Data Multiplication for Cross-Document Event Coreference with Large Language Models

Anonymous ACL submission

Abstract

Creating a Cross-Document Event Coreference (CDEC) dataset is complex and labor-intensive. As a result, existing CDEC data sets are small in the size of event mentions and limited in the number of event types that are covered. This is a substantial hurdle in training robust CDEC systems. In this paper, we propose to leverage large language models (LLMs) to address this bottleneck. Specifically, to enrich trigger variety and word order variation, we introduce two *Data Multiplication (DM)* techniques that employ GPT-4 to generate realistic synthetic training data, effectively increasing the volume of existing annotated CDEC data sets with high-quality annotations. We demonstrate the effectiveness of our approach by conducting experiments on the ECB+ and Aylien Covid datasets, and show that adding LLM-generated CDEC data improves the performance on the two benchmarks by up to 1.8 and 3 points respectively in CoNLL F1. We believe that our method is generally applicable to other tasks as well and underscores the potential of LLMs in addressing data scarcity challenges in natural language processing tasks. All the data and source code are publicly available. [1]

1 Introduction

Cross-Document Event Coreference (CDEC) is the task of identifying and linking event mentions across multiple documents that refer to the same real-world event. Recent approaches in CDEC involve training an transformer-based scorer to assess the likelihood of a pair of event mentions referring to the same event. This requires training data meticulously annotated by human annotators who compare each event mention against all others across documents, a costly and time-consuming process. To make it feasible, the scope of the annotation often needs to be restricted by selecting specific topics and predefined event types (Cybulska and Vossen, 2014). As a result, the resulting annotated data sets tends to be small in size and limited in event type. There also have been efforts to automatically generate CDEC datasets by exploiting hyperlinks in Wikipedia articles, but the resulting data sets tend to focus more on referential events that are Wikipedia worthy, as opposed to unlimited descriptive events frequently found in news reports (Eirew et al., 2021). These limitations make it challenging to apply models trained on these datasets effectively to real-world scenarios. The collection and curation of diverse, high-quality data sets are crucial for enhancing the performance of CDEC models and their application in downstream tasks.

In this paper, we introduce two innovative Data Multiplication (DM) methods for generating CDEC-annotated data using GPT-4 (OpenAI, 2023), and demonstrate their effectiveness in extending both the ECB+ (Cybulska and Vossen, 2014) and Aylien Covid datasets (Zhao et al., 2023). Specifically, we prompt GPT-4 to generate paraphrases of the event mentions in the two data sets without changing their referential properties and multiply the resulting coreferent event mention pairs by exploiting their transitive properties. We show that models trained on just the generated data points are almost as good as that trained on the original data sets, and models trained on combined (original and generated) data sets yield significant improvements over that trained on the original data sets alone. Our approach differs from previous data augmentation approaches in that rather than predicting annotation on additional unannotated data, we generate additional new synthetic data in zero-shot setting for small human annotated data sets by taking advantage of the strong language generation capability of LLMs. We believe that our approach is not limited to CDEC and can be generalized to other tasks. The implication of our research is that the impact of human-annotated data can be magnified by generating similar data points with LLMs,

[1] Anonymized for reviewing.

thus reducing the need for human annotated data.

The rest of the paper is organized as follows. In §2, we discuss related work on data augmentation using LLMs, CDEC data creation and modeling. §3 describes our DM methods. §4 describes the construction of new data by applying DM on two existing CDEC datasets. We present our experimental results on two datasets in §5, we discuss error analysis in §6 and conclude in §7.

2 Related Work

The utilization of data generated by language models has become increasingly prevalent across various NLP tasks (Chowdhury and Chadha, 2023; Su et al., 2023; Mekala et al., 2022). Recently, LLMs have been particularly effective in generating in-context training data for sentence-level tasks such as slot tagging (Lee et al., 2021) intent classification (Sahu et al., 2022), and paraphrasing (Tu et al., 2023). These techniques have been successful in augmenting training sets and enhancing the performance of text classification models. However, they often struggle to simultaneously generate text and accurate labels, typically requiring postprocessing for data cleaning, human supervising for relabelling, filtering, additional word alignment model or controlling the outputs to adhere to specific slot types. Zhao et al. (2023) employ GPT-4 to label event coreference pairs automatically and comparing with human annotations. However, the generated dataset label accuracy is not comparable with gold data. Our approach is different because it requires minimal postprocessing while achieving data quality comparable to gold-standard data in terms of text authenticity and label accuracy. Instead of using LLMs for predicting labels, our DM method strategically manipulates text while retaining high-quality labels, thereby streamlining the process and enhancing data accuracy.

