Cross-document Event Identity via Dense Annotation

Adithya Pratapa, Zhengzhong Liu, Kimihiro Hasegawa,
Linwei Li, Yukari Yamakawa, Shikun Zhang, Teruko Mitamura
Language Technologies Institute, Carnegie Mellon University
{vpratapa,zhengzhl,kimihiro,linweil,yukariy,shikunz,teruko}@andrew.cmu.edu

Abstract

In this paper, we study the identity of textual events from different documents. While the complex nature of event identity is previously studied (Hovy et al., 2013), the case of events across documents is unclear. Prior work on cross-document event coreference has two main drawbacks. First, they restrict the annotations to a limited set of event types. Second, they insufficiently tackle the concept of event identity. Such annotation setup reduces the pool of event mentions and prevents one from considering the possibility of quasi-identity relations. We propose a dense annotation approach for cross-document event coreference, comprising a rich source of event mentions and a dense annotation effort between related document pairs. To this end, we design a new annotation workflow with careful quality control and an easy-to-use annotation interface. In addition to the links, we further collect overlapping event contexts, including time, location, and participants, to shed some light on the relation between identity decisions and context. We present an open-access dataset for cross-document event coreference, CDEC-WN, collected from English Wikinews and open-source our annotation toolkit to encourage further research on cross-document tasks.[1]

1 Introduction

Coreference resolution is the task of identifying events (or entities) that refer to the same underlying activity (or objects). Accurately resolving coreference is a prerequisite for many NLP tasks, such as question answering, summarization, and dialogue understanding. For instance, to get a holistic view of an ongoing natural disaster, we need to aggregate information from various sources (newswire, social media, public communication, etc.) over an extended period. Often this requires resolving

(October 23, 2010) Nearly *200 people* are confirmed dead and approximately *2600* are ill in a central Haitian cholera **outbreak**.

(October 26, 2010) At least *259 people* are dead and over *3000 people* have been infected in the Haitian cholera **outbreak**.

(October 28, 2010) The Haitian cholera **outbreak** has killed *292 people* and infected over *4000*, according to the Haitian government.
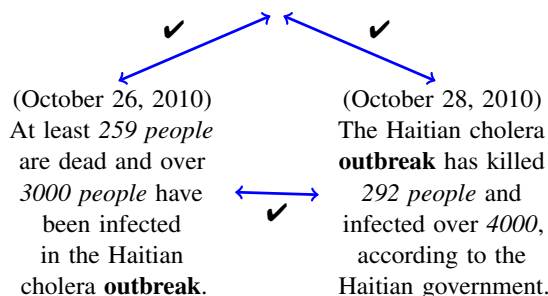
Figure 1: An illustration of the quasi-identity nature of events. The event [Haitian cholera] 'outbreak' is expressed by instances with varying counts of infections and deaths. The identity of this event continuously evolves over space and time, attributed to a new type of quasi-identity, spatiotemporal continuity.

coreference between mentions across documents.[2]

Recasens et al. (2011) defines coreference as "identity of reference". Therefore, modeling event coreference requires understanding the extent of the shared identity between event mentions. Numerous factors determine this identity, including the semantics of the event mention, its arguments, and the document context. Resolving coreference across documents is more challenging, as it requires modeling identity over a much longer context. To this end, we identify two major issues with existing cross-document event coreference (CDEC) datasets that limit the progress on this task. First, many prior datasets often annotate coreference *only on a restricted set of event types*, limiting the coverage of mentions in the dataset. Second, many datasets and models *insufficiently tackle the concept of event identity*. As highlighted by Hovy et al. (2013), the decision of whether two mentions refer

---

[1]Data and code are available at https://github.com/adithya7/cdec-wikinews.

[2]A mention is a linguistic expression in text that denotes a specific instance of an event.

to the same event is often non-trivial. Occasionally, event mentions only share a partial identity (*quasi-identity*). In this work, we present a new dataset for CDEC that attempts to overcome both issues.

Earlier efforts on CDEC dataset collection were limited to specific pre-defined event types, restricting the scope of event mentions that could be studied. In this work, we instead annotate mentions of all types, i.e., open-domain events (Araki and Mitamura, 2018), and provide a *dense annotation* (Cassidy et al., 2014) by checking for coreference relationship between every mention pair in all underlying document pairs. We compile documents from the publicly available English Wikinews.[3] To facilitate our goal of dense annotation of mentions and their coreference, we develop and release a new easy-to-use annotation tool that allows linking text spans across documents. We crowdsource coreference annotations on Mechanical Turk.[4]

Prior work has attributed the quasi-identity behavior of events to two specific phenomena, membership and subevent (Hovy et al., 2013). However, its implications in cross-document settings remain unclear. In this work, we specifically focus on a cross-document setup. As highlighted by Recasens et al. (2012), a direct annotation of quasi-identity relations is hard because annotators might not be familiar with the phenomenon. Therefore, we propose a new annotation workflow that allows for easy determination of quasi-identity links. To this end, we collect evidence for time, location, and participant(s) overlap between corefering mentions. We also collect information regarding any potential inclusion relationship between the mention pair.

Our workflow allowed us to empirically identify a new type of quasi-identity, *spatiotemporal continuity*, in addition to the existing types defined by Hovy et al. (2013). Figure 1 illustrates this phenomenon using the case of [Haitian cholera] outbreak. The event gradually evolves over space and time, leading to cases of partial coreference. Additionally, traditional coreference annotations cluster mentions together. However, this methodology can be misleading when dealing with cases of quasi-identity (see §5). To overcome this limitation, we frame our annotation task as a (cross-document) mention pair linking. The proposed task simplifies the annotation process by avoiding merging quasi-identical mentions into a single cluster.

The main contributions of our work can be summarized as follows,

- We present an empirical study of the quasi-identity of events in the context of CDEC. In addition to providing evidence for previously studied types of quasi-identity (membership, subevent), we identify a novel type relating to the spatiotemporal continuity of events.

- We release a densely annotated CDEC dataset, CDEC-WN, spanning 198 document pairs across 55 subtopics from English Wikinews. The dataset is available under an open license. To serve as a benchmark for future work, we provide two baselines, lemma-match, and a BERT-based cross-encoder.

- To efficiently collect evidence for quasi-identity, we develop a novel annotation workflow built upon a custom-designed annotation tool. We deploy the workflow to crowdsource CDEC annotations from Mechanical Turk.

In the upcoming sections, we first position our work within the existing CDEC literature (§2). We then describe our methodology for preparing the source corpus (§3), and our crowdsourcing setup for collecting coreference annotations on this corpus (§4). In §5, we present a study of quasi-identity of events in our dataset. Finally, in §6, we present two baselines models for the proposed dataset.

