Abstractive Event Analysis of Armed Conflicts: Introducing the UCDP-AEC Dataset

Étienne Simon[*1] **and Helene Bøsei Olsen**[*1] **and Ramón Carreño**[1] **and Rahul Mishra**[1]
Nikolay Arefyev[1] **and Mert Can Yilmaz**[2] **and Lilja Øvrelid**[1] **and Erik Velldal**[1]
[1] University of Oslo, [2] Uppsala University

Abstract

This paper introduces a new dataset of document-level event annotations in the domain of armed conflict. By augmenting the event database from the Uppsala Conflict Data Program (UCDP) with source documents identified in public web archives, we create the *UCDP Abstractive Event analysis Corpus* (UCDP-AEC). While a large part of research on information extraction is focused on extracting text spans, real-world use cases often require inferring more high-level information that is not necessarily explicitly mentioned in texts. UCDP-AEC differs from traditional event extraction datasets in that the document-level annotations do not correspond to mere text spans of the input, but capture expert-interpreted and often implicit information. With more than 10 000 documents, UCDP-AEC is of comparable size to the largest human-annotated traditional event extraction datasets. We also report preliminary experimental results for various generative approaches, by fine-tuning both decoder models and existing event argument extraction models that require minimal adaptation to our abstractive formulation of the task.

1 Introduction

For several decades, the Uppsala Conflict Data Program (UCDP) has compiled and released datasets on quantitative armed conflict event data (Sundberg and Melander, 2013; Davies et al., 2024), designed primarily to enable research on the causes, dynamics, and consequences of organised violence, as well as supporting research projects such as predicting conflict escalation (Hegre et al., 2022). UCDP's event database, the Georeferenced Event Dataset (GED), encodes rich information about incidents related to organised violence, including casualty counts, involved actors, location, time, and more. An example of (a relevant subset of) the data fields and corresponding values for a given input document is shown in Figure 1. Carried out through manual coding by human experts, the process is highly resource-demanding and time-consuming, making it challenging to scale to larger data sources and update in near real-time. Moreover, the continual exposure to accounts of violence can cause emotional and psychological distress for those involved in manual coding. For several reasons, automated machine-coding of events could offer valuable assistance in the annotation process.

2014-05-25 Islamabad: At least eight militants and two soldiers were killed in a gun battle in Pakistan's restive northwest tribal area, a security official said. The incident occurred late yesterday in Landikotal area of Khyber tribal agency. "Two soldiers embraced shahadat (martyrdom) and three others injured this evening in an exchange of fire with terrorists in Landikotal, Khyber," The Dawn said. Three soldiers were also injured, according to the security official. Yesterday was a deadly day for forces as six soldiers and a civilian were killed in three separate bombings in Pakistan, including twin blasts in the heavily guarded capital. The soldiers were killed in Mohmand agency which borders Khyber.

Actors	**Side A:** Government of Pakistan **Side B:** TTP
Date	**Start:** 2014-05-24 **End:** 2014-05-24
Location	**Country:** Pakistan **Subregion:** Khyber agency **Region:** Federally Administered Tribal Areas **Point:** Landi Kotal town
Deaths	**Side A:** 2 **Side B:** 8 **Civilian:** 0 **Unknown:** 0 **Low:** 10 **High:** 10

Figure 1: Sample document with UCDP annotations. The document reports on an armed conflict event and includes the publication date and location. The annotations specify the two conflict actors, the start and end dates of the event, four levels of location detail, and death counts for each side, civilians and unknown individuals. Estimated total deaths are given as low (conservative) and high (maximum reliable) figures.

The UCDP GED example above illustrates a

[*]Equal contribution.

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 104–119, 2025

more general challenge acknowledged in other recent work on event extraction (EE) and information extraction (IE) for real-world applications, which is the need to move beyond span-based annotations. For example, Olsen et al. (2024) refer to an 'abstraction gap' when discussing the differences between the text-bound annotations common in most NLP datasets for EE versus the higher-level information documented in socio-political event databases. In such databases, coders aim to represent what actually happened in the real world, drawing on contextual understanding and domain expertise. As demonstrated in Figure 1, important event information often can not be found as a single continuous text span. Instead they must often be inferred from indirect references (e.g., "soldiers" or "militants" as proxies for specific groups), resolved through temporal reasoning (e.g., interpreting "yesterday" relative to the publication date of the source document), aggregated from scattered mentions (e.g., casualty counts reported across multiple sentences), or produced through numerical reasoning (e.g., combining the casualty counts across parties). Locations and actors can be described at varying levels of granularity and may require entity-linking to canonical forms, and the dynamic nature of conflicts means new actors and conflicts continually emerge, introducing important temporal variations. On the basis of considerations like this, Simon et al. (2024) advocate for a shift from an "extractive" to an "abstractive" view of the EE task, proposing to use generative approaches for capturing more higher-level event information. Similarly, Sharif et al. (2024) also show how event information can be scattered across a document or only implicitly stated, and propose a generative formulation of event argument extraction.

While the recognition of the challenges associated with real-world applications is growing, the majority of traditional event extraction datasets within Natural Language Processing (NLP) are annotated in the extractive paradigm, capturing only text-bound facts (Doddington et al., 2004; Song et al., 2015; Li et al., 2021). On the other hand, while there is an abundance of high-quality event databases developed within social and political sciences (Sundberg and Melander, 2013; Raleigh et al., 2010; Chenoweth et al., 2019; Salehyan et al., 2012; Turchin, 2012), the source documents are generally not made available (Olsen et al., 2024), representing a missed opportunity for training machine-learning models for domain-specific applications. As a case in point, the event annotations of the UCDP GED are publicly available, but the corresponding source documents used by the human coders are not licensed for redistribution in the original database.

