et al. (2020),

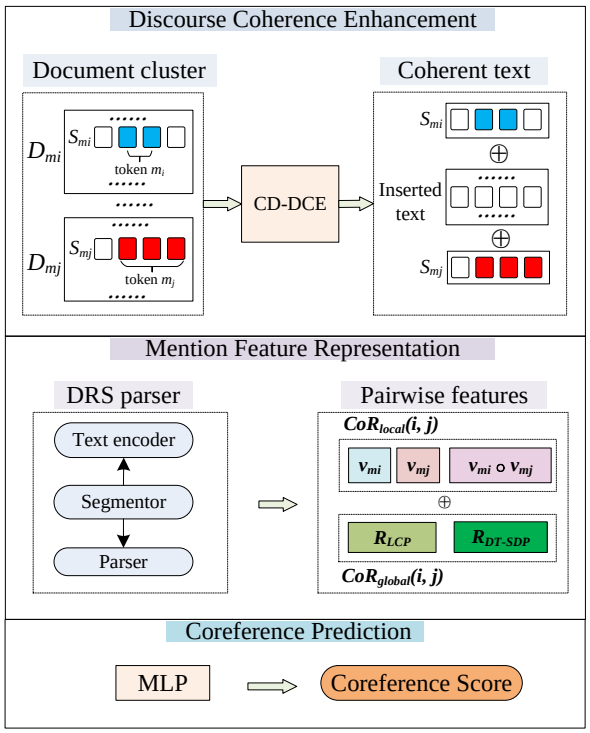

Figure 2: High level overall framework.

Yu et al. (2022) and Caciularu et al. (2021) conducted event and entity coreference resolution simultaneously. Our work is inspired by Chen et al. (2023). Different from Chen et al. (2023) directly concatenating the documents of cross-document event mentions which lacks coherence, we adopt the CD-DCE task to enhance coherence. This task includes steps such as document preprocessing, constructing training instances, and specific training and inference manner to select inserted texts. Our research focuses on resolving cross-document mention problems through discourse coherence enhancement to improve the performance of CDCR.

3 Methodology

Formally, CDCR receives a given set of documents $D = \{D_1, D_2, ..., D_{|D|}\}$ as input, the set of mentions $M = \{m_1, m_2, ..., m_{|M|}\}$ is a set of events or entities mentions distributed in multiple documents from D, and is output with the prediction of the correlation relation of the mention pair (m_i, m_j), and then gather the coreferent mentions into the same clusters according to the results.

Following prior research (Chen et al., 2023), we similarly partition D into distinct subtopics and focus solely on resolving cross-document coreferent mentions within these subtopics to prevent low recall. Specifically, we employ the document clustering method[2] (Barhom et al., 2019) to assign subtopics to all documents. Those mention pairs sharing the same subtopic are considered as candidate coreferent pairs.

Figure 2 shows three main steps of our CDCR model. First, the Cross-Document Discourse Coherence Enhancement (CD-DCE) module is utilized to select one or more coherent sentences from a subtopic containing the specific cross-document mention pair (m_i, m_j). These selected sentences are subsequently inserted into the mention context sequences. Next, the Mention Feature Representation (MFR) module processes the coherent text as input and derives local and global information representations from the discourse tree generated by the discourse rhetorical structure (DRS) parser. Lastly, the Coreference Prediction (CP) module outputs the likelihood of two mentions being coreferent. For within-document mention pairs, the entire document is treated as coherent text and directly fed into the DRS parser. It should be noted that the data of event and entity mention are processed separately in the following stages.

3.1 CD-DCE

Our CD-DCE aims at enhancing the coherence between cross-document mention pairs by inserting a text T relevant to the mention pair (m_i, m_j) between their sentences S_{m_i} and S_{m_j}.

Task definition In a specific subtopic C, given two cross-document mention sentences S_{m_i} and S_{m_j} that containing the span of the mentions m_i and m_j, respectively, and a set of candidate insertion sentences C_{sent}[3]. CD-DCE first inputs S_{m_i}, S_{m_j}, and C_{sent} into a coherence evaluation model M, where M's inference stage is designed to take a given text order O as input and then assign a coherence score (denoted as $Coh(O)$) for O. Next, CD-DCE search for n sentences $T = [s_{r1}, s_{r2}, ..., s_{rn}]$ and finally outputs the concatenated text X=[S_{m_i}, T, S_{m_j}], which is regarded as coherent text. X satisfies the condition that the overall coherence score $\sum_{i=1}^{n+1} Coh([X_i, X_{i+1}])$[4] is maximized.

[2]Almost 99% of coreferent mentions are in the same subtopic using this method.

[3]We first extract all sentences from the documents within the same subtopic C. Then, we deduplicate the candidate sentences, and use RoBERTa to encode the sentences and calculate their cosine similarity with the mention sentences. We finally select the top 50% as candidate insertion sentence set (excluding the mentioned sentences themselves).

[4]The coherence scores for each pair of (X_i, X_{i+1}) are pre-computed and stored in a two-dimensional array for the search.

Document preprocessing Motivated by Jia et al. (2023), we assess text coherence using a graph-based model. Specifically, we transfer documents into a graph structure G_{doc}, incorporating sequential edges, skip edges and document-to-sentence edges (Jia et al., 2023) as depicted by the blue, green, and purple arrows in Figure 3(a), respectively. The graph nodes correspond to the information representations of the sentences within the document. Each sentence is independently encoded using RoBERTa, and the hidden state of the [CLS] token serves as the node representation. Furthermore, the entire document is encoded by RoBERTa to obtain the document-level representation, denoted as doc-node, enabling the positional embeddings to inherently capture the ordering information.

For edge construction, unlike Jia et al. (2023), we specially consider the edge whose arc head or tail contains the mention spans in the above three kinds of edges (e.g., the document-to-sentence edge $\langle$doc-node, $s_1\rangle$ and the skip edge $\langle s_6, s_{10}\rangle$) as our objective to improve the coherence between the documents containing the mentions.

Training instance construction The coherence evaluation model M is designed as a pairwise ranking manner, with training instances constructed in the form (t^+, t^-), where t^+ represents a more coherent text compared to t^-.

