ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240937

Enhancing Discourse Coherence to Improve Cross-Document Event Coreference Resolution

Xinyu Chen[a], Sheng Xu[a], Peifeng Li[a,*] and Qiaoming Zhu[a]

[a]School of Computer Science and Technology, Soochow University, Suzhou, China

Abstract. Cross-Document Event Coreference Resolution (CD-ECR) is a task of grouping event mentions across multiple documents that refer to the same real-world events. In contrast to within-document event mentions, which are linked by rich, coherent contexts, cross-document event mentions lack such contexts, making it challenging for the model to establish a connection between two event mentions in different documents. To address this issue, we propose a novel mechanism of enhancing discourse coherence to boost CD-ECR. Specifically, we introduce a new task, ECD-CoE (Event-oriented Cross-Document Coherence Enhancement), which selects coherent sentences that form a coherent text for two cross-document event mentions. We then use this coherent text to represent the event mentions and resolve coreferent events. Experimental results on both the ECB+ and GVC datasets indicate that our proposed method outperforms several state-of-the-art baselines.

1 Introduction

Event coreference resolution (ECR) aims to group together multiple event mentions that refer to the same real-world event into the same event cluster. This task is of increasing importance as it benefits many downstream tasks in natural language processing (NLP), such as information extraction [42], topic detection [37], and question answering [35]. ECR can be further divided into within-document (WD-ECR) and cross-document (CD-ECR) event coreference resolution depending on whether the event mentions are in the same document. This paper focuses on the cross-document task.

Events mainly consist of triggers and arguments (entities involved in an event). Since triggers are the main words that can most clearly express the occurrence of events, each event mention (a phrase or sentence within which an event is described) can be represented as its corresponding trigger. Consider the following two event sentences S1 and S2 as examples.

S1: *An allegedly intoxicated driver who tried to flee after* ***striking*** *and fatally injuring a woman in Queens has been charged in her murder.* (D1)

S2: *Police say a 59-year-old woman died after being* ***struck*** *by a vehicle in Queens Friday night.* (D2)

The event triggers of S1 in the document D1 and S2 in the document D2 are “striking” and “struck”, respectively. Although these two triggers have different forms, they both refer to the same event ontology. Therefore, “striking” in S1 and “struck” in S2 have a coreference relation and can be aggregated to form a coreferent chain.

* Corresponding Author. Email: pfli@suda.edu.cn

Document D_1:

(T1)Smith, 26, who played a young political researcher in the show, will become the biggest star of all after **{winning}**$_{e1}$ the role of the 11th Doctor.
(T2)Speaking to The Guardian, Buchan said his old co-star would make an excellent Doctor Who.
(T3)``It's a sublime bit of **{casting}**$_{e2}$. He's got that huge hair, a twinkle in his eye - Matt's the king of geek chic. He is possibly going to be one of the best Doctors we've ever had."

Document D_2:

(T4)26-year-old Matt Smith has been cast as the next incarnation of the Doctor. Users on the Facebook Doctor Who forum that I frequent mostly had the same reaction : `` Who 's Matt Smith ? "
(T5)The guy is relatively unknown and the skeptics wondered if the right person was **{chosen}**$_{e3}$.

Figure 1. Examples of with-document and cross-document event mentions.

WD-ECR researchers typically model ECR as a pairwise similarity problem and begin by utilizing pre-trained language models (e.g., Xu et al. [40] and Liu et al. [26]) to encode event mentions and their contexts. In addition to the event mentions itself, its coherent contexts can establish a connection between the event mentions, which is a critical cue for ECR. In this mode, the encoder can more easily understand coherent document content. This improves the model’s ability to comprehend event contexts and mine relations between event mentions, ultimately aiding in the resolution of coreferent events.

However, there is a significant disparity between WD-ECR and CD-ECR. There is no direct textual connection between the event mentions derived from different documents, while the within-document event mentions can be linked by the text between them. As shown in Figure 1, the within-document event mention pair (e_1, e_2) is bridged by the text T2, while there is no text between the cross-document pair (e_1, e_3). Hence, for the candidate coreferent pair (e_1, e_2), it is easy to understand their document D_1 and accurately extract event mention features to predict whether they are coreferent due to the bridge text T2 makes the whole text [T1, T2, T3] more coherent according to the discourse coherence theory [17, 19]. However, the cross-document instances, such as (e_1, e_3), cannot benefit from this due to the nonexistence of coherent text between their contexts. This issue is ignored in previous work.

To address the above issue, we aim to enhance the discourse coherence among cross-document event mentions. According to discourse coherence theory [17, 19], coherent discourse lightens the burden of

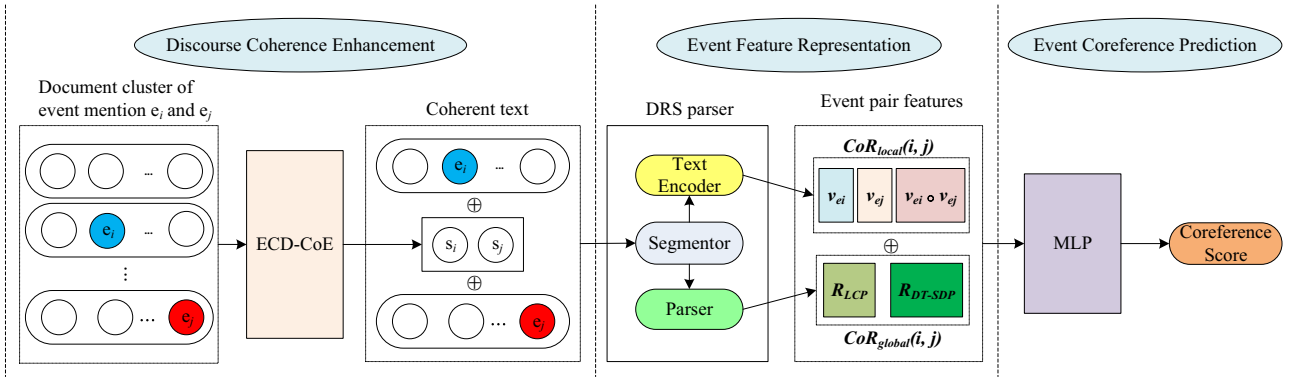

Figure 2. The high level overall framework of our CD-ECR model.

comprehension and enhances the likelihood of being understood. As a result, successive sentences should convey a high degree of overlapping information, including entities, which are an important component of events. Moreover, the cross-document event mentions exhibit gaps in the different ways and perspectives of event description from various describers. This gap amplifies the difference between two cross-document event mentions, which is not conducive to the CD-ECR task.