Contributions To bridge NLP modelling with socio-political event coding, we introduce the new dataset UCDP-AEC, a conflict event dataset for abstractive event analysis built from UCDP GED, augmented with pointers to the underlying source texts.[1] The dataset covers 11 363 documents describing armed conflict events and links expert-annotated event records to their corresponding source documents. UCDP-AEC focuses on a single event type, and every document in the dataset contains a single, relevant event of the same type, where each event is annotated with a fixed set of 14 fields. Reflecting the complexities of real-world event reports, the annotations often demand numerical and temporal reasoning, world- and domain knowledge, contextual integration, and entity linking to canonical forms. We refer to this challenging task of structured prediction as abstractive event analysis.

We show how a substantial subset of the documents from the UCDP GED database can be identified and retrieved from the public web archives of HPLT v2 (High Performance Language Technologies; Burchell et al., 2025), and that distributing the corresponding document IDs provides a way to establish an open dataset for event analysis that can be used by the NLP community, ultimately also benefitting the peace and conflict studies. Finally, we report on the first experimental results for a suite of generative approaches on the new dataset.

2 Related Work

Relation to other NLP tasks Automatically extracting UCDP-AEC event structures combines elements from multiple NLP tasks involving representing and understanding event-centric information in text. The task includes aspects of temporal reasoning, text summarisation, and entity-linking, but is best understood as a form of event extraction (or perhaps more narrowly the associated subtask of event argument extraction) though in a way that diverges from typical NLP formulations.

While event extraction is a widely studied task in NLP, the current annotation paradigm and bench-

[1]The dataset alongside evaluation scripts and baseline models are available at https://github.com/ltgoslo/ucdp-aec.

mark datasets do not correspond well to the types of events typically encoded in socio-political event databases. Even though both tasks aim to derive structured event information from unstructured text, existing NLP datasets typically contain sentence-level event information that corresponds to substrings in the input text (Doddington et al., 2004; Song et al., 2015), while socio-political event recordings typically capture higher-level event information inferred from the entire document, possibly in combination with domain knowledge. For instance, in the news article shown in Figure 1, Side B is described in the text as "militants" and "terrorists". However, a socio-political event database is expected to convey the specific group involved ("TTP" in this case) based on contextual information such as time and location.

The datasets from the Automatic Content Extraction (ACE; Doddington et al., 2004) program have been highly influential in the development and evaluation of event extraction systems in NLP, inspiring further development of several datasets with more detailed processing of entities and events in the ERE (Entities, Relations and Events) dataset (Song et al., 2015), document-level event extraction from Wikipedia articles in the WikiEvents dataset (Li et al., 2021), from news articles referenced in Wikipedia articles in the DocEE dataset (Tong et al., 2022), and domain-specific content (Sun et al., 2022; Satyapanich et al., 2020).

Some recent works have taken a step towards removing the reliance on surface forms for event extraction. For the task of Chinese financial event extraction, Zheng et al. (2019) do not extract trigger words, but still rely on text spans for argument extraction. More recently, Sharif et al. (2024) challenge the traditional span-based approach and introduce DiscourseEE, an online health discourse dataset annotated with explicit, implicit, and scattered arguments. Apart from DiscourseEE, a related work to ours is the early version of the task introduced at the Message Understanding Conferences (MUC; Sundheim, 1992),with the most well-known being the MUC-4 dataset, which is annotated with rich, high-level, and fine-grained information in the form of event templates at the document level. Recent work has renewed focus on the MUC-4 data, however, existing work is largely limited to argument roles that correspond directly to explicit text spans from the source document (Du et al., 2021a,b; Gantt et al., 2024).

Concurrent to our work Semnani et al. (2025) introduce LEMONADE, a large-scale multilingual dataset for abstractive event extraction based on a re-annotated subset of the ACLED database (Raleigh et al., 2010). It spans 25 closely related socio-political event types over a 13-month period and includes an abstractive entity linking task to a curated domain-specific entity database. Each event type is associated with a specific schema, which supports a variety of argument types, including categorical fields, booleans, and integers, with many roles limited to values in a pre-defined database, and some re-annotated to only reflect information explicitly present in the text. In contrast, UCDP-AEC focuses on a single domain, fatal armed conflicts, grounded in the expert coding of UCDP GED, with a fixed 14-field schema that is always filled, often requiring inference over implicit or scattered evidence. UCDP-AEC contains multiple open domain roles, such as exact event dates or fine-grained location details, without requiring that the gold value appear verbatim in the document.

Largely independently of the NLP community, the political science community has seen the development of their own event extraction systems (Schrodt et al., 1994; Norris et al., 2017; Halterman et al., 2023). However, these systems tend to rely on older rule-based models prone to overcounting events and producing numerous false positives, which has limited their use in political science research (Raleigh et al., 2023). Not relying on learned models and annotated datasets also makes most systems less flexible and adaptable. Some initial work making training data available is worth mentioning though, such as GLOCON (Global Contentious Politics Dataset; Duruşan et al., 2022) and CEHA (Conflict Events in the Horn of Africa region; Bai et al., 2025), with the latter reaching 500 events, but with annotations for event detection and classification only.