For within-document instances, we directly treat the sequence of sentences sorted in their original order within the document as a coherent text. Specifically, for a $\mid D_i \mid$-sentence document D_i=$\{s_{i1}$,s_{i2},...,$s_{i|D_i|}\}$, in order to construct training instances in the form of (t^+,t^-) and make it directly comparable, we first selected a sequence of $k-1$ ($k \geq 2$) sentences, denoted as com = [s_{a+1}, s_{a+2},...,s_{a+k-1}], as the common subsequence for both t^+ and t^- from all possible combinations of sentence orders. We then concatenate the sentence s_a before com (or s_{a+k} after com) to represent t^+ and the sentence $s_{\neq a}$ before com (or $s_{\neq(a+k)}$ after com) to represent t^-.

The construction of cross-document sentence order is to select the sentences from different documents to build t^+ and t^-. However, for such instances, we cannot directly determine who is t^+ and who is t^- by subscript as we do in within-document instances. Based on the similarity between the Next Sentence Prediction (NSP) task and our CD-DCE task, we fine-tune the NSP task on BERT to assess the coherence of sentence orders in cross-document instances. Specifically, we first label within-document sentence orders that can serve as the t^+ order as "IsNext", and the others as "NotNext", for fine-tuning the NSP task. We then input the cross-document instances into the fine-tuned BERT, followed by transformation through a Linear layer and a Sigmoid activation function, treating the output as the coherence score (as shown in Figure 3(b)). Finally, The sentence order with a higher coherence score is considered t^+, while the one with a lower score is t^-.

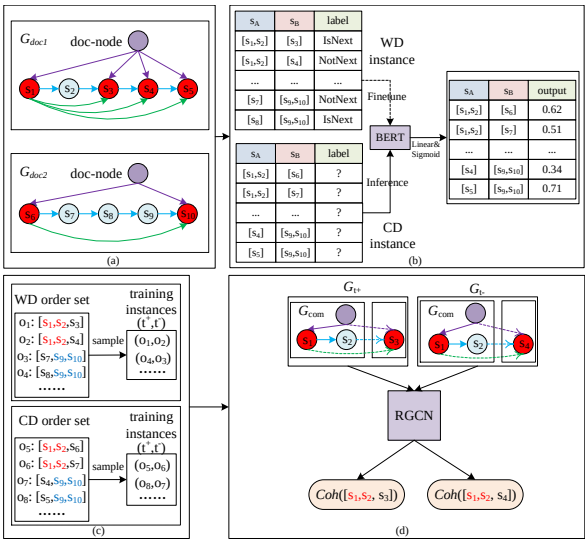

Figure 3: Framework of CD-DCE, where the red and blue nodes represent sentences with and without mentions, respectively.

Figure 3(b) and (c) show the process of within and cross document training instance construction for the case of $k = 3$. The within document sentence order o_1 is more coherent than o_2 due to o_1 has consecutive indices, while the cross-document sentence order o_5 is regarded as more coherent than o_6 according to coherence confidence score. It is worth mentioning that the order pair contains order without mention sentences, such as ([s_7, s_8], [s_7, s_9]), are excluded in the finally training set.

Training During training, we use graph structure to represent documents, which allow our model to understand semantic coherence and capture complex relations between sentences, rather than simply ranking them based on their position. Specifically, for an input pair (t^+, t^-), we initially extract their common sub-graph G_{com} from the full document graph G_{doc} based on the longest common sub-order of t^+ and t^- (e.g., the common sub-order [s_1, s_2] of o_1 and o_2 in Figure 3(d)). Subsequently, the nodes corresponding to the remaining sentences (e.g., s_3 and s_4 in Figure 3(d)) are

connected to their respective positions in the subgraph via directed edges according to the positions in the order. This yields the graph structures G_{t^+} and G_{t^-} for t^+ and t^-, respectively. These graphs are then fed into the relational graph convolutional networks (RGCN), which accumulate relational evidence from the neighborhood around a given node v_i over multiple inference steps as follows:

$$h_i^{(l+1)} = ReLU(\sum_{r\in R}\sum_{j\in N_i^r}\frac{W_r^{(l)}h_j^{(l)}}{\mid N_i^r \mid}) + W_0^{(l)}h_i^{(l)}, \quad (1)$$

where $h_i^{(l)}$ represents the hidden state of the node v_i in the l-th layer. $h_i^{(0)}$ is the embedding of the i-th sentence node obtained from RoBERTa. $r \in R = \{\text{sequential, skip , document-to-sentence edges}\}$ is one of the edge types and N_i^r represents the nodes set connected to v_i through the edge type r. W_r is the parameter matrix for r and W_0 is the parameter matrix for the self-connection edge, which is an extra type in addition to R. We map final representations of all nodes of G_{t^+} and G_{t^-} to a coherence score as follows:

$$Coh(t) = sigmoid(FFN(\sum_{v\in V_{G_t}} h_v)), \quad (2)$$

where FFN is a feed-forward neural network. We update model parameters by the following loss function where $\tau = 0.1$ is the margin.

$$L_{coh} = \max(0, \tau - Coh(t^+) + Coh(t^-)). \quad (3)$$

Inference During inference, we only compute coherence scores for given orders. Specifically, for each cross-document mention pair (m_i, m_j), we first select n sentences $T = [s_{r1}, s_{r2}, ..., s_{rn}]$ from their document cluster (subtopic) to concatenate their corresponding sentences to obtain two candidate coherent texts A and B, denoted as $CTA = [S_{m_i}, s_{r1}, ..., s_{rn}, S_{m_j}]$ and $CTB = [S_{m_j}, s_{r1}, ..., s_{rn}, S_{m_i}]$. Then, we compute their overall coherence as follows:

$$\begin{aligned} Sum_1 &= \sum_{i=1}^{n+1} Coh([CTA_i, CTA_{i+1}]), \\ Sum_2 &= \sum_{i=1}^{n+1} Coh([CTB_i, CTB_{i+1}]). \end{aligned} \quad (4)$$

The n selected sentences that maximize $\max\{Sum_1, Sum_2\}$ are ultimately selected for bridging the mention sentences S_{m_i} and S_{m_j}. It is noteworthy that n is set to 2 in the event-level task and to 1 in the entity-level, which are tuned on the development set. The concatenate text $X_{inp} = CTA$ (or CTB) is sent to discourse rhetorical structure parser.

3.2 Mention Feature Representation

MFR takes the coherent text X_{inp} as input and aims to derive the feature representation of mention pairs. Following Chen et al. (2023), we capture the mention features from both local and global perspectives. Unlike them directly concatenating the documents, resulting in a lack of coherence, we provide more coherent text for DRS parser to construct discourse tree.

