$2 * n$ is better than n^2: Decomposing Event Coreference Resolution into Two Tractable Problems

Shafiuddin Rehan Ahmed[1] **Abhijnan Nath**[2] **James H. Martin**[1] **Nikhil Krishnaswamy**[2]

[1]Department of Computer Science, University of Colorado, Boulder, CO, USA
{shah7567, james.martin}@colorado.edu

[2]Department of Computer Science, Colorado State University, Fort Collins, CO, USA
{abhijnan.nath, nkrishna}@colostate.edu

Abstract

Event Coreference Resolution (ECR) is the task of linking mentions of the same event either within or across documents. Most mention pairs are not coreferent, yet many that are coreferent can be identified through simple techniques such as lemma matching of the event triggers or the sentences in which they appear. Existing methods for training coreference systems sample from a largely skewed distribution, making it difficult for the algorithm to learn coreference beyond surface matching. Additionally, these methods are intractable because of the quadratic operations needed. To address these challenges, we break the problem of ECR into two parts: a) a heuristic to efficiently filter out a large number of non-coreferent pairs, and b) a training approach on a balanced set of coreferent and non-coreferent mention pairs. By following this approach, we show that we get comparable results to the state of the art on two popular ECR datasets while significantly reducing compute requirements. We also analyze the mention pairs that are "hard" to accurately classify as coreferent or non-coreferent[1].

1 Introduction

Event coreference resolution (ECR) is the task of finding mentions of the same event within the same document (known as “within-document coreference resolution,” or *WDCR*) or across text (known as “cross-document coreference resolution,” or *CDCR*) documents. This task is used for knowledge graph construction, event salience detection and question answering (Postma et al., 2018).

Traditionally, ECR is performed on pairs of event mentions by calculating the similarity between them and subsequently using a clustering algorithm to identify ECR relations through transitivity. The pairwise similarity is estimated using a supervised machine learning method, where an algorithm is trained to distinguish between positive and negative examples based on ground truth. The positive examples are all pairs of coreferent mentions, while the negative examples are all pairs of non-coreferent mentions. To avoid comparing completely unrelated events, the negative pairs are only selected from documents coming from the set of related topics.

Many coreferent pairs are similar on the surface, meaning that the event triggers (the words or phrases referring to the event) have the same lemma and appear in similar sentences. We can use these features in a heuristic to further classify the positive (P^{+}) and negative (P^{-}) pairs into four categories:

1. $\text{P}^{+}_{\texttt{easy}}$: coreferent/positive mention pairs with high surface similarity.
2. $\text{P}^{+}_{\texttt{FN}}$: coreferent/positive mention pairs with low surface similarity.
3. $\text{P}^{-}_{\texttt{hard}}$: non-coreferent/negative mention pairs with high surface similarity.
4. $\text{P}^{-}_{\texttt{TN}}$: non-coreferent/negative mention pairs with low surface similarity

As shown in Figure 1, $\text{P}^{+}_{\texttt{easy}}$ represents coreferent mention pairs that can be correctly identified by the heuristic, but $\text{P}^{-}_{\texttt{hard}}$ are non-coreferent pairs that might be difficult for the heuristic to identify. Similarly, $\text{P}^{-}_{\texttt{TN}}$ (True Negatives) are non-coreferent pairs that the heuristic can correctly infer, but $\text{P}^{+}_{\texttt{FN}}$ (False Negatives) require additional reasoning (that *Indianapolis* is coreferent with *Colts*) to make the coreference judgement.

Most mention pairs are non-coreferent, comprising all pairs corresponding to $\text{P}^{-}_{\texttt{hard}}$ and $\text{P}^{-}_{\texttt{TN}}$. However, we observe that that the distribution of the three categories ($\text{P}^{+}_{\texttt{easy}}$, $\text{P}^{-}_{\texttt{hard}}$, and $\text{P}^{+}_{\texttt{FN}}$) is fairly similar across most ECR datasets, with $\text{P}^{-}_{\texttt{TN}}$ causing the imbalance between positive and negative pairs. Previous methods do not differentiate between these four categories and randomly select

[1] code repo: `github.com/ahmeshaf/lemma_ce_coref`

Findings of the Association for Computational Linguistics: ACL 2023, pages 1569–1583
July 9-14, 2023 ©2023 Association for Computational Linguistics

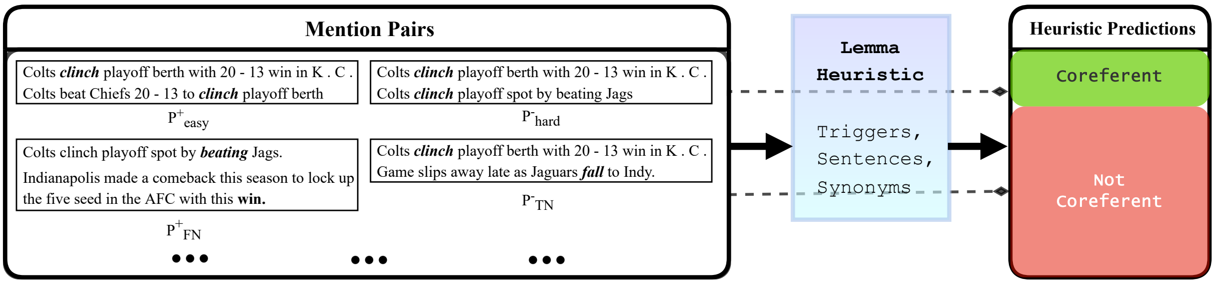
Mention Pairs
Colts clinch playoff berth with 20 - 13 win in K . C .
Colts beat Chiefs 20 - 13 to clinch playoff berth
P+easy
Colts clinch playoff berth with 20 - 13 win in K . C .
Colts clinch playoff spot by beating Jags
P-hard
Colts clinch playoff spot by beating Jags.
Indianapolis made a comeback this season to lock up the five seed in the AFC with this win.
P+FN
Colts clinch playoff berth with 20 - 13 win in K . C .
Game slips away late as Jaguars fall to Indy.
P-TN
Lemma Heuristic
Triggers, Sentences, Synonyms
Heuristic Predictions
Coreferent
Not Coreferent

Figure 1: In this approach, we use a lemma-based heuristic to identify coreference, or the relationship between two mentions in a text that refer to the same event. We compare the similarity between the event trigger, which is highlighted in bold and italic, and the lemmas, or base forms, of the sentences. The heuristic classifies the mention pairs "$P^{+}_{\mathtt{easy}}$" and "$P^{-}_{\mathtt{hard}}$" as coreferent, and "$P^{+}_{\mathtt{FN}}$" and "$P^{-}_{\mathtt{TN}}$" as not coreferent. "$P^{+}_{\mathtt{easy}}$" and "$P^{-}_{\mathtt{TN}}$" are correct predictions, meaning they are classified correctly as coreferent and not coreferent. "$P^{-}_{\mathtt{hard}}$" and "$P^{+}_{\mathtt{FN}}$" are incorrect predictions, meaning they are misclassified as coreferent and not coreferent.

the positive and negative pairs to train their coreference systems from this heavily skewed distribution. This makes it challenging for the coreference algorithm to identify coreferent links among a large number of non-coreferent ones. Furthermore, as ECR is performed on n^2 number of mention pairs, where n is the number of mentions in the corpus, these methods can become intractable for a large corpus.