2 Related Work

Event Coreference: Widely studied in the literature, with datasets curated for both within and cross-document tasks. ACE 2005 (Walker et al., 2006), OntoNotes (Weischedel et al., 2013), and TAC-KBP (Mitamura et al., 2017) are commonly used benchmarks for within-document coreference. For cross-document coreference, ECB+ (Cybulska and Vossen, 2014) is a widely popular benchmark and is an extended version of the original ECB dataset (Bejan and Harabagiu, 2008). ECB+ suffers from a major limitation with coreference annotations restricted to only the first few sentences in the documents. However, CDEC is a long-range phenomenon, and there is a need for more densely annotated datasets.

Many other datasets have since been curated for the task of CDEC. Some related works include, MEANTIME (Minard et al., 2016), Event hoppers (Song et al., 2018), Gun Violence Corpus (GVC)

(Vossen et al., 2018), Football Coreference Corpus (FCC) (Bugert et al., 2020), and Wikipedia Event Coreference (WEC) (Eirew et al., 2021). However, most CDEC systems are still evaluated primarily on ECB+. Additionally, all of these datasets do not account for the quasi-identity nature of events.

Though compiled from Wikinews, CDEC annotations in the MEANTIME corpus were limited to events with participants from a pre-defined list of 44 seed entities. While the FCC corpus was also crowdsourced, the annotation unit was an entire sentence instead of a single event mention. WEC corpus uses hyperlinks from Wikipedia but primarily handles referential events. In this work, we use open-domain events and treat an event mention as our annotation unit. We collect coreference links across all the mention pairs from all the underlying document pairs.

Event Identity: Recasens et al. (2011) postulated entity coreference as a continuum, with identity, non-identity and near-identity relations. In a follow-up work (Recasens et al., 2012), they identify near-identity relations using the disagreement between annotators. They say subjects are not fully aware of the near-identity behavior, therefore making direct annotation collection hard. The continuum idea has since extended to events (Hovy et al., 2013). Determining if two event mentions are identical is not a trivial decision. It depends on the arguments of the mentions (often underspecified in the local context), the semantics of the mention, and the document context. In this work, we are specifically interested in cross-document coreference. Wright-Bettner et al. (2019) studied the impact of the subevent relationship on quasi-identity, but a more general annotation framework is missing. Accurately capturing event identity is critical to CDEC dataset construction and the subsequent modeling. Therefore, we qualitatively study this phenomenon by collecting supplementary information with each coreference link.

3 Corpus Preparation

In our goal of curating a CDEC dataset, we first needed to identify documents that exhibit cross-document coreference. We now describe our document collection process and our methodology for annotating event mentions in these documents.

Document Selection: To facilitate the redistribution of the documents under an open license,

we prioritized collecting the documents from publicly available news sources. We chose Wikinews for three key reasons. First, the news articles were sourced from trusted news outlets and reported impartially. Second, these articles are available under an open license (CC BY 2.5), allowing easy redistribution. Finally, each article is human-labeled with categories (e.g., Disaster and accidents, Health, Sports, etc.),[5] as we describe later, this meta-information plays a significant role in our dataset collection. We use the July 1st, 2020 dump of English Wikinews, which contains a total of 21k titles (or articles/documents). These news articles are timestamped from November 2004 to July 2020. Annotating coreference between every document pair in Wikinews is infeasible. Therefore, we first identify groups of related news articles. Articles within a given group usually describe a part of a developing news story or storyline.

Identifying Storylines: To identify these latent storylines, we first construct an undirected Wikinews graph (W) with articles as nodes and add an edge between two nodes if one is mentioned under the "Related News" section in the other. We then identify cliques (C_W) (i.e., fully connected sub-graphs) in the Wikinews graph, which constitute our potential set of storylines. While the articles within each clique are related, we also want to minimize the relatedness of articles across cliques. Therefore, we construct a new graph (M), where each clique ($\in C_W$) is a node, and an edge is added between two nodes if the two cliques are not disjoint or if any two articles in the two cliques share an edge in the Wikinews graph (W). Finally, we extract maximal independent sets from M that correspond to separate storylines. Among the multiple feasible maximal independent sets, we optimize for maximum overlap in Wikinews categories of articles within each clique.

This algorithm satisfies two requirements of a CDEC dataset. First, within each storyline, all articles are related to each other. Second, articles from different storylines aren't adjacent in the Wikinews graph (W); thereby, they are very likely unrelated.

For this work, we narrow our focus only to articles in the "Disaster and Accidents" category on Wikinews.[6] Following the terminology of prior

---

[5]https://en.wikinews.org/wiki/Wikinews:Categories_and_topic_pages
[6]https://en.wikinews.org/wiki/Category:Disasters_and_accidents

498

# topics	1
# subtopics	55
# documents	176
# sentences per doc (avg.)	14.6
# tokens per doc (avg.)	344
# event mentions	7220
# mentions per doc (avg.)	41
# document pairs	198
# CDEC links	4282
# CDEC links per document pair	21.6
# full coreference links	2914
# partial coreference links	1368

Table 1: An overview of the compiled CDEC dataset.

work, our dataset constitutes of a single topic (Disaster and accidents) and 55 subtopics (individual storylines). We restrict CDEC annotations to subtopics that contain 3 or 4 documents. Our algorithm aims for completeness of the CDEC dataset by maximizing for intra-subtopic and minimizing inter-subtopic coreference.

Event Mention Identification: To annotate the event mentions in the above-collected documents, we first run a combination of mention detection systems. Specifically, we use the OpenIE system (Stanovsky et al., 2018) from AllenNLP (Gardner et al., 2018) and an open-domain event extraction system (Araki and Mitamura, 2018). The former is effective at extracting verbal events, whereas the latter is good at nominal events. In contrast to most prior work, we do not restrict the mentions to specific event types or salient events. We believe it is important to study all underlying events to achieve a complete understanding of the corpus. Since the quality of mention identification is critical to our CDEC dataset, we ask an expert to go through the automatically identified mentions and add/edit/delete mentions using the Stave annotation tool (Liu et al., 2020).[7]

Table 1 presents the overall statistics of our document corpus. Our documents are $\sim$14.6 sentences long, comparable to prior work, ECB+ (16.6), GVC (19.2), and FCC (34.4). However, our documents are significantly more dense in terms of event mentions. Our documents contain $\sim$41 mentions (on avg.), much higher compared to prior work, ECB+ (15.3), GVC (14.3), FCC (5.8). Given the dense

nature of our documents, we appropriately design our annotation task and interface.

4 Annotating Coreference via Crowdsourcing

Corefering event mentions share their identity. However, the extent of sharing for them to be considered coreferential is unclear. To empirically study this behavior, we crowdsource annotations on Mechanical Turk. We use the crowd workers' responses to analyze the influence of quasi-identity on coreference decisions.

4.1 Annotation Task

The input to our annotation task constitutes a pair of documents, with all event mentions pre-identified. Annotator iterates through every mention on the left document and select corefering mentions from the right document. We also provide the document titles and publication dates to help set the context for the articles. Note that we focus solely on cross-document coreference in this work and leave the addition of within-document links to future work.