Specifically, we employ the DRS parser (Zhang et al., 2021) to construct discourse trees. The parser first receives the coherent text X_{inp} and divides it into a set of Elementary Discourse Units (EDU) sequences, which are then fed to a encoder to obtain word embeddings and extract the mention tokens of the event (or entity) m_i and m_j for local information representation, denoted as $CoR_{local}(i,j) = [v_{m_i}, v_{m_j}, v_{m_i} \circ v_{m_j}]$, where v_m represents the trigger/entity span token embedding of the event or entity mention, and $\circ$ is element-wise multiplication.

To derive the global information representation of the mention pair, the EDU sequences are passed to the parser to construct a discourse tree. We extract the representation of the lowest common parent node R_{LCP} and the shortest dependency path on the discourse tree R_{DT-SDP} (Chen et al., 2023) (detailed in Appendix A) and concatenate them as $CoR_{global}(i,j) = [R_{LCP}, R_{DT-SDP}]$.

3.3 Coreference Prediction

After obtaining the two representations $CoR_{global}(i,j)$ and $CoR_{local}(i,j)$ of the mention pair (m_i, m_j) on the coherent text, we concatenate them as the input of multi-layer perceptron (MLP) and sigmoid activation function is used for scoring the coreference confidence of (m_i, m_j) as follows:

$$\theta = MLP(CoR_{global}(i,j), CoR_{local}(i,j)), \quad (5)$$

$$S = Sigmoid(\theta). \quad (6)$$

3.4 Training and Inference

During training, we train our CDCR model on balanced train sets of ECB+ obtained by the lemma heuristic method (Ahmed et al., 2023) (detailed in Appendix B). We apply dropout in MLP networks, and the training objective is to minimize the binary cross-entropy loss L_{cr} as follows:

$$L_{cr} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log\hat{y}_i + (1-y_i)log(1-\hat{y}_i)], \quad (7)$$

where N is the size of mention pair samples and $y \in \{\text{Coref, Non\_Coref}\}$ is a pairwise label.

During inference, we first apply the topic predictor (Barhom et al., 2019) to cluster the test set documents, and the mention pairs with the same subtopic are considered as candidate coreferent pairs. We then send these pairs to our CD-ECR model to obtain the coreference score. Finally, we perform best-first clustering (Huang et al., 2019) on the pairwise predictions to cluster mentions. It is important to note that the entity and event coreference resolution tasks do not train on a multitasking framework simultaneously. On the contrary, they are carried out separately using the same model.

4 Experimentation

4.1 Experimental Settings

Datasets and metrics We evaluate our model on three CDCR benchmarks: Event Coreference Bank Plus (ECB+) (Cybulska and Vossen, 2014), WEC (Eirew et al., 2021) and ECB+META (Ahmed et al., 2024). ECB+ is the most widely used CDCR dataset which annotated different but similar event and entity mentions as subtopics for each ECB topic. We use gold mentions for both training and evaluation following previous work (Chen et al., 2023; Ahmed et al., 2023). Unlike ECB+, the coreference event mentions in WEC are mostly noun phrases, while ECB+ predominantly uses verbs. ECB+META is divided into ECB+META1 and ECB+METAm, which is the harder variation of ECB+ with more rich lexicon.

We follow the data split of Cybulska and Vossen (2014), Eirew et al. (2021), and Ahmed et al. (2024) for ECB+, WEC, and ECB+META, respectively. It is worth mentioning that we only evaluate event-level coreference resolution on WEC and ECB+META due to WEC only annotates event-level coreference relations and ECB+META only extended ECB+ from the event level.

Following previous work (Caciularu et al., 2021; Chen et al., 2023), we use MUC (Vilain et al., 1995), B^3 (Bagga, 1998), and $CEAF_e$ (Luo, 2005) to evaluate the performance of our model and also report the overall CoNLL score, which is the average of the above three metrics.

Hyper parameters In the stage of discourse coherence enhancement, we adopted the BERT-base model for fine-tuning, where the batch size, epochs, dropout and learning rate were set to 8, 3, 0.1 and 10^{-5}, respectively. During training coherence model, the batch size, epochs, dropout and learning rate were set to 8, 10, 0.2 and 10^{-5}, respectively. The RGCN layers was set to 2. During inference, the number of selected sentences n is set to 2 and 1 in the event-level and entity-level, respectively.

In the stage of coreference resolution, we utilized LongFormer-base to embed the mentions with 768 dimensions. The batch size, epochs, dropout and learning rate were set to 8, 10, 0.1 and 10^{-5}, respectively.

Adam optimizer was used for above all stages.

Baselines We conduct experiments on the following baselines: **Caciularu et al. (2021)**, **Yu et al. (2022)**, **Chen et al. (2023)**, **Ahmed et al. (2023)**, **Nath et al. (2024)**, **Eirew et al. (2021)** and **Gao et al. (2024)** and **Ahmed et al. (2024)**. Besides, "w/o DCE" represents our model without DCE. Please refer to C for details.

4.2 Experimental Results

Table 1 shows the performance of all models on three benchmarks, which shows that our model significantly (P<0.01, t-test) outperforms the SOTA baselines. These results indicate the effectiveness of discourse coherence enhancement in resolving cross-document coreferent mentions.

Compare with the baseline w/o DCE, which represents our model without discourse coherence enhancement (i.e., we directly concatenate S_{m_i} and S_{m_j} and then send them to MFR), the results show that the discourse coherence enhancement leads to an increase of 1.6 and 1.0 on event-level and entity-level tasks in CoNLL on ECB+, respectively, and the significant improvement of 5.7, 3.6 and 5.7 in CoNLL on the WEC, ECB+META1 and ECB+METAm datasets, respectively, indicating that it enables the discourse tree to provide more effective interaction information for cross-document event and entity mentions, thereby improving the performance of CDCR.