To improve the efficiency of the ECR process while achieving near sate of the art (SOTA) results, we divide the problem into two manageable subtasks: a) a heuristic to efficiently and accurately filter out a large number of $P^{-}_{\mathtt{TN}}$ as a way of balancing the skewed distribution, and b) an ECR system trained on the balanced set of coreferent and non-coreferent mention pairs ($P^{+}_{\mathtt{easy}}$ and $P^{-}_{\mathtt{hard}}$). This approach also eases the analysis of some of the mention pairs that are difficult to classify with an ECR system, which we present in this paper.

2 Related Work

Pre-Transformer Methods Pre-Transformer language model-related works in event coreference such as Kenyon-Dean et al. (2018) trained neural models with customized objective (loss) functions to generate richer representations of mention-pairs using "static" embeddings such as contextual Word2Vec (Mikolov et al., 2013) as well as document-level features such as TF-IDF and heuristically-motivated features like mention-recency, word overlap, and lemma overlap, etc. As such, they improved upon the baselines established by Cybulska and Vossen (2015) on the ECB+ corpus. Similarly, works such as Barhom et al. (2019) suggest both disjoint and joint-clustering of events mentions with their related entity clusters by using a predicate-argument structure. In this, their disjoint model surpassed Kenyon-Dean et al. (2018) by 9.5 F1 points using the CoNLL scorer (Pradhan et al., 2014) whereas their joint model improved upon the disjoint model by 1.2 points for entities and 1 point for events.

Transformer-based Cross-encoding Most recent works (Meged et al., 2020; Zeng et al., 2020; Cattan et al., 2021; Allaway et al., 2021; Caciularu et al., 2021; Held et al., 2021; Yu et al., 2022a) in CDCR have shown success in using pairwise mention representation learning models, a method popularly known as cross-encoding. These methods use distributed and contextually-enriched "non-static" vector representations of mentions from large, Transformer-based language models like various BERT-variants to calculate supervised pairwise scores for those event mentions. At inference, such works use variations of incremental or agglomerative clustering techniques to form predicted coreference links and evaluate their chains on gold coreference standards. The methods vary with the context they use for cross-encoding. Cattan et al. (2021) use only sentence-level context, Held et al. (2021) use context from sentences surrounding the mentions, and Caciularu et al. (2021) use context from entire documents.

In our research, we have focused on the CDLM model from Caciularu et al. (2021) and their methodology, which uses a combination of enhanced pretraining using the global attention mechanism inspired by Beltagy et al. (2020) as well as finetuning on a task-specific dataset using pretrained special tokens to generate more semantically-enhanced embeddings for mentions.

Beltagy et al. (2020) and Caciularu et al. (2021) cleverly use the global attention mechanism to linearly scale the oft-quadratic complexity of pairwise scoring of mentions in coreference resolution while also accommodating longer documents (up to 4,096 tokens). Previous works such as Baldwin (1997), Stoyanov and Eisner (2012), Lee et al. (2012), and Lee et al. (2013) also reduce computation time by strategically using deterministic, rule-based systems along with neural architectures.

Recently, pruning $\mathrm{P}^{-}_{\mathtt{TN}}$ for ECR has been shown to be effective by Held et al. (2021). They create individual representations for mentions and use them in a bi-encoder method to retrieve potential coreferent candidates, which are later refined using a cross-encoder trained on hard negative examples. In contrast, our approach utilizes a computationally efficient pruning heuristic and trains the cross-encoder on a smaller dataset. We also conduct an error analysis on all hard examples that are misclassified by the cross-encoder, which is made feasible by the heuristic.

3 Datasets

We experiment with two popular ECR datasets distinguished by the effectiveness of a lemma heuristic on the dataset.

3.1 Event Coreference Bank Plus (ECB+)

The ECB+ corpus (Cybulska and Vossen, 2014) is a popular English corpus used to train and evaluate systems for event coreference resolution. It extends the Event Coref Bank corpus (ECB; Bejan and Harabagiu (2010)), with annotations from around 500 additional documents. The corpus includes annotations of text spans that represent events, as well as information about how those events are related through coreference. We divide the documents from topics 1 to 35 into the training and validation sets[2], and those from 36 to 45 into the test set, following the approach of Cybulska and Vossen (2015).

3.2 Gun Violence Corpus (GVC)

The Gun Violence Corpus (Vossen et al., 2018) is a recent English corpus exclusively focusing on event coreference resolution. It is intended to be a more challenging dataset than ECB+ which has a very strong lemma baseline (Cybulska and Vossen, 2014). It is a collection of texts surrounding a single topic (gun violence) and various sub-topics. Since it does not have coreference links across sub-topics, we only consider mention pairs within the sub-topics. We use the data split by Bugert et al. (2021). Table 1 contains the statistics for ECB+ and GVC corpora.

	ECB+			GVC		
	Train	Dev	Test	Train	Dev	Test
T/ST	25	8	10/20	1/170	1/37	1/34
D	594	196	206	358	78	74
M	3808	1245	1780	5313	977	1008
C	1464	409	805	991	228	194
S	1053	280	623	252	70	43

Table 1: ECB+ and GVC Corpus statistics for event mentions. T/ST = topics/sub-topics, D = documents, M = event mentions, C = clusters, S = singletons.

4 System Overview

There are two major components in our system: the heuristic and the discriminator (cross-encoder) trained on the output of the heuristic.

4.1 Lemma Heuristics (**`LH`**, **`LH`$_{\mathtt{Ora}}$**)

A key feature of ECR is its high baseline achieved by comparing the lemmas of mention triggers and sentences. To leverage this feature, we incorporate it as the first step in our coreference resolution system. We utilize spaCy[3] to extract the lemmas, a widely-used tool for this task. In addition to matching lemmas of triggers, we also create and utilize a set of synonymous[4] lemma pairs that commonly appear in coreferent mention pairs in our training set. This approach allows us to identify coreferent mention pairs that have different triggers and improve the overall recall. The heuristic, `LH`, only utilizes the synonymous lemma pairs from the training set. We also evaluate the performance of `LH`$_{\mathtt{Ora}}$, which uses synonymous lemma pairs from the entire dataset which means it uses the coreference information of the development and test sets to create synonymous lemma pairs.

