Beyond Benchmarks: Building a Richer Cross-Document Event Coreference Dataset with Decontextualization

Anonymous ACL submission

Abstract

Cross-Document Event Coreference (CDEC) annotation is challenging and difficult to scale, resulting in existing datasets being small and lacking diversity. We introduce a new approach leveraging large language models (LLMs) to decontextualize event mentions, by simplifying the document-level annotation task to sentence pairs with enriched context, enabling the creation of Richer EventCorefBank (RECB), a denser and more expressive dataset annotated at faster speed. Decontextualization has been shown to improve annotation speed without compromising quality and to enhance model performance. Our baseline experiment indicates that systems trained on RECB achieve comparable results on the EventCorefBank (ECB+) test set, showing the high quality of our dataset and its generalizability on other CDEC datasets. In addition, our evaluation shows that the strong baseline models are still struggling with RECB comparing to other CDEC datasets, suggesting that the richness and diversity of RECB present significant challenges to current CDEC systems. All the data and source code are publicly available.[1]

1 Introduction

Cross-Document Event Coreference (CDEC) annotation is a complex and labor-intensive process. Many restrictions have to be applied to make it feasible. As a result, existing data sets tend to be small and sparsely annotated. In the widely used benchmark dataset for CDEC, EventCorefBank (ECB+) (Cybulska and Vossen, 2014), 95% of annotated events are non-coreferential (Vossen et al., 2016). In 88% of all sentences, no events are annotated (Cybulska, 2021).

Moreover, existing CDEC datasets are often unrepresentative and lack expressiveness. Many of these datasets show lower referential diversity and ambiguity that allow high lemma match (CoNLL F1 61.9 for ECB+) (Bugert and Gurevych, 2021). ECB+ deliberately adds ambiguity to *admit to rehab* by adding *Lindsey Lohan admitted to rehab* as a new event along side the existing *Tara Reid admitted to rehab*. While this approach artificially increases diversity, which is necessary for a more representative dataset, there are still only two persons involved with *admit to rehab*. This does not align with the natural distribution of diverse events in a collection of news articles. For example, articles discussing protests in Hong Kong might reference numerous distinct protests, involving various people or organizations across different districts over the span of several months.

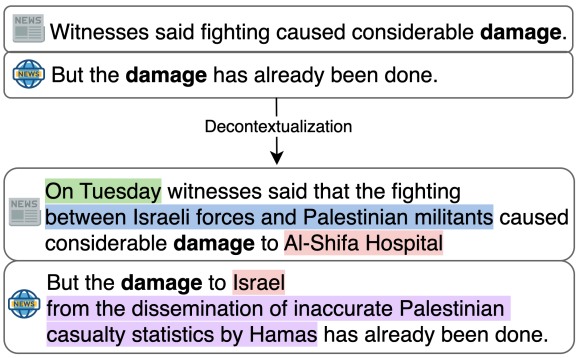

Figure 1: A pair of sentences is decontextualized to improve event coreference evaluation.

Improved decontextualization accuracy by large language models (LLMs) paves the way for a more efficient approach to CDEC annotation, which we use to create a more representative, and expressive CDEC dataset Richer EventCorefBank (RECB) in significantly less time. Unlike traditional CDEC annotation methods, which require annotators to read entire articles to gather context for evaluating coreference, our approach streamlines the process by reducing the context to a single, self-contained sentence. This simplifies the annotator's task, as they now only need to evaluate pairs of sentences,

[1] Anonymized for reviewing.

with participant, time, and location information all directly included in the sentence, rather than dispersed throughout the article. As shown in Figure 1, by minimizing the amount of text to process, we significantly reduce the cognitive load on the annotator and increase annotation speed. As a result, RECB provides a more scalable and efficient solution for generating large-scale CDEC datasets.

We demonstrate the utility of this new dataset by showing that models trained on it can generalize effectively to other test sets. Our experiments reveal that CDEC models trained on RECB achieve comparable performance on the ECB+ test set to models trained on the ECB+ training set. Our experiments also show the RECB dataset presents a more difficult challenge for current CDEC systems. Fewer restrictions are imposed on RECB data selection compared to other datasets. Unlike more curated datasets that focus on one topic or a limited set of events, RECB encompasses a broader range of event types, temporal and location variations, and participant roles. This results in greater lexical diversity and more referential ambiguity, making it harder for models to rely on simple cues like event trigger word.

We also highlight the effectiveness of decontextualization in both facilitating efficient CDEC annotation and enhancing model performance in experiments with RECB. Decontextualization enables the creation of a larger, more scalable dataset, allowing models to be trained on a wider variety of examples with rich representations, thereby improving their robustness and generalization across different domains and tasks.

The main contributions of this paper include: (1) The creation of RECB, a novel CDEC dataset with greater representativeness and expressiveness; (2) A new scalable methodology leveraging LLMs for decontextualization in CDEC annotation and modeling. The rest of the paper is organized as follows. In §2, we discuss related work on current datasets and modeling, and recent research on decontextualization. In §3, we describe the methods used in our data preparation. In §5, we describe and analyze our created data. In §6, we describe our experiments and results. We discuss our experimental results in §7, and conclude in §8.