Recent advancements in CDEC modeling uses 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 use agglomerative clustering for the formation of coreference event clusters. Former state-of-the-art pairwise model (Yu et al., 2022) marked a significant innovation by shifting the focus to the representation learning of mention pairs rather than individual mentions. Building on this, Caciularu et al. (2021b) set a new standard on the ECB+ dataset. Their approach involved 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. Our experiments evaluate the effectiveness of the DM approaches on sentence level, therefore we adopted the pairwise cross encoder settings in Yu et al. (2022).

Recent efforts in creating CDEC datasets, including ECB+ (Cybulska and Vossen, 2014), MEANTIME (Minard et al., 2016), EER (Hong et al., 2016), and RED (O'Gorman et al., 2016) face similar challenges primarily due to the intensive labor required for annotation. Annotators are tasked with meticulously comparing event mentions across various documents to determine coreference relations, a process that often results in smaller dataset sizes, shorter articles, limited trigger word ambiguity and diversity. Additionally, the sparsity of co-referring events in texts frequently leads to the necessity of restricting annotations to certain topics and event types. To address these scalability challenges, semi-automatic or automatic methods have been adopted for CDEC annotation. For instance, the GVC database (Pavlick et al., 2016) employs a structured, semi-automatic approach for marking event references, though it is exclusively focused on gun violence events. Similarly, HyperCoref (Bugert and Gurevych, 2021) and WEC-Eng (Eirew et al., 2021) have leveraged Wikipedia hyperlinks for automated data generation. However, this method might suffer from inconsistencies due to relying on the hyperlinks and generally focus on events that are significant enough for Wikipedia entries, and often overlook a range of descriptive or anecdotal events that are prevalent in news reports. These constraints underscore the difficulties in producing extensive and authentic CDEC datasets that accurately reflect the complexity and diversity of real-world events. In this work, we utilize GPT-4 to generate authentic text and coreference labels together at scale and requiring minimal human efforts.

3 Method

3.1 Motivation

The analysis of error distributions in latest CDEC models (Caciularu et al., 2021b; Yu et al., 2022), reveals that these systems still struggle with several types of errors. Common issues include the lack of direct evidence, events that are identical but share minimal contextual overlap, contextually similar but distinct events, and insufficient argument matches, among others. We hypothesize that enriching the model's training data with a broader array of examples that specifically address these challenges will lead to enhanced performance. Our strategy focuses on addressing the three false negative error types listed in Table 1 by employing DM techniques to directly tackle and mitigate these identified issues.

The errors observed in rows 1 and 2 can be attributed to the fact that idiom trigger words or abstract trigger words are less seen in training set. To mitigate this, we leverage the capabilities of GPT-4 to produce paraphrased sentences based on the source sentence, incorporating synonyms of the trigger words with nuanced differences in granularity, concreteness, or perspective. The usage of GPT-4 ensures the generation of coherent and grammatically sound sentences, rendering the artificially generated data more lifelike compared to mere synonym substitutions.

In cases of errors in row 3, Sentence 1 is in active voice, with "Sudan" as the subject performing the action of *Bombs*. Sentence 2 is in passive voice. The subject, "a refugee camp", is the recipient of the action *bombed*, and the doer of the action is mentioned later in the sentence. Missing the coreference link between the two *bombing* events could stem from the under-representation of specific argument structures or word orders in the training data. To address this type of error, we utilize GPT-4 to generate sentences covering a spectrum of different word orders that reflect changes in voice, emphasis, or stylistic nuances within the sentences.

3.2 Prompting for Data Multiplication

GPT-4 demonstrates a remarkable ability to generate text that is both realistic and coherent, especially when guided by specific constraints. Leveraging this capability, we utilize GPT-4 to create two types of synthetic data to augment the diversity of our source datasets: one through variations in trigger words and the other by modifying word orders. The text produced by GPT-4 closely resembles the style, structure, and patterns found in actual data, thereby effectively creating scenarios that were underrepresented in the original datasets. To guarantee that this generated data aligns precisely with the unique demands of our task, we carefully craft the prompts guiding GPT-4's text generation.