For a mention pair (A, B), with triggers (t_A, t_B), head lemmas (l_A, l_B) and for a given synonymous lemma pair set ($\mathrm{Syn_P}$), we consider only lemma pairs that pass any of the following rules:

- $(l_A, l_B) \in \mathrm{Syn_P}$
- $l_A == l_B$
- t_B *contains* l_A

[2] Validation set includes documents from the topics 2, 5, 12, 18, 21, 34, and 35

[3] https://spacy.io/ model `en_core_web_md` v3.4

[4] The words need not be synonyms in strict definitions, but rather appear in coreference chains.

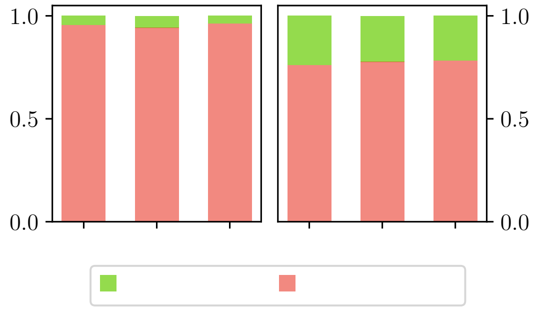

Figure 2: Coreferent vs. non-coreferent mention pairs ratio across datasets.

- *t_A contains l_B*

For mentions that have matching trigger lemmas/triggers or are synonymous, we proceed by comparing the context of the mentions. In this work, we only compare the mention's sentence to check for similarities between two mentions. To further refine our comparison, we remove stop words and convert the tokens in the text to their base form. Then, we determine the overlap between the two mentions and predict that the pair is coreferent if the overlap exceeds a certain threshold. We tune the threshold using the development sets.

4.1.1 Filtering out $P^{-}_{\mathtt{TN}}$

Cross-document coreference systems often struggle with a skewed distribution of mention pairs, as seen in Figure 2. In any dataset, only 5-10% of the pairs are corefering, while the remaining 90% are non-coreferent. To address this, we use the heuristic to balance the distribution by selectively removing non-coreferent pairs ($P^{-}_{\mathtt{TN}}$), while minimizing the loss of coreferent pairs ($P^{+}_{\mathtt{FN}}$). We do this by only considering the mention pairs that the heuristic predicts as coreferent, and discarding the non-coreferent ones.

4.1.2 $P^{-}_{\mathtt{hard}}$, $P^{+}_{\mathtt{easy}}$, and $P^{+}_{\mathtt{FN}}$ Analysis

$P^{+}_{\mathtt{easy}}$ and $P^{-}_{\mathtt{hard}}$: As defined earlier, $P^{+}_{\mathtt{easy}}$ are the mention pairs that the heuristic correctly predicts as coreferent when compared to the ground-truth, and $P^{-}_{\mathtt{hard}}$ are the heuristic's predictions of coreference that are incorrect when compared to the ground-truth. In §4.2.1, we go through how we fix heuristic's $P^{-}_{\mathtt{hard}}$ predictions while minimizing the errors introduced in terms of $P^{+}_{\mathtt{easy}}$.

$P^{+}_{\mathtt{FN}}$: We define a pair as a $P^{+}_{\mathtt{FN}}$ only if it cannot be linked to the true cluster through subsequent steps.

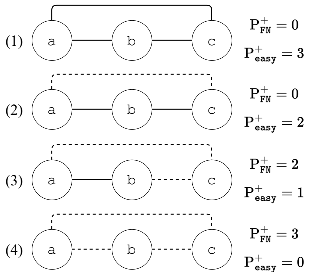
(1)
a
b
c
$P^{+}_{\mathtt{FN}} = 0$
$P^{+}_{\mathtt{easy}} = 3$
(2)
a
b
c
$P^{+}_{\mathtt{FN}} = 0$
$P^{+}_{\mathtt{easy}} = 2$
(3)
a
b
c
$P^{+}_{\mathtt{FN}} = 2$
$P^{+}_{\mathtt{easy}} = 1$
(4)
a
b
c
$P^{+}_{\mathtt{FN}} = 3$
$P^{+}_{\mathtt{easy}} = 0$

Figure 3: Counting size of mention pairs ($P^{+}_{\mathtt{FN}}$ and $P^{+}_{\mathtt{easy}}$) in a true cluster {a, b, c} using heuristic's coreferent predictions (solid line) and non-coreferent predictions (dotted line). We count $P^{+}_{\mathtt{FN}}$ after performing transitive closure, resulting in a size of 0 (instead of 1) in (2).

As shown in Figure 3, if a true cluster is {a, b, c} and the heuristic discards one pair (a, c), it will not be considered as a $P^{+}_{\mathtt{FN}}$ because the coreference can be inferred through transitivity. However, if it discards two pairs {(a,c), (b,c)}, they will both be considered as $P^{+}_{\mathtt{FN}}$. We hypothesize that an ideal heuristic is one that maintains a balance between $P^{+}_{\mathtt{easy}}$ and $P^{-}_{\mathtt{hard}}$ while minimizing $P^{+}_{\mathtt{FN}}$, and therefore, we tune the heuristic's threshold accordingly using the development sets of the corpora.

We evaluate the heuristics `LH` and $\mathtt{LH}_{\mathtt{Ora}}$ by plotting the distributions $P^{+}_{\mathtt{easy}}$, $P^{-}_{\mathtt{hard}}$, and $P^{+}_{\mathtt{FN}}$ generated by each for the two corpora. From Figure 4, We observe similar distributions for the test and development sets with the chosen threshold value from the development set. We also observe that `LH` causes a significant number of $P^{+}_{\mathtt{FN}}$, while $\mathtt{LH}_{\mathtt{Ora}}$ has a minimal number of $P^{+}_{\mathtt{FN}}$. Minimizing the count of $P^{+}_{\mathtt{FN}}$ is important as it directly affects

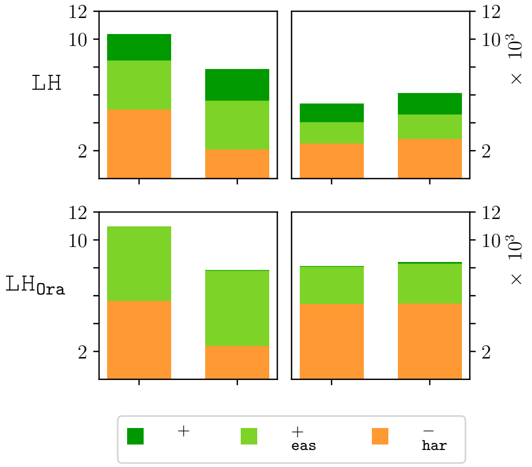

Figure 4: `LH` and $\mathtt{LH}_{\mathtt{Ora}}$ Distributions of $P^{-}_{\mathtt{hard}}$, $P^{+}_{\mathtt{easy}}$, and $P^{+}_{\mathtt{FN}}$ for ECB+ and GVC corpora. $\mathtt{LH}_{\mathtt{Ora}}$ ensures no (or negligible) loss in $P^{+}_{\mathtt{FN}}$.