2 Related Work

CDEC Dataset Creation Previous work on CDEC datasets include ECB+ (Cybulska and Vossen, 2014), FCCT (Bugert et al., 2020), MEANTIME (Minard et al., 2016), EER (Hong et al., 2016), and RED (O'Gorman et al., 2016). When annotating such datasets, annotators must exhaustively compare each event mention in the dataset against all other event mentions across documents to establish coreference relations. This is a labor-intensive process and as a result, existing datasets are all relatively small. In our work, by representing events as decontextualized sentences that can stand alone, there is the potential to create CDEC datasets on a much larger scale, as annotators only need to examine a pair of sentences instead of two articles to make coreference decisions.

A lot of effort is put into circumventing the scalability issue of manually created data by creating auto- or semi-automatically annotated CDEC datasets. Gun Violence Corpus (GVC) (Vossen et al., 2018) marks event references using a structured database of known gun violence events in a semi-automatic fashion. It considerably improves annotation efficiency and event variation compared to ECB+, but it does not apply to broader data topics other than gun violence. HyperCoref (Bugert and Gurevych, 2021) and WEC-Eng (Eirew et al., 2021) leveraged article hyperlinks in Wikipedia data to create data automatically. However, there is no guarantee that the events marked by the Wikipedia contributors will be consistent. Moreover, they mainly consist of Wikipedia-entry worthy or what Eirew et al. (2021) call referential event mentions, but do not cover descriptive or anecdotal events that arise in news reports.

CDEC Systems Recent advancements in CDEC modeling use neural cross encoders for event mention pairwise classification (Yu et al., 2022; Held et al., 2021; Caciularu et al., 2021b; Zeng et al., 2020; Cattan et al., 2020; Meged et al., 2020; Barhom et al., 2019). These methods typically encompass preprocessing steps of document topic modeling and event arguments labeling, and then the use of neural classifiers to analyze pairs of event mentions. This classification involves developing a scoring system based on the distance between event mentions within specific topics, which then uses agglomerative clustering for the formation of coreference event clusters.

Recent CDEC state-of-the-art systems use representation learning of mention pairs (Caciularu et al., 2021a; Held et al., 2021; Chen et al., 2023; Ding et al., 2024). Caciularu et al. (2021b)'s ap-

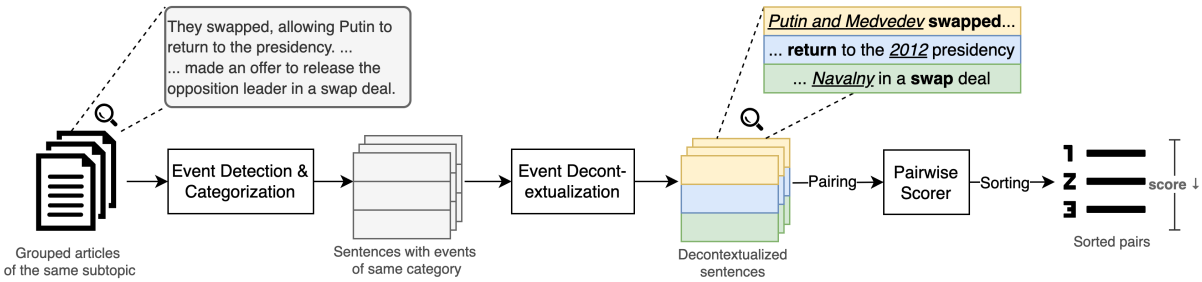

Figure 2: RECB data preparation pipeline.

proach involves pretraining the model on documents within the same topic to facilitate learning of cross-document relations. Moreover, they implemented a bigger context window to cross-encode and classify pairs of event mentions on document level. Held et al. (2021) leverages discourse coherence theory to limit candidate mentions to those within a learned latent embedding space, sampling hard negatives to train a classifier. Chen et al. (2023) incorporates global discourse structure, using rhetorical tree structures and shortest dependency paths to model interactions between event mentions. Ding et al. (2024) induce the model to learn through rationale-centric counterfactual data augmentation.

Decontextualization Event descriptions often span multiple sentences, making it essential for CDEC to understand how a sentence represents an event within a larger context. Important semantic information, such as participants, time, or location, may not always be explicitly stated in the immediate text. Including entire documents containing candidate event mention pairs in annotation can be inefficient, especially when they are too lengthy for annotation tasks or computational models (Vossen et al., 2018). To overcome this, we utilize event decontextualization, a method that makes events interpretable outside of their original document context while preserving their intended meaning (Zhao et al., 2023). We are the first to leverage LLMs for event summarization to decontextualize events for CDEC.

Recent research has explored various strategies in text generation, such as "decontextualization" (Choi et al., 2021), "enrichment" (Elazar et al., 2022), and "dense paraphrasing" (Tu et al., 2023; Rim et al., 2023). These methods leverage fine-tuned models such as T5 and SpanBERT to restore missing, elliptical, or underspecified content, enhancing the overall coherence and clarity of the text. Event components are rarely found within a single sentence, often requiring inference from the broader document context (Bugert and Gurevych, 2021). To address this challenge, we are the first to leverage zero-shot LLMs for event decontextualization, effectively reducing the need for lengthy context and enhancing the scalability of our method for creating the dataset.

Topic	Source	
SHIFA	Al Arabiya News (AAN)	Israel National News (INN)
PUTIN	Sputnik News (SN)	Google News (GN)
HONGKONG	China Daily (CD)	Google News (GN)
RITTENHOUSE	The Federalist (TF)	Google News (GN)

Table 1: Media sources of the articles from each topic.