Trigger Variation Example

Instruction
Paraphrase the sentence in three ways by changing the marked words. Keep the original brackets
Query
Microsoft has just <released> a new security update for Internet Explorer in order to <patch> a flaw <affecting> all versions of the company - s in - house browser , including IE11 on Windows 8 . 1 RTM . While Microsoft <claims> that it has received reports of <attacks> aimed at Internet Explorer 8 and 9 , it appears that the security flaw <affects> all the other versions of the browser.
Output
Microsoft has recently <launched> a new security enhancement for Internet Explorer to <address> a vulnerability <impacting> every iteration of the company's proprietary browser, including IE11 on Windows 8.1 RTM. Although Microsoft <states> they've gotten feedback about <breaches> on Internet Explorer 8 and 9, it seems the security issue <influences> the rest of the browser versions.

Word Order Variation Example

Instruction
Paraphrase the sentence in three ways by changing the word order. Don't change the marked words and keep the original brackets
Query
With Web <attacks> <increasing> , Microsoft <fixes> IE bug
Output
The IE bug gets <fixes> by Microsoft due to the <increasing> Web <attacks>.

Figure 1: Examples of GPT-4 prompt and output for generating trigger variation dataset and word order variation dataset.

Trigger Variation Prompt We utilize GPT-4 to paraphrase sentences from the ECB+ and Aylien Covid training datasets, specifically focusing on creating sentences with distinct trigger words. Figure 1 illustrates the process we apply to all the training data. The clusters for these generated events can be easily identified, as they belong to the same clusters as their corresponding source event mentions in the original dataset. In the example provided, the event mention represented by trigger word *launched* in the generated sentence is associated with the same cluster as "Microsoft *released* security update" in the original sentence. This approach ensures that paraphrased sentences effectively retain the coreference links of the source

ERROR TYPE	EXAMPLES IN TEST SET	EXPLANATION
idiomatic trigger	1. Advanced Micro Devices will acquire server vendor SeaMicro in an attempt to <make a run> at Intel in the microserver market. 2.AMD Buys SeaMicro, <Enters> Server Hardware Business.	fail to identify the coref link with a less seen trigger <make a run> in training data
abstract trigger	1. In a <move> that will expand its services division, Hewlett - Packard will acquire EYP Mission Critical Facilities, a New York company that offers data center consulting services. 2. Hewlett - Packard to <buy> consulting firm EYP Mission Critical Facilities news	fail to identify the coref link between a concept-instance relationship: <move> is a abstract term of corporate actions, whereas <buy> focuses specifically on the act of acquisition, which is an instance of <move>.
voice change	1. Breaking News: Sudan <Bombs> Yida Refugee Camp in South Sudan 2. A refugee camp in South Sudan's Unity state was <bombed> on Thursday, South Sudan officials and witnesses said	fail to identify the coref link between two <bombing> events with underrepresented argument structure.

Table 1: Persisting errors in SOTA models we want to address with Data Multiplication.

event mentions, albeit with generated different trigger words.

Word Order Variation Prompt Similarly, We utilize GPT-4 to create varied arguments order for all sentences in both training datasets. Figure 1 demonstrates this paraphrasing process involving altering the word order but carefully maintaining the original trigger words. Identifying the clusters for the generated events is also straightforward, as they are part of the same clusters as their respective source event mentions in the original dataset. The majority of these paraphrased sentences successfully preserve the original event mentions, represented by the same trigger words. For example, in the source sentence, "Microsoft" is the active subject addressing the "IE bug". Conversely, in the generated sentence, "The IE bug" becomes the subject, with the sentence now adopting a passive voice. This structure implies that the IE bug is the entity being affected, with Microsoft executing the action. Refer to Appendix A.2 for additional discussions regarding the impact of word order alterations on argument structure change.

4 Datasets

We show the utility of DM by applying it on ECB+ and Aylien Covid (hereon, *Covid*), two exisiting CEDC datasets. For the construction of the augmentation data with variated triggers and word order, we apply GPT-4 with different prompts to paraphrase the training sentences from the datasets.

To create training pairs, we pair up all the event mentions under the same topic. Figure 2 illustrates how the event mentions are paired across the original and DM datasets. DM_{TV} and DM_{WOV} only consists of pairs from the DM datasets (purple lines in Figure 2). In order to use the DM datasets to do continued training on models trained on original pairs, we also create $ECB+_{TV}$ and $ECB+_{WOV}$ datasets to combine all the pairs. $ECB+_{TV}$ and $ECB+_{WOV}$ consists of pairs from the DM datasets (purple lines), and pairs between mentions in original ECB+ and DM datasets (blue lines). The Covid datasets pairs are created the same way.