Prior work has highlighted the difficulty in capturing event coreference, specifically in cases where the mentions are only quasi-identical (Hovy et al., 2013). Notably, Recasens et al. (2012) found direct annotation of partial identity to be a difficult task. Therefore, we propose to analyze this behavior by collecting supplementary information from the annotators. For each coreference link created by an annotator, we ask them four *follow-up questions*, 1. overlap in location, 2. overlap in time, 3. overlap in participants, and 4. potential inclusion relationship.[8] Annotators implicitly consider these aspects when making a coreference decision; therefore, responding to these questions won't increase the annotators' cognitive load significantly. As we show in §5, the responses to these questions help us tease apart the cases of partial identity.

Unlike within-document coreference, disjoint narratives between documents often complicate CDEC annotation tasks. Wright-Bettner et al. (2019) analyzed this behavior in detail and proposed a new `contains-subevent` label for within-document links that improved annotator agreement and reduced inconsistencies. However, they rely on experts to create the within-doc `contains-subevent` label beforehand. Instead, we

---

[7]the expert annotator is an author of this work.

[8]see Table 14 in Appendix for the exact formulation of these follow-up questions.
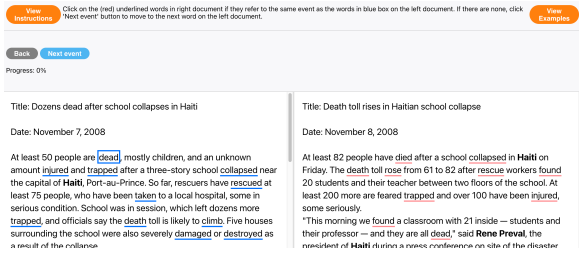
Figure 2: Tool for annotating cross-document event coreference. The two documents are shown side-by-side, with event mentions pre-highlighted. We provide on-screen instructions as well as dedicated pages for viewing detailed instructions and examples. As seen in the example here, we allow annotation of every pair of mentions in the given document pair. In our annotation effort, we present every pair of related documents on this tool, leading to a *densely* annotated dataset.

focus solely on cross-document links and frame the task as a simple pair-wise classification. Our framing allows non-expert annotators to make decisions without concern for complex granularity issues. Our follow-up question regarding inclusion facilitates a post hoc analysis of the event granularities in our dataset.

To ensure completeness of our CDEC dataset, we collect annotations for each pair of documents in a given subtopic (§3). As highlighted earlier, the quasi-identity of events may or may not allow for the application of transitivity property. Therefore, in our dataset, we cannot expand coreference links using transitivity. So collecting annotations between each pair in a given subtopic is necessary.

Annotation Guidelines: Events are commonplace in the newswire; therefore, it is feasible to explain the concept of events and their coreference via simple example-based guidelines. In our guidelines, we first define *events* and then provide numerous examples of identical and non-identical event mentions, with detailed explanations. Following prior work (Song et al., 2018), we rely on the annotator's intuition to decide coreference.[9]

4.2 Annotation Tool

To efficiently crowdsource annotations, we require a tool that is both easy-to-use and customizable to our workflow. For this purpose, we build upon the Forte[10] and Stave[11] toolkits (Liu et al., 2020). We extend both the toolkits to support cross-document

linking as required by our annotation task. Figure 2 presents a snapshot of our annotation interface. We highlight event mentions in both the documents and allow the annotator to iterate through each mention on the left document. In addition to dedicated links to instructions and examples, we provide on-screen instructions to assist the annotator in real-time. We also use an English NER tool (Ma and Hovy, 2016) to highlight the named entities in the documents. These entities help the annotator keep track of various event participants in the two documents.

We utilize this tool for our entire dataset collection. While we show an application of our annotation tool for CDEC, we believe it's adaptable to other cross-document tasks like entity coreference and event/entity relation labeling tasks. We will release our toolkit to encourage future work on cross-document NLP tasks.

4.3 Collecting CDEC annotations

We crowdsource annotations for CDEC using Amazon Mechanical Turk (MTurk). Each Human Intelligence Task (HIT) constitutes annotating cross-document links for one pair of documents. We obtained IRB approval and set our HIT price based on preliminary studies.[12] On MTurk, we restricted our HITs to crowd workers from the US and set our qualification thresholds for % HITs, and total HITs approved as 95% and 1000 respectively. We paid a fair compensation of $10.9/hour on average.[13] Our annotation task requires proficiency in English, as well as a good understanding of event coreference. To this end, we attach a qualification test with eight yes/no questions regarding event coreference, with a qualification threshold of 75%.[14]

For each document pair, we collected annotations from three different crowd workers. In each task, crowd workers go through the two documents and develop a high-level understanding of the news story. They then iterate through the mentions in the left document, in the narrative order, to identify potential cross-document coreference links. From our preliminary studies, we found that annotators spend considerable time reading the two documents. Therefore, to make the best use of the crowd workers' time and effort, we group HITs that constitutes document pairs from the same subtopic.

---

[9] see A.2 in Appendix for complete guidelines.

[10] https://github.com/asyml/forte

[11] https://github.com/asyml/stave

[12] see A.1 in Appendix for more details.

[13] The median pay was slightly higher at $16.3/hour. Both mean and median pay are above the current minimum wage requirements in the United States.

[14] see A.4 in Appendix for the test format and the questions.

This way, if the crowd worker chooses to, they can annotate the entire subtopic in one sitting, sharing their understanding of a document from one HIT to the next. In total, we collected annotations for 198 document pairs, spanning 176 unique documents and 55 subtopics from 46 crowd workers.

Inter Annotator Agreement (IAA): For each pair of documents, we collect annotations from three crowd workers. Our setup allows the annotator to decide coreference for every mention pair. To measure IAA, we associate a value to each mention pair (corefering or non-corefering) and compute Krippendorff's α. For coreference links, we observed an α of 0.46, indicating moderate agreement (Artstein and Poesio, 2008).[15] Additionally, we compare the impact of the quasi nature of coreference on the annotator agreement. In our dataset, 31% of the full-coreference links have a perfect majority (3/3 annotators). However, only 13% of the partial-coreference links have the same (see section 5 for the methodology used to determine partial coreference). This sharp contrast illustrates the difficulty in capturing partial coreference links.

Selecting CDEC links: For each pair of mentions, we take a majority vote on the three crowd-sourced annotations. In our preliminary analysis, we found many valid coreference links annotated by just one crowd worker. While we encourage the crowd workers to annotate every pair of corefering mentions, they occasionally miss links. Therefore, to ensure completeness of our dataset, we use an adjudicator to go through the single-annotator links to decide if they are in-fact corefering or not.[16]

Table 1 presents an overview of the compiled CDEC dataset. Unlike prior work, we do not create mention clusters by expanding the links via transitive closure. As we show in §5, quasi-identity of events warrants the need to analyze coreference at the level of mention pairs instead of clusters.