In this paper, we select a text T that overlaps information from two cross-document event mentions e_i in the document D_1 and e_j in D_2, and then insert T between these two documents or event mentions to form a new coherent text C (i.e., [D_1, T, D_2] or [e_i, T, e_j]). Thus, our CD-ECR model can benefit from the coherent text C when predicting the coreference relation between two cross-document event mentions e_i and e_j.

To achieve this goal, we introduce a new task called **E**vent-oriented **C**ross-**D**ocument **Co**herence **E**nhancement (ECD-CoE). The task involves selecting coherent sentences and inserting them between two documents containing cross-document event mentions are located, in order to form a coherent text. The resulting coherent text is then parsed using a discourse rhetorical structure parser to represent it as a discourse tree. The interaction information is extracted from the discourse tree to enhance the representation of the event mention pairs. Finally, a multi-layer perceptron is used to predict the coreference relation between event mentions, and coreferent event mentions are grouped into the same cluster. The contributions of this paper are as follows:

- We introduce a new task, ECD-CoE, to enhance CD-ECR, which selects and inserts coherent sentences between two documents or two event mentions to form a new coherent text.
- Experimental results on both ECB+ and GVC indicate that our proposed method outperforms several state-of-the-art baselines.

2 Related work

Research on event coreference resolution mainly draws on the entity coreference resolution method, which focuses on resolving noun phrases/mentions for entities [34, 30, 16, 25, 22]. Event coreference resolution is a more challenging task than entity coreference resolution due to the more complex structures of event mentions [43].

Most previous studies on both WD-ECR and CD-ECR model the event coreference resolution task as a pairwise similarity problem, they typically take event mention pairs with various features as input and then a binary classifier is used to assign a score to represent feature similarity to measure whether two event mentions are coreferent [12, 36, 32, 9, 28, 27, 23]. To this end, discriminative features of event mentions are extracted to represent specific event mention pair, such as linguistic features [4], sentential features [24] and argument features [45, 10]. Some studies employ data augmentation methods to improve the quality of the dataset [31, 13, 20, 3, 18, 1] and increase the performance of event coreference resolution. Besides, recent neural methods focus more on ECR-aware event representation [41, 11] representing event mentions from both local and global perspectives.

In the research field of CD-ECR, argument information has been introduced into event representations [3, 45, 44]. For example, Barhom et al. [3] jointly learn entity and event coreference resolution and leveraged predicate-argument structures. Zeng et al. [45] integrate event-specific paraphrases and argument-aware semantic embeddings. Caciularu et al. [7] pretrain a cross-document language model via sets of related document. Held et al. [18] extract event mentions features from the local perspective and train a fine-grained classifier. Yu et al. [44] augment the pairwise representation with structured argument features. Ahmed et al. [1] use a lemma heuristic method to balance set of coreferent and non-coreferent event mention pairs and a LongFormer-based [5] cross encoder. Chen et al. [11] introduce discourse rhetoric structure and then extract local and global information from this structure to represent event mentions.

Our research is based on Chen et al. [11], which introduces discourse structure to extract global information representation for event mention pairs and proposes a strategy for constructing the cross-document discourse tree. Different from them, we focus on enhancing the discourse coherence, which improves the quality of input data for discourse rhetorical structure parser and make the discourse tree provide more useful information for the cross-document task.

3 Methodology

Formally, given a set of documents $D = \{D_1, D_2, ..., D_{|D|}\}$ and a set of event mentions $E = \{e_1, e_2, ..., e_{|E|}\}$ in D, CD-ECR receives an event mention pairs (e_i, e_j) as input to predict their label $y \in Y (Y = \{\text{Coref, Non\_Coref}\})$, and then organize all the event mentions in D into event clusters according to the predictions.

Following previous work [3, 18, 11], we also cluster D into different subtopics and resolve only cross-document coreferent event mentions in the same subtopics to avoid low recall. Specifically, we use the document clustering method [3] to predict subtopics for all

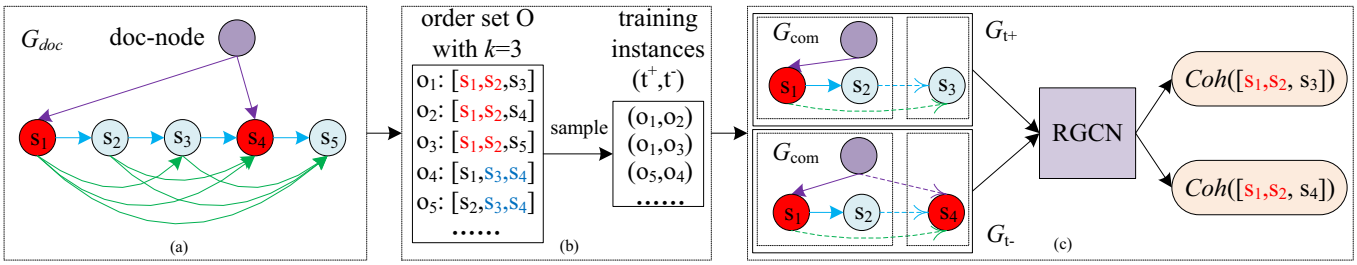

Figure 3. The framework of ECD-CoE, where the red nodes represent event sentences, and the light blue nodes represent sentences without event triggers.

documents. The event mention pairs with the same subtopic are regarded as candidate coreferent pairs.