3 Methodology

The RECB data preparation process requires multiple steps to convert a collection of documents into candidate sentence pairs that are ready for the annotation (Figure 2). This section describes the technical details involved in the pipeline for data preparation.

3.1 Source Article Collection

We collect English news articles on four contentious international and domestic topics: Al-Shifa Hospital Raid (SHIFA), Putin's 2024 Election Win (PUTIN), Hong Kong July 1 Protests (HONGKONG) and Kyle Rittenhouse Acquittal (RITTENHOUSE). As shown in Table 1, each topic includes articles from media outlets with differing political stances. The SHIFA topic, for instance, draws articles from Al Arabiya News and Israel National News, representing Arabic and Israeli perspectives. In the PUTIN topic, Sputnik News rarely discusses deaths of political opponents, while Western media outlets seldom highlight Russia's economic progress under Putin. In the following example from the RITTENHOUSE topic, events with similar meanings are conveyed through different ex-

pressions by sources with opposing political views.

(1) *Young man* ***acquitted*** *on all charges, acted in self-defense during Kenosha riots.*

Teen vigilante ***found not guilty*** *after fatally shooting two during racial justice protests.*

Our goal is to create the RECB dataset, which captures the lexical diversity of event mentions in real-world news articles. News sources with differing stances often emphasize unique aspects of a topic, highlighting a variety of events beyond the central, seminal event. Additionally, contentious news articles tend to generate extensive debate, covering multiple facets of the topic and providing a rich set of events for our task.

3.2 Event Extraction

CDEC involves identifying and grouping textual mentions across multiple documents that refer to the same event. These mentions include descriptive event mentions, typically expressed through verbs or nominalizations (e.g., *fighting*) to introduce new information, and referential event mentions, generally represented by noun phrases (e.g., *Israel–Hamas war*) to establish a reference point (Eirew et al., 2021).

We extract both types of event mentions from the source articles using the event extraction model proposed by Yao et al. (2021). We follow the definition of TimeML (Pustejovsky et al., 2003) to extract action verbs (e.g., *attacked*), aspectuals (e.g., *continue*), causative verbs (e.g., *cause*), static verbs indicating moral judgment (e.g., *protesters deserve recognition*), and verbs referring to generic events with unspecified time or location (e.g., *People discuss politics*).

3.3 Sentence Decontextualization

Decontextualization task is proposed by Choi et al. (2021), which enriches each sentence with missing contexts through name completion, bridging, etc. Inspired by this, we apply the GPT model to decontextualize each article sentence based on the context from the whole article. We use the OpenAI API with version `o1-preview`. In the prompt, we instruct the model to generate a sentence with richer context based on the original one containing the target event along with its full article. We include the prompts in the Appendix A.2.

For example in Figure 1, the relation between the two *damage* event mentions is ambiguous. Decontextualization clarifies the context by adding specific actors (*Israeli forces, Palestinian militants, Hamas, and Israel's reputation*), locations (*Al-Shifa Hospital*), and time (*On Tuesday*), as well as explaining the cause through prepositional phrases. With this additional context, it becomes clear that the two ***damage*** mentions are not coreferential.

Decontextualization effectively adds missing components such as participants, time, and locations to events from the article's context. This step allows for CDEC annotation to be conducted solely on sentence pairs, eliminating the need to refer to the entire article, which significantly improves our annotation speed on RECB.

3.4 Pairwise Scoring

The CDEC task is typically approached by scoring event pairs and then clustering based on these scores across all pair combinations. However, the number of event pairs increases quadratically with the number of events, making it infeasible to annotate or model large datasets (Vossen et al., 2018). Previous work, such as ECB+, addresses this issue by including more topics, each with fewer documents, thus reducing the number of in-topic pairs. Additionally, ECB+ only annotates sentences containing seminal events, averaging 1.87 sentences per document.

We address this issue by generating pairs only from articles with similar content and the same event action type. Within each subtopic (one news source from a topic), we first group articles based on their textual similarity using agglomerative clustering, leveraging both BERT-based similarity (Reimers and Gurevych, 2019) and TF-IDF scores.

Observing that events of some different action types rarely corefer (e.g., the reporting event *said* rarely corefers with the perception event *feel*), we employ a customized verb list to classify action types and restrict pairings to those of the same type within the grouped articles. The verb list includes common reporting verbs, copulas, aspectual verbs, and verbs related to mental activities.

We then apply the pairwise cross-encoder (Yu et al., 2022) to score all generated pairs from our data, ranking them from highest to lowest based on the score. The top-ranked pairs typically exhibit similar or ambiguous meanings, whereas the lower-ranked pairs are more likely to be non-coreferential.

4 RECB Annotation

We describe the process for constructing the RECB dataset, and discuss the annotation results and statistics.

4.1 Near-Identity Labels

Existing CDEC tasks typically consider only binary relations between events. However, events in real-world data often exhibit more complex relationships beyond just coreferential or non-coreferential. Reducing event labels to binary classifications can introduce noise into the annotation process and overlook these richer relationships.

To address this, we introduce additional labels to capture directional near-coreference relations as a middle ground, beyond the standard IDENTITY, NOT-RELATED, and CANNOT-DECIDE categories. These include CONCEPT-INSTANCE(-INV), WHOLE-SUBEVENT(-INV), and SET-MEMBER(-INV) relations. Annotators are asked to consider why pairs may feel coreferential yet differ, and to determine the true relationship instead of making a binary decision based purely on intuition.