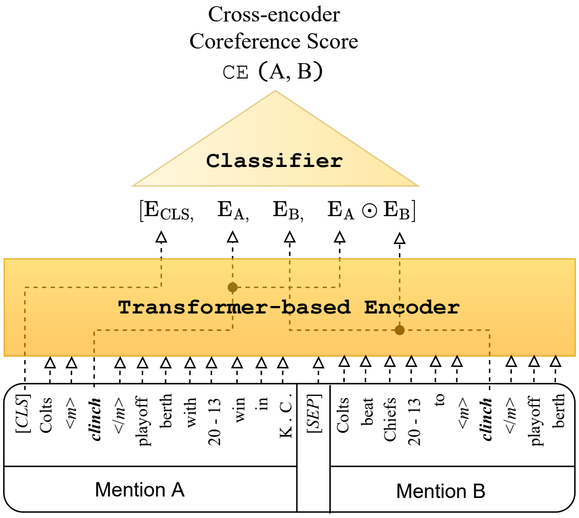

Figure 5: The cross-encoding technique to generate the coreference score between the mention pair (A, B). This involves adding special tokens, <*m*> and </*m*>, around the event triggers, and then combining and processing the two mentions through a transformer-based language model. Certain outputs of the transformer (E_{CLS}, E_{A}, E_{B}) are then concatenated and fed into a classifier, which produces a score between 0 and 1 indicating the degree of coreference between the two mentions.

the system's recall. The distributions of $\text{P}^{+}_{\texttt{easy}}$ and $\text{P}^{-}_{\texttt{hard}}$ remain balanced across all datasets except when $\texttt{LH}_{\texttt{Ora}}$ is used in GVC where there are double the number of $\text{P}^{-}_{\texttt{hard}}$ to $\text{P}^{+}_{\texttt{easy}}$. $\text{P}^{-}_{\texttt{hard}}$ should be minimized as it can affect the system's overall precision.

4.2 Cross-Encoder

A common technique to perform ECR is to use Transformer-based cross-encoding (CE) on the mention pair (A, B). This process, depicted in Figure 5, begins by surrounding the trigger with special tokens (<*m*> and </*m*>). The mentions are then combined into a single input for the transformer (e.g., RoBERTa). The pooled output of the transformer (E_{CLS}) and the output corresponding to the tokens of the event triggers (E_{A} and E_{B}) are extracted.[5] E_{CLS}, E_{A}, E_{B}, and the element-wise product of the mention embeddings ($\text{E}_{\text{A}} \odot \text{E}_{\text{B}}$) are all concatenated to create a unified representation of the mention pair. This representation is used, with a classifier, to learn the coreference score, CE (A, B), between the pair after finetuning the transformer.

4.2.1 $\text{P}^{+}_{\texttt{easy}}$ & $\text{P}^{-}_{\texttt{hard}}$ Discriminator (D)

The cross-encoder's encoding is non-symmetric, meaning, depending on the order in which the mentions are concatenated, it will give different coreference scores. In reality, the order should not matter for predicting if the two events are the same or not. We propose a symmetric cross-encoding scorer where we take the average of the scores predicted from both combinations of concatenation. So for a mention pair, $p = (\text{A}, \text{B})$, the symmetric cross-encoder coreference scorer (D) is given as:

$$\texttt{D}(p) = \frac{\texttt{CE}(\text{A},\text{B}) + \texttt{CE}(\text{B},\text{A})}{2} \tag{1}$$

We employ a cross-encoder with a symmetric scorer, as outlined in Equation 1, as the discriminator for $\text{P}^{+}_{\texttt{easy}}$ and $\text{P}^{-}_{\texttt{hard}}$. We conduct experiments utilizing two different Transformer models, RoBERTa ($\texttt{D}_{\text{small}}$) and Longformer ($\texttt{D}_{\text{long}}$), which vary in their maximum input capacity.

5 Experimental Setup

We describe our process of training, prediction, and hyperparameter choice in this section.

5.1 Mention Pair Generation

We use the gold mentions from the datasets. Following previous methods, we generate all the pairs (P_{all}) of mentions (M^{v}) from documents coming from the same topic. We use gold topics in the training phase and predicted topics through document clustering in the prediction phase (Bugert et al., 2021).

5.2 Training Phase

During the training phase, we leverage LH to generate a balanced set of positive and negative samples, labeled as $\text{P}^{+}_{\texttt{easy}}$ and $\text{P}^{-}_{\texttt{hard}}$, respectively. These samples are then used to train our models, $\texttt{D}_{\text{small}}$ and $\texttt{D}_{\text{long}}$ separately, using the Binary Cross Entropy Loss (BCE) function as follows:

$$L = \sum_{\substack{p_{+} \in\, \text{P}^{+}_{\texttt{easy}},\\ p_{-} \in\, \text{P}^{-}_{\texttt{hard}}}} \log \texttt{D}(p_{+}) + \log\left(1 - \texttt{D}(p_{-})\right)$$

Unlike traditional methods, we do not rely on random sampling or artificial balancing of the dataset. Instead, our heuristic ensures that the positive and negative samples are naturally balanced (as depicted in Figure 6). A side-effect of adopting this approach is that some of the positive samples are

[5] E_{A} and E_{B} represent the sum of the output embedding of each token for event triggers with multiple tokens.