Figure 2 presents an overview of our CD-ECR model, which consists of three main steps. First, Discourse Coherence Enhancement (DCE) is employed to obtain two selected sentences from a subtopic that includes a specific cross-document event mention pair (e_i, e_j). These two selected sentences are then inserted between the documents. Second, Event Feature Representation (EFR) takes the coherent text as input and extracts local and global information representations from the discourse tree constructed by the discourse rhetorical structure (DRS) parser. Finally, Event Coreference Prediction (ECP) predicts the probability of two event mentions being coreferent. For the within-document event mention pairs, we directly regard the entire document as coherent text and use it as input for the DRS parser.

3.1 Discourse coherence enhancement

We introduce a new task "Event-oriented Cross-Document Coherence Enhancement" (ECD-CoE) to enhance the coherence between the documents of cross-document event mention pairs, which inserts a text S_{group} related to the specific event pair (e_i, e_j) between their corresponding documents D_{e_i} and D_{e_j}, so that the overall coherence of [D_{e_i}, S_{group}, D_{e_j}] is improved.

Task definition The ECD-CoE task is formulated as: Given a cross-document event mention pair (e_i, e_j) and their corresponding documents D_{e_i} and D_{e_j}, where D_{e_i} and D_{e_j} belong to the same subtopic C. C_{sent}={C_{s_i} | $1 \leq i \leq$| C_{sent} |} is a set of sentences in cluster C. M is a text coherence evaluation model that is trained in the pairwise ranking mode, and receives an order [D_e, C_{s_i}] as input during inference to compute coherence score $Coh([D_e, C_{s_i}])$, where $Coh(\cdot)$ is a function in M for computing coherence score for input text orders. The final goal of ECD-CoE is to search for two sentences s_i and s_j, where s_i is used to concatenate after/before D_{e_i} while s_j is used to concatenate before/after D_{e_j}, and then the obtained concatenate text [D_{e_i}, s_i, s_j, D_{e_j}] (or [D_{e_j}, s_j, s_i, D_{e_i}]) is regarded as coherent text.

Document preprocessing Inspired by Jia et al. [21], we evaluate text coherence with a graph-based model. Specifically, we represent documents as the graph structure G_{doc} with sequential edges, skip edges and document-to-sentence edges [21] as shown by the blue, green, and purple arrows in Figure 3(a), respectively, and the nodes are the information representation of the sentences in a document. We use RoBERTa to encode each sentence independently and the hidden state of the [CLS] token is used as the node representation. Additionally, we let RoBERTa encode the entire document to get the document-level representation doc-node so that the positional embeddings naturally encode the ordering information of the document.

In terms of constructing edges, different from Jia et al. [21], we only consider the information of edges between doc-node and event sentence nodes (as shown by the red circle in Figure 2(a)) since our goal is to enhance the coherence between the documents containing event mentions.

Training instance construction We design a coherence evaluation model as a pairwise ranking training manner, and the training instances are constructed as pair (t^+, t^-) form, where t^+ is a more coherent text than t^-. Specifically, for a specific document D={s_1,s_2,...,s_n} with n sentences, we select k ($k \geq 2$) different sentences from D to form a sentence order o, where both t^+ and t^- are instances of o. In order to make the coherence degree of t^+ and t^- directly comparable, we constrain that:

1) t^+ and t^- have the same sentence number k;

2) t^+ and t^- have a common sub-order of length greater than 1, and the longest length of the common sub-order is $k-1$;

3) The common sub-order is coherent, sentence orders with consecutive indices are considered as coherent text (e.g., [s_1, s_2, s_3]).

The common sub-order can be denoted as com = [s_{a+1}, s_{a+2},...,s_{a+k-1}], and we construct training instances pair (t^+, t^-) by concatenating 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^-. Figure 3(b) shows the process of training instance construction for the case of $k = 3$, where o_1 is more coherent than o_2. It is worth mentioning that the order pair contains order without event sentences, such as ([s_2, s_3], [s_2, s_4]), are excluded in the finally training set.

Training During training, given an input pair (t^+, t^-), we first extract their common sub-graph G_{com} from the full document graph G_{doc} according to 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(b)), and then the nodes corresponding to the other sentences (e.g., s_3 and s_4 in Figure 3(b)) are connected to the corresponding positions in the subgraph by directed edges according to the position in the order. The graph structures G_{t+} and G_{t-} of t^+ and t^- are obtained, respectively. Then, we feed G_{t+} and G_{t-} to the relational graph convolutional networks (RGCN), which can accumulate relational evidence from the neighborhood around a given node v_i in multiple inference steps and formalized as follows.

$$h_i^{(l+1)} = \sigma\Big(\sum_{r\in R}\sum_{j\in N_i^r}\frac{W_r^{(l)}h_j^{(l)}}{|\,N_i^r\,|}\Big) + W_0^{(l)}h_i^{(l)}, \tag{1}$$

where $h_i^{(l)}$ represents the hidden state of the node v_i in the l-th layer. $h_i^{(0)}$ is embedding of the i-th sentence node obtained from RoBERTa. $r \in R(R = \{\text{sequential, skip , document-to-sentence edges}\})$ is one of the edge types and N_i^r represents the set of nodes connected to v_i through 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. $\sigma(\cdot)$ is the activation function $ReLU(\cdot)$. After getting the final representations of all nodes of G_{t^+} and G_{t^-}, we map them to a coherence score as follows.

$$Coh(\mathrm{t}) = sigmoid(FFN(\textstyle\sum_{v \in V_{G_{\mathrm{t}}}} h_v)), \tag{2}$$

where FFN is a feed-forward neural network.

We update model parameters using the following loss function.

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

where $\tau = 0.1$ is the margin.