Generative approaches to event extraction In recent years, the NLP field has seen a transition from traditional sequence labelling approaches for event extraction (Wadden et al., 2019; Nguyen et al., 2016) towards generative methods where the task is framed as a structured text generation task (Lu et al., 2021; Li et al., 2021). Generative methods present a promising path to move beyond span-based annotations by allowing models to generate structured event representations directly from text, and may be suitable for tasks requiring inference from underspecified information and long document contexts,

such as for UCDP-AEC.

However, the current landscape of generative approaches to event analysis still remains closely tied to the extractive paradigm (Simon et al., 2024). As the architecture and training objective of generative models are intended for producing free-form text, they provide a less natural fit to the extractive setting, where the goal is to identify and reproduce exact text spans from the source text. As a consequence, implementations of generative event extraction models often include additional constraints to ensure that the generated string appears in the input, such as constraint decoding in the Text2Event model (Lu et al., 2021), or the pointer mechanism in the BART-Gen model (Li et al., 2021).

Recently, several decoder-only models have been developed for generative event extraction, such as DeepStruct (Wang et al., 2022), InstructUIE (Wang et al., 2023), YAYI-UIE (Xiao et al., 2023), a Baichuan2 model (Yang et al., 2023), and LLMEE (Chen et al., 2024). Notably, except for DeepStruct and YAYI-UIE, most works, including LLMEE, do not involve any task-specific fine-tuning, relying solely on in-context learning. While these models, in addition to some encoder-decoder models such as DEGREE (Hsu et al., 2022), do not explicitly constrain their outputs to input tokens during prediction, they remain in the extractive paradigm through evaluation, where the predictions are converted into text spans for evaluation against benchmarks such as ACE (Doddington et al., 2004).

3 The UCDP GED Conflict Database

An event in the UCDP GED database is defined as an incident where armed force was used by an organised actor resulting in at least one direct death at a specific location and date. In this section, we provide some background on the UCDP GED annotations, starting with the annotation process itself, followed by a description of the relevant data fields that we include in the derived UCDP-AEC dataset described in Section 4.

3.1 Annotation Process

The UCDP GED database follows a rigorous, multi-step process to document each instance of fatal organised violence as a distinct event (Sundberg and Melander, 2013). Data collection begins with global search queries in the Dow Jones Factiva aggregator, yielding tens of thousands of news reports annually. These reports are then supplemented with local and social media sources, as well as information from NGOs and international organisations. Reports are meticulously examined by a team of around 15 analysts, who hold advanced degrees in peace and conflict studies or related fields and possess extensive regional and conflict-specific expertise. They identify and code events according to strict methodological criteria, assigning detailed spatial, temporal, and actor identifiers (Högbladh, 2023). The individual event entries are subsequently categorised into three mutually exclusive types of violence: state-based, non-state, or one-sided.

A candidate event is only included in the final event database after several stages of validation, including both manual review and automated consistency tests. In sum, this comprehensive coding approach aims to provide a high level of accuracy and reliability in capturing global patterns of organised violence, and updated versions of UCDP GED are published annually.[2] The current full GED contains more than 500 000 events with a good coverage of all world conflicts from 1989 onwards.

3.2 Data Fields

The UCDP GED database contains a wide range of different data fields, which, beyond event information, include source metadata, estimates of annotation reliability, identifiers linking to other databases, and more. In this work, however, we limit the description to the 14 fields selected for the new UCDP-AEC dataset as shown in Figure 1 and in Appendix C.[3] We group the event fields into the following four categories, each with associated challenges.

Actors The two fields "Side A" and "Side B" corresponds to the two parties of a conflict. "Side A" is always an organised actor, while "Side B" can be either an organised actor or "Civilians" when the first side kills people indiscriminately. All mentions of the same actor are labelled with a unique canonical name, rather than how they are referred to in the text.

Dates The two fields "Start Date" and "End Date" reference a time range during which the event occurred. As UCDP aims to record events with a day-level precision, the start and end dates will typ-

[2] https://ucdp.uu.se/downloads/

[3] For the complete list and description of the fields in UCDP GED, see Högbladh (2023).

ically be identical for most events.[4] Both relative time references – also known as temporal deixis – and the publication date and time are used to identify the start and end date of an event, as shown in Figure 1.

Locations The location of the event is described by four fields with an increasing level of precision. The first field, country, is always provided and indicates where the event took place. Each country is divided into region and subregion, which may be left empty if the location does not align with administrative areas. The point field refers to a city-level location, such as a town or a specific site, such as "Jabal al Akrad mountain", and may also be empty if the location is unclear.

Deaths Six distinct fields are used to describe the casualties associated with the conflict event. Specifically, the fields represent the number of casualties on each side of the conflict, the number of civilian casualties, and the number of deaths that cannot be attributed to any of those categories. In the case of one-sided violence – when "Side B" is "Civilians" – the number of civilian casualties is reported in the field "Deaths Civilians". Uncertainty in casualty reporting is reflected in the fields "Deaths Low" and "Deaths High", representing the most conservative estimate and the highest reliable estimate, respectively, for the total number of fatalities reported in the source text.

4 The UCDP-AEC Dataset

In this section, we describe the construction of the new UCDP-AEC dataset designed for machine learning applications based on the socio-political database UCDP GED. We begin by outlining the event filtering process, which ensures a one-to-one correspondence between events and documents. We then outline how we used publicly available web archives to make the data available. Finally, we present some descriptive statistics of UCDP-AEC, followed by a discussion of relevant challenges.