While this paper focuses on events with binary labels for the CDEC task, our annotated near-identity event sets are useful for many downstream tasks such as information retrieval. Understanding event relations enables systems to retrieve more relevant information. E.g., querying for a broader event (such as *protest*) can also return related subevents (such as *speeches* or *clashes*), enhancing the granularity of the results (Guan et al., 2024).

4.2 Stopping Criteria

Annotators are tasked with classifying ranked event pairs with decontextualized sentences into one of the predefined event relation labels, starting from the top-ranked pairs. For each subtopic, they stop annotation after encountering 200 consecutive non-coreferential pairs. The goal of this process is to maximize the recall of positive pairs while maintaining feasibility, as these pairs become increasingly sparse as annotation progresses, with varying thresholds across topics.

4.3 Annotation Process

We hire 4 researchers and graduate students from the linguistics and computer science departments of a US-based university for the annotation work. Each annotator is familiar with the definition of the CDEC task and the annotation guidelines which include ambiguous cases from each event category. We include the the full guideline in Appendix A.3.

Annotators are paired for each subtopic, with each pair tasked with iteratively annotating 200 event pairs. During this "burn-in" phase, they work closely together to resolve discrepancies and clarify any misunderstandings in the guidelines. These 200 pairs are designed to include edge cases, exposing annotators to a range of complexities in the data. At this stage, confusion between identity and near-identity relations is most common due to their shared participants and overlapping meanings. We train our annotators to assess event granularity and refer to specific cases in the guidelines to resolve these ambiguities.

After the burn-in stage, each pair of annotators continues to double-annotate 400 mention pairs per subtopic and jointly adjudicates the results. We assess the annotation quality using Cohen's κ, which yields a score of 0.70 for all labels and 0.78 for binary labels (with near-identity labels mapped to NOT-RELATED), demonstrating strong agreement between annotators.

The remaining pairs are divided among the annotators for single annotation. The annotation process took place over the course of one month. On average, each mention pair took 15 seconds to annotate, with the four annotators collectively spending 254 hours to complete the annotation after reaching the stopping criteria for the RECB dataset. Finally, we represent the annotated pairs with binary labels as graphs and map them to event coreference clusters. Any conflicts in relation links are resolved through majority voting on the graph edges.

Table 2 presents statistics of the RECB dataset. On average, each topic contains 136 documents. At the sentence level, RITTENHOUSE-GN contains the most sentences, indicating that documents within this topic are longer and potentially more complex, whereas SHIFA has the fewest sentences per document. Decontextualization enriches the sentences by about 12% in terms of token count. We also provide the number of mention pairs annotated for each topic, among which approximately 20% have near-identity relations. Finally, we present the number of clusters for each topic, with PUTIN having the highest number.

Topic	Source	Docs	Sentences	Ori / Decont. tokens	Mentions	Pairs	Near-Identity Pairs	Clusters
SHIFA	AAN	74	643	17k / 19k	1,267	6,834	406	353
	INN	58	692	17k / 20k	1,082	4,933	303	311
PUTIN	SN	77	1,047	29k / 32k	2,096	12,796	3,610	1,075
	GN	77	1,164	31k / 35k	2,346	12,197	3,690	1,094
HONGKONG	CD	76	868	22k / 26k	1,324	3,281	333	788
	GN	78	897	25k / 29k	1,677	5,226	368	1,046
RITTENHOUSE	TF	40	684	18k / 20k	1,025	1,679	364	493
	GN	64	1,340	34k / 36k	2,567	9,219	1,438	794
Total		**544**	**7,335**	**195k / 220k**	**13,384**	**51,665**	**10,512**	**5,954**

Table 2: Data statistics overview of RECB dataset. The number of articles, sentences, and tokens from each subtopic are reported after the data collection. We also report the number of event mentions, annotated pairs and cluster numbers from the human evaluation.

	RECB	ECB+	GVC
Docs	588	982	510
Sentences	7,335	15,812	9,782
Annot. sentences	7,121	1,840	4,604
Mentions	13,384	6,833	7,298
Clusters	5,954	2741	1,411
Non-singleton Clusters	2,358	1,958	1,048
Positive Pairs	26,756	26,712	50,799
Lemma-cluster Ratio	3.3	2.1	2.6
Cluster-lemma Ratio	5.6	3.5	2.0

Table 3: Comparison of the statistics on the RECB, ECB+, and GVC datasets.

5 Data Analysis

We compare RECB with the other two widely used CDEC datasets. ECB+ (Cybulska and Vossen, 2014) extends the original ECB (Bejan and Harabagiu, 2010) by adding more similar and ambiguous events, enhancing the coverage and diversity of ECB. Only seminal events and a small number of other event mentions are annotated. The GVC (Vossen et al., 2018) dataset contains events involved in gun violence incidents (e.g., location, time, victim details) with related news articles. Only event mentions related to gun violence are annotated.

Table 3 compares the key statistics between RECB, ECB+ and GVC. ECB+, although contains with twice number of the sentence than RECB, only annotates around 10% of the sentences. Since the annotation on RECB does not filter specific event mentions, almost all the sentences (97%) with events from the dataset are annotated, indicating a high density of annotations relative to its overall sentence count. This annotation strategy also improves the efficiency on CDEC annotation by reducing the total number sentences required to be read (e.g., all sentences from ECB+ need to be inspected to annotation the 10% sentences).