Inference During inference, we only compute coherence scores for given orders. Specifically, for each cross-document event mention pair $(\mathrm{e}_i, \mathrm{e}_j)$ in the documents $\mathrm{D}_{\mathrm{e}_i}$ and $\mathrm{D}_{\mathrm{e}_j}$, respectively, we first select two different sentences s_i and s_j from their document cluster (subtopic) to concatenate their corresponding document $[\mathrm{D}_{\mathrm{e}_i}, \mathrm{s}_i, \mathrm{s}_j, \mathrm{D}_{\mathrm{e}_j}]$ (or $[\mathrm{D}_{\mathrm{e}_j}, \mathrm{s}_j, \mathrm{s}_i, \mathrm{D}_{\mathrm{e}_i}]$). Then, we compute two kinds of coherence values sum $Sum_1 = Coh([\mathrm{D}_{\mathrm{e}_i}, \mathrm{s}_i]) + Coh([\mathrm{s}_i, \mathrm{s}_j]) + Coh([\mathrm{s}_j, \mathrm{D}_{\mathrm{e}_j}])$ and $Sum_2 = Coh([\mathrm{D}_{\mathrm{e}_j}, \mathrm{s}_j]) + Coh([\mathrm{s}_j, \mathrm{s}_i]) + Coh([\mathrm{s}_i, \mathrm{D}_{\mathrm{e}_i}])$. The two sentences s_i and s_j that the maximization $\max\{Sum_1, Sum_2\}$ are ultimately selected for bridging documents $\mathrm{D}_{\mathrm{e}_i}$ and $\mathrm{D}_{\mathrm{e}_j}$. The concatenate text $X_{inp} = [\mathrm{D}_{\mathrm{e}_i}, \mathrm{s}_i, \mathrm{s}_j, \mathrm{D}_{\mathrm{e}_j}]$ (or $[\mathrm{D}_{\mathrm{e}_j}, \mathrm{s}_j, \mathrm{s}_i, \mathrm{D}_{\mathrm{e}_i}]$) is sent to discourse rhetorical structure parser.

3.2 Event feature representation

EFR takes the coherent text X_{inp} as input and aims to obtain the feature representation of event mention pairs. Following Chen et al. [11], we represent event mention feature from both local and global perspectives. Chen et al. [11] directly concatenates the documents of cross-document event mention pairs, which lacks coherence. Different from them, we provide more coherent text for DRS parser to construct discourse tree.

Specifically, we utilize DRS parser system [46] to construct discourse trees, which consists of two main components: segmentor and parser. The segmentor receives coherent text X_{inp} and splits it into a set of elementary discourse units (EDU) sequences. These EDU sequences are sent to the text encoder (e.g., RoBERTa and Longformer) to obtain word embeddings and extract the trigger tokens of the event mentions e_i and e_j to get local information representation, denoted as $CoR_{local}(i,j) = [v_{\mathrm{e}_i}, v_{\mathrm{e}_j}, v_{\mathrm{e}_i} \circ v_{\mathrm{e}_j}]$, where v_{e} is the trigger token embedding of the event mention e, and $\circ$ is element-wise multiplication.

To obtain the global information representation of event mention pair, the EDU sequences are sent to the parser to construct a discourse tree. We extract the representation of the lowest common parent node R_{LCP} and shortest dependency path on discourse tree R_{DT-SDP} [11] and concatenate them as $CoR_{global}(i,j) = [R_{LCP}, R_{DT-SDP}]$.

3.3 Event coreference prediction

We have obtained the global and local information representations $CoR_{global}(i,j)$ and $CoR_{local}(i,j)$ of the event mention pair $(\mathrm{e}_i, \mathrm{e}_j)$ on the coherent text. We concatenate the two features and send them to the multi-layer perceptron (MLP) and sigmoid activation function to obtain the coreference score S as follows.

$$\theta = MLP(CoR_{global}(i,j), CoR_{local}(i,j)), \tag{4}$$

$$S = Sigmoid(\theta). \tag{5}$$

3.4 Training and inference

During training, we train our CD-ECR model on balanced train sets of ECB+ and GVC obtained by the lemma heuristic method [1]. We apply dropout in MLP networks, and the training objective is to minimize the binary cross-entropy loss L as follows.

$$L_{ecr} = -\frac{1}{N}\textstyle\sum_{i=1}^{N}[y_i log\hat{y}_i + (1-y_i)log(1-\hat{y}_i)], \tag{6}$$

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

During inference, we first apply the topic predictor [3] to cluster the test set documents, and event 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 [20] on the pairwise predictions to cluster event mentions.

4 Experimentation

In this section, we first introduce the experimental settings and then report the results.

4.1 Experimental settings

Dataset Following the previous work [18, 1], we use two popular CD-ECR datasets ECB+ [14] and GVC [39] to train and test our model. ECB+ is extended from ECB [4], which annotated different but similar events as subtopics for each ECB topic. GVC is a recent English corpus exclusively focusing on event coreference resolution, which is a collection of texts surrounding a single topic (gun violence). We use gold event mentions for both training and evaluation following previous work [11, 1]. The detailed statistics of ECB+ and GVC datasets are shown in Table 1. For the ECB+ dataset, we follow the data split by Cybulska and Vossen [15]: train: 1, 3, 4, 6-11, 13-17,19-20, 22, 24-33; dev: 2, 5, 12, 18, 21, 23, 34, 35; test: 36-45. For the GVC dataset, we use the data split by Bugert et al. [6].

Metrics Following the previous work [3, 8, 44], we use MUC [38], B^3 [2], and $CEAF_e$ [29] to evaluate the performance of our model and also report the overall CoNLL score, which is the average of the above three metrics. Among them, MUC is based on event links to evaluate the performance of the model, B^3 compensates for MUC's neglect of non-coreferent events by using event nodes as the computational target. $CEAF_e$ is similar to B^3 , adding entities to evaluate the performance of event coreference resolution. CoNLL, the comprehensive use of the above three metrics, can more objectively measure model performance.

Hyper parameters We use pre-trained language models, RoBERTa_{LARGE} and LongFormer_{BASE}, to embed event mentions with 1024 and 768 dimensions, respectively. The training epoch of

Table 1. ECB+ and GVC statistics. T=topics, D=documents, EM=event mentions, EC=event clusters, ES=event singletons, WD=within-document pairs, and CD=cross-document pairs. Event clusters include singletons.

	ECB+			GVC		
	Train	Dev	Test	Train	Dev	Test
T	25	8	10	1	1	1
D	574	196	206	358	78	74
EM	3808	1245	1780	5313	977	1008
EC	1527	409	805	991	228	194
ES	1116	280	623	157	70	43
WD	15695	4685	10751	52212	7345	9193
CD	169798	51849	83191	60142	9193	10660

Table 2. Performance comparision of different models on the ECB+ and GVC datasets, where "*" indicates that the models use LongFormer as their encoders and the other models use BERT/RoBERTa as their encoders.