	Train pairs	Dev pairs	Test pairs
ECB+	185,493	56,534	93,878
Aylien Covid	18,867	2,358	2,358

Table 2: Data splits statistics of original ECB+ and Aylien Covid Datasets

	ECB+	DM_{TV}	DM_{WOV}	$ECB+_{TV}$	$ECB+_{WOV}$
Mentions	3,808	2,674	2,561	6,482	6,369
Pairs	185,493	96,756	89,144	360,838	341,703
	Covid	DM_{TV}	DM_{WOV}	$Covid_{TV}$	$Covid_{WOV}$
Mentions	560	552	523	1,112	1,083
Pairs	18,867	18,004	17,962	44,457	43,859

Table 3: Augumented training sets of ECB+ and Covid through different DM prompts. $_{TV}$ =Trigger Variation, $_{WOV}$ =Word Order Variation.

4.1 DM for ECB+

ECB+ comprises 45 unique topics, each split into two similar yet distinct subtopics. For generating original training pairs, we pair all the events within

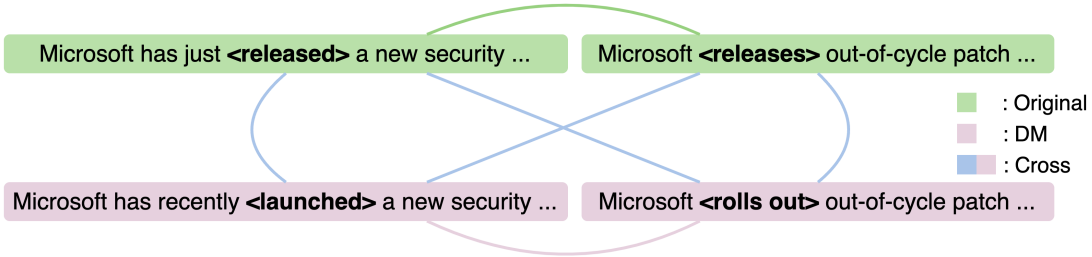

Figure 2: Generate pairs among original and DM datasets. The sentence pair in green boxes are from original dataset. The sentence pair in purple boxes are in GPT-4 generated DM datasets. Green lines are coreference links of original datasets. Purple lines are coreference links of DM datasets. Purple lines + blue lines are coreference links of ECB+$_{\text{TV}}$, ECB+$_{\text{WOV}}$, Covid$_{\text{TV}}$ or Covid$_{\text{WOV}}$ datasets.

each of the 25 topics in training. Following the data split proposed by Cybulska and Vossen (2014), we use the ECB+ training set, as shown in table 2, as our source dataset to create two DM datasets: Trigger Variation (TV) dataset (DM$_{\text{TV}}$) and Word Order Variation (WOV) dataset (DM$_{\text{WOV}}$), as detailed in Table 3. These methods augment the dataset with additional sentences and new event mentions derived from the original sentences and mentions. The new mentions are added in original clusters, thus the number of clusters stays the same.

During paraphrasing, some event trigger words might be altered or omitted. For instance, "nominated" is rephrased as "candidate", which is much less eventive, leading to the loss of the original event mention "nominating" in the generated sentence. As a result, our DM$_{\text{TV}}$ dataset gains an additional 2,674 event mentions and 96,756 pairs. Similarly, some trigger words were lost during word order paraphrasing, such as "being arrested" changing to "in jail". Consequently, we add 2,561 more event mentions and 89,144 training pairs to the DM$_{\text{WOV}}$ dataset .

4.2 DM for Aylien Covid Dataset

Our evaluation of the DM method also employs the Aylien Covid dataset (Zhao et al., 2023), which comprises 4,000 coreference pairs preselected by CDLM event coreference model (Caciularu et al., 2021b) across 10 topics that are covid related and later manually annotated for CDEC. This dataset contains coreference pairs with positive (coreference) ratio of 47.5%, markedly higher than the 8.1% ratio observed in ECB+, and too evenly balanced for realistic scenarios where the coreference pairs are sparse. We add additional negative samples from lower-ranked pairs from the same topics to align the positive ratio more closely with ECB+. Table 2 shows the statistics of our modified Covid Dataset. The Covid dataset annotates six event relations. We categorize Identity, Concept-instance, and Whole-Subevent relations into coreference relation, while Not-Related, Cannot-Decide, and Set-Member relations were grouped into non-coreference relation.

Following the original Covid data split of 8:1:1, we utilize the training set as the source dataset to generate two DM datasets. Details are in Table 3. Like the ECB+ paraphrasing, some event trigger words are omitted. Consequently, we have 552 additional event mentions and 18,004 additional pairs in DM$_{\text{TV}}$, and an increase of 523 event mentions and 17,962 event pairs in DM$_{\text{WOV}}$.