RECB has the most number of clusters among all three datasets, when the positive pairs count is comparable to ECB+. This suggests that RECB has a more fine-grained or densely annotated coreference structure. GVC, on the other hand, has more positive pairs that coming from less diverse clusters.

Lemma-cluster ratio measures the number of different lemmas from the same cluster. RECB having a higher number of lemmas indicates that the clusters are more lexically diverse, capturing a broader range of expressions for each event.

Cluster-lemma Ratio refers to the referential diversity or ambiguity of the dataset. RECB having the highest clusters per lemma indicates a broader spectrum of event distinctions for the same lemma. This suggest more complexity in how it handles event distinctions, making it potentially better suited for tasks requiring nuanced CDEC.

6 Experiments

We evaluate our proposed RECB dataset on two strong CDEC baselines including the setup of lemma matching and pairwise encoding. We compare the model performance on RECB and existing CDEC datasets such as GVC and ECB+ under the cross-evaluation setting.

To keep the integrity of the event coreference clusters, we separate the documents by their topics for train and test splits. Specifically, we select the documents from three topics for training, and the documents from one subtopic of the last topic for testing; and from the other subtopic for validation. We rotate the train split with different topics during the evaluation.

6.1 Lemma Matching

Lemma matching is a heuristic-based method that links two event mentions as co-referential if they have the same lemmatized mention head. Cybul-

Test Split	CoNLL F1	Pairwise F1
ECB+	61.9	9.5
GVC	33.8	36.4
SHIFA	32.3	6.2
PUTIN	39.2	5.5
HONGKONG	48.2	**4.9**
RITTENHOUSE	**30.0**	5.9

Table 4: Lemma matching results on the test split of CDEC datasets. Pairwise F1 is based on the scores from all the sentence pairs, while CoNLL F1 is based on final event clusters.

ska and Vossen (2014) shows that the CDEC system performance largely stems from mentions with overlapped lemmas. Due to the nature of the task, and the way those datasets are constructed, lemma matching is considered a strong baseline and an effective approach to show the event diversity in the dataset (Choubey and Huang, 2017; Bugert et al., 2021) .

Table 4 shows the lemma matching results on different CDEC datasets. GVC performs much worse than ECB+ by obtaining a higher CoNLL F1. This may due to the fact the Vossen et al. (2018) intentionally incorporates ambiguous event mentions to the GVC dataset (e.g., event *shot* appears frequently in almost all the event types). However, the high pairwise F1 on GVC suggests that the coreference patterns from certain clusters could be more easily learnt by systems during training (e.g., large coreference clusters of events all with the same lemma). compared to GVC and ECB+, RECB has both low CoNLL and Pairwise F1, indicating the event complexity in our dataset. Among all topic, HONGKONG has the highest CoNLL F1. Through inspection, this is due to the type invariability of certain event mentions (e.g., *ceremony* always refers to the *flag-raising ceremony* in this topic).

6.2 Pairwise Encoding

We apply PAIRWISERL, a pairwise representation learning method proposed by Yu et al. (2022) as another baseline for the CDEC task. PAIRWISERL is a cross-encoder with $\text{RoBERTa}_{\text{LARGE}}$ as the base model. It is applied on each sentence pair with event mentions, followed by the agglomerative clustering to form the corefernece clusters. PAIRWISERL shows near state-of-the-art results on ECB+, and serves as the basic structure for more recent CDEC systems (Chen et al., 2023; Ding et al., 2024). We provide the model details in the Appendix A.1.

Table 5 shows the cross-evaluation results from pairwise encoders on different CDEC datasets. When evaluating on the ECB+ test split, models trained with RECB show much close results with the ECB+ train split (82.9 from $\text{RECB}_{\text{-w/o HongKong}}$), suggesting the high quality of our dataset and its generalizability on other CDEC dataset.

Model trained on the $\text{RECB}_{\text{-w/o Rittenhouse}}$ split shows lower performance compared to models trained on other RECB split configuration. We observed an increase in the number of coreference errors under the topic *find guilty of killing pregnant partner*, which has similar event patterns in the RITTENHOUSE split that is not in the training under this setting. As a result, the model struggles with coreference resolution in the test set without access to these familiar event patterns. This suggests that domain overlap and event similarity play an important role in model performance, and removing key event types from training data can hinder the model's ability to generalize to similar events in the test set.

Models trained on either ECB+ or GVC perform worse on GVC (20 points lower). This may be due to the narrower domain and highly ambiguous events from the dataset. Correspondingly, GVC train split also cannot be generalized well on the other two datasets with a significant performance drop (30 points lower on ECB+).

Comparing to the results from training and testing on the same dataset, the model trained on RECB on every subtopic performs worse than ECB+ or GVC (7.5 to 19.1 points lower than ECB+). This indicates the complexity of our dataset and poses greater challenge to current and furture CDEC systems.

7 Analysis

7.1 RECB Baselines

We describe the result patterns and common errors from the model baselines trained and tested on RECB (Table 5). Models on ECB+ and GVC datasets outperform those trained and tested on any of the RECB test split. This suggests that RECB introduces more diverse and challenging event patterns, including temporal, location changes and participant variations.