4.4 Dataset Validation

To facilitate benchmarking future coreference resolution models, we split our dataset into train and test. Of the 55 subtopics, 40 are for model training and development, and 15 are for the unseen test set. Given the importance of the test set quality, we perform expert validation on a randomly selected
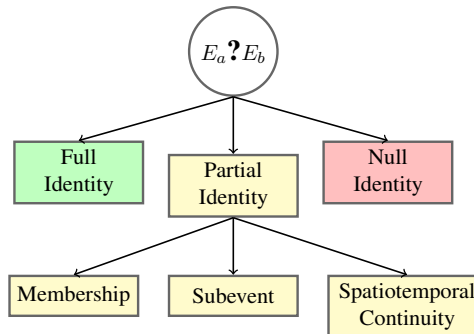

Figure 3: A taxonomy of event identity. While full and null identities are well understood, the definition of partial identity is still evolving. We present the three types of partial identity found in our dataset.

subset of 18 document pairs from our test set. The expert inspected the annotated coreference links in the subset and found 97.5% precision (549/563 were corefering). On the other hand, measuring the recall is hard due to a large number of mention pairs. Therefore, we specifically focus on two types of potentially missing coreference links, 1. mention pairs that share the same head lemma (but not annotated as corefering), 2. mention pairs that are part of a non-transitive triplet.[17] Upon inspection by the expert, we find that majority of lemma-match links are non-corefering (50/565 were corefering), while a majority of non-transitive pairs are corefering (149/173 were corefering). This result indicates the scope for improvement in tackling missing coreference links. We leave this extension to future work.

5 Studying Quasi-Identity of Events

Numerous factors determine the identity of an event mention, including the semantics of the mention, arguments (place, time, and participants), and the overall document context. Therefore, overlap in these factors determines the extent of coreference between two given mentions. This overlap leads to cases of partial (quasi-) identity. Our annotation workflow allows for empirical investigation of this phenomenon, and we summarize our observations through a taxonomy of event identity in Figure 3. Except for Wright-Bettner et al. (2019), prior CDEC datasets do not account for the partial identity during the annotation process. Hovy et al. (2013) have previously proposed two types of partial identity, membership, and subevent. In ad-

---

[15]It's important to note that we compute IAA on our entire dataset. Our IAA score is comparable to those of quasi-relations from Hovy et al. (2013).

[16]the adjudicator is an author of this paper.

[17](E_A, E_B, E_C) is a non-transitive event triplet if E_A corefers with E_B, E_B corefers with E_C, but E_A and E_C are non-corefering.

Membership	
1a	The **fire** has burned about 4400 acres so far and 15 homes have been lost, however there have been no reported injuries or deaths.
1b	Reports say that the amount of people fleeing from their homes in California located in the United States due to **wildfires** has reached the 1,000,000 mark as the fires continue to grow.
2a	Several aftershocks have rocked the same area, the latest measuring 7.1, had a depth of 10 km. It was first reported to be a 7.3 **aftershock**.
2b	Some smaller **aftershocks** with magnitudes between 5.2 and 5.7 were also reported in the region.
2c	That quake was followed by as many as 60 **aftershocks** for at least a week, with some ranging as high as magnitude 7.8.
Subevent	
3a	A freight train in Lviv, Ukraine derailed, caught fire, and spilled a toxic chemical, releasing dangerous fumes into the air early Tuesday morning (local time), and people who live near the site of the **crash** are still becoming sick.
3b	The available information about the phosphorous cloud following the railway **accident** in the Ukraine last Monday is becoming more and more cryptic.
4a	During the fifteen days of the trial, the prosecutors called 92 witnesses to testify as to the chaotic scenes following the **bombing**.
4b	Two **explosions** within seconds of each other tore through the finish line at the Boston Marathon, approximately four hours after the start of the men's race.
Spatiotemporal Continuity	
5a	Tropical **storm** Richard is nearing hurricane strength with winds of 70 mph (115 kph) as it lashes Honduras with heavy rains
5b	**Hurricane** Richard made landfall in Belize about 20 mi (35 km) south-southeast of Belize City with winds of 90 mph (150 kph) at approximately 6:45 local time (0045 UTC) according to the National Hurricane Center (NHC)

Table 2: An illustration of quasi-identity of event mentions across documents. These examples cover the three identified types of quasi-identity, membership, subevent, and spatiotemporal continuity.

dition to providing evidence for these two types in our dataset, we also identify a novel type of partial identity termed as *spatiotemporal continuity*.

Collecting Partial Identity: We use the responses to follow-up questions for qualitatively analyzing cases of partial identity. We consider a link to be a case of partial identity if a strict majority of annotators indicate one of the following. First, there is an inclusion relationship between corefering mentions. Second, the two overlap in place, time, or participants. With this screening methodology, we found ∼32% of the total CDEC links to be candidates for partial identity (Table 1). We qualitatively analyze the dataset and identify three types of partial identity, 1. Membership, 2. Subevent, and 3. Spatiotemporal continuity. Table 2 illustrates each type with examples from our compiled dataset.

Membership: An event mention E_a is a member of event mention E_b. Consider the two sentences, 1a, and 1b. The mention 'fire' (1a) denotes a specific wildfire, whereas 'wildfires' (1b) denotes a group of wildfires, including the one in 1a. The concept of partial identity often challenges the transitivity assumption of coreference. For instance, the mentions [smaller] 'aftershocks' (2b) and [7.1] 'aftershock' (2a) share no identity, thereby, non-coreferential. However, both the mentions partially corefer with [60] 'aftershocks' from 2c.

Subevent: An event mention E_a is a subevent of event mention E_b. This behavior can be seen in the coreference between the 'crash' event from 3a, and the 'accident' event from 3b. While the 'accident' event involves many individual events, derailed, caught fire, spill chemical, and release fumes, it partially corefers with the event 'crash' from 3a that likely refers only to the derailment.

Similarly, consider the case of the Boston Marathon Bombing in examples 4a and 4b. The 'bombing' event from 4a refers to the whole incident, whereas the 'explosions' in 4b refers to specific subevents of the 'bombing'.

Spatiotemporal Continuity: The identity of an event can *continuously* evolve over space and time. Consider the two mentions, 'storm' and 'Hurricane' from Table 2 (5a, 5b). At a high level, these mentions are corefering because they denote the same event (storm Richard). However, the expressions of this event differ slightly across the two documents. In the former, it's a storm (with 70mph winds) having an impact in Honduras, whereas, in the latter, it's a hurricane (with 90mph winds) impacting Belize. Similar behavior is visible with the [Haitian cholera] 'outbreak' event from Figure 1. The outbreak gradually evolves, with growing infection ($2600 \rightarrow 3000 \rightarrow 4000$) and deaths ($200 \rightarrow 259 \rightarrow 292$). In both of these examples, we observe the event changes gradually and is always continuous in both space and time dimensions.[18]

In line with prior work on entities (Recasens et al., 2011), we believe identity and coreference of events to be a continuum. Our dataset already includes many instances of partial identity to support this hypothesis. The above-described cases of partial identity (membership, subevent, and spatiotemporal continuity) will pose new challenges to future dataset collection efforts. We believe our annotation workflow and guidelines will be of use to future work.