Chen et al. (2023) introduced discourse structure to represent event mention pairs from both local and global perspectives, which is similar to our paper in terms of feature representation. However, when they constructed the discourse tree for cross-document mentions, they ignored the coherence of input text. Comparing with them, our model improves the CoNLL score on all datasets. This discrepancy in performance indicates that the coherent text can facilitate the model in more effectively extracting the representations of event mention pairs from the discourse tree, especially in scenarios with weak semantic relationships or large gaps between documents, such as WEC and ECB+META, high-

Dataset	System	CoNLL
ECB+(event)	Caciularu et al. (2021)	85.6
	Chen et al. (2023)	86.4
	Nath et al. (2024)	86.4
	Ours(non-oracle)	**87.4**
	Ahmed et al. (2023)*	87.4
	w/o DCE*	86.9
	Ours(oracle)*	**88.5**
ECB+(entity)	Caciularu et al. (2021)	82.9
	Chen et al. (2023)	83.2
	w/o DCE	83.1
	Ours	**84.1**
WEC(event)	Eirew et al. (2021)	62.3
	Chen et al. (2023)	61.3
	Gao et al. (2024)	65.0
	w/o DCE	60.2
	Ours	**65.9**
ECB+META1 (event)	Chen et al. (2023)*	69.2
	Ahmed et al. (2024)*	71.4
	w/o DCE*	68.4
	Ours(oracle)*	**72.0**
ECB+METAm (event)	Chen et al. (2023)*	51.8
	Ahmed et al. (2024)*	55.6
	w/o DCE*	50.5
	Ours(oracle)*	**56.2**

Table 1: F1 scores of CoNLL on the ECB+, WEC and ECB+META datasets, where "*" refers to using the oracle setting. The comparisons on WEC is conducted under RoBERTa, while the ECB+ is based on the Longformer encoder. It is worth mentioning that the comparison on ECB+META1 (Ahmed et al. (2024) used RoBERTa) and ECB+METAm (Ahmed et al. (2024) used GPT-4) is a best-reported comparison regardless of the underlying model architecture. For the detail scores of MUC, B^3 and $CEAF_e$, please refer to Appendix D.

lighting the importance of text coherence.

The coreference event mentions in WEC are mostly noun phrases, while ECB+ predominantly uses verbs. ECB+META, as a variant of the ECB+ corpus, incorporates more complex annotation structures and richer metadata, which significantly increases the task's difficulty. Hence, there is a gap in the event context descriptions between the three corpora. The improvements on all the three datasets further indicate the effectiveness and generalization of our model in dealing with different forms of texts.

Our model also outperforms the two entity-level baselines on ECB+, which indicates that coherent text can also enhance the information interaction between entity mentions.

Dataset	Rand	Subopt	Coh
ECB+(event)	86.2	86.8	**88.5**
ECB+(entity)	81.9	83.0	**84.1**
WEC	62.1	64.1	**65.9**
ECB+META1	69.5	69.9	**72.0**
ECB+METAm	52.1	54.2	**56.2**

Table 2: CoNLL-F1 scores of using different input text, and the details of the other three metrics are shown in Appendix E.

4.3 Analysis on Coherence Enhancement

Impact of coherent text To validate the effect of coherent text on coreference resolution, we inserted some randomly chosen sentences, which were randomly sampled from the same topic as the related text, between the two mention sentences. In addition, we also select suboptimal solutions to perform insertion according to Eq. 4. Specifically, we rank the multiple groups of results calculated by Eq. 4, and regard the results corresponding to the median as suboptimal insertion. The results are shown in Table 2, where "Rand", "Subopt" and "Coh" represent sentences selected based on randomness, suboptimality, and coherence, respectively.

The experimental results demonstrate that using random-selected and suboptimal-selected sentences for insertion between cross-document mentions leads to significantly performance drop compared to using coherence-selected sentences. Specifically, the insertion of the random-selected and suboptimal-selected sentences resulted in a noticeable decline across all evaluation metrics for the coreference resolution task in Table 9 of Appendix E. This clearly indicates that the performance improvement is not merely due to the addition of extra textual features but is instead attributable to the semantic coherence of the inserted text, which better captures the relationships between cross-document mentions.

Impact of coherence evaluation methods To verify the effectiveness of our CD-DCE method, we compared CD-DCE with the other three coherence evaluation methods (BERT Score (BS), Sentence-BERT (Sent-B) and LDA) across all datasets. The results in Table 3 demonstrate that CD-DCE consistently outperforms all baselines.

Training instances construction For comparison with our proposed construction strategy, we designed two strategies "-enden" and "-evden" for event and entity coreference resolution, respectively. "-enden" indicates the removal of the sen-

Dataset	BS	Sent-B	LDA	Ours
ECB+(event)	83.5	83.7	82.9	**88.5**
ECB+(entity)	82.3	82.7	81.5	**84.1**
WEC	63.4	63.7	62.4	**65.9**
ECB+META1	69.3	69.7	68.2	**72.0**
ECB+METAm	53.4	53.8	52.1	**56.2**

Table 3: CoNLL-F1 scores of using different coherence evaluation methods.

Event-level	MUC	B^3	$CEAF_e$	CoNLL
Ours	90.3	89.0	**86.3**	**88.5**
-enden	-2.6	**+1.5**	-2.7	-1.2
NLI	-2.0	-2.8	-1.1	-1.9
Entity-level	MUC	B^3	$CEAF_e$	CoNLL
Ours	**90.7**	83.7	**77.9**	**84.1**
-evden	-3.8	**+5.4**	-4.3	-0.9
NLI	-2.3	-2.1	-3.9	-2.8

Table 4: Performance comparison of different aspect on the ECB+ dataset.

tences in the set of the candidate insertion sentences C_{sent} with a high density of entity information (fewer than 4 annotated entities, accounting for 37.20% of the total). "-evden" indicates the removal of the sentences in C_{sent} with a high density of event information (fewer than 3 annotated triggers, accounting for 37.63%). These two strategies are used to verify the benefit of entity to event and that of event to entity, respectively. The results of "-enden" and "-evden" in Table 4 suggest that entities/events have a substantial impact on the performance of event/entity coreference resolution.

Fine-tuning tasks We incorporated NLI models to compare with our NSP-finetuned model, which could more accurately capture logical and thematic coherence across non-adjacent sentences in a cross-document setting. Specifically, we fine-tuned the NLI model[5] on three NLI datasets SNLI, MNLI and RTE. We mapped "entailment" to coherent and "contradiction/neutral" to incoherent, keeping other steps unchanged. Results in Table 4 shows that our NSP-based fine-tuning strategy outperforms NLI, indicating NSP-enhanced coherence is more suitable for our task. However, we also believe that NLI has the potential to enhance cross-document coherence with the use of more suitable datasets for fine-tuning.