ECB+										
System	MUC			B^3			$CEAF_e$			CoNLL
	P	R	F1	P	R	F1	P	R	F1	F1
Zeng et al. [45]	85.6	89.3	87.5	77.6	89.7	83.2	84.5	80.1	82.3	84.3
Cattan et al. [8]	81.9	85.1	83.5	82.7	82.1	82.4	78.9	75.2	77.0	81.0
Held et al. [18]	88.1	87.0	87.5	87.7	85.6	86.6	85.8	80.3	82.9	85.7
Yu et al. [44]	85.1	88.1	86.6	84.7	86.1	85.4	79.6	83.1	81.3	84.4
Caciularu et al. [7]*	89.2	87.1	88.1	87.9	84.9	86.4	81.2	83.3	82.2	85.6
Chen et al. [11]*	87.2	89.4	88.3	86.4	88.3	87.3	84.0	83.2	83.6	86.4
Ahmed et al. [1]*	87.9	93.7	90.7	79.6	94.1	86.3	88.7	81.6	85.0	87.4
Ours(RoBERTa)	86.7	89.7	88.2	84.2	90.4	87.2	85.3	81.7	83.5	86.3
Ours(LongFormer)*	86.9	95.1	**90.8**	82.1	95.3	**88.2**	90.9	80.6	**85.4**	**88.1**
GVC										
System	MUC			B^3			$CEAF_e$			CoNLL
	P	R	F1	P	R	F1	P	R	F1	F1
Held et al. [18]	91.2	91.8	91.5	83.8	82.2	83.0	77.9	75.5	76.7	83.7
Ahmed et al. [1]*	91.1	84.0	87.4	76.4	79.0	77.7	52.5	69.6	59.9	75.0
Ours(RoBERTa)	91.4	92.6	92.0	81.4	89.4	85.2	81.9	78.7	80.3	85.8
Ours(LongFormer)*	92.0	94.5	**93.2**	81.4	92.3	**86.5**	85.9	79.0	**82.3**	**87.3**

our CD-ECR model is set to 10, the learning rate is set to 10^{-5}, and Adam optimizer is used to update the parameters. The layer of RGCN is set to 2.

4.2 Experimental results

To verify the effectiveness of our model, we select seven strong baselines as follows.

1) Zeng et al. [45] incorporate event-specific paraphrases and argument semantic embeddings;

2) Cattan et al. [8] develop an end-to-end baseline for resolving coreferent events;

3) Held et al. [18] extract event mention features from the local perspective and trains a fine-grained classifier.

4) Yu et al. [44] augment pairwise representation with structured argument features.

5) Caciularu et al. [7] pretrain a language model via a set of related documents, which use a stronger text encoder LongFormer.

6) Chen et al. [11] resolve coreference events by local and global information on discourse tree.

7) Ahmed et al. [1] propose a simple heuristic paired with a cross-encoder.

Table 2 shows the performance of the baselines and our model on ECB+ and GVC with the encoders RoBERTa and LongFormer, which shows that our model Ours(LongFormer) significantly (P<0.01) outperforms the SOTAs Ahmed et al. [1] and Held et al. [18] on ECB+ and GVC, respectively. Our model Ours(RoBERTa) also outperforms the four BERT/RoBERTS-based baselines both on ECB+ and GVC. These results indicate the effectiveness of our proposed model using discourse coherence enhancement in resolving cross-document coreferent events.

Among the four baselines [45, 44, 8, 18] using BERT or RoBERTa as encoder, Held et al. [18] adopt a pruning method to improve the quality of training data, and uses the context of sentences surrounding event mentions, outperforming the other four baselines. This indicates the importance of training data quality for CD-ECR. Comparing with them, our model Ours(RoBERTa) improves the CoNLL by 0.6 and 2.1 on ECB+ and GVC datasets, respectively, due to the gain of coherence on cross-document instances.

Caciularu et al. [7] and Chen et al. [11] use a stronger document-level encoder LongFormer. The former pretrains a cross-document language model via a set of related documents and outperforms the four BERT/RoBERTa-based models. The latter introduces discourse structure to represent event mention pairs from both local and global perspectives. However, when they construct the discourse tree for cross-document instances, they ignore the coherence of the input text. In comparison with Caciularu et al. [7] and Chen et al. [11], our Ours(LongFormer) improves CoNLL score by 2.5 and 1.7 on ECB+. Ours(RoBERTa) also achieves the CoNLL score of 86.3 on ECB+ dataset, which is competitive with the LongFormer-based model [11].

Ahmed et al. [1] resolve coreferent events by a LongFormer-based cross-encoder and a lemma heuristic method, which achieves the best performance on the ECB+ dataset. Compared with this model, our model Ours(LongFormer) increases CoNLL score by 12.3 and 0.7 (P<0.01) on the GVC and ECB+ datasets, where the improvement on GVC is much higher than that in ECB+. Ahmed et al. [1] only achieve a CoNLL score of 75.0 on the GVC dataset. This is due to the fact that the GVC dataset has a lower proportion of singletons event mentions, meaning fewer non-coreferent samples. Since their heuristic method focuses on non-coreferent samples with similar lemmas in the training data, it does not have a significant advantage on the GVC dataset. The substantial discrepancy in performance between Ahmed et al. [1] and our model Ours(LongFormer) indicate that coherent text can facilitate the CD-ECR model in more effectively extracting the representations of event mention pairs from the discourse tree, thereby enhancing the performance of CD-ECR. From the perspective of the other three metrics, the improvement of $CEAF_e$ is the most significant. This can be attributed to the fact that the cross-document discourse tree based on coherent text can better distinguish non-coreferent event mentions with similar features, thereby improving the prediction of non-coreferent mention pairs.

Advanced instruction-tuned large language models (e.g., ChatGPT [1]) has presented possibilities for the task of CD-ECR. However, our preliminary experimental results show that the CoNLL score of ChatGPT-4 is only 44 in a few-shot manner, which is much lower than the models in Table 2.