4.1 Event Selection

Since the UCDP GED is annotated for political science research rather than machine learning applications, some filtering was necessary to adapt the data for automated event analysis. In particular, some events are referenced by multiple source articles, some of which reference multiple events. In practice, the annotators are aware of which parts of a document have already been coded in the database, ensuring that there is no confusion. However, from an NLP perspective, this would require modelling the problem as mapping a set of documents to a set of events without a simple one-to-one correspondence between them. To simplify the problem, we only include events described by a single source and sources that describe a single event. Furthermore, because of the limited amount of non-English documents in the selected subset, we restricted UCDP-AEC to English-language documents. Additional details on data filtering are provided in Appendix A.

4.2 Web Archive Identification of Sources

HPLT matching The documents used to create the GED are not public. Since UCDP relies in part on large news networks for its source documents, we cannot indiscriminately share them. Instead, we distribute only the IDs of documents that can be found freely on the web. To this end, we make use of the recently released HPLT v2 dataset,[5] which consists of texts extracted from web crawls provided by two major web crawling initiatives, Common Crawl and the Internet Archive. It contains 21 billion documents released under a CC0 licence. Using the MinHash algorithm (Broder, 1997), we selected the UCDP documents having an approximate Jaccard similarity of at least 0.5 with an HPLT document.

Human evaluation of HPLT matching The search for documents in HPLT web crawls introduces some biases. In particular, the older a document is, the more likely it is to have been crawled. We recovered only 11 documents from 2023 and none from 2024, despite thousands being annotated in UCDP during those years.

Furthermore, the MinHash document matching used to replace the original UCDP documents with their HPLT version is not exact. It is not uncommon for news providers to update their articles as new information becomes available. These different versions of the same article have a high syntactic similarity – as measured by their Jaccard index – but can convey different event information – for example, when the number of deaths is updated. To assess the quality of the HPLT substitution, we performed a manual comparison of HPLT documents with their original UCDP documents.

[4] In some cases, the precise date remains uncertain due to vague temporal references in the text, such as "last week". In such cases, the start and end dates will differ by a week.

[5] https://hplt-project.org/datasets/v2.0

Split	Train	Validation	Test
Documents	10 064	651	648
Period	< 2021	2021	≥ 2022

Table 1: Number of documents and years in each split.

Three annotators assessed a total of 130 documents and found 96.9% documents were a perfect match, with 98.5% containing the same information for 3 out of the 4 categories of fields in the event records (Actors, Date, Location, Deaths).

4.3 Dataset Statistics and Analysis

UCDP-AEC consists of 11 363 document–event pairs, where each document is annotated with a single event record with the 14 fields described in Section 3. These fields vary in how constrained their value spaces are, a property we call domain openness. In this subsection, we describe the dataset splits and document length, as well as challenges associated with each field with a particular focus on how open or closed the value space is for each role.

Domain openness We define a closed-domain field as one whose admissible values make up a relatively small, and mostly fixed set that recurs frequently in the corpus (e.g., country names). In contrast, an open-domain field is characterised by a large or evolving set of admissible values, often including many items not seen during training (e.g., event-specific dates). To quantify this property, we report for each field in Table 2 the three following metrics: *value density* $\frac{|\mathcal{D}|}{|\mathcal{Y}|}$ which denotes the mean number of instances per unique value. Here, lower values indicate a more open domain. *Unique value overlap* $\frac{|\mathcal{Y}_{\text{test}} \cap \mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|}$ denotes the proportion of unique test values that also occur in the training set. Finally, *instance-level overlap* $\frac{|\mathcal{D}_{\text{test}} \cap \mathcal{D}_{\text{train}}|}{|\mathcal{D}_{\text{test}}|}$ measures the proportion of test instances whose value is observed during training. We describe these metrics in more detail in Appendix B.

Standard splits To mimic the complexity of modelling a dynamic domain, where new conflicts and violent groups arise and relationships between entities change, we provide standard splits based on temporal information, considering both the occurrence and reporting times of events. Events from 2021 are taken for validation, older events are used for training, while the most recent events, from 2022 and later, are reserved for the test split. See Table 1 for the size of the respective splits.

Field	Density	Unique Overlap	Instance Overlap
Source article	1.00	0.00%	0.00%
Source date	1.76	0.00%	0.00%
Side A	27.31	71.74%	94.91%
Side B	26.24	60.24%	80.25%
Date Start	3.17	0.00%	0.00%
Date End	3.15	0.00%	0.00%
Location Country	129.12	93.18%	99.54%
Location Region	18.78	81.09%	83.33%
Location Subregion	6.49	51.30%	60.80%
Location Point	2.75	27.76%	40.74%
Deaths Side A	183.27	83.33%	99.54%
Deaths Side B	155.66	100.00%	100.00%
Deaths Civilian	195.91	94.74%	99.85%
Deaths Unknown	241.77	100.00%	100.00%
Deaths Low	113.63	90.00%	99.54%
Deaths High	97.96	90.24%	99.38%

Table 2: Per field statistics showing the value density, the unique value overlap between test and train set, and the instance-level overlap between test and train.

Document length The average length of the 11 363 documents in UCDP-AEC is 315 words (white-space separated / non-tokenised). However, there is a significant variation in document length, ranging from as few as 26 words to 1 781 words. This wide range reflects the variation in source materials used to collect armed conflict events, from short summaries to longer regional reports.