Substitutability of LLMs We explored the potential of using LLMs for discourse coherence enhancement to investigate whether LLMs can be used to replace the auxiliary task CD-DCE. Specifically, we select ChatGPT-3.5 and LLaMA for comparison. The results show that our CD-DCE outperforms two LLMs. The prompt, performance and detailed discussion are all shown in Appendix F.

[5]transformers.BertForSequenceClassification

Training	Test		
Dataset	ECB+	WEC	GVC
ECB+	86.9 | 88.5	50.7 | 55.6	60.3 | 65.9
WEC	58.4 | 63.6	60.2 | 65.9	68.1 | 72.4
GVC	55.8 | 67.6	40.9 | 45.3	85.3 | 87.3

Table 5: Evaluation results on out-of-domain datasets, where the values before and after "|" refer to the model w/o DCE and w/ DCE, respectively.

Model	MUC	B^3	$CEAF_e$	CoNLL
Held et al. (2021)$^+$	91.5	83.0	76.7	83.7
Ding et al. (2024)$^+$	91.3	**85.8**	76.0	84.4
Ours$^+$	**92.0**	85.2	**80.3**	**85.8**
Ahmed et al. (2023)$^\#$	86.6	85.4	81.3	84.4
Nath et al. (2024)$^\#$	92.9	84.3	71.7	83.0
Ours$^\#$	**93.2**	**86.5**	**82.3**	**87.3**

Table 6: Comparison with the baselines on the GVC dataset, where "+" and "#" refer to using RoBERTa and LongFormer as encoder, respectively.

4.4 Out-of-domain Evaluations

We conducted experiments on the model's cross-domain generalization capability to reveal how well it handles diverse real-world data. We incorporated the GVC (Vossen et al.) dataset into the experiments due to its marked difference in comparison with ECB+ and WEC. The training set was randomly selected from ECB+, WEC and GVC, with the model being tested on the other two datasets. The CoNLL F1 scores are shown in Table 5.

It is evident that, in all cases, the performance with DCE (w/ DCE) is superior to that without DCE (w/o DCE). This tendency remains consistent irrespective of the dataset utilized for training and the dataset employed for testing. The incorporation of DCE has been demonstrated to enhance the model's capacity to generalize and adapt, even when dealing with datasets that may possess varying event types and vocabularies. This finding suggests that DCE assists the model in more effectively capturing the underlying patterns and relationships in the data, thereby enabling it to perform more efficiently across diverse datasets.

In Table 6, we also compare the in-domain experimental results on GVC with existing work using

RoBERTa and LongFormer, which also shows the effectiveness of our proposed method.

However, we still need to note the differences between using out-of-domain and in-domain training data. Lemma diversity is usually an important factor contributing to this difference so that we also attempt to incorporate paraphrased or synthetic examples to increase lemma diversity during training. A detailed discussion is provided in Appendix G.

4.5 Case Study

We provide an example to analyze the effectiveness of discourse coherence. In reference to the example presented in Figure 1, our model concentrates on the cross-document event mention pair (v_2, v_4). Using our CD-DCE, the following two sentences s_i and s_j are the selected insertion sentences by the ECD-CoE task.

s_i: *Matt Smith, 26, will make his debut in 2010, replacing David Tennant, who leaves at the end of this year.*

s_j: *When the 26-year-old unknown was unveiled as the 11th Doctor on Saturday evening, it took most viewers by surprise.*

It can be seen that the concatenated text [D_1,s_i,s_j,D_2] has a high degree of coherence, which revolve around the topic of "Matt Smith being chosen as the 11th Doctor", and are arranged in temporal order. The phrase "his debut in 2010, replacing David Tennant" in s_i describes the development of event mention v_2 and s_j further describes its impact. D_2 provides more information about the audience's reaction to event mention v_2. Hence, this text can be easily understood by the encoder due to its clear logical relationship between sentences, and will result in a more enriched representation of event mention in the next step of event feature representation, finally improving the accuracy of coreference resolution.

If we do not enhance the coherence between D_1 and D_2 , the lack of background information, time clues, and character consistency will cause the subsequent DRS parser to be unable to accurately capture the interactive information between "winning" and "chosen", fail to provide useful clues for coreference resolution, and the model will mistakenly predict them as non-coreferent.

4.6 Error Analysis

Errors of DCE We manually select 100 coherent and 100 incoherent samples from the test set to evaluate the effectiveness of DCE. Among these coherent samples, 52% correct the wrong results using the model w/o DCE, 38% still maintain the results, and 10% cause additional errors. Although DCE may cause additional errors, the proportion of errors corrected by using DCE is higher (52% vs 10%). This indicates that these coherent sentences are beneficial for the CR task. Among these incoherent sentences, almost 87% still maintain the results because many incoherent sentences just paraphrase the event sentences. Moreover, 11% cause additional errors and 2% can correct the wrong results, which indicates that the harm of these incoherent sentences is relatively small.

Errors in DRS parser Since the evaluated CD-ECR dataset does not include annotated discourse relations, we take the cross-document event coreference dataset ECB+ as an example and manually select 20 high-quality discourse tree samples and 20 low-quality samples. Here, high-quality discourse trees are defined as those instances where the proportion of correctly predicted rhetorical relations on the shortest path between the EDUs containing the two mentions is greater than or equal to 60%, while low-quality discourse trees are those where this proportion is less than 60%.

In the high-quality samples, 65% of cases result in the correction of erroneous results, 20% maintain the status of results as they were, and 15% result in the occurrence of additional errors. Although discourse trees may introduce additional errors, the proportion of errors that are corrected by using discourse trees is higher (65% vs. 15%). This finding suggests that the discourse tree is beneficial for the CR task. Conversely, in the low-quality samples, 70% maintain the results, 25% introduce additional errors, and 5% correct incorrect results. This outcome suggests that the adverse impact of these low-quality samples is relatively negligible.

5 Conclusion

In this paper, we first proposed a novel task called Cross-Document Discourse Coherence Enhancement (CD-DCE) and then introduce it to improve cross-document event and entity coreference resolution by enhencing the coherence between two event or entity mentions. Experimental results on the ECB+, WEC, and ECB+META dataset show that our proposed method outperforms several SOTA baselines. Our future work will focus on how to introduce LLM-generated coherent text to improve our tasks.