5 Analysis

In this section, we first analyze our proposed model on the impact of discourse coherence enhancement, training instances on coherence

[1] https://chat.openai.com/chat

and inserted text selection number. Then, we give the case study and error analysis.

5.1 Impact of discourse coherence enhancement

To further evaluate the mechanism of discourse coherence enhancement on CD-ECR, we perform the ablation experiments and the F1 scores of Pairwise, MUC, B^3, $CEAF_e$, and CoNLL are shown in Table 3, where w/ DCE and w/o DCE refer to with and without discourse coherence enhancement.

Table 3. Ablation study for discourse coherence enhancement on the ECB+ and GVC datasets.

ECB+					
System	Pairwise	MUC	B^3	$CEAF_e$	CoNLL
w/ DCE	**77.6**	**90.8**	**88.2**	**85.4**	**88.1**
w/o DCE	68.7	88.9	86.9	84.5	86.8
GVC					
System	Pairwise	MUC	B^3	$CEAF_e$	CoNLL
w/ DCE	**83.5**	**93.2**	**86.5**	**82.3**	**87.3**
w/o DCE	73.0	91.3	86.0	78.7	85.3

The metric pairwise is used to evaluate whether the pairwise classifier can correctly distinguish coreferent and non-coreferent cross-document event mention pairs. Note that the pairwise F1 scores in Table 3 are calculated only on the cross-document mention pairs. The significant improvement (+8.9 and +10.5 on ECB+ and GVC datasets respectively) of Pairwise in Table 3 shows that the coherent texts are helpful to distinguish coreferent and non-coreferent cross-document event mention pairs. This result verifies the effectiveness of our mechanism of discourse coherence enhancement on resolving cross-document coreferent events.

The results show that discourse coherence enhancement leads to an increase of 1.3 and 2.0 on the ECB+ and GVC datasets, respectively, in CoNLL, indicating that it enables the discourse tree to provide more effective interaction information for cross-document event mentions, thereby improving the performance of CD-ECR.

On the ECB+ dataset, the improvement of 1.9 in MUC is the most significant based on the three different evaluation metrics. The precision of coreferent sample prediction improved due to the introduction of two selected sentences that added overlapping event participant information, enriched the event expression, and completed the event information in the shortest dependency path of the discourse tree. However, the model's lack of ability to disambiguate pronouns could result in some pronoun entities being mistakenly associated with the document. This can lead to incorrect "Coref" predictions and slow down the improvement of the entity-sensitive metric $CEAF_e$.

On the GVC dataset, the improvement of 3.6 in $CEAF_e$ is the most significant. GVC dataset focuses on the topic of gun violence, typically has event mentions triggered by words strongly related to "kill", "death", and "shot". Similarity models can easily misidentify non-coreferent samples as coreferent in this dataset, especially in cross-document instances with different contexts. Our proposed method for enhancing cross-document coherence enables the model to better understand the event text, making it easier to distinguish the features of participants in different gun violence events and improve the performance of event coreference resolution.

5.2 Impact of training instances on coherence

To verify the effectiveness of our training instance construction strategy for ECD-CoE, we also perform experiments on all instances during the training stage of ECD-CoE. Table 4 reports the CD-ECR performance on different training instances for the ECD-CoE model, where "Event-related" represents that ECD-CoE is trained on texts containing event mentions, and "All" represents that the other texts without event mentions are included in the training set.

Table 4. Performance comparison of different training instances on the ECB+ and GVC datasets.

ECB+				
System	MUC	B^3	$CEAF_e$	CoNLL
Event-related	**90.8**	**88.2**	**85.4**	**88.1**
All	89.4	86.6	84.3	86.7
GVC				
System	MUC	B^3	$CEAF_e$	CoNLL
Event-related	**93.2**	**86.5**	**82.3**	**87.3**
All	91.2	86.3	78.0	85.2

The results suggest that using the "All" strategy negatively impacts performance with CoNLL, resulting in a decrease of 1.4 and 2.1 on the ECB+ and GVC datasets in CoNLL, respectively. This is due to the fact that this training strategy increases the probability of considering some sentences without event or entity information as candidate insertion sentences, such as a simple monologue like *"I've done it"*. Even if this sentence achieves a high coherence score, it cannot provide effective information to enrich the event representation.

5.3 Impact of inserted text selection number

To verify the effectiveness of inserted text selection number for ECD-CoE, we design different inserted text selection number for comparison. The comparison results are shown in Table 5, where ITS-N represents the number of inserted sentences is N and means N sentence(s) is (are) selected from document cluster to maximize the sum of coherence score. ITS-2 is our coherence enhancement strategy. Additionally, we also design the strategy ITS-Summ that summaries the two documents D_{e_i} and D_{e_j} by T5 [33] and then insert the two summaries between the two documents. That is, use transitional sentences to link two documents. In ITS-Summ, the text selection number is also set to 2.

Table 5. Performance comparison of different numbers of inserted texts on the ECB+ and GVC datasets.

ECB+				
System	MUC	B^3	$CEAF_e$	CoNLL
ITS-1	89.8	87.8	85.1	87.6
ITS-2(Ours)	**90.8**	**88.2**	**85.4**	**88.1**
ITS-3	89.7	87.9	85.8	87.8
ITS-4	89.6	87.4	85.3	87.5
ITS-Summ	88.3	88.6	83.6	86.9
GVC				
System	MUC	B^3	$CEAF_e$	CoNLL
ITS-1	92.5	85.9	81.7	86.7
ITS-2(Ours)	**93.2**	**86.5**	**82.3**	**87.3**
ITS-3	91.8	87.4	80.0	86.4
ITS-4	91.2	86.9	78.8	85.6
ITS-Summ	92.2	85.3	80.8	86.1

Table 5 shows that our strategy outperforms the other three strategies on both the ECB+ and GVC datasets. Compared to ITS-2 and other ITS-N, ITS-1 has the lowest computational complexity, but it is difficult for a single sentence to make two documents with different expression styles coherent simultaneously, and the effective information it can provide is also limited. Though ITS-3 increases the possibility of introducing more coherent sentences, the computational

complexity is increased, and it also introduces more data noise. Using strategy ITS-4, the performance on both datasets has decreased. This indicates that the more sentences fail to effectively highlight the key information of the document, leading to information overload.