Location distribution The dataset covers events that occur in 88 countries. Figure 2 shows the temporal shifts in the geographical distribution between the splits. For example, the training set is dominated by events occurring in Syria, the validation set (2021) shows increased violence in countries such as the Philippines and Myanmar, and the test set reflects the heightened violence in Nigeria alongside the further invasion of Ukraine. With respect to openness, illustrated in Table 2, the set of possible country values is, to a large degree, limited and fixed. With only a few countries in the test set that are not present in the training set, the country field should be considered a relatively closed domain. There is, however, a relationship between location granularity and domain openness, where Subregion can be regarded as semi-open, while Region and Point tend to be highly event-specific, with a test

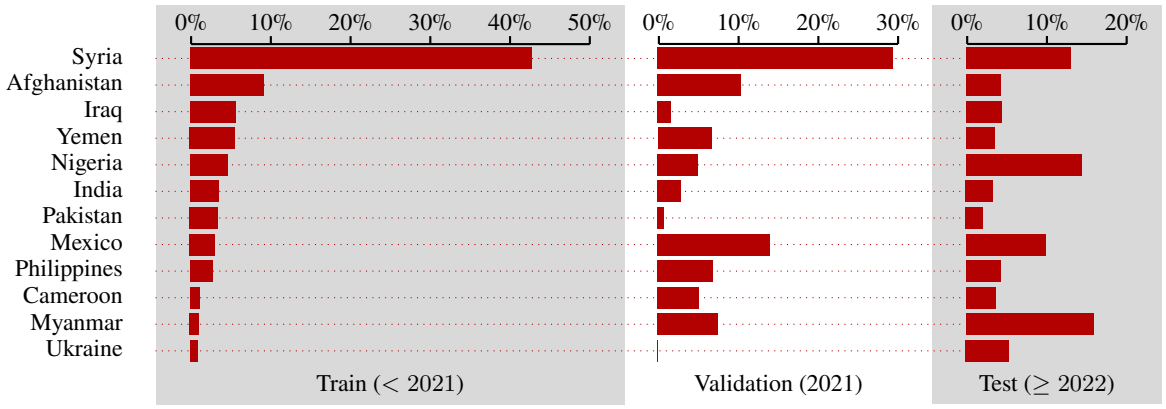

Figure 2: Repartition of events per country in each split for the 12 most eventful countries. A clear increase in conflict events can be seen in Nigeria and Myanmar, showcasing how the dataset reflects the changing dynamics of real-world conflict over time.

set containing many unseen values.

Actors The Actor roles often feature governments, organisations, or group names that are moderately recurrent, but also include many entities that appear only in specific events. Due to the dataset's temporal splits, a notable proportion of actor names in the test set do not appear in the training set, mirroring the involvement of new actors in ongoing and emerging armed conflicts. For example, the test set includes events involving the Indigenous People of Biafra (IPOB), a Nigerian separatist group that started engaging in violent attacks in 2021.[6] Several test documents labelled with IPOB contain references to members of the Eastern Security Network (ESN), which is the paramilitary wing of IPOB. Identifying the two surface forms as references to the same entities is a particularly difficult challenge.

Date Start date and end date are by design fully open (0% unique-value overlap) due to the temporal partitioning of the splits. Since UCDP-AEC only includes single-day events, 91.4% of events have matching start and end dates as expected. When the start and end dates differ, it indicates uncertainty or imprecision in the source document about the exact date of the event.

Deaths Death-related fields are largely a closed domain with high density and overlap. However, identifying the correct values from text may still be challenging. Death counts can be expressed in various ways, such as explicit numbers, ranges (e.g., "between 15 and 20"), or approximations (e.g., "dozens killed"), and require mapping to the affected individuals or parties. In the deaths category, the low and high estimates are identical in 87.1% of the cases, again indicating that the exact number is often reported. However, for 7.3% of the events, the best estimate is equal to 0 – that is the sum of deaths on side A, side B, civilians and unknown – implying that the annotator questioned the truthfulness of the reported event.

5 Experiments

This section reports on the first preliminary experiments on training and evaluating models for abstractive event analysis on UCDP-AEC.

5.1 Evaluation Metrics

Unlike classical event extraction, which only assigns roles to arguments that are explicitly mentioned in the input text, all fields in UCDP-AEC are assigned a value in all instances. Even when not explicitly stated in the document, argument values are deducted from the context. Since this is an information extraction task where the goal is to abstract away surface form differences, we evaluate model predictions using exact match accuracies for each field. This contrasts with the usual F_1 metrics reported in the classical task, where roles are not always attached to an event, as extractive models may predict excess or insufficient roles.[7] We average the accuracies of different fields over the same category. For example, we report location accuracy by averaging the four accuracies over country, region, subregion and point. Furthermore, we report

[6] https://ucdp.uu.se/actor/6287

[7] For a broader discussion on why traditional F_1 is not applicable and supplementary alternatives we considered for different categories see Appendix F.

a global accuracy aggregate across categories.[8] In the case of actors, the aggregate accuracy is invariant to a permutation of "Side A" and "Side B". If those were predicted in the opposite order of the gold annotation, the values of "Deaths Side A" and "Deaths Side B" are also inverted.

While predicting a unique symbolic representation is central to most information extraction tasks, generative models can predict semantically correct output that does not conform to the expected output (e.g. "USA" instead of "Government of United States of America"), and the abstractive nature of UCDP-AEC particularly exacerbates this. To evaluate how much of the error can be attributed to the selection of the gold surface form instead of a failure of the models' reasoning abilities, we introduce a semantic evaluation in line with Sharif et al. (2024). Given a field, all values appearing in any of the splits are embedded using BERT. When evaluating a model, its prediction for a given field is in turn embedded using BERT. The prediction is considered correct if it has the highest cosine similarity with the gold value amongst all possible values for this field. We report this value as "precision at top 1" (P@1), following the information retrieval convention. Since the identity maximises cosine similarity, the P@1 will always be greater than or equal to the accuracy, providing an upper bound to real model performance.