Algorithm 1 Training Phase

```
Require: $D$: training document set
    $T$: gold topics
    $M^v$: gold event mentions in $D$
    $S^v$: sentences of the mentions
    $D^v$: documents of the mentions
    $G$: gold mention cluster map

    $P \leftarrow \text{TopicMentionPairs}(M^v, T)$
    $\text{Syn}_\text{P} \leftarrow \text{SynonymousLemmaPairs}(P, G)$
    $\text{P}^+_\texttt{easy}, \text{P}^-_\texttt{hard}, \text{P}^+_\texttt{FN}, \text{P}^-_\texttt{TN} \leftarrow \texttt{LH}(P, G, \text{Syn}_\text{P}, S^v)$
    $\text{D}_\text{long} \leftarrow \text{TrainCrossEncoder}(\text{P}^+_\texttt{easy}, \text{P}^-_\texttt{hard}, D^v)$
    $\text{D}_\text{small} \leftarrow \text{TrainCrossEncoder}(\text{P}^+_\texttt{easy}, \text{P}^-_\texttt{hard}, S^v)$
    return $\text{Syn}_\text{P}, \text{D}_\text{long}, \text{D}_\text{small}$
```

Algorithm 2 Prediction Phase

```
Require: $D$: testing document set
    $T$: gold/clustered topics
    $M^v$: gold event mentions in $D$
    $S^v$: sentences of the mentions
    $\text{Syn}_\text{P}$: synonymous lemma pairs from training
    $\text{D}_\text{small}, \text{D}_\text{long}$: trained CE discriminators

    $P \leftarrow \text{TopicMentionPairs}(M^v, T)$
    $\text{A}_\text{H}, \text{P}^+ \leftarrow \texttt{LH}(P, \text{Syn}_\text{P}, S^v)$
    $\text{A}_\text{P} \leftarrow \text{D}_\text{small}(\text{P}^+) > 0.5$
    $\text{A}_\text{P} \leftarrow \text{D}_\text{long}(\text{P}^+) > 0.5$
    return ConnectedComponents($\text{A}_\text{H}$),
           ConnectedComponents($\text{A}_\text{P}$)
```

excluded in training. We do this to keep the training and prediction phases consistent and, to ensure the cross-encoder is not confused by the inclusion of these hard positive examples.

Additionally, for D with Longformer, we utilize the entire document for training, while for D with RoBERTa, we only use the sentence containing the mention to provide contextual information. We employ the Adam optimizer with a learning rate of 0.0001 for the classifier and 0.00001 for fine-tuning the Transformer model. This entire process is illustrated in Algorithm 1.

To ensure optimal performance, we train our system separately for both the ECB+ and GVC training sets. We utilize a single NVIDIA A100 GPU with 80GB memory to train D_long with the Longformer model, and a single NVIDIA RTX 3090 GPU (24 GB) for training D_small with the RoBERTa-BASE model. We train each system for 10 epochs, with each epoch taking approximately one hour for the Longformer model and 15 minutes for the RoBERTa model.

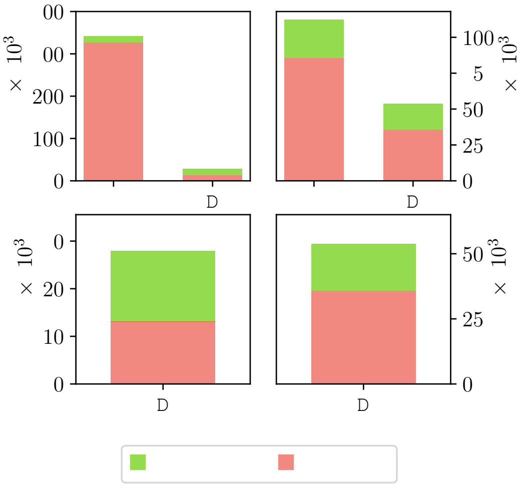

Figure 6: Training Samples of previous methods vs. ours. The heuristic creates a balanced and significantly smaller training set for ECB+. For GVC, the heuristic discards half of the negative samples while somewhat balancing the dataset.

5.3 Prediction Phase

In the prediction phase, we first pass the mention pairs through the heuristic and create an adjacency matrix called A_H based on its coreferent predictions. The ones predicted not coreferent by the heuristic are discarded. This step is crucial in terms of making the task tractable. Next, we pass the mention pairs that are predicted to be coreferent by the heuristic through D_small and D_long separately. Using the subsequent coreferent predictions from these models, we generate another adjacency matrix A_P. To create event clusters, we use these matrices to identify connected components.

As a baseline, we use the matrix A_H to generate the clusters. We then use A_P to assess the improvements made by using D_small and D_long over the baseline. This process is illustrated in Algorithm 2. The process takes between 6-10 minutes to run the Longformer model and between 1-2 minutes to run the RoBERTa one.

6 Results

We evaluate the event clusters formed using the standard coreference evaluation metrics (MUC, B^3, $CEAF_e$, LEA and CoNLL F1—the average of MUC, B^3 and $CEAF_e$ Vilain et al. (1995); Bagga and Baldwin (1998); Luo (2005); Luo et al. (2014); Pradhan et al. (2014); Moosavi et al. (2019)). We

	CoNLL F_1	
Methods	ECB+	GVC
Bugert et al. (2021)	-	59.4
Cattan et al. (2021)	81.0	-
Caciularu et al. (2021)	85.6	-
Held et al. (2021)	**85.7**	**83.7**
LH	76.4	51.8
LH + D$_{\text{small}}$	80.3	73.7
LH + D$_{\text{long}}$	81.7	75.0
LH$_{\texttt{Ora}}$	81.9	53.4
LH$_{\texttt{Ora}}$ + D$_{\text{small}}$	85.9	75.4
LH$_{\texttt{Ora}}$ + D$_{\text{long}}$	**87.4**	**76.1**

Table 2: Results on within and cross-document event coreference resolution on ECB+ and GVC test sets.

run the baseline results (LH and LH$_{\texttt{Ora}}$) and the combination of each heuristic with the two discriminators (LH/LH$_{\texttt{Ora}}$+ D$_{\text{small}}$/D$_{\text{long}}$). We compare to previous methods for ECB+ and GVC as shown in Table 2. Bold indicates current or previous SOTA and our best model.

CoNLL F1 scores show that LH and LH$_{\texttt{Ora}}$ are strong baselines for the ECB+ corpus, where LH$_{\texttt{Ora}}$ surpasses some of the previous best methods. From this, we can say that making improvements in the heuristic by better methods of finding synonymous lemma pairs is a viable solution for tackling ECB+ with a heuristic. However, the heuristics fall short for GVC, where LH$_{\texttt{Ora}}$ is only marginally better than LH. This may be due to the lower variation in lemmas in the GVC corpus. We hypothesize methods that can automatically detect synonymous lemma pairs will not be beneficial for GVC, and LH itself is sufficient as a heuristic here.

The discriminators consistently make significant improvements over the heuristics across both datasets. For ECB+, D$_{\text{long}}$ is nearly 2 points better than D$_{\text{small}}$ in terms of the CoNLL measure. Both D$_{\text{small}}$ and D$_{\text{long}}$ when coupled with LH$_{\texttt{Ora}}$ surpass the state of the art for this dataset. LH +D$_{\text{long}}$ beats Cattan et al. (2021) but falls short of SOTA, albeit by only 4 points. On GVC, both fall short of SOTA (Held et al., 2021) by only 8-9 points on CoNLL F1, with substantially fewer computations. In terms of computational cost-to-performance ratio, as we elaborate in §7.1, our methods outperform all the previous methods.