Compared to ITS-Summ, which attempts to bridge two documents in the form of summary-transition, where the coherence of intra-summary is ignored. Hence, the comparison between ITS-Summ and ITS-2 proves the effectiveness of intra-selected-sentence coherence (i.e., $Coh([s_i, s_j])$).

5.4 Case study

We provide an example to analyse the effectiveness of discourse coherence on the ECB+ dataset. In reference to the example presented in Figure 1, our model concentrates on the cross-document event mention pair (e_1, e_3). Using our ITS-2 strategy, 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 e_1 and s_j further describes its impact. D_2 provides more information about the audience's reaction to event mention e_1. 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.

5.5 Error analysis

We perform a qualitative analysis of the major sources of error made by our model, which mainly come from the following aspects: 1) errors of discourse coherence enhancement ; 2) incorrect coreference prediction due to the lack of document clues, although the inserted text can enhance text coherence; 3) errors in the DRS parser, which introduces noise into the subsequent event feature representation.

The coherence enhancement errors mainly come from the lack of diversity in the construction of training instances in the ECD-CoE task. Although Table 4 proves that using only event sentences to construct training instances is better than using all sentences, there are also no-trigger sentences that may provide useful information. For example, the no-trigger sentence *"The victim's name hadn't been released as of Wednesday morning."* contains a time clue phrase "Wednesday morning", using it to construct a training instance can provide time clue information for related events, which is an important component of an event.

For error 2), taking the following two documents D_3 and D_4 with two selected insertion sentences s_k and s_l as an example.

D_3:*Connecticut State Police: 16-yea-old boy dies after accidental shooting Killingly (AP)- .Police say they got a call for a* ***{shooting}***$_{e_4}$ *around 7 p.m. Tuesday at a home in Killingly . They found a boy unresponsive in an upstairs bedroom . He was taken to Day Kimball Hospital in Putnam , where he was later pronounced dead .*

s_k:*Authorities say 22-year-old Kyle Carney of Killingly had been pointing a rifle in the boy's direction when it accidentally discharged.*

s_l:*Carney was taken into custody and charged with manslaughter and reckless endangerment. He's being held on a $500,000 bond and was due to appear in court Wednesday.*

D_4:*Vigil held for teenager shot and killed in Killingly WFSB 3 Connecticut Friends and family were on hand on Friday night to remember the life of a teenager from Killingly, who was killed earlier this month. A special vigil for 16-year-old Matthew Regula was held at 8 p.m. at Davis Park in Danielson. Police said 22-year-old Kyle Carney* ***{shot}***$_{e_5}$ *and killed Regula inside a home on Kenneth Drive on Aug.5.*

The event mention pair (e_4, e_5) is non-coreferent, and the model is easily misled to make incorrect predictions. First, the triggers "shooting" and "shot" have high lexical similarity, and both sentences contain the location argument "home", which makes it easy for the model to assign them high similarity scores. Second, the inserted sentences both contain the participant "Carney", which further increases the similarity of these two non-coreferent events. Third, the lack of key no-trigger sentences in the training instances causes the inserted sentences to be unable to provide effective clues.

Finally, the discourse tree obtained by the DRS parser contains numerous errors. Parsing errors primarily involve incorrect predictions of rhetorical relations. For example, two EDUs (elementary discourse units) connected by a rhetorical relation of "Summary" are mistakenly predicted to have a rhetorical relation of "Cause". "Summary" implies a process where the same event is mentioned from complex to simple, whereas "Cause" involves mentioning two non-coreferent events. Consequently, such parsing errors may lead to incorrect non-coreferent predictions, which can adversely affect model performance.

6 Conclusion

We improve cross-document event coreference resolution through a novel task called Event-oriented Cross-Document Coherence Enhancement. First, we enhance the coherence of two documents containing cross-document event mention pairs through this task. Then, we send the coherent text to DRS parser to obtain the event representations. Finally, we perform coreference relation prediction. Experimental results on both the ECB+ and GVC datasets show that our method outperforms several SOTA baselines.

Our proposed model still suffers from the following shortcomings. First, we only perform event coreference resolution on golden event mentions following previous work. However, the upstream task event extraction is also important for event coreference resolution. Second, the tasks of ECD-CoE and CD-ECR are performed in a pipeline, which will lead to the accumulation of cascading errors. Finally, We directly use all discourse rhetoric relation type without filtering the useless type for predicting coreferent relation, which is also an important and challenging issue.

In the future, we will focus on how to use generation methods to generate more coherent texts to link two cross-document event mentions, and how to jointly train the two tasks with entity coreference resolution.

Acknowledgements

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

[1] S. R. Ahmed, A. Nath, J. H. Martin, and N. Krishnaswamy. 2*n is better than n^2: Decomposing event coreference resolution into two tractable problems. In *ACL (Findings)*, pages 1569–1583. Association for Computational Linguistics, 2023.

[2] A. Bagga. Evaluation of coreferences and coreference resolution systems. In *LREC*, pages 563–572, 1998.

[3] S. Barhom, V. Shwartz, A. Eirew, M. Bugert, N. Reimers, and I. Dagan. Revisiting joint modeling of cross-document entity and event coreference resolution. In *ACL*, pages 4179–4189, 2019.

[4] C. A. Bejan and S. M. Harabagiu. Unsupervised event coreference resolution with rich linguistic features. In *ACL*, pages 1412–1422, 2010.

[5] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020.

[6] M. Bugert, N. Reimers, and I. Gurevych. Generalizing cross-document event coreference resolution across multiple corpora. *Comput. Linguistics*, 47(3):575–614, 2021.

[7] A. Caciularu, A. Cohan, I. Beltagy, M. E. Peters, A. Cattan, and I. Dagan. CDLM: cross-document language modeling. In *EMNLP (Findings)*, pages 2648–2662, 2021.

[8] A. Cattan, A. Eirew, G. Stanovsky, M. Joshi, and I. Dagan. Cross-document coreference resolution over predicted mentions. In *ACL/IJCNLP (Findings)*, pages 5100–5107, 2021.