5.2 Baseline Models

To establish preliminary results for UCDP-AEC, we here report on experimental results for five different generative models, including both encoder-decoder and decoder-only architectures, all fine-tuned on the UCDP-AEC training set. The baselines are based on relatively small, openly available generative models to make sure the experiments are both reproducible and accessible to a broad range of researchers. These baselines are intended to explore and illustrate the specific challenges of UCDP-AEC, as well as provide a lower bound against which future work, potentially using larger and more specialised models, can be fairly compared.

In prior work, generative approaches to EE have predominantly been applied in an extractive setting, where the text generation is often constrained to the vocabulary present in the input text, as described in Section 2. Two of our baselines are based on adapting approaches originally proposed for such an extractive formulation, namely Text2Event and DEGREE, both widely used. As further detailed below, for our experiments, we make minimal modifications to the code provided with the original model papers to make them work with our abstractive formulation of the task, in which gold argument values may not appear as spans in the source text.

Text2Event (Lu et al., 2021) is using a T5 encoder-decoder architecture (Raffel et al., 2020) to translate text into S-expressions representing events. In the original setup, the model uses constrained decoding, which masks the output softmax to enforce valid S-expression structure and role names, and restricts argument generation to text spans in the input sentence. We adapt Text2Event to the abstractive setting by removing the constraint on argument generation. Additionally, we increase the input context window to 512 tokens to fit UCDP-AEC source documents, as Text2Event is originally designed for sentence-level event extraction.

DEGREE (Hsu et al., 2022) relies on the BART encoder-decoder transformer (Lewis et al., 2020) as a backbone, and the event extraction task is approached as a sequence-to-sequence problem between the input text and a natural language template to be filled. Since the UCDP-AEC events cover more fields than what DEGREE was originally designed for, we introduce a fuzzy-template matching algorithm to ensure that the model respects the template. The algorithm finds the argument assignment that minimises the Levenshtein distance between the predicted text and template, thus ignoring small variations in the natural language template output of DEGREE.

In addition, we fine-tune the Flan-T5 large model (Chung et al., 2022) using a slightly simpler formal template than that of Text2Event. Finally, we fine-tune two text-only Llama-3.2-Instruct-based models (1B and 3B, Grattafiori et al., 2024), using the Alpaca prompt format. The instruction prompt is illustrated in Appendix H.

5.3 Results

We report the performance of the baselines for each category and the aggregate accuracy in Table 3. The full per-role accuracies are given in Appendix D. We can see that for all models except Llama 3B, the date and location categories prove more challenging than the actor and deaths categories. We posit this is due to both actor and deaths having a smaller set of admissible values compared to date and location, which are closer to open-domain problems. While

[8] That is the average of 4 values, one for each category, not the average of the 14 fields' accuracies.

Model	#param	Exact match (Accuracy %)					Semantic (Precision@1 %)	
		Actors	Date	Location	Deaths	Aggregate	Actors	Location
Text2Event	223M	70.1	38.2	14.1	53.4	43.9	70.4	24.8
DEGREE	406M	73.5	64.4	**64.5**	**81.3**	70.9	**74.0**	**66.6**
Flan-T5 large	783M	**73.9**	63.0	62.8	81.0	70.2	**74.0**	65.6
Llama-3.2-Instruct	1B	71.0	70.5	59.9	76.4	69.4	71.3	61.3
Llama-3.2-Instruct	3B	73.0	**73.8**	63.9	78.5	**72.3**	73.2	65.5

Table 3: Baseline accuracies and P@1 on the UCDP-AEC test set, also showing the the number of model parameters.

the Llama 3B model demonstrates the strongest overall performance, obtaining the highest accuracy in the date category, the DEGREE and Flan-T5 large models are strong contenders, achieving the highest accuracies in the other categories.

Looking at the semantic evaluation, the P@1 follow the accuracies closely except for very low values. This seems to indicate that using single symbolic representations for the fields is not an important source of error for the most performant models.

5.4 Error-analysis

Looking into details, models struggle to predict new actors that were not observed during training. Around 5% of side A actors in the test set do not appear as actors in the training set. Out of those, Llama-3B was able to identify 25.8% correctly, while Flan-T5 large identified 32.3% correctly. A similar pattern can be observed for side B, suggesting that Flan-T5 large is better suited for previously unseen actors.

All models handle temporal uncertainty poorly, when the start and end date differ (in 8.6% of samples, usually as a result of expressions such as "over the weekend" or "last week"), date accuracy falls to 37.5% for Llama-3B and 33.8% for DEGREE, with models often failing to recognise the uncertainty and predict the same start and end date. Similarly, models tend to rely too heavily on the publication date, with a loss of performance when the event date differs from the publication date. DEGREE achieves a 35.9% accuracy in those cases, and Llama-3B a 41.5% accuracy.

6 Summary

This paper has introduced the UCDP Abstractive Event analysis Corpus – a large-scale dataset of high-quality document-level annotations of rich event representations in the domain of armed conflict, built from the UCDP GED event database and the HPLT v2 web archive. By coupling the event records to source documents identified in a public web archive, we are able to create a complete and open dataset for document-level event analysis. It makes the UCDP GED conflict annotations machine-learnable and available to the larger research community for experimentation. We believe the approach described here should also be applicable to other event databases where the underlying source documents are intact but currently not made openly available.