In this section, we establish a clear case for tackling partial identity within the coreference resolution task. However, in practical settings, the boundaries between full, partial, and null identities remain fuzzy. As seen in our analysis on the inter-annotator agreement, humans find it hard to identify cases of partial coreference. In the downstream coreference resolution task, users are primarily interested in knowing if two given mentions share an identity or not. Therefore, we propose to view both full and partial identity under a single coreference label ('coreference') and contrast them against cases with no shared identity ('non-coreference'). Compared to prior datasets, this presents new challenges in tackling partial identity

---

[18]We borrow the term spatiotemporal continuity from the Philosophy literature. It describes the properties of well-behaved objects (Wiggins et al., 1967). A similar treatment for entities is presented in Recasens et al. (2011).

within the 'coreference' label.

6 Baselines

We define the task as a mention pair classification problem. Due to the quasi-identity nature of event mentions (§5), we do not cluster mentions in coreference groups. Additionally, we consider both full and partial identity under the coreference label. We present two baseline models, lemma-match, and a cross-encoder model. We split the dataset of 55 subtopics into train and test, with 40 subtopics for training and development, and 15 subtopics for the held-out test set. For our experiments, we assume gold mentions and subtopic information.[19]

Lemma-match: For our first baseline, we implement the traditional lemma-match baseline. We use spacy's large model[20] to extract the head lemma of the event mentions, and consider two mentions corefering if the lemma's match. Following Upadhyay et al. (2016), we also experiment with a Lemma-δ baseline. In our experiments, we found the best dev performance with δ=0, resolving to a simple lemma baseline. This could be due to our assumption of access to gold subtopic information.

Cross-Encoder: As a second baseline, we implement BERT-based cross-encoder model. The input consists of a pair of sentences with both mentions highlighted using special tokens to indicate the start and end of mention spans (<E>, </E>). We first concatenate the two event-tagged sentences (with [SEP] token) and pass it through a bert-base-uncased encoder. We then perform mean pooling on the event start tags (<E>), and pass the pooled embedding through a linear classification layer to predict coreference vs. non-coreference. For training the cross-encoder, in addition to the positive coreference pairs, we generate two types of negative mention pairs. For the first type, we collect non-coreference mention pairs from sentences that have a coreference link between a different mention pair. For the second type, we extract non-coreference mention pairs from random sentence pairs between the documents. During training, we use a dataset ratio of 1:5:5 (positive:negative-I:negative-II). We use huggingface transformers (Wolf et al., 2020), and train the model using AdamW (Loshchilov and Hutter, 2019) with an initial learning rate of 2e-5. We also use a linear warmup scheduler, with

---

[19]topic-level performance (Cattan et al., 2021)
[20]en_core_web_lg from https://spacy.io

Model	Dev			Test		
	P	R	F1	P	R	F1
Lemma-match	46.6	54.9	49.9	42.3	56.0	48.2
Cross-Encoder	43.1	75.4	54.3	45.9	77.3	57.6
	± 0.6	± 0.5	± 0.5	± 0.8	± 1.1	± 0.6

Table 3: Baseline results on development and test sets. For cross-encoder, we report the average scores and their standard deviation across five runs.

10% of training steps for warmup. We finetune the # epochs and positive:negative dataset ratio during the development stage (5-fold cross-validation) and use the best configuration when training on the entire train set.

Results: Table 3 presents the results of our baselines. For model development, we perform 5-fold cross-validation on the training set (40 subtopics). To report the results on the held-out test set (15 subtopics), we train the model's best configuration on the entire training set. We report precision, recall, and F1 scores of the coreference label averaged on five different runs. The lemma baseline only achieves an F1 score of 48.2, indicating that the proposed dataset is lexically diverse. The cross-encoder improves upon the lemma baseline, especially on the recall. Upon inspection of development set predictions, we observe two possible error cases for the cross-encoder model. First, the model struggles at the cases of partial identity ('explosion' vs. 'incident' and 'evacuate' vs. 'evacuations'). This drawback of cross-encoder indicates that the model requires a deeper understanding of event identity. Second, the cross-encoder model is often limited by the information available in a single sentence. It is known the event arguments are often underspecified in the local context (Ebner et al., 2020); therefore, increasing the context to a paragraph or the entire document might help improve the performance.

7 Conclusion & Future Work

In this work, we present a study of the identity of events through annotation of cross-document event coreference. We use a custom-designed annotation tool to collect coreference annotations on a subset of English Wikinews articles. We release our dataset, CDEC-WN, under an open license to encourage further research on event coreference. By collecting evidence for the extent of shared identity between events, we identify three types of partial-identity, membership, subevent, and spatiotemporal continuity. To serve as a benchmark for future coreference resolution systems, we provide results on two baseline models, lemma-match and BERT-based cross-encoder. We believe that our work will encourage further research on the identity of events in the context of CDEC. Potential future directions include expanding CDEC-WN to include within-document coreference links, designing coreference resolution systems that account for cases of partial identity between mentions, and expanding the study of the partial identity of event coreference to new domains.

Acknowledgements

References

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Cosmin Bejan and Sanda Harabagiu. 2008. A linguistic resource for discovering event structures and resolving event coreference. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Michael Bugert, Nils Reimers, Shany Barhom, Ido Dagan, and Iryna Gurevych. 2020. Breaking the

subtopic barrier in cross-document event coreference resolution. In *Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts co-located with 42nd European Conference on Information Retrieval, Text2Story@ECIR 2020, Lisbon, Portugal, April 14th, 2020 [online only]*, volume 2593 of *CEUR Workshop Proceedings*, pages 23–29. CEUR-WS.org.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Realistic evaluation principles for cross-document coreference resolution. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 143–151, 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).

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

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.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28, Atlanta, Georgia. Association for Computational Linguistics.

Zhengzhong Liu, Guanxiong Ding, Avinash Bukkittu, Mansi Gupta, Pengzhi Gao, Atif Ahmed, Shikun

Zhang, Xin Gao, Swapnil Singhavi, Linwei Li, Wei Wei, Zecong Hu, Haoran Shi, Xiaodan Liang, Teruko Mitamura, Eric Xing, and Zhiting Hu. 2020. A data-centric framework for composable NLP workflows. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 197–204, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

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).

Teruko Mitamura, Zhengzhong Liu, and Eduard H Hovy. 2017. Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track. In *TAC*.

Marta Recasens, E. Hovy, and M. A. Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121:1138–1152.

Marta Recasens, M. Antònia Martí, and Constantin Orasan. 2012. Annotating near-identity from coreference disagreements. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 165–172, Istanbul, Turkey. European Language Resources Association (ELRA).