[9] C. Chen and V. Ng. Joint inference over a lightly supervised information extraction pipeline: Towards event coreference resolution for resource-scarce languages. In *AAAI*, pages 2913–2920, 2016.

[10] X. Chen, S. Xu, P. Li, and Q. Zhu. Sentence rewriting with few-shot learning for document-level event coreference resolution. In *ICONIP (1)*, volume 13108 of *Lecture Notes in Computer Science*, pages 152–164. Springer, 2021.

[11] X. Chen, S. Xu, P. Li, and Q. Zhu. Cross-document event coreference resolution on discourse structure. In *EMNLP*, pages 4833–4843. Association for Computational Linguistics, 2023.

[12] Z. Chen and H. Ji. Graph-based event coreference resolution. In *ACL*, pages 54–57, 2009.

[13] P. K. Choubey and R. Huang. Improving event coreference resolution by modeling correlations between event coreference chains and document topic structures. In *ACL*, pages 485–495, 2018.

[14] A. Cybulska and P. Vossen. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *LREC*, pages 4545–4552, 2014.

[15] A. Cybulska and P. Vossen. Translating granularity of event slots into features for event coreference resolution. In *EVENTS@HLP-NAACL*, pages 1–10, 2015.

[16] G. Durrett and D. Klein. Easy victories and uphill battles in coreference resolution. In *EMNLP*, pages 1971–1982, 2013.

[17] B. J. Grosz. Focusing in dialog. In *TINLAP*, pages 96–103, 1978.

[18] W. Held, D. Iter, and D. Jurafsky. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *EMNLP*, pages 1406–1417, 2021.

[19] J. R. Hobbs. Coherence and coreference. *Cogn. Sci.*, 3(1):67–90, 1979.

[20] Y. J. Huang, J. Lu, S. Kurohashi, and V. Ng. Improving event coreference resolution by learning argument compatibility from unlabeled data. In *NAACL-HLT*, pages 785–795, 2019.

[21] S. Jia, W. Song, J. Gong, S. Wang, and T. Liu. Sentence ordering with a coherence verifier. In *ACL (Findings)*, pages 9301–9314. Association for Computational Linguistics, 2023.

[22] M. Joshi, O. Levy, L. Zettlemoyer, and D. S. Weld. BERT for coreference resolution: Baselines and analysis. In *EMNLP/IJCNLP*, pages 5802–5807, 2019.

[23] K. Kenyon-Dean, J. C. K. Cheung, and D. Precup. Resolving event coreference with supervised representation learning and clustering-oriented regularization. In **SEM@NAACL-HLT*, pages 1–10, 2018.

[24] S. Krause, F. Xu, H. Uszkoreit, and D. Weissenborn. Event linking with sentential features from convolutional neural networks. In *CoNLL*, pages 239–249, 2016.

[25] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In *EMNLP*, pages 188–197, 2017.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[27] J. Lu and V. Ng. Joint learning for event coreference resolution. In *ACL*, pages 90–101, 2017.

[28] J. Lu, D. Venugopal, V. Gogate, and V. Ng. Joint inference for event coreference resolution. In *COLING*, pages 3264–3275, 2016.

[29] X. Luo. On coreference resolution performance metrics. In *HLT/EMNLP*, pages 25–32, 2005.

[30] V. Ng. Supervised noun phrase coreference research: The first fifteen years. In *ACL*, pages 1396–1411, 2010.

[31] T. H. Nguyen, A. Meyers, and R. Grishman. New york university 2016 system for KBP event nugget: A deep learning approach. In *TAC*. NIST, 2016.

[32] H. Peng, Y. Song, and D. Roth. Event detection and co-reference with minimal supervision. In *EMNLP*, pages 392–402, 2016.

[33] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21: 140:1–140:67, 2020.

[34] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. D. Manning. A multi-pass sieve for coreference resolution. In *EMNLP*, pages 492–501, 2010.

[35] G. Ramesh, M. N. Sreedhar, and J. Hu. Single sequence prediction over reasoning graphs for multi-hop QA. In *ACL (1)*, pages 11466–11481. Association for Computational Linguistics, 2023.

[36] S. Sangeetha and M. Arock. Event coreference resolution using mincut based graph clustering. *International Journal of Computing and Information Sciences*, pages 253–260, 2012.

[37] S. Vahidnia. *Deep and Temporal Ontology Guided Clustering Methods and Representation Learning for Topic Detection and Tracking*. PhD thesis, University of New South Wales, Sydney, Australia, 2023.

[38] M. B. Vilain, J. D. Burger, J. S. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *MUC*, pages 45–52, 1995.

[39] P. Vossen, F. Ilievski, M. Postma, and R. Segers. Don't annotate, but validate: a data-to-text method for capturing event data. In *LREC*. European Language Resources Association (ELRA), 2018.

[40] H. Xu, B. Liu, L. Shu, and P. S. Yu. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *ACL*, pages 2324–2335, 2019.

[41] S. Xu, P. Li, and Q. Zhu. Improving event coreference resolution using document-level and topic-level information. In *EMNLP*, pages 6765–6775. Association for Computational Linguistics, 2022.

[42] H. Yan, Y. Sun, X. Li, Y. Zhou, X. Huang, and X. Qiu. UTC-IE: A unified token-pair classification architecture for information extraction. In *ACL (1)*, pages 4096–4122. Association for Computational Linguistics, 2023.

[43] B. Yang, C. Cardie, and P. I. Frazier. A hierarchical distance-dependent bayesian model for event coreference resolution. *Trans. Assoc. Comput. Linguistics*, 3:517–528, 2015.

[44] X. Yu, W. Yin, and D. Roth. Pairwise representation learning for event coreference. In **SEM@NAACL-HLT*, pages 69–78, 2022.

[45] Y. Zeng, X. Jin, S. Guan, J. Guo, and X. Cheng. Event coreference resolution with their paraphrases and argument-aware embeddings. In *COLING*, pages 3084–3094, 2020.

[46] L. Zhang, F. Kong, and G. Zhou. Adversarial learning for discourse rhetorical structure parsing. In *ACL/IJCNLP*, pages 3946–3957, 2021.