For ECR, where context is key, we would expect better performance from encoders with longer context. D$_{\text{long}}$ and D$_{\text{small}}$ show this trend for both ECB+ and GVC datasets. However, the gain we get from using the entire document is not substantial for the amount of additional computation required. An interesting line of future work would to automatically detect the core sections in the document that contribute to coreference and then only use that as context for ECR.

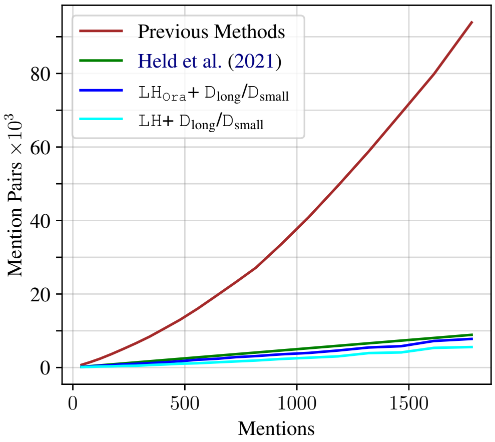

Figure 7: Prediction Phase Time Complexity in terms of Mention Pair Encoding.

7 Discussion

7.1 Time Complexity Analysis

The heuristic is a very fast process that scales linearly with the number of mentions in a corpus. Specifically, by hashing the lemma pairs and sentence token lemmas, this step performs linear comparisons of mention pairs at prediction. The mention pair cross-encoding with Transformer is a computationally intensive process. A method that encodes all mention pairs in a large corpus can become intractable. Our method, however, is linear in complexity with the number of mentions, as shown in Figure 7, and outperforms previous methods in terms of computational efficiency. While Held et al. (2021)'s cross-encoding at prediction is linear (5*n), their pruning step is quadratic. They rely additionally on training a bi-encoder and a mention neighborhood detector step that requires GPUs.

7.2 Synonymous Lemma Pairs

We have established an upper limit for ECR using the LH$_{\texttt{Ora}}$+ D$_{\text{long}}$ method for ECB+. Previous methods such as Held et al. (2021), use an oracle coreference scorer after their pruning step. In other words, their oracle assumption involves using a perfect cross-encoder. In contrast, we only use the oracle for pruning by assuming a perfect set of synonymous lemma pairs. This means that improved pruning methods can lead to better ECR performance. We believe that it is possible to create a more effective synonymous pair detector than $\texttt{LH}_{\texttt{Ora}}$ by adopting recent work on predicate class detection (Brown et al., 2014, 2022) that use VerbNet (Schuler, 2005). In future research, we aim to enhance the process of generating synonymous pairs through the use of cross-encoding or additional steps such as word sense disambiguation with the Proposition Bank (Palmer et al., 2005; Pradhan et al., 2022). Identifying the sense of the trigger will help refine the lemma pairs that appear in coreference chains. Additionally, annotating the sense of the trigger is a straightforward process that can be easily incorporated into annotation procedures for new datasets, which is more efficient than coreference annotations.

7.3 Qualitative Error Analysis

We carry out a comprehensive analysis on errors the discriminator makes after the heuristic's predictions. Unlike previous methods (Barhom et al., 2019) where they sample a subset of mentions to carry out the error analysis, we do so for the entire dataset. By efficiently discarding the large number of $\text{P}^{-}_{\texttt{TN}}$, we are able to isolate the shortcomings of the crossencoder, analyze them and offer solutions. Table 6 in Appendix C lists the various kinds of errors (incorrect and missing links) made by $\texttt{D}_{\text{small}}$ on the ECB+ and GVC dev sets.

We find error categories like same-sentence pronouns, weak temporal reasoning, ambiguity due to coreferring entities, misleading lexical similarity, and missed set-member coreferent links. Table 6 in the appendix presents examples of each.

Incorrect links due to same-sentence pronouns like “it” and “this” can be avoided by refining the heuristics-based mention-pair generation process to exclude same-sentence pronouns. Similarly, ambiguous temporal contexts like "Saturday" and "New Year's Day" that refer to the day of occurrence of the same event in articles published on different dates can be resolved by leveraging more temporal context/metadata where available. Also, errors in lexically-different but semantically similar event mention lemmas can be reduced by leveraging more-enriched contextual representations.

By using the Oracle for pruning, we can focus on where $\texttt{D}_{\text{small}}$ falls short in terms of false positives. We first sort the final event clusters based on purity (number of non-coreferent links within the cluster compared to ground truth). Next, we identify pairs that the discriminator incorrectly predicted to be coreferent within these clusters, specifically focusing on highly impure clusters. We look for these pairs in highly impure clusters and analyze the mention sentences. Our findings are as follows:

- Problems caused when two big clusters are joined through very similar (almost adversarial) examples, e.g., “British hiker” vs. “New Zealand hiker.” This error can be fixed by performing an additional level of clustering, such as, K-means.
- Problems with set-member relations, such as "shootings" being grouped with specific "shooting" events. The sets often include many non-coreferent member events. To address this issue, we can identify whether an event is plural or singular prior to coreference resolution.
- Contrary to the notion that singleton mentions cause the most errors, we found that singletons appear in the *least* impure clusters. This means the cross-encoder discriminator is good in separating out singletons.

8 Conclusion & Future work

We showed that a simple heuristic paired with a crossencoder does comparable ECR to more complicated methods while being computationally efficient. We set a upper bound for the performance on ECB+ suggesting improvement with better synonyms pairs detection we can achieve better results. Through extensive error analysis, we presented the shortcomings of the crossencoder in this task and suggested ways to improve it.

Future research directions include applying our method to the more challenging task of cross-subtopic event coreference (e.g., FCC (Bugert et al., 2020)) where scalability and compute-efficiency are crucial metrics, making the current heuristic-based mention pair generation process “learnable” using an auxiliary cross-encoder, and incorporating word-sense disambiguation and lemma-pair annotations into the pipeline to resolve lexical ambiguity. An exciting direction for future work made tractable by our work is to incorporate additional cross-encoding features into the pipeline, especially using the latest advancements in visual transformers (Dosovitskiy et al., 2021; Bao et al., 2021; Liu et al., 2021; Radford et al., 2021). Another important direction is to test our method on languages with a richer morphology than English.

Limitations

The most evident limitation of this research is that is has only been demonstrated on English coreference. Using a lemma-based heuristic requires using a lemmatization algorithm in the preprocessing phase and for more morphologically complex languages, especially low-resourced ones, lemmatization technology is less well-developed and may not be a usable part of our pipeline. Application to more morphologically-rich languages is among our planned research directions.

In addition, all our experiments are performed on the gold standard mentions from ECB+ and GVC, meaning that coreference resolution is effectively independent of mention detection, and therefore we have no evidence how our method would fare in a pipeline where the two are coupled.