6 Limitations

Our method still suffers from several shortcomings, which will be addressed in our future work. First, we only perform coreference resolution on golden mentions. The upstream task span detection is also important for coreference resolution. Second, the tasks of CD-DCE and CDCR are performed in a pipeline, which will lead to the accumulation of cascading errors. In the future, we will jointly train the two tasks with entity coreference resolution. Due to our model only works on monolingual and single-modal datasets, we also plan to extend discourse coherence enhancement method to multilingual and multimodal scenarios of CDCR task, exploring methods to establish coherence relations across different languages and modalities (e.g., text and images).

Acknowledgments

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177 and 62376181), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin, and Nikhil Krishnaswamy. 2023. 2*n is better than n^2: Decomposing event coreference resolution into two tractable problems. In *ACL (Findings)*, pages 1569–1583.

Shafiuddin Rehan Ahmed, Zhiyong Eric Wang, George Baker, Kevin Stowe, and James H. Martin. 2024. Generating harder cross-document event coreference resolution datasets using metaphoric paraphrasing. In *ACL (Short Papers)*, pages 276–286.

Amit Bagga. 1998. Evaluation of coreferences and coreference resolution systems. In *LREC*, pages 563–572.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL*, pages 79–85.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *ACL*, pages 4179–4189.

Cosmin Adrian Bejan and Sanda M. Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *ACL*, pages 1412–1422.

Cosmin Adrian Bejan and Sanda M. Harabagiu. 2014. Unsupervised event coreference resolution. *Comput. Linguistics*, 40(2):311–347.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *EMNLP*, pages 294–303.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *ACL*, pages 47–57.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E. Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: cross-document language modeling. In *EMNLP (Findings)*, pages 2648–2662.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling. *CoRR*, abs/2009.11032.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. In *ACL/IJCNLP (Findings)*, pages 5100–5107.

Xinyu Chen, Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2023. Cross-document event coreference resolution on discourse structure. In *EMNLP*, pages 4833–4843.

Zheng Chen and Heng Ji. 2009. Graph-based event coreference resolution. In *ACL*, pages 54–57.

Prafulla Kumar Choubey and Ruihong Huang. 2017. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *EMNLP*, pages 2124–2133.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *EMNLP*, pages 2256–2262.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *LREC*, pages 4545–4552.

Bowen Ding, Qingkai Min, Shengkun Ma, Yingjie Li, Linyi Yang, and Yue Zhang. 2024. A rationale-centric counterfactual data augmentation method for cross-document event coreference resolution. *CoRR*, abs/2404.01921.

Sourav Dutta and Gerhard Weikum. 2015. Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. *Trans. Assoc. Comput. Linguistics*, 3:15–28.

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: deriving a large-scale cross-document event coreference dataset from wikipedia. In *NAACL-HLT*, pages 2498–2510.

Eraldo R. Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *EMNLP-CoNLL Shared Task*, pages 41–48.

Qiang Gao, Bobo Li, Zixiang Meng, Yunlong Li, Jun Zhou, Fei Li, Chong Teng, and Donghong Ji. 2024. Enhancing cross-document event coreference resolution by discourse structure and semantic information. In *LREC-COLING*, pages 5907–5921.

Chung Heong Gooi and James Allan. 2004. Cross-document coreference on a large scale corpus. In *HLT-NAACL*, pages 9–16.

Barbara J. Grosz. 1978. Focusing in dialog. In *TINLAP*, pages 96–103.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *EMNLP*, pages 1406–1417.

Jerry R. Hobbs. 1979. Coherence and coreference. *Cogn. Sci.*, 3(1):67–90.

Yin Jou Huang, Jing Lu, Sadao Kurohashi, and Vincent Ng. 2019. Improving event coreference resolution by learning argument compatibility from unlabeled data. In *NAACL-HLT*, pages 785–795.

Sainan Jia, Wei Song, Jiefu Gong, Shijin Wang, and Ting Liu. 2023. Sentence ordering with a coherence verifier. In *ACL (Findings)*, pages 9301–9314.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S. Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *EMNLP/IJCNLP*, pages 5802–5807.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *ACL-IJCNLP*, pages 14–19.

Sebastian Krause, Feiyu Xu, Hans Uszkoreit, and Dirk Weissenborn. 2016. Event linking with sentential features from convolutional neural networks. In *CoNLL*, pages 239–249.

Heeyoung Lee, Marta Recasens, Angel X. Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *EMNLP-CoNLL*, pages 489–500.

Jing Lu and Vincent Ng. 2017. Joint learning for event coreference resolution. In *ACL*, pages 90–101.

Jing Lu and Vincent Ng. 2021. Span-based event coreference resolution. In *AAAI*, pages 13489–13497.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *HLT-EMNLP*, pages 25–32.

Qingkai Min, Qipeng Guo, Xiangkun Hu, Songfang Huang, Zheng Zhang, and Yue Zhang. 2024. Synergetic event understanding: A collaborative approach to cross-document event coreference resolution with large language models. In *ACL*, pages 2985–3002.

Abhijnan Nath, Shadi Manafi, Avyakta Chelle, and Nikhil Krishnaswamy. 2024. Okay, let's do this! modeling event coreference with generated rationales and knowledge distillation. *CoRR*, abs/2404.03196.

Vincent Ng and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING*.

Gowtham Ramesh, Makesh Narsimhan Sreedhar, and Junjie Hu. 2023. Single sequence prediction over reasoning graphs for multi-hop QA. In *ACL*, pages 11466–11481.

Sameer Singh, Amarnag Subramanya, Fernando C. N. Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *ACL*, pages 793–803.

Hieu Minh Tran, Duy Phung, and Thien Huu Nguyen. 2021. Exploiting document structures and cluster consistencies for event coreference resolution. In *ACL-IJCNLP*, pages 4840–4850.

Sahand Vahidnia. *Deep and Temporal Ontology Guided Clustering Methods and Representation Learning for Topic Detection and Tracking*. Ph.D. thesis.

Marc B. Vilain, John D. Burger, John S. Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *MUC*, pages 45–52.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. Don't annotate, but validate: a data-to-text method for capturing event data. In *LREC*, pages 3034–3042.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *HLT-NAACL*, pages 994–1004.

Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2022. Improving event coreference resolution using document-level and topic-level information. In *EMNLP*, pages 6765–6775.

Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2023. Coref-prompt: Prompt-based event coreference resolution by measuring event type and argument compatibilities. In *EMNLP*, pages 15440–15452.