For the HONGKONG test split, One of the fre-

Train Split	Test Split (CoNLL F1)					
	ECB+	GVC	SHIFA	PUTIN	HONGKONG	RITTENHOUSE
ECB+	**82.9**	64.9	59.5	71.4	67.1	63.6
GVC	50.2	**84.4**	53.6	64.1	63.7	63.1
RECB$_{\text{-w/o Shifa}}$	80.2	62.9	**63.8**	-	-	-
RECB$_{\text{-w/o Putin}}$	82.4	64.8	-	**75.4**	-	-
RECB$_{\text{-w/o HongKong}}$	**82.9**	65.1	-	-	**68.3**	-
RECB$_{\text{-w/o Rittenhouse}}$	78.8	64.1	-	-	-	**68.5**

Table 5: Cross-evaluation results on the test split of CDEC datasets with pairwise-encoding.

quent error lemmas is *attack*, which spanned 24 different clusters with a broad and even distribution. In example 2, the pair of event actions is very similar, showing that identifying the temporal difference is important to evaluate coreference.

(2) *- The protesters tried to **attack** riot police with iron bars during an unauthorized assembly on July 27, 2019.*

*- During the Sheung Shui demonstration on July 13, radical protesters beat and **attacked** police officers with iron bars.*

In contrast, the PUTIN split largely centers around the lemma *election*, which spanned only 6 clusters, most of which focused on the *2024 Putin election* event cluster. This indicates that events in HONGKONG are more temporally and geographically dispersed, posing a greater challenge for the model to learn.

While models achieve high results on ECB+ and GVC, CDEC is far from being solved. Datasets like RECB, which offer greater representativeness and diversity, are crucial for advancing the field and providing the needed space for models to improve and better handle the complexities of real-world scenarios.

7.2 CDEC with Decontexualization

We evaluate the effect of decontextualization on the CDEC task with our dataset. Table 6 shows the CoNLL F1 results for different test splits of the RECB dataset, comparing performance of the PAIRWISERL model trained and evaluated on the original or decontextualized sentence pairs.

Comparing to the original sentences, the overall results demonstrate that using decontextualized sentence pairs improves performance across all test splits (1.2 - 3.2 higher). It indicates that reducing context to a single, self-contained sentence

Test Split	Original	Decontextualization
SHIFA	63.8	**65.5**
PUTIN	75.4	**78.4**
HONGKONG	68.3	**71.5**
RITTENHOUSE	68.5	**69.7**

Table 6: CoNLL F1 Results on the test split of RECB evaluated with original and decontextualized sentence pairs.

can offer the model with more contextual information, leading to better generalization and more accurate coreference resolution. This echoes with our previous observations on the robustness of the decontextualization step to enrich sentences with contexts that in turn helps both human annotators and models better capture coreference relationships and nuanced event patterns.

8 Conclusion

In this paper, we create a novel CDEC dataset RECB leveraging decontextualization to streamline and scale the annotation process. Through our experiments, we demonstrated that RECB provides a more representative and expressive dataset compared to traditional datasets like ECB+ and GVC. Models trained on RECB achieve competitive performance. It indicates RECB's high quality and generalizability on other CDEC dataset. Comparing to ECB+ and GVC, our dataset also post bigger challenges to current systems to identify coreferential events, providing a benchmark and directions for future model development. Our findings also suggest that decontextualization is a highly effective method for enhancing both annotation speed and model performance, offering a scalable solution to the growing demands of CDEC data in complex real-world scenarios.

9 Limitations

Our approach relys on O1-PREVIEW for high quality decontextualization. It comes with financial costs. This reliance on a large-scale, pre-trained language model limits the scalability of the method, particularly for large datasets. A potential area for improvement lies in exploring more cost-effective alternatives, such as smaller, locally-deployable models or fine-tuned models that specifically focus on the task of decontextualization. This exploration would make our methodology more accessible and scalable, especially in resource-constrained environments.

Another limitation is the relatively small number of topics in our dataset, which affects the overall effectiveness of CDEC task, which would benefit from a large and diverse set of topics to fully capture the range of event variations across different domains. With a limited number of topics, the variety of event types, temporal shifts, and participant variations is restricted, potentially leading to overfitting or poor generalization in the models. To address this, future work should focus on expanding the dataset by introducing more diverse topics, which would provide better training data for models to learn CDEC in more realistic and challenging scenarios.

10 Ethical issues

References

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden. Association for Computational Linguistics.

Michael Bugert and Iryna Gurevych. 2021. Event coreference data (almost) for free: Mining hyperlinks from online news. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 471–491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Bugert, Nils Reimers, Shany Barhom, Ido Dagan, and Iryna Gurevych. 2020. Breaking the subtopic barrier in cross-document event coreference resolution. In *Text2Story@ECIR*.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021a. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E. Peters, Arie Cattan, and Ido Dagan. 2021b. Cdlm: Cross-document language modeling.

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

Xinyu Chen, Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2023. Cross-document event coreference resolution on discourse structure. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4833–4843, Singapore. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Prafulla Kumar Choubey and Ruihong Huang. 2017. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2133, Copenhagen, Denmark. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Agata Katarzyna Cybulska. 2021. Event coreference in the news: Who, what, where and when?

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.

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 2498–2510, Online. Association for Computational Linguistics.

Yanai Elazar, Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2022. Text-based NP enrichment. *Transactions of the Association for Computational Linguistics*, 10:764–784.

Yong Guan, Dingxiao Liu, Jinchen Ma, Hao Peng, Xiaozhi Wang, Lei Hou, and Ru Li. 2024. 1. event gdr: Event-centric generative document retrieval.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 1406–1417. Association for Computational Linguistics.

Yu Hong, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer. 2016. Building a cross-document event-event relation corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 1–6, Berlin, Germany. Association for Computational Linguistics.