The annotations in UCDP-AEC have some important characteristics that set them apart from traditional event extraction datasets in NLP: Rather than sentence-level annotations tied to particular text spans with explicit information to be extracted, the document-level event representations in UCDP-AEC require inference from implied or otherwise under-specified information, possibly piecing together evidence that is scattered throughout a document. This requires a more "abstractive" approach to analysing events, which we argue may be particularly well-suited for generative approaches.

We have reported on a first suite of experiments with fine-tuning of both encoder-decoder and decoder-only models for generating event representations. While showing encouraging results and demonstrating the viability of the approach, we believe there is ample room for improvement in future work, in particular with respect to a more principled handling of event dates, and also exploring a larger space of decoder model tuning. We hope the UCDP-AEC dataset will be an important building block for future research on abstractive event analysis within both NLP and peace and conflict studies.

Limitations

While a defining feature of the UCDP-AEC dataset is that it reflects the temporal dynamics of real-world events, it offers a somewhat skewed view

of global conflict dynamics compared to the underlying UCDP GED data. This misalignment is mostly an artefact of the process for deriving the dataset, as described in Section 4.2. Firstly, UCDP-AEC is a small subset of UCDP GED, representing 3.3% of the events in the database due to filtering and availability of source documents. Furthermore, because of the source quality filter – described in Appendix A – 84% of events in the dataset occurred in the last 10 years, thereby under-representing or missing many earlier conflicts.

The misalignment results in a substantial shift in the country coverage when compared to UCDP GED. One such example is Brazil, which has more than 12 000 recorded events in UCDP GED, but is represented by only four events in UCDP-AEC, 0.03% of the original. Similarly, while 40% of the UCDP-AEC events occur in Syria, only 19% of the UCDP GED events are recorded there.

The dataset we release is built from two sources: UCDP GED and HPLT. Despite providing all the code used in the creation of the dataset, since UCDP GED is partly private – the source documents cannot be shared – it is impossible to recreate the dataset without privileged access that we negotiated with UCDP. Similarly, the evaluation of the HPLT substitution described in Section 4.2 cannot be reproduced without access to the original UCDP documents.

The UCDP GED database incorporates sources from multiple languages, as sources closer to conflict zones can often provide more precise information and contribute to a broader and more representative basis for event recordings. By restricting the UCDP-AEC dataset to English-language sources, we may not only reinforce existing biases in the data, but also contribute to the ongoing over-representation of the English language in research and resources in NLP. After the event filtering and substitution with HPLT documents, described in Section 4, only a small portion of non-English sources was identified, which we decided to exclude.

All of the encoder-decoder transformers used in the baseline models have a maximum input context window of 512 tokens. Larger UCDP-AEC documents cannot fit inside this limit. Furthermore, we are using the Text2Event and DEGREE models outside of their intended extractive setting, which creates an unfair comparison with the Flan-T5 and the decoder-only baselines.

Ethics

Event data on armed conflict, especially when human lives have been lost, warrants a heightened focus on ethical implications, and particularly on the potential bias in the domain and on the bias introduced by our work. As the field advances towards event representations based on more high-level and implied information, combined with the use of large language models, there might be an increased risk of harm, with the potential of generating non-factual events.

A large proportion of the sources in the dataset consist of news articles, which present various types of bias, particularly when reporting on conflict events. Accurately documenting casualty numbers in conflict zones poses challenges due to a variety of factors, many of which are described in Seib (2021), and should be interpreted with caution. Moreover, not all armed conflicts receive media coverage, and the framing of the reported conflicts can be skewed, particularly when relying on sources in a single language (Chojnacki et al., 2012). This bias can manifest in several forms, for example, as unbalanced representations of certain actors as aggressors, the use of loaded terminology, or reliance on reports from countries with restricted freedom of the press and/or freedom of speech.

Some of the source documents in the dataset may contain names of individuals. Although the documents in the dataset consist of publicly available sources, when combined with event extraction systems, it is possible to imagine scenarios such as the surveillance of individuals, targeted monitoring of groups or other forms of malicious and harmful applications. The released dataset, however, does not include metadata or annotations related to individuals, and we suspect a greater challenge lies in the potential risks of unintentional misuse.

The substitution with HPLT document IDs was done with the intention of removing news articles covered by proprietary licences. HPLT relies on a crawler that respects the `robots.txt` convention and has a procedure to request the take down of documents. Even though HPLT has yet to receive any claims, we cannot guarantee that no such documents are included in the HPLT corpus. The HPLT project has the stated goal of providing training data for NLP applications, as such, we consider our use of their data as fair.

UCDP-AEC is released for the sole purpose of training machine learning models rather than

analysing armed conflict events directly in the scope of socio-political research. The dataset presented is a simplified subset of the UCDP database and does not preserve the original distribution of events found in UCDP. For readers interested in research on socio-political events related to armed conflicts, Olsen et al. (2024) provide a comprehensive overview of available datasets.

Given the challenges with bias outlined, event extraction systems trained on UCDP-AEC are not intended to replace human experts but serve as a tool in the coding / annotation process. We stress that human evaluation is essential before automatically extracted event records can be used for further analysis or application.

The research presented in this paper, as well as the released dataset, is created to advance research aimed at better understanding why, how, and when armed conflict arises, with the main overarching goal of finding solutions to achieve a more peaceful world. We disapprove of using this dataset for research or applications that do not align with this vision and strongly discourage such usage.