5 Experiments

To evaluate the effectiveness of our DM methods, we conduct the experiments on the CDEC task with ECB+ and Covid datasets. Following recent approaches for CDEC (Caciularu et al., 2021a; Yu et al., 2022), we apply a cross-encoder based pairwise scorer on each event mention pair, followed by the agglomerative clustering to form the coreference clusters. We use pairwise F1 score and four common coreference resolution metrics to evaluate the model's performance: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF$_e$ (Luo, 2005), and the cumulative CoNLL F1 score.

5.1 CDEC on ECB+

Experimental Settings We use PAIRWISERL, the end-to-end CDEC system described in Yu et al. (2022) as our baseline model.[2] PAIRWISERL

[2]The CDEC models from Caciularu et al. (2021a) and Held et al. (2021) had a better performance than PAIRWISERL on the ECB+ (85.6 and 85.7 CoNLL F1 respectively). However, Caciularu et al. (2021a) utilized the whole document as the event context to train the pairwise scorer, and Held et al. (2021) proposed a two-step method that optimized the clustering with gold event mentions, both of which are not

Model	MUC			B^3			$CEAF_e$			CoNLL	Pairwise
	R	P	F1	R	P	F1	R	P	F1	F1	F1
PAIRWISERL (Yu et al., 2022)	91.6	83.1	**87.2**	89.4	81.1	85.1	70.5	85.5	79.9	84.0	N/A
PAIRWISERL (reproduced)	84.2	84.6	84.4	84.1	84.9	84.5	79.7	79.2	79.5	82.8	75.66
PAIRWISERL + $ECB+_{TV}$	89.4	84.5	86.9	88.3	83.4	**85.8**	78.3	84.2	**81.1**	**84.6**	77.69
PAIRWISERL + $ECB+_{WOV}$	88.8	84.6	86.6	85.9	84.3	85.1	78.4	83.5	80.9	84.2	77.09
PAIRWISERL + both	89.8	83.4	86.5	88.7	82.6	85.5	76.1	83.9	79.9	84.0	**77.73**
$RoBERTa_{LARGE}$ + DM_{TV}	89.2	82.4	85.7	88.3	80.7	84.3	74.7	83.0	78.6	82.9	72.72
$RoBERTa_{LARGE}$ + DM_{WOV}	87.4	81.6	84.4	86.1	80.6	83.3	73.9	80.9	77.3	81.7	72.11

Table 4: CDEC performance of unstructured pairwise model on ECB+. Top section shows the baseline results from models trained on original ECB+ dataset. Middle section shows the results from continued training on the reproduced baseline model with DM data. Bottom section shows the results from training on the derived DM datasets only.

Model	MUC			B^3			$CEAF_e$			CoNLL	Pairwise
	R	P	F1	R	P	F1	R	P	F1	F1	F1
$PAIRWISERL_{COVID}$	73.4	86.3	79.3	78.0	87.8	82.6	83.1	70.5	76.3	79.4	71.37
$PAIRWISERL_{COVID}$ + $Covid_{TV}$	85.5	83.7	**84.6**	84.5	83.0	**83.8**	77.9	80.1	**79.0**	**82.4**	**73.79**
$PAIRWISERL_{COVID}$ + $Covid_{WOV}$	80.2	84.8	82.4	82.6	84.6	83.6	80.3	75.4	77.8	81.3	72.36
$PAIRWISERL_{COVID}$ + both	83.9	83.8	83.5	83.3	83.5	83.4	78.0	78.1	78.0	81.7	73.61
$RoBERTa_{LARGE}$ + DM_{TV}	79.6	84.6	82.0	82.2	85.1	83.6	80.4	75.0	77.6	81.1	70.89
$RoBERTa_{LARGE}$ + DM_{WOV}	73.5	85.7	79.1	77.9	87.3	82.3	82.6	70.5	76.1	79.2	70.65

Table 5: CDEC performance of unstructured pairwise model on Aylien Covid Dataset. Top section displays baseline results from the model trained on original Covid dataset. Middle section displays evaluation results of continued training on baseline model with DM datasets. Bottom section displays evaluation results on models trained on only DM datasets.

learns local contexts by using transformer models to encode the concatenated sentences pair with marked event mentions. Trigger token representations is then aggregated into a unified feature vector for pairwise classification and event clustering. We use the unstructured version of PAIRWISERL with $RoBERTa_{LARGE}$ as the pair encoder. All the other model settings are the same with those reported in (Yu et al., 2022).[3]

For the experiments, we start by training the PAIRWISERL model from scratch with the original ECB+ data as our baseline results. Then we conduct a sequence of continued trainings on the PAIRWISERL with $ECB+_{TV}$, $ECB+_{WOV}$ or both. We also train the PAIRWISERL model solely on the DM_{TV} and DM_{WOV} datasets created from ECB+ to evaluate the quality and reliability of the GPT-4 generated DM data.