Zhiyi Song, Ann Bies, Justin Mott, Xuansong Li, Stephanie Strassel, and Christopher Caruso. 2018. Cross-document, cross-language event coreference annotation using event hoppers. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, and Dan Roth. 2016. Revisiting the evaluation for cross document event coreference. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1949–1958, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. *LDC2006T06, Philadelphia, Penn.: Linguistic Data Consortium*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. OntoNotes Release 5.0. *LDC2013T19, Philadelphia, Penn.: Linguistic Data Consortium*.

David Wiggins et al. 1967. *Identity and spatio-temporal continuity*. Blackwell Oxford.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kristin Wright-Bettner, Martha Palmer, Guergana Savova, Piet de Groen, and Timothy Miller. 2019. Cross-document coreference: An approach to capturing coreference without context. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 1–10, Hong Kong. Association for Computational Linguistics.

A Appendix

A.1 Ethical Considerations

In our dataset construction, we follow the standard norms for ethical research involving human participants. We obtained IRB approval before starting our study. Our pilot study indicated that each HIT takes ~10-15 minutes; therefore, we set the price of individual HIT to be $2.3. Overall, we paid a fair compensation of $10.9/hour (with median pay of $16.3/hour). For each HIT, the crowd workers on Mechanical Turk have signed the informed consent form before starting the task (see A.3 in Appendix). We provided clear instructions for using our annotation tool, both within and through an instructional video. We provide positive and negative examples to illustrate event coreference to the crowd workers (see A.2 in Appendix). Our dataset is limited to the English language, specifically for text documents relating to Disasters and accidents. While we have taken specific steps to improve the quality of our dataset, there might be incorrect or missing coreference links. However, we believe that such incorrect/missing links will not create additional risks to the models trained on our dataset.

A.2 Annotation Guidelines

To explain the task of cross-document event coreference to crowd workers on Mechanical Turk, we present detailed example-based guidelines (Table 6, Table 7). Additionally, we provide crowd workers with detailed instructions to our annotation interface (Table 4, Table 5). Workers view these instructions before the start of each task and optionally during the task. In our HIT, we also link to a 1-minute video tour of our annotation interface.

In our guidelines, we only present examples of full and null coreference. While we consider membership a form of coreference (partial), we don't train the crowd workers on full and partial identity.

A.3 MTurk Consent Form

A consent form is attached to the start of each HIT. Crowd workers are required to go through the form and provide their consent before starting the task. Anonymized version of the consent form is presented in Table 8 and Table 9. We anonymize the document for the conference review process.

A.4 MTurk Qualification Test

To identify high-quality crowd workers, we design a qualification test and add it as an additional re-

quirement to solving our HITs.

A.4.1 Test Questions

In the qualification test on MTurk, we randomly select eight questions from a pool of 20 questions. Table 10 and Table 11 list all the questions.

A.4.2 Test Format

Table 12 presents the format of the qualification test used for screening crowd workers.

A.5 HIT Template

Table 13 presents our HIT layout. Our layout is simple, and all of our annotations are collected using our custom-designed annotation tool.

A.6 Follow-up Questions

Table 14 lists the four follow-up questions. We present these questions for each coreference link annotated by the crowd worker.

Instructions for using the tool

This tool can be used to select events that are the same across the two given documents.

How to open instructions
<embedded GIF>
How to annotate one pair of events
<embedded GIF>
How to delete previous annotations
<embedded GIF>
How to proceed to the next event
<embedded GIF>

At any point during the task, you can click on the "View Instructions" button to read these instructions.

What is this task about?

- Two related documents are presented side-by-side on the tool.

- A few words in both the documents are underlined and these are referred to as events.

- The task is to select events from the right document that are the same as the currently highlighted event in the left document.

How should I solve this task?

- When you first start the task, make sure you read through both the left and right documents to get an overall understanding of the two documents.

- At each step, an event is highlighted in a blue box on the left document (aka. target event). Now, your goal is to identify underlined events from the right document that are the same as the target event from the left document.

- Once you select an event from the right document (an annotation), you are presented a few follow-up questions. Make sure you answer these questions to the best of your knowledge.

- If you change your mind while answering the questions, you can click the "Cancel" button to remove your annotation.

- After you have identified all possible same events from the right document (if any), please use the "Next event" button to move to the next target event on the left document.

Table 4: Instructions as shown to the annotators on the interface.

Instructions for using the tool (contd.)

FAQs

Q: I made a mistake and incorrectly marked two events as the same. How do I correct this?
If you are still answering the follow-up questions, you can just click on the "Cancel" button. If you have already moved to the next target event, you can use the "Back" button to move back the previously finished target events.

Q: I am not sure how to respond to the follow-up questions. How should I proceed?
The follow-up questions help us understand more about your decision that two events are the same. It is important to note that the response to these questions need not always be "Yes". In fact, in many cases, you may not have enough information to respond with a definite "Yes" or "No", then please feel free to select "Not enough information".

Q: How do I decide if two events are the same or different?
We understand that this decision is not always easy. To help you with this, we compiled a bunch of examples. You can quickly glance through them using the "View Examples" button on the tool.

Q: How do I contact the authors of the task?
For any comments, feedback and/or suggestions, please use this form (XXXX). We strive to make this a great experience for you.

Table 5: Instructions as shown to the annotators on the interface. (contd)

Examples

Goal of the Task
You will help us identify the same events from different documents.

What is an event?
People use text to describe what happen(ed) in the world. These are called events in text. We often use verbs, sometimes even (pro)nouns, and adjectives as <u>events</u>. For example:

It <u>rained</u> a lot yesterday.
There was a <u>fire</u> last night.
He <u>got sick</u>.

How do we know that the two events are the same?
In the following examples (1 to 5), two events are the same.

1. When two events refer to the same thing, they should be the same in terms of meaning, or semantically identical.
   - Taken as a whole, the evidence suggests that the plan to <u>bomb</u> the Boston Marathon took shape over three months.
   - Dzhokhar Tsarnaev apologized for suffering caused by the Boston Marathon <u>bombing</u>.

2. When two events are the same, one event may be the synonym for the other.
   - A 16-year-old southern Utah boy was <u>accused</u> of bringing a homemade bomb to his high school.
   - The teen was <u>charged</u> Monday with attempted murder and use of a weapon of mass desctuction, both first-degreen felonies.

3. Sometimes one event may be the pronoun (e.g.,it) or the anaphora (e.g., this, that) of the other, when they are the same.
   - Both drones carried explosives, and no YPF ("People's Defence Units") fighters were injured in the <u>incident</u>.
   - <u>This</u> would not be the first terrorist drone strike.

4. The same events do not have to take place at the same time. In the following example, one event ("go") would happen in the future, while the other ("went") did occur.
   - The couple had been planning to <u>go</u> to Paris for a long time.
   - They finally <u>went</u> there last month.

5. Sometimes the same events are described from different perspectives. The following example refers to the exchange of the gift from two perspectives.
   - John <u>gave</u> a gift to Mary.
   - Mary <u>received</u> a gift from John.