Yehudit Meged, Avi Caciularu, Vered Shwartz, and Ido Dagan. 2020. Paraphrasing vs coreferring: Two sides of the same coin.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness, and James Pustejovsky. 2023. The coreference under transformation labeling dataset: Entity tracking in procedural texts using event models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12448–12460.

Jingxuan Tu, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky. 2023. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics (IWCS)*, Nancy, France. Association for Computational Linguistics.

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. 2016. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85.

Piek Vossen and Agata Cybulska. 2017. Identity and granularity of events in text.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: a data-to-text method for capturing event data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jiarui Yao, Haoling Qiu, Jin Zhao, Bonan Min, and Nianwen Xue. 2021. Factuality assessment as modal dependency parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1540–1550, Online. Association for Computational Linguistics.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2022. Pairwise representation learning for event coreference. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 69–78, Seattle, Washington. Association for Computational Linguistics.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jin Zhao, Nianwen Xue, and Bonan Min. 2023. Cross-document event coreference resolution: Instruct humans or instruct GPT? In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 561–574, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Model Details

We use OpenAI API to run GPT models for the decontextualization. The pricing for the models used in the paper is on the OpenAI website.[2]. We fine-tune PAIRWISERL with configurations on four 40GB Nvidia RTX A6000 GPU. We use the hyperparameter setting with epoch 10, batch size 32, input length 128. All the other hyperparameters remain default. Following the previous work (Yu et al., 2022), we evaluate on the test split with the model that achieves the best pairwise F1 on the dev split. The training time is around 3 hours for each model run.

A.2 More on Decontextualization

A.2.1 Prompting

The objective of our approach is to automatically generate decontextualized sentences using GPT-based models, making sentences that contain events interpretable without the need for surrounding context.

We design a prompt template that included a full article with all events tagged using the marker _EVENT following the event trigger word. The instruction given to `o1-preview` was to add the necessary contextual information for each tagged event based on the article's content, reducing the article context in the sentence and ensuring that the resulting sentence could stand alone. The phrasing of the instruction was refined through multiple iterations, incorporating human evaluations to optimize the clarity and quality of the outputs. While `o1-preview` was allowed to edit other parts of the sentence to improve fluency and naturalness, we explicitly instructed it to preserve the tagged event triggers to ensure traceability.

Input article in prompt below contains multiple sentences formatted as following original, and output article contains corresponding decontextualized sentences. The prompt is shown in Figure 3.

A.2.2 Human Evaluation

We sampled 100 sentences and annotated all edit categories as defined in (Choi et al., 2021). Table 7 presents the decontextualized edits `o1-preview` makes for the sampled 100 sentences.

shows the helpfulness of decontextualization edits in determining coreference relations between events in sentence pairs. The second column indicates the number of sentence pairs that would have been mislabeled as CANNOT-DECIDE without decontextualization. Overall, decontextualization has proven effective in helping human annotators make more accurate judgments in the CDEC task.

[2] https://openai.com/api/pricing/

Edit Category	Counts
Pronoun/NP swap	32
Name Completion	12
Global Scoping	22
Dm Removal	0
Bridging	40
Global Scoping	22
Addition	21

Table 7: Counts of Decontextualization Edit Categories.

True Label	Helpfulness
Identity	20/46
Confirmation	0/10
Subevent	2/3
Subset	1/11
Not-Related	16/30

Table 8: Decontextualization Helpfulness.

A.3 Annotation Guidelines

A.3.1 Identity

Understanding Identity Strict Identity can be identified through several linguistic mechanisms: Repetition: Involves using the same word or phrase multiple times to refer to the same entity or event, making coreference easy to establish.

> "The protest_EVENT continued for hours." "The protest_EVENT was peaceful."

Anaphora: Uses pronouns or expressions to refer back to a previously mentioned entity or event, requiring resolution of the pronoun to its antecedent.

> "The protest_EVENT continued for hours. " "It_EVENT was peaceful. "

Synonymy: Refers to different words or phrases with similar meanings used to refer to the same entity or event, which requires recognizing semantic equivalence.

> "The protest_EVENT" "The demonstration_EVENT"

Disjunction (Negative Indicator): Presents alternatives using"or", complicating coreference resolution by introducing multiple potential referents.

```
You are an intelligent text rewriter that can add contexts to events from each sentence based on the full article.

I will provide you with a news article, each sentence in the article is indicated by a numerical identifier [].
Each event trigger is tagged as trigger_EVENT in all the sentences.

{article_list}

Rewrite each sentence by adding contexts from the article to the tagged events.
Each rewritten sentence should provide full contexts about its tagged events without the need to read the whole article.
The tagged events should be kept unchanged in the rewrite.
The output format should be a list of rewritten sentence, each indicated by a numerical identifier [], e.g.,

 [1] rewritten sentence 1
 [2] rewritten sentence 1
 [3] rewritten sentence 1

The number of rewritten sentences should be the same as the number of sentences in the article. Do not combine or merge sentences in the output.
```

Figure 3: GPT prompt for decontextualization.

> "The protest_EVENT or rally_EVENT continued for hours."

Perspective: Refers to differences in word choice that reflect pragmatic use while referring to the same entity or event. For instance, "aggressors" vs. "liberators" or "liberation" vs. "invasion", where the coreferential link is influenced by framing.

Labeling Identity: Event A is identiy with Event B.