Results Table 4 outlines the evaluation results on the ECB+ and derived DM datasets. We show PAIRWISERL baseline results from both our reproduced model and reported model in Yu et al. (2022) (82.8 and 84.0 CoNLL F1 respectively). Reproduction details and performance difference is discussed in the Appendix A.1. All the continued trainings are applied on the produced PAIRWISERL model for consistency.

Overall, our DM methods lead to notable performance enhancements, with improvements ranging from 1.4 to 1.8 in CoNLL F1 over the reproduced PAIRWISERL, and gains of 0.2 to 0.6 over the reported PAIRWISERL. Comparing to the continued training with $ECB+_{WOV}$, $ECB+_{TV}$ has a better performance, indicating that paraphrasing event triggers are more effective than swapping word order.

Combining both $ECB+_{WOV}$ and $ECB+_{TV}$ does not show better performance than each individual dataset, suggesting that too much data variation and context repetition might bring in more noise to the event clustering. However it achieves the

directly comparable to our work under the sentence-level DM methods.

[3]Structured PAIRWISERL used additional annotation of event arguments as model input.

highest pairwise F1 score, showing its capability to make the best binary coreference/non-coreference decisions through the training from the additional DM pairs.

5.2 CDEC on Covid

Experimental Settings Similar to the previous experiment, we use the unstructured PAIRWISERL as the CDEC model. Since there is no previous results reported on the Covid dataset, we train the PAIRWISERL model with the original Covid data as the baseline. We then continue the training of the model with Covid_{TV}, $\text{Covid}_{\text{WOV}}$ and both.

Results Table 5 shows the CDEC results on the Covid and its derived DM datasets. $\text{PAIRWISERL}_{\text{COVID}}$ is the baseline result from training on the original Covid dataset. Similar to the previous experiment, our DM methods improve the baseline results through continued training on the additional data. Comparing to the ECB+, $\text{PAIRWISERL}_{\text{COVID}}$ + Covid_{TV} shows the most significant improvement across all evaluation metrics, boosting the CoNLL F1 by 3 and the pairwise F1 by 2.42. This may be due to the much smaller size of the Covid dataset, and additional DM data has bigger positive impact on the results.

5.3 CDEC on DM Datasets

As shown in Table 4 and 5, for both ECB+ and Covid, training solely with GPT-generated DM datasets (DM_{TV} and DM_{WOV}) yields comparable or even better results than the baseline trained on the original gold dataset (+0.1 CoNLL F1 on ECB+ and +1.7 on Covid). Changing the triggers and the word order do change the style, focus or emphasis of the sentence, but it usually does not change the transitive properties of the events, indicating the high quality and reliability of the DM datasets and potential their potential usage as "near gold" CDEC datasets in the future.

We further evaluate the quality of the DM data by sampling 50 paraphrased sentences from each of the four DM datasets. Out of a total of 200 samples, we only find 12 noise occurrences: 8 cases of unnatural sounding sentences such as "He was <found> <dead>" is paraphrased into "He was <uncovered> <expired>...", *uncovered expired* is not a natural English sentence and does not form a coherent meaning with *found dead*. 4 cases of changing trigger causes event change, such as "John Jenkins, <charged> with murder..." is paraphrased into "John Jenkins, <arraigned> for ...". To *arraign* somebody is a legal process that occurs after a person has been *charged*. They are falsely considered as coreferential. Overall, the consistent improvements achieved by the two augmentation approaches across these DM datasets underscore their effectiveness and potential applicability in diverse settings.

6 Analysis

To delve deeper into the underlying factors contributing to the performance enhancements observed with the DM datasets, we conduct an error analysis on the models trained with additional TV data, which are proven to be the most effective in our experiments.

Increased Trigger Distribution In the original ECB+ training set, the average count of unique trigger lemma types per cluster was 1.36, which increased to 3.30 upon incorporating the trigger variation dataset. A comparative error analysis between the reproduced model and the PAIRWISERL + ECB+_{TV} on the ECB+ test set reveals that the latter successfully corrected 883 error pairs involving 166 different trigger word lemma types initially made by the baseline model. Notably, 56 trigger lemma types appear in a greater number of clusters in the combination of ECB+ and ECB+_{TV} comparing to that in ECB+, and 15 trigger lemma types are exclusively found in the combination of ECB+ and ECB+_{TV} data. In Covid training set, the average count of unique trigger lemma types rise from 2.40 to 3.40. The the $\text{PAIRWISERL}_{\text{COVID}}$ + Covid_{TV} model rectified 45 errors involving 15 lemmas, 13 of which are present in more clusters in the combination of covid and Covid_{TV} than in the original data. Through the trigger variation method, the models learn to better distinguish the triggers through being exposed to more diversified trigger distribution over clusters.