Table 6: Examples for coreference and non-coreference, as shown to the annotators on the interface.

Examples (contd.)

In the following examples (6 to 8), two events are not the same.

6. When one event is a part of the other larger event, they are not the same.

- Following the trial of Mahammed Alameh, the first suspect in the bombing, investigators discovered a jumble of chemicals, chemistry implements and detonating materials.
- The explosion killed at least five people. ("bombing" refers to the entire process which starts with making a bomb and ends with destructions, damages and injuries, while "explosion" is a smaller event that occurs in that processes)

7. Two events are not the same even if they are the same semantically. The first example refers to the general bomb-making process, while the second one indicates a particular bomb-making event that took place in the garage.

- They obtained the online manual of bomb-making. (general bomb-making process)
- They made a bomb in the garage. (specific bomb-making event that happened in the specific place)

8. When one event consists of, or is a member of the other event, they are not the same. The first example refers to the specific death of a 44-year-old man, while the second one refers to the deaths of 305 people.

- The government announced that a 44-year-old man died from the COVID. (death of a 44-year-old man)
- There are more than 14,300 confirmed COVID cases, and 305 people have died. (deaths of 305 people)

Table 7: Examples for coreference and non-coreference, as shown to the annotators on the interface. (contd)

Consent From

This task is part of a research study conducted by XXX at XXX and is funded by XXX.

Purpose
The goal of this study is to collect datasets of coreference-labeled pairs sampled from public online news articles through the help of crowd workers.

Procedures
You will be directed to a website implemented by the research team to complete the task. You will be asked to read upto 3 pairs of articles. For each pair of articles, you will need to label pieces of text that refer to the same event, and answer additional questions about your labeling. Labeling one pair of articles whose length sums up to 40 sentences is expected to take around 15 minutes.

Participant Requirements
Participation in this study is limited to individuals age 18 and older, and native English speakers.

Risks
The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during other online activities.

Benefits
There may be no personal benefit from your participation in the study but the knowledge received may be of value to humanity.

Compensation & Costs
For this task, you will receive between $2 to $3 for annotating each pair of articles. The exact reward for each pair depends on the length of corresponding articles. You will not be compensated if you provide annotations of poor quality.
There will be no cost to you if you participate in this study.

Future Use of Information and/or Bio-Specimens
In the future, once we have removed all identifiable information from your data (information or bio-specimens), we may use the data for our future research studies, or we may distribute the data to other researchers for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Confidentiality
The data captured for the research does not include any personally identifiable information about you except your IP address and Mechanical Turk worker ID.
By participating in this research, you understand and agree that XXX may be required to disclose your consent form, data and other personally identifiable information as required by law, regulation, subpoena or court order. Otherwise, your confidentiality will be maintained in the following manner:

Table 8: Consent Form attached to each of our HITs. We anonymize the document for the conference review process.

Consent From (contd.)

Confidentiality (contd.)
Your data and consent form will be kept separate. Your consent form will be stored in a secure location on XXX property and will not be disclosed to third parties. By participating, you understand and agree that the data and information gathered during this study may be used by XXX and published and/or disclosed by XXX to others outside of XXX. However, your name, address, contact information and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by XXX. Note that per regulation all research data must be kept for a minimum of 3 years. The Federal government offices that oversee the protection of human subjects in research will have access to research records to ensure protection of research subjects.

Right to Ask Questions & Contact Information
If you have any questions about this study, you should feel free to ask them by contacting the Principal Investigator now at XXX, XXX, or by phone at XXX, or via email at XXX. If you have questions later, desire additional information, or wish to withdraw your participation please contact the Principal Investigator by mail, phone or e-mail in accordance with the contact information listed above.
If you have questions pertaining to your rights as a research participant; or to report concerns to this study, you should contact the XXX at XXX. Email: XXX. Phone: XXX or XXX.

Voluntary Participation
Your participation in this research is voluntary. You may discontinue participation at any time during the research activity. You may print a copy of this consent form for your records.

I am age 18 or older. ☐ Yes ☐ No

I have read and understand the information above. ☐ Yes ☐ No

I want to participate in this research and continue with the task. ☐ Yes ☐ No

Table 9: Consent Form attached to each of our HITs. We anonymize the document for the conference review process. (contd)

#	Text	Answer	Type
1	A 500lb bomb packed in the Cavalier is <u>detonated</u> with a remote trigger. The <u>explosion</u> tears through Market Street.	yes	Synonym
2	The <u>death</u> toll of the Omagh bomb blast in Northern Ireland has risen to 29 following the <u>death</u> of a man in hospital.	no	Member
3	Ahmed al-Mughassil was <u>arrested</u> in Beirut and transferred to Riyadh, the Saudi capital, according to the Saudi newspaper Asharq Alawsat. The Saudi Interior Ministry and Lebanese authorities had no immediate comment on the <u>capture</u>.	yes	Synonym
4	The blast didn't cause the <u>destruction</u> its planners intended. But it <u>opened up</u> a multi-story crater in the building, injured more than 1,000 people and ultimately killed six.	no	Member
5	March 4, 1998 - Four defendants, Salameh, Ayyad, Abouhalima, and Ajaj, are convicted. They are <u>sentenced</u> to prison terms of 240 years each. In 1998, the sentences were vacated. In 1999, the men were <u>re-sentenced</u> to terms of more than 100 years.	no	Unrelated
6	Perhaps the only early clues to emerge on an early quiet second day of the Boston Marathon bombing <u>investigation</u> - from the ATF and the FBI and the Boston police, from anonymous law enforcement officials and doctors pulling ball bearings out of victims limbs - concern the Boston bombs themselves. A similar scene played out in the Boston suburb of Newton, where a bomb used a robot to <u>investigate</u> a suspicious object that turned out to be a circuit board.	no	Member
7	As of Tuesday morning, jurors began reviewing evidence and witness testimony, which will play a role in helping them divide Dzhokhar Tsarnaev's <u>guilt</u> on each of the 30 charges he faces. A key issue for jurors - both in the guilt phase and later the penalty phase if Tsarnaev is <u>convicted</u> - will be whether the jurors see Tsarnaev as an equal partner with his old brother, Tamerian Tsarnaev, in the Boston Marathon bombing and the violent events that followed.	yes	Synonym
8	Though <u>building</u> the bomb was relatively easy, the experts say, it was not by any means free of danger. The bulkiest part of the bomb, they say, was extremely stable and could only have been touched off with a tremendous kick, like that provided by nitroglycerine. Making the nitroglycerin, blending some of the chemicals, was the trickiest part of the <u>process</u>.	yes	Synonym
9	An ongoing Somali <u>offensive</u>, backed by the U.S. and an African Union peace-keeping force has recaptured territory from al Shabaab in south-central So-malia, but has not eliminated al Shabaab's ability to conduct VBIED attacks. U.S.-backed Somali ground <u>operations</u> along with improved counter-VBIED capabilities among Somali forces may have slightly decreased VBIED attacks between November 2017 and January 2018.	yes	Synonym
10	According to the United Nations, more than 2.3 million Venezuelans have left their country in recent years. Increasingly they are leaving with no money and are <u>traveling</u> on foot across South American countries like Colombia, Ecuador and Peru, in dangerous <u>journeys</u> that can take several weeks.	no	Member

Table 10: Examples used with the qualification test on Mechanical Turk. For each paragraph with two highlighted events, we ask the question, "In the above paragraph, are the highlighted events the same?". The crowd worker has to select one of the "Yes" or "No" options.