A further limitation is that training of the cross-encoders still requires intensive usage of GPU hardware (the GPU used for training Longformer is particularly high-end).

Ethics Statement

We use publicly-available datasets, meaning any bias or offensive content in those datasets risks being reflected in our results. By its nature, the Gun Violence Corpus contains violent content that may be troubling for some.

We make extensive use of GPUs for training the discriminator models as part of our pipeline. While this has implications for resource consumption and access implications for those without similar hardware, the linear time complexity of our solution presents a way forward that relies less overall on GPU hardware than previous approaches, increasing the ability to perform event coreference resolution in low-compute settings.

Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers whose insightful comments and constructive feedback helped to greatly improve the quality of this paper. We gratefully acknowledge the support of U.S. Defense Advanced Research Projects Agency (DARPA) FA8750-18-2-0016-AIDA – RAMFIS: Representations of vectors and Abstract Meanings for Information Synthesis. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or the U.S. government. Finally, we extend our thanks to the BoulderNLP group and the SIGNAL Lab at Colorado State for their valuable input and collaboration throughout the development of this work.

References

Emily Allaway, Shuai Wang, and Miguel Ballesteros. 2021. Sequential cross-document coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4659–4671, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Breck Baldwin. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

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.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv e-prints*, pages arXiv–2004.

Susan Windisch Brown, Julia Bonn, Ghazaleh Kazeminejad, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2022. Semantic representations for nlp using verbnet and the generative lexicon. *Frontiers in artificial intelligence*, 5.

Susan Windisch Brown, Dmitriy Dligach, and Martha Palmer. 2014. Verbnet class assignment as a wsd task. *Computing Meaning: Volume 4*, pages 203–216.

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*, pages 23–29.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. 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.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107, Online. 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 Cybulska and Piek Vossen. 2015. Translating granularity of event slots into features for event coreference resolution. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2018. Resolving event coreference with supervised representation learning and clustering-oriented regularization. *arXiv preprint arXiv:1805.10985*.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational linguistics*, 39(4):885–916.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500, Jeju Island, Korea. Association for Computational Linguistics.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 25–32, USA. Association for Computational Linguistics.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 2014. An extension of BLANC to system mentions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29, Baltimore, Maryland. Association for Computational Linguistics.

Yehudit Meged, Avi Caciularu, Vered Shwartz, and Ido Dagan. 2020. Paraphrasing vs coreferring: Two sides of the same coin. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4897–4907, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4168–4178, Florence, Italy. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Marten Postma, Filip Ilievski, and Piek Vossen. 2018. SemEval-2018 task 5: Counting events and participants in the long tail. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 70–80, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wrightbettner, and Martha Palmer. 2022. PropBank comes of Age—Larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and*

Computational Semantics, pages 278–288, Seattle, Washington. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Veselin Stoyanov and Jason Eisner. 2012. Easy-first coreference resolution. In *Proceedings of COLING 2012*, pages 2519–2534.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, page 45–52, USA. Association for Computational Linguistics.

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)*.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2022a. 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.

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

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.

A Ablation Study of Global Attention

Table 3 compares $\texttt{D}_{\text{long}}$ performance with and without Longformer global attention on the ECB+ and GVC dev sets. This shows a dataset-specific contrast vis-à-vis sequence length where performance with global attention on GVC dev set is only *marginally* better than without, while the reverse is seen on the ECB+ dev set. More specifically, this suggests that perhaps the "relevant" or "core" context for ECR lies closer to the neighborhood of event lemmas (wrapped by trigger tokens) than the CLS tokens (that use global attention) in both corpora, albeit more so in ECB+. As such, applying global attention to the CLS tokens here encodes more irrelevant context. Therefore, $\texttt{D}_{\text{long}}$ with Longformer global attention performs less well on ECB+ while being almost comparable to $\texttt{D}_{\text{long}}$ without global attention on GVC.

Features	**ECB+**	**GVC**
w/o global attn.	85.0	76.5
w/ global attn.	82.9	77.0

Table 3: Table showing the CoNLL F1 scores from the $\texttt{D}$ Encoder with and without Longformer Global Attention on GVC and ECB+ dev sets.

B Full Results

Table 4 shows complete results for all metrics from all models for within and cross-document coreference resolution on the GVC test set. Table 5 shows complete results for all metrics from all models on the ECB+ test set.

C Qualitative Error Examples

Table 6 presents an example of each type of error we identified in the output of our discriminator ($\texttt{D}_{\text{small}}$).

	MUC			B^3			$CEAFe$			LEA			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	F_1
Bugert et al. (2021)	78.1	66.3	71.7	73.6	49.9	59.5	38.2	60.9	47.0	56.5	38.2	45.6	59.4
Held et al. (2021)	91.8	91.2	**91.5**	82.2	83.8	**83.0**	75.5	77.9	**76.7**	79.0	82.3	**80.6**	**83.7**
LH	94.8	82.0	87.9	90.1	28.5	43.3	16.3	47.8	24.3	85.1	23.9	37.4	51.8
LH$_{\text{Ora}}$	95.2	82.3	88.3	91.2	29.1	44.1	18.6	54.7	27.8	86.4	24.9	38.6	53.4
LH + D$_{\text{small}}$	87.0	89.6	88.3	82.3	67.9	74.4	62.0	55.2	58.4	77.6	57.8	66.2	73.7
LH$_{\text{Ora}}$ + D$_{\text{small}}$	89.1	90.2	**89.6**	85.0	68.0	75.6	62.7	59.6	61.1	80.6	59.5	68.5	75.4
LH + D$_{\text{long}}$	84.0	91.1	87.4	79.0	76.4	77.7	69.6	52.5	59.9	74.1	63.9	68.6	75.0
LH$_{\text{Ora}}$ + D$_{\text{long}}$	84.9	91.4	88.0	80.4	77.4	78.9	70.5	54.3	**61.3**	75.7	65.5	**70.2**	**76.1**

Table 4: Results on within and cross-document event coreference resolution on GVC test set. Bolded F1 values indicate current or previous state of the art according to that metric as well as our best model.