Increased Trigger Appearance For example, trigger words with lemma *disclose* only appears in cluster "Lohan undisclosed why in rehab" in original data. In DM data, *disclose* also appears in clusters "Isna agency reports earthquake", "hollywood reports movie", "publicist statement on Reid", "speaker confirms death". PAIRWISERL + ECB+_{TV} rectifies 11 errors related to trigger word *disclose* without introducing any new error.

Confusion from Unseen Triggers While the DM_{TV} is able to help the models learn better the trigger distribution, it can introduce new errors through coreference pairs from unseen triggers in the original training set. For example, trigger *acquire* is absent in any clusters in the original training set, but DM introduces *acquire* in the cluster related to "gaining control of Windows machines". In the test set, it helps the model correct 36 error pairs with different trigger words as shown in Example 1. However it also generate 19 new errors. Example 2 show a new error pair that is predominantly due to falsely established coreference links between *acquisition* and nominal trigger words like *deal*, *it*, and *offer* that are paraphrased events of *deal* in the DM data.

(1) **Sentence 1**: If the deal is completed , it would be HP 's biggest <acquisition> since it bought Compaq Computer Corp. for $ 19 billion in 2002.
Sentence 2: Hewlett-Packard is negotiating to <buy> technology services provider Electronic Data Systems in a deal that could help the world 's largest personal computer maker snap up more data management and consulting contracts.
Gold / Prediction: Coref. / Coref.

(2) **Sentence 1**: Hewlett - Packard Engineers <Deal> for EYP.
Sentence 2: Extending its reach into the ripening green - consulting space, HP today announced the <acquisition> of EYP Mission Critical Facilities, a consulting company specializing in strategic technology planning, design and operations support for large-scale datacenters.
Gold / Prediction: Non-Coref. / Coref.

Ambiguous Triggers Highly ambiguous triggers remain challenging to models trained with DM data. In reviewing triggers in both the corrected errors and newly introduced errors, no model demonstrates an enhanced capability in resolving errors associated with ambiguous triggers. For example, the false positive errors involving trigger lemma *Indonesia earthquake* in ECB+ and false negative *assets freeze* in Covid dataset still persist.

In the ECB+ testing data, the term *earthquakes* appears only in two clusters, however, these pairs are especially hard because they refer to the "6.1 Indonesia earthquakes in 2009" and "6.1 Indonesia earthquake in 2013". It presents a significant disambiguation challenge even only between two clusters. If the time information is not in local context, it is difficult to tell which year's earthquake struck Indonesia or caused house damage. In the following example 3, the context similarity is high, but the pair is non-coreferential. $ECB+_{TV}$ enriches trigger *earthquake* to *temblor*, *quake* and *shake*, but it does not help in this case.

(3) **Sentence 1**: Strong <earthquake> hits Indonesia's Aceh province.
Sentence 2: A powerful <6.1-magnitude earthquake> struck the Indonesian province of Aceh with no tsunami warning issued.
Gold / Prediction: Non-Coref. / Coref.

In contrast, the Covid dataset features the same trigger word across numerous clusters, distributed more uniformly. For instance, the trigger word *freeze* is associated with a variety of different money freezing events: "Isabella's bank accounts freeze", "couple's company holding freeze", "Isabella asset freeze", "board director bank account freeze", "Isabella brother asset freeze". In example 4, two sentences are focus on the different aspects related to the same asset *freeze* event. Little overlap in context still poses great ambiguities for the $Covid_{TV}$ model to correctly establish the coreference link.

(4) **Sentence 1**: The asset <freeze> follows an injunction application by the government, which is seeking to recover around $1 billion of funds that it says it is owed by Isabel dos Santos and her associates.
Sentence 2: Angolan court orders <seizure> of ex-president's daughter's assets in graft probe.
Gold / Prediction: Coref. / Non-Coref.