A.3.2 Subevent

Understanding Subevent Relations Subevents refer to smaller, component events that are part of a larger, overarching event. The goal of subevent annotation is to determine whether two event mentions describe the same event at different levels of granularity or detail.

Tests for Subevent Annotation When annotators are unsure if two event mentions are in a subevent relation, the following questions should help guide the decision:

Granularity Test: Do the two mentions describe events at different levels of detail? For example, does one describe a specific action while the other describes a broader series of actions or outcomes?

> "The protest turned_EVENT violent." (high granularity, overarching)
> "Protesters clashed_EVENT with the police" (fine granularity, subevent)

Part-Whole Test (Meronymy): Does one event form part of another event? Consider whether the two events have a part-whole relationship, where one is a component of the larger process.

> "The election_EVENT was held in November." (overaching event)
> "Voters in the eastern district began casting_EVENT ballots at 7 AM." (subevent)

Temporal Perspective: Event durations shift from short, specific moments to longer periods.

> "The protest_EVENT lasted two hours." (short duration)
> "Unrest continued_EVENT for months." (extended duration)

Labeling Subevent-Whole: Event 1 in the first sentence is a necessary stage/phase of Event 2 in the second sentence.
Whole-Subevent: Event 1 in the second sentence is a necessary stage/phase of Event 2 in the first sentence.

A.3.3 Confirmation

Understanding Confirmation Relations In a confirmation relation, one event is a concrete, specific example or instance of a broader, more abstract event. The abstract event provides a general or high-level description, while the more specific event confirms or details that abstract event.

Test for Confirmation Annotation When annotators are uncertain whether two event mentions are in a confirmation relation, they should apply the Hyponymy Test:

Hyponymy (Type-of) Test: Is one event a more specific instance or type (hyponym) of a broader event? The broader event implies the existence of more specific instances, and the more concrete event provides a confirmation of the broader concept. In this following case, the action of signing the law confirms the broader event of legislation passing, establishing a confirmation relation.

> "The government passed_EVENT the legislation." (broad event)
> "The law was signed_EVENT by the president."(specific event confirming the broader legislation process)

Temporal Perspective Hypothesis: The closer the text is to the event's occurrence, the more specific and concrete the event descriptions will be. As the temporal distance increases, descriptions of the same event become broader, more general, and more abstract. Thus, the more abstract event descriptions can often be confirmed by specific, earlier mentions of the event.

> "On November 1st, the new trade agreement was signed_EVENT by the representatives." (specific, closer to event)
> "Last year, several key trade deals were finalized_EVENT."

Labeling Concept-Instance: Event 1 in the first sentence is an abstract generalization and Event 2 in the second sentence is an instance of the abstract generalization. Instance-Concept: Event 1 in the first sentence is an instance of an abstract generalization and Event2 in the second sentence is an abstract generalization.

A.3.4 Subset

Understanding Subset Relations This relation occurs when one event mention refers to a collection of events, while the other event mention refers to a subset of this collection. The subset may involve specific members or occurrences within the broader set of events. Quantifiers Clues and Contextual phrases like "one of the" or "first" can be strong indicators that a subset relation exists between event mentions.

Tests for Subset Annotaion Totality vs. Partiality test: Does one event refers to a group of events (e.g., multiple arrests, several observations), while the other focuses on one or a few events from that group?

> "The concert series in New York ran for three consecutive nights." (Set of concerts) "The opening night of the concert was a major success." (Subset: one concert)

Time or Participant Specificity test: Does subset event often have additional details specifying time, participants, or other features that distinguish it from the full set?

> "Several protests erupted in the city." (Participants: ndefined protesters) "The student protest drew significant media attention." (Participant: the student)

Labeling Set-Member: Event 1 in the first sentence is a collection of events and Event 2 in the second sentence is a subset from the larger collection. Member-Set: Event 1 in the first sentence is a subset from the larger collection and Event 2 in the second sentence is a collection of events.

A.3.5 Not Related

The two event mentions refer to distinct actions or occurrences involving different participants and/or taking place at different times or locations.

In cases where the distinction is ambiguous and annotators are hesitant to classify the events as "Not Related," Topical Relations(Vossen and Cybulska, 2017) may be applied. Topical relations refer to events that are related by a common theme but remain distinct. For example, two protests on the same issue occurring in different locations can be considered topically related

> "Diana joined_EVENT Noon Against Putin protest. "
> "Dimitry attended_EVENT the Noon protest. "

Additionally, there are Causal Relations, where one event leads to or is the result of another. For example, "victory_EVENT" may result from "people voted_EVENT." However, these events are not considered identical.

"Putin's landslide victory_EVENT in 2024." "Almost 90% Russian people voted_EVENT for Putin"

We do not annotate other relations like topical and causal relations for this task.

Labeling Not-Related: two events are not related in any of the ways described above.

A.3.6 Cannot Decide

This is usually due to lack of sufficient context. Choose this option if you cannot decide if the participants, location, or time are the same, due to the lack of context.

> "Eran Bendheim , an Israeli photographer and web developer living in New York City , captured air traffic in the night sky in April 2019 and again in April 2020. "
> "The photo was taken last month by him. „

The second sentence indicates the photo was taken in 2019, which overlaps the time of the capturing

event in the first sentence, but we don't know who "him" in the second refers to. Therefore we cannot decide if the two refer to the same event or not.

Labeling Cannot-Decide: cannot decide due to lack of sufficient context.