Hang Yan, Yu Sun, Xiaonan Li, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. 2023. UTC-IE: A unified token-pair classification architecture for information extraction. In *ACL*, pages 4096–4122.

Bishan Yang, Claire Cardie, and Peter I. Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *Trans. Assoc. Comput. Linguistics*, 3:517–528.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2022. Pairwise representation learning for event coreference. In **SEM@NAACL-HLT*, pages 69–78.

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021. Adversarial learning for discourse rhetorical structure parsing. In *ACL-IJCNLP*, pages 3946–3957.

Yilun Zhu, Siyao Peng, Sameer Pradhan, and Amir Zeldes. 2024. SPLICE: A singleton-enhanced pipeline for coreference resolution. In *LREC-COLING*, pages 15191–15201.

A Details of DT-SDP

DT-SDP stands for "Discourse Tree Shortest Dependency Path", which is a representation extracted from the discourse tree to capture global information between mention pairs. The computation of R_{DT-SDP} is as follows:

1) Using a DRS parser to segment the text into Elementary Discourse Units (EDUs).

2) Feeding Edus to a DRS parser to Construct a discourse tree.

3) Searching the tree to find a shortest path between the EDUs where the two event mentions are located.

4) Encoding all the nodes on the path using a Bidirectional LSTM (Bi-LSTM), and the output of the final hidden layer is R_{DT-SDP}.

In this way, R_{DT-SDP} captures the global dependency relation between the event mention pairs.

B Details of Lemma Heuristic Method

The lemma heuristic method (LH) filters non-coreferent samples (P^{-}_{TN}) through the following steps:

1) Extracting Lemmas: The spaCy tool is used to extract the lemmas of event triggers and the words in the sentences.

2) Synonym Pair Matching: A set of synonymous lemma pairs (Syn_p) is created, which frequently appear in coreferent event pairs in the training set. For each event pair (A, B) with the trigger pair (t_A, t_B), their trigger lemmas (l_A, l_B) are checked to see if they satisfy any of the following conditions:

a. (l_A, l_B) $\in$ Syn_p (the lemma pair is in the synonym set)

b. l_A == l_B (the lemmas are identical)

c. t_B contains l_A (the trigger of B contains the lemma of A)

d. t_A contains l_B (the trigger of A contains the lemma of B)

3) Context Comparison: For event pairs that satisfy the above conditions, their contexts (sentences) are further compared. Specifically, the stop words are removed, and the words in the sentences are converted to their lemma forms. The overlap between the two sentences is then calculated. If the overlap exceeds a certain threshold, the event pair is predicted as coreferent.

4) Filtering Non-Coreferent Samples: Through the above steps, LH efficiently filters out a large number of non-coreferent samples (P^{-}_{TN}) while minimizing the loss of coreferent samples (P^{+}_{FH}). Specifically, LH only retains event pairs predicted as coreferent and discards those predicted as non-coreferent.

By using this method, LH can maintain high accuracy while significantly reducing the number of event pairs that need further processing, thereby improving the efficiency of event coreference resolution.

C Details of Baselines

To verify the effectiveness of our model, we select several strong baselines for comparison. For ECB+, we select five baselines as follows.

1) Caciularu et al. (2021) pretrained a language model via a set of related documents, which used a stronger text encoder LongFormer;

2) Chen et al. (2023) resolved coreference events by local and global information on discourse tree.

3) Ahmed et al. (2023) proposed a simple heuristic paired with a cross-encoder;

4) Nath et al. (2024) implemented knowledge distillation methods for event coreference scoring.

5) "w/o DCE" represents our model without discourse coherence enhancement.

It should be noted that only the baseline Caciularu et al. (2021) reported the results of entity coreference resolution on ECB+ dataset. Since WEC has only two baselines, Eirew et al. (2021) and Gao et al. (2024), we selected them for this dataset comparison. Because ECB+META is the latest cross-document coreference resolution dataset, we can only compare it with Ahmed et al. (2024).

D Experimental Results on All Metrics

Table 7 shows all the results of MUC, B^3, $CEAF_e$, and CoNLL on the ECB+, WEC and ECB+META datasets, where "*" refers to using the oracle setting. The comparisons on WEC is conducted under

ECB+				
Event-level	MUC	B^3	$CEAF_e$	CoNLL
Caciularu et al. (2021)	88.1	86.4	82.2	85.6
Chen et al. (2023)	88.3	87.3	83.6	86.4
Nath et al. (2024)	87.9	86.8	84.5	86.4
Ours(non-oracle)	89.9	87.5	86.3	87.4
Ahmed et al. (2023)*	**90.7**	86.3	85.0	87.4
w/o DCE*	88.7	87.2	84.9	86.9
Ours(oracle)*	90.3	**89.0**	**86.3**	**88.5**
Entity-level	MUC	B^3	$CEAF_e$	CoNLL
Caciularu et al. (2021)*	89.9	82.1	76.8	82.9
Chen et al. (2023)*	90.0	82.2	77.5	83.2
w/o DCE*	89.4	82.7	77.1	83.1
Ours(oracle)*	**90.7**	**83.7**	**77.9**	**84.1**
WEC				
Event-level	MUC	B^3	$CEAF_e$	CoNLL
Eirew et al. (2021)	80.7	60.2	45.9	62.3
Chen et al. (2023)	79.3	58.7	45.9	61.3
Gao et al. (2024)	81.8	65.8	47.3	65.0
w/o DCE	79.1	57.9	43.6	60.2
Ours	**82.5**	**67.1**	**48.2**	**65.9**
ECB+META1				
Event-level	MUC	B^3	$CEAF_e$	CoNLL
Ahmed et al. (2024)*	-	-	-	71.4
Chen et al. (2023)*	70.1	72.0	65.5	69.2
w/o DCE*	69.2	69.8	66.3	68.4
Ours(oracle)*	74.1	73.2	68.7	**72.0**
ECB+METAm				
Event-level	MUC	B^3	$CEAF_e$	CoNLL
Ahmed et al. (2024)*	-	-	-	55.6
Chen et al. (2023)*	54.3	52.8	48.3	51.8
w/o DCE*	47.8	52.6	51.2	50.5
Ours(oracle)*	58.8	57.3	52.4	**56.2**

Table 7: Coreference resolution results on the three datasets, where "-" refers to the scores that were not reported.