7 Conclusion

The development of CDEC datasets is challenging due to the inherent complexity and demanding labor requirements of the process, often resulting in datasets that are constrained in both size and scope. We this data these bottleneck by effectively utilizing the advanced capabilities of GPT-4 to enhance existing CDEC datasets. By introducing two innovative DM techniques, we successfully generate "near gold" synthetic data that is both realistic and of high quality, significantly enriching the volume and diversity of annotations within CDEC datasets. This method offers a substantial advantage in terms of swift data generation, eliminating the need for extensive manual collection or detailed curation. Evaluations on the CDEC models training on the combination of our DM ECB+ and Aylien Covid datasets show improvement over baseline model and demonstrate the effectiveness of our DM methods, underscoring the potential of our approach in enhancing CDEC datasets. In future work, we would like to evaluate our methods on more datasets, languages and model architectures, and extend our methods to a broader range of NLP tasks beyond CDEC.

8 Limitations

We propose the data multiplication methods with LLMs to enrich the existing CDEC datasets. Given the enrichment is sentence-level and only applied to the source sentence with event mentions, one limitation is the new enriched datasets cannot be intuitively used to train the state-of-the-art model like CDLM (Caciularu et al., 2021a) that takes the whole documents as model input. Future work includes the accommodation of the DM datasets to other CDEC model structures.

9 Ethical issues

There are potential copyright issues when generating new points based on existing data points, and we selected data sets governed by the Creative Commons Attribution 3.0 Unported License. This license explicitly permits the redistribution and modification of the data, thereby providing a legal and ethical foundation for our work. In creating our generated dataset, we are committed to adhering to the terms of this license, which includes proper attribution and ensuring that any modifications or derivative works are also shared under the same or compatible terms.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

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.

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.

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.

Arijit Ghosh Chowdhury and Aman Chadha. 2023. Generative data augmentation using llms improves distributional robustness in question answering. *arXiv preprint arXiv:2309.06358*.

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).

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.

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*. 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.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

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

Dheeraj Mekala, Tu Vu, Timo Schick, and Jingbo Shang. 2022. Leveraging qa datasets to improve generative data augmentation.

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.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. 2016. The gun violence database: A new task and data set for NLP. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1024, Austin, Texas. Association for Computational Linguistics.

Gaurav Sahu, Pau Rodriguez, Issam H. Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models.

Dan Su, Mostofa Patwary, Shrimai Prabhumoye, Peng Xu, Ryan Prenger, Mohammad Shoeybi, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2023. Context generation improves open domain question answering.

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*, pages 39–49, Nancy, France. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

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.

Setting	Pairwise F1 (Dev)	Pairwise F1 (Test)	CoNLL F1 (Test)
Titan XP, 2 GPUs, bs=8	80.63	75.66	82.8
V100, 2 GPUs, bs=32	80.09	74.11	81.8
V100, 1 GPU, bs=32	79.35	76.01	82.7

Table 6: Different reproduced unstructured PAIRWISERL models (Yu et al., 2022). bs = batch size.

A Appendix

A.1 Model Reproduction Attempts

In table 6, We display the evaluation on different PAIRWISERL models we trained with the same hyperparameters except the training platform, number of gpus and batch size. We choose the best performing model for our continued training in the paper.

A.2 Word Order Variation Types

In DM_{WOV} dataset, GPT-4 creates a more varied collection of sentence structures by manipulating word order as shown in Example 5:

(5) **Voice change**:
The software patch has <closed> a loophole ... / A loophole ... has been <closed> by a software patch
Predicate nominalization:
After three submarine cables were <damaged>... / The <damaged> state of three submarine cables ...
Fronting:
Klitschko <stopped> Thompson in the sixth round to retain his title belts .../ To retain his title belts, Klitschko <stopped> Thompson ...
Causative construction:
Several <die> in south Iran quake / Quake in south Iran results in several <deaths>

The first sentence pair involves the voice change from active to passive. In the original sentence, the agent *patch* precedes the predicate *close*, and the recipient *loophole* follows it. The DM_{WOV} reverses this argument order.

The second pair shows the transformation from predicate adjective to nominal adjective constructions. Originally, *damaged* functions as a predicate adjective following the linking verb *are* and describes *cables*, then it is tranformed to be an attributive adjective, modifying *cables* directly by preceding it, thereby reversing the order of the patient *cables* and the predicate *damaged*.

The third pair makes syntactic rearrangement by fronting the goal of *stopped* being "to retain his title belts", altering the emphasis focus of the goal argument.

The introduction of "results in" in the four pair adds a causal relationship between the earthquake and the event *deaths*.

By introducing such diverse sentence structures, the likelihood of underrepresented argument orders in each cluster is reduced, thereby enhancing the robustness of the model.