	MUC			B^3			$CEAFe$			LEA			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	F_1
Barhom et al. (2019)	78.1	84.0	80.9	76.8	86.1	81.2	79.6	73.3	76.3	64.6	72.3	68.3	79.5
Meged et al. (2020)	78.8	84.7	81.6	75.9	85.9	80.6	81.1	74.8	77.8	64.7	73.4	68.8	80.0
Cattan et al. (2021)	85.1	81.9	83.5	82.1	82.7	82.4	75.2	78.9	77.0	68.8	72.0	70.4	81.0
Zeng et al. (2020)	85.6	89.3	87.5	77.6	89.7	83.2	84.5	80.1	82.3	-	-	-	84.3
Yu et al. (2022b)	88.1	85.1	86.6	86.1	84.7	85.4	79.6	83.1	81.3	-	-	-	84.4
Allaway et al. (2021)	81.7	82.8	82.2	80.8	81.5	81.1	79.8	78.4	79.1	-	-	-	80.8
Caciularu et al. (2021)	87.1	89.2	**88.1**	84.9	87.9	86.4	83.3	81.2	82.2	76.7	77.2	**76.9**	85.6
Held et al. (2021)	87.0	88.1	87.5	85.6	87.7	**86.6**	80.3	85.8	**82.9**	74.9	73.2	74.0	**85.7**
LH	85.1	75.6	80.1	83.2	72.2	77.3	66.2	78.1	71.7	67.3	62.6	64.9	76.4
LH$_{\text{Ora}}$	99.1	79.6	88.3	97.9	67.7	80.0	65.9	93.7	77.4	85.1	63.8	72.9	81.9
LH + D$_{\text{small}}$	76.2	86.9	81.2	77.8	85.7	81.6	83.9	73.0	78.1	68.7	71.5	70.1	80.3
LH$_{\text{Ora}}$ + D$_{\text{small}}$	89.8	87.6	88.7	90.7	80.2	85.1	82.5	85.1	83.8	83.3	72.2	77.3	85.9
LH + D$_{\text{long}}$	80.0	87.3	83.5	79.6	85.4	82.4	83.1	75.5	79.1	70.5	73.3	71.9	81.7
LH$_{\text{Ora}}$ + D$_{\text{long}}$	93.7	87.9	**90.7**	94.1	79.6	**86.3**	81.6	88.7	**85.0**	86.8	73.2	**79.4**	**87.4**

Table 5: Results on within and cross-document event coreference resolution on ECB+ test set with gold mentions and predicted topics. Bolded F1 values indicate current or previous state of the art according to that metric as well as our best model.

Category	Snippet
Adversarial/Conflicting	British climber <m> dies </m> in New Zealand fall.....The first of the <m> deaths </m> this weekend was that of a New Zealand climber who fell on Friday morning.
Adversarial/Conflicting	British climber <m> dies </m> in New Zealand fall.....Australian Ski Mountaineer <m> Dies</m> in Fall in New Zealand.
Adversarial/Conflicting	..Prosecutor Kym Worthy announces charges against individuals involved in the gun violence <m> deaths </m> of children in Detroit ..... Grandparents charged in 5-year - old 's shooting <m> death </m> Buy Photo Wayne County Prosecutor Kym Worthy announces charges against individuals involved in the gun violence deaths of children...
Pronoun Lemmas	This just does not happen in this area whatsoever . <m> It </m>'s just unreal , " said neighbor Sheila Rawlins....<m> This </m> just does not happen in this area whatsoever . It 's just unreal , " said neighbor Sheila Rawlins .
Set-Member Relationship	On Friday , Chicago surpassed 700 <m> homicides </m> so far this year . ....<m> **Homicide** </m> Watch Chicago Javon Wilson , the teenage grandson of U.S. Rep. Danny Davis , was shot to death over what police called an arugment over sneakers in his Englewood home Friday evening .
Weak Temporal Reasoning	Police : in an unrelated <m> incident </m> a man was shot at 3:18 a.m. Saturday in North Toledo ....Toledo mother grieves 3-year - old 's <m> **shooting**</m> death | Judge sets bond at 580,000 USD for Toledo man accused of rape , kidnapping | Toledo man sentenced to 11 years in New Year 's Day shooting
Incomplete, Short Context	Ellen DeGeneres to <m> Host </m> Oscars....It will be her second <m> **stint** </m> in the job , after hosting the 2007 ceremony and earning an Emmy nomination for it .
Similar context, Different event times	near Farmington Road around 9 p.m. There they found a 32-year - old unidentified man with a <m> gunshot </m> wound outside of a home ....The family was driving about 8:26 p.m. Sunday in the 1100 block of South Commerce Street when <m> gunshots were fired </m> from a dark sedan that began following their vehicle...
Same Lemma, Ambiguous Context	Police : Man Shot To Death In Stockton Related To 3-Year - Old <m> Killed </m> By Stray Bullet 2 p.m. UPDATE : Stockton Police have identified the man shot and killed on ....Police : Man Shot To Death In Stockton Related To 3-Year - Old Killed By Stray Bullet 2 p.m. UPDATE : Stockton Police have identified the man shot and <m> killed </m> on Tuesday night.
Lexically different, Semantically same	One man is dead after being <m> shot </m> by a gunman ....Employees at a Vancouver wholesaler were coping Saturday with the death of their boss , who was <m> **gunned down** </m> at their office Christmas party .
Misc.	Baton Rouge Police have charged 17-year - old Ahmad Antoine of Baton Rouge with Negligent Homicide in the city 's latest shooting <m> death </m> .....Tagged Baton Rouge , <m> homicide </m>.

Table 6: Qualitative Analysis on the hard mention pairs incorrectly linked (or missed) by our Discriminator ($\mathtt{D}_{\text{small}}$) in the ECB+ and GVC dev set: Underlined and bold-faced mentions surrounded by trigger tokens respectively indicate incorrect and missing assignments. Underlined spans without trigger tokens represents the category-specific quality being highlighted. The miscellaneous category (Misc.) refers to other errors including (reasonable) predictions that are either incorrect annotations in the gold data or incomplete gold sentences.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We talk about the limitations of our approach in the "Limitations" section after the Conclusion.

☑ A2. Did you discuss any potential risks of your work?
We discuss the risks in the Ethics statement that is a part of our paper.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

CoVal: A coreference evaluation tool for CoNLL datasets cited in Section 6.

☑ B1. Did you cite the creators of artifacts you used?
CoVal: Cited in Section 6 and the Longformer Model in Section 2

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The authors of the scientific artifact have included the BibTeX citation in their open-source platform. As such, we have given full credit to this and cited them with the provided Bibtex citation.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Since the artifacts are popularly used in the AI/NLP/CL domain (both academia and industry) with proper citations, we have taken all measures to ensure that our usage is consistent with their intended use.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We use publicly-available datasets like the ECB+ corpus and the Gun-Violence Corpus and we discuss the risks in the ethics statement.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We do not create or collect any data in this experiment and have used publicly available datasets that are often cited and used in this field.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, we describe them in section 3 and in table .

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

section 5.2

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
section 5.2

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
section 5.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
section 6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
section 4.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

No human annotators were used in our experiments.

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No human annotators were used in our experiments.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No human annotators were used in our experiments.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No human annotators were used in our experiments.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human annotators were used in our experiments.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No human annotators were used in our experiments.