System	MUC	B^3	$CEAF_e$	CoNLL
Cattan et al. (2021)	83.5	82.4	77.0	81.0
Held et al. (2021)	87.5	86.6	82.9	85.7
Yu et al. (2022)	86.6	85.4	81.3	84.4
Gao et al. (2024)	87.8	86.3	82.4	85.5
Ours	**88.2**	**87.2**	**83.5**	**86.3**

Table 8: Results using RoBERTa on the ECB+ dataset.

RoBERTa, while the ECB+ is based on the Longformer encoder. It is worth mentioning that the comparison on ECB+META1 (Ahmed et al. (2024) used RoBERTa) and ECB+METAm (Ahmed et al. (2024) used GPT-4) is a best-reported comparison regardless of the underlying model architecture.

In addition, we also include the comparison with existing baselines using RoBERTa on the ECB+ dataset in Table 8.

E Ablation on Different Input Text

Table 9 shows the results of all metrics of using random selected sentences and coherent sentences selected by our CD-DCE for mention feature representation. The results show that our CD-DCE outperforms the random strategy on all four metrics on all datasets.

F LLMs Prompts and Performance

The prompt used in ChatGPT and LLaMA for discourse coherence enhancement is as follows. Ta-

Event-level	ECB+			
System	MUC	B3	CEAFe	CoNLL
Random	89.3	81.8	74.6	81.9
Ours	**90.7**	**83.7**	**77.9**	**84.1**
Entity-level	ECB+			
System	MUC	B3	CEAFe	CoNLL
Random	89.3	81.8	74.6	81.9
Ours	**90.7**	**83.7**	**77.9**	**84.1**
Event-level	WEC			
System	MUC	B3	CEAFe	CoNLL
Random	78.5	63.0	44.9	62.1
Ours	**82.5**	**67.1**	**48.2**	**65.9**
Event-level	ECB+META1			
System	MUC	B3	CEAFe	CoNLL
Random	71.1	70.9	66.6	69.5
Ours	**74.1**	**73.2**	**68.7**	**72.0**
Event-level	ECB+METAm			
System	MUC	B3	CEAFe	CoNLL
Random	53.9	54.2	48.1	52.1
Ours	**58.8**	**57.3**	**52.4**	**56.2**

Table 9: All scores of using different input text for mention feature representation.

ble 10 shows the comparison results.

Instruction: Here are two sentences S1 and S2 from difference documents, and CIT is a set of candidate insertion sentences. Please select 1-3 sentences to form a sentence group G from CIT and insert them between S1 and S2, to make the whole text Xinp=[S1, G, S2] more coherent, please output with Xinp.
Input:
S1: It's a sublime bit of casting. He's got that huge hair, a twinkle in his eye.
S2: The guy is relatively unknown and the skeptics wondered if the right person was chosen.
CIT: ["Smith, 26, who played a young political researcher in the show, will become the biggest star of all after winning the role of the 11th Doctor.", "Speaking to The Guardian, Buchan said his old co-star would make an excellent Doctor Who", "The guy is relatively unknown and the skeptics wondered if the right person was",....]

Table 10 shows that ChatGPT-3.5 and LLaMA do not achieve comparable performance to our DCE. DCE has been designed to enhance coherence between cross-document mentions and to facilitate understanding of event and entity coreference relations, while ChatGPT-3.5 and LLaMA was not optimized for our task. It is worth mentioning that it is quite challenging for LLMs to truly comprehend this task, and we acknowledge that there are differences in the level of understanding of coherence models between our model and LLMs.

Event-level	MUC	B^3	$CEAF_e$	CoNLL
Ours	90.3	89.0	**86.3**	**88.5**
ChatGPT-3.5	-2.4	-5.6	-3.7	-3.9
LLaMA	-1.0	-2.9	-5.1	-3.3
Entity-level	MUC	B^3	$CEAF_e$	CoNLL
Ours	**90.7**	83.7	**77.9**	**84.1**
ChatGPT-3.5	-6.1	-2.3	-3.4	-3.9
LLaMA	-7.6	-2.9	-3.0	-4.5

Table 10: Performance comparison of LLMs and our model.

Training	Test		
Dataset	ECB+	WEC	GVC
ECB+ | $ECB+_{para}$	88.5 | 88.2	55.6 | 55.7	65.9 | 66.1
WEC | WEC_{para}	63.6 | 64.1	65.9 | 65.6	72.4 | 72.2
GVC | GVC_{para}	67.6 | 67.1	45.3 | 44.9	87.3 | 87.0

Table 11: Evaluation results on lemma diversity increment.

G Impact of Lemma Diversity

We use BART to generate paraphrased versions of the origin sentence in our datasets. The number of paraphrased are set to 5. We conduct both in-domain and out-of-domain evaluation on the ECB+, WEC and GVC datasets, and the results are shown in Table 11.

From the perspective of out-of-domain, comparing the results before and after paraphrasing, there is no significant improvement in the performance of out-of-domain training for coreference resolution. Although paraphrasing increases lemma diversity, it fails to address the deep-seated semantic distribution differences in out-of-domain data. It is difficult for the model to learn effective coreference resolution patterns, thus unable to significantly improve the out-of-domain training performance. From the perspective of in-domain , comparing the results before and after paraphrasing, the performance of the three datasets after paraphrasing has decreased. This is because the coreference resolution task relies heavily on context. Increasing lemma diversity may change the coherence and semantic relationships of the context. The new lemma may locally change the semantic associations between sentences, causing the model to lose the originally

Dataset	ROUGE-L	BERTScore
ECB+(event)	0.57 | 0.71	0.52 | 0.65
ECB+(entity)	0.51 | 0.66	0.50 | 0.61
WEC	0.45 | 0.56	0.40 | 0.52
ECB+META1	0.55 | 0.70	0.49 | 0.63
ECB+METAm	0.56 | 0.68	0.49 | 0.61

Table 12: Coherence scores before and after enhancement on the ECB+, WEC, ECB+META1 and ECB+METAm datasets.

clear clues when predicting coreference relations.

H Coherence Metrics Study

We used ROUGE and BERTScore metrics to measure coherence scores before and after enhancement. Specifically, we calculated the average coherence scores of CD mention pair of all data sets before and after using the above two coherence metrics, and the results are shown in Table 12 where the values before and after "|" refer to the model w/o and w/ enhancement mechanism, respectively. This shows that through our coherence enhancement mechanism, the coherence between mention sentence has been effectively strengthened.