#	Text	Answer	Type
11	Spain's King Juan Coarlos and Queen Sofia traveled to their summer residence in Majorca Saturday just two days after a <u>bombing</u> blamed on Basuqe separatists <u>killed</u> two policemen on the resort island.	no	Member
12	Yahoo Inc. is preparing to <u>lay off</u> between 600 and 700 workers in the latest shakeup triggered by the Internet company lackluster growth. Employees could be notified of the <u>job cuts</u> as early as Tuesday, according to a person familiar with Yahoo's plans.	yes	Synonym
13	A man shot and killed by police officers during a burglary here early Monday was identified by law enforcement authorities as the suspect in a string of five shooting <u>deaths</u> in South Carolina over the last 10 days. Sheriff Bill Blanton of Cherokee Country, S.C., where the <u>killings</u> took place, confirmed Monday evening that the authorities had been seeking the man killed in the burglary, Patrick T. Burris, a felon with a long record who had served seven years in prison and was paroled in April.	yes	Synonym
14	Staff Sgt. Robert Bales offered a tearful <u>apology</u> Thursday for gunning down 16 unarmed Afghan civilians inside their homes but said he still could not explain why he had carried out one of the worst U.S. war crimes in years. The unsworn <u>statement</u> from Bales, 40, came on the third day of hearing to determine whether he should ever be eligible for parole in the March 2012 Massacre.	yes	Synonym
15	In January two men were <u>hanged</u> after being convicted of involvement in protests, and in May, four Iranian Kurds and another man accused of terrorism were <u>executed</u>.	no	Unrelated
16	The Dow Corning Corporation filed for <u>bankruptcy</u> protection in a Federal court in Bay City, Michigan. Dow Corning said that seeking the protection of the <u>bankruptcy</u> court was the only way it could devise an enforceable plan to deal with the claims against it.	no	Realis
17	The UN report accused both Israel and Palestinian armed groups of commiting <u>war</u> crimes during the three-week <u>war</u> in Gaza that erupted on December 27, killing some 1,400 Palestinians and 13 Israelis.	no	Realis
18	A judge has ordered the surviving children of the Rev. Martin Luther King Jr. and Coretta Scott King to hold a shareholder's <u>meeting</u> to discuss their father's estate. The three siblings are the sole shareholders, directors and officers of a company that manages their father's intellectual property, but they have not <u>met</u> for an annual shareholder's meeting since 2004.	no	Realis
19	The first <u>attack</u> was a failure, but if the report is accurate, then it signals a dangerous new terror threat. The report showed pictures of the remains of a homemade <u>attack</u> drone.	no	Realis
20	A key issue for jurors - both in the guilt phase and later in the penalty phase if Tsarnaev is convicted - will be whether the jurors see Tsarnaev as an equal partner with his older brother, Tamerlan Tsarnaev, in the Boston Marathon <u>bombing</u> and the violent events that followed. Taken as a whole, the evidence suggests that the plan to <u>bomb</u> the Boston Marathon took shape over three months.	yes	Realis

Table 11: Examples used with the qualification test on Mechanical Turk. For each paragraph with two highlighted events, we ask the question, "In the above paragraph, are the highlighted events the same?". The crowd worker has to select one of the "Yes" or "No" options. (contd)

Screening Test

In this test, we ask you to identify whether two events (**highlighted** in each paragraph) indicate the same thing or not. Read each paragraph carefully and answer the question by selecting the appropriate option, *Yes* or *No*.

In total, you are presented with 8 questions and the time limit for this test is 20 minutes.

Note: It is important you do this test on your own because our HITs are similar to the questions presented in this test. For your reference, we provide five examples below,

He **died** of injuries from the accident. His friends were all saddened to hear his **death**.
Question: In the above paragraph, are the highlighted events the same?
Answer: Yes (both words, **died** and **death** indicate the person's death)

The suspect was **shot** and killed in the **raid** by the armed officers.
Question: In the above paragraph, are the highlighted events the same?
Answer: No (**shot** happened during the **raid**)

The couple had been planning to **go** to Paris for a long time. They finally **went** there last month.
Question: In the above paragraph, are the highlighted events the same?
Answer: Yes (The two events do not have to take place at the same time. Here, **go** would happen in the future, and **went** did occur.)

John **gave** a gift to Mary. Mary **received** a gift from John.
Question: In the above paragraph, are the highlighted events the same?
Answer: Yes (Same events described from different perspectives.)

Following the trial of Mahammed Alameh, the first suspect in the **bombing**, investigators discovered a jumble of chemicals, chemistry implements and detonating materials. The **explosion** killed at least five people.
Question: In the above paragraph, are the highlighted events the same?
Answer: No (One event is part of the other larger event. **bombing** refers to the entire process which starts with making a bomb and ends with destructions, damages and injuries, while **explosion** is a smaller event that occurs in that processes.)

Q1. ....
☐ Yes ☐ No
Q2. ....
☐ Yes ☐ No
...

Table 12: The template used in the qualification test to screen annotators. In addition to instructions and examples, we present eight yes/no questions.

Annotating Event Coreference in News Articles

In this HIT, you will be using our tool to perform the task. For a short tutorial on using our interface, see this 1 minute video: XXX. This HIT contains the following two steps,

- Visit the URL provided below to perform the task.

- At the end of the task, you will be provided a secret code. To submit this HIT, copy the secret code and paste it into the box provided below. Note that the secret code is unique for each task.

Link to the task: XXX

Fill in the secret code

Paste the secret code provided at the end of the task into the text box (*required)

```
                                                            
                                                            
                                                            
```

Table 13: The template used for each Human Intelligence Task (HIT) on Mechanical Turk.

Place: Do you think the two events happen at the same place?

☐ Exactly the same   ☐ The places overlap   ☐ Not at all   ☐ Cannot determine

Time: Do you think the two events happen at the same time?

☐ Exactly the same   ☐ They overlap in time   ☐ Not at all   ☐ Cannot determine

Participants: Do you think the two events have the same participants?

☐ Exactly the same   ☐ They share some participants   ☐ Not at all   ☐ Cannot determine

Inclusion: Do you think one of the events is part of the other?

☐ Yes, the left event is part of right one   ☐ Yes, the right event is part of left one
☐ No, they are exactly the same   ☐ Cannot determine

Table 14: Follow-up questions used for each annotated coreference link.