References

Rui Bai, Di Lu, Shihao Ran, Elizabeth M. Olson, Hemank Lamba, Aoife Cahill, Joel Tetreault, and Alejandro Jaimes. 2025. CEHA: A dataset of conflict events in the horn of Africa. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1475–1495, Abu Dhabi, UAE. Association for Computational Linguistics.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. An expanded massive multilingual dataset for high-performance language technologies (hplt). *Preprint*, arXiv:2503.10267.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17772–17780.

Erica Chenoweth, Jonathan Pinckney, and Orion A. Lewis. 2019. NAVCO 3.0 Dataset.

Sven Chojnacki, Christian Ickler, Michael Spies, and John Wiesel. 2012. Event data on armed conflict and security: New perspectives, old challenges, and some solutions. *International Interactions*, 38(4):382–401.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Shawn Davies, Garoun Engström, Therése Pettersson, and Magnus Öberg. 2024. Organized violence 1989–2023, and the prevalence of organized crime groups. *Journal of Peace Research*, 61(4):673–693.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinya Du, Alexander Rush, and Claire Cardie. 2021a. GRIT: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 634–644, Online. Association for Computational Linguistics.

Xinya Du, Alexander Rush, and Claire Cardie. 2021b. Template filling with generative transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 909–914, Online. Association for Computational Linguistics.

Fırat Duruşan, Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Çağrı Yoltar, Burak Gürel, and Alvaro Comin. 2022. Global contentious politics database (GLOCON) annotation manuals. *Preprint*, arXiv:2206.10299.

William Gantt, Shabnam Behzad, Hannah An, Yunmo Chen, Aaron White, Benjamin Van Durme, and Mahsa Yarmohammadi. 2024. MultiMUC: Multilingual template filling on MUC-4. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 349–368, St. Julian's, Malta. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Andrew Halterman, Philip A. Schrodt, Andreas Beger, Benjamin E. Bagozzi, and Grace I. Scarborough. 2023. Creating custom event data without dictionaries: A bag-of-tricks. *arXiv preprint arXiv:2304.01331*.

Håvard Hegre, Paola Vesco, and Michael Colaresi. 2022. Lessons from an escalation prediction competition. *International Interactions*, 48(4):521–554.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Stina Högbladh. 2023. UCDP georeferenced event dataset codebook version 23.1. *Department of Peace and Conflict Research*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Clayton Norris, Philip A Schrodt, and John Beieler. 2017. PETRARCH2: Another event coding program. *J. Open Source Softw.*, 2(9):133.

Helene Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict and unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 40–53, St. Julians, Malta. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Clionadh Raleigh, Roudabeh Kishi, and Andrew Linke. 2023. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanities and Social Sciences Communications*, 10(1):74.

Clionadh Raleigh, Andrew Linke, Havard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660.

Idean Salehyan, Cullen S Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. 2012. Social conflict in africa: A new database. *International Interactions*, 38(4):503–511.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. CASIE: Extracting cybersecurity event information from text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8749–8757.

Philip A. Schrodt, Shannon G. Davis, and Judith L. Weddle. 1994. Political science: KEDS—a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561–587.

Philip Seib. 2021. *Information at war: journalism, disinformation, and modern warfare*. John Wiley & Sons.

Sina Semnani, Pingyue Zhang, Wanyue Zhai, Haozhuo Li, Ryan Beauchamp, Trey Billing, Katayoun Kishi, Manling Li, and Monica Lam. 2025. LEMONADE: A large multilingual expert-annotated abstractive event dataset for the real world. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25813–25852, Vienna, Austria. Association for Computational Linguistics.

Omar Sharif, Joseph Gatto, Madhusudan Basak, and Sarah Masud Preum. 2024. Explicit, implicit, and scattered: Revisiting event extraction to capture complex arguments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12061–12081, Miami, Florida, USA. Association for Computational Linguistics.

Étienne Simon, Helene Olsen, Huiling You, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2024. Generative approaches to event extraction: Survey and outlook. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 73–86, Miami, Florida, USA. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ralph Sundberg and Erik Melander. 2013. Introducing the UCDP georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. Association for Computational Linguistics.

Peter Turchin. 2012. Dynamics of political instability in the united states, 1780–2010. *Journal of Peace Research*, 49(4):577–591.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, and 1 others. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. 2023. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction. *arXiv preprint arXiv:2312.15548*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.

A Details of Data Subset Selection

The details of some filters might be hard to understand without knowledge of UCDP design detailed in the codebook (Högbladh, 2023). Starting from the UCDP candidate events dataset, we apply the following filters:

- the UCDP annotator marked the event as "clear",
- the event occurred on a single day (but this day might not be exactly known, thus the start date $\neq$ end date),
- the same event was not split into multiple entries (no "deaths split"),
- the source document uses the new post-2014 source format (ensuring more accurate publication dates),
- the event is described by a single source document (when using multiple sources, UCDP annotators might pick some of the information from one of the sources, distrust some specific roles from another, transforming the task into a weakly-supervised problem),
- the source document describes a single event (multi-event documents tend to include long tables of battle deaths without enough context to fully code the events),
- the start date, end date and publication date all occur in the same split,

- the deaths estimates are ordered properly: best (the sum of side A, side B, civilians and unknown) is in-between low and high,
- the split cut-off time are widened by 12 hours to avoid time zone reporting issues,
- the source document is in English,
- the length of the source document is between 100 and 10 000 characters.

All but the last two filters are applied on the UCDP database, resulting in 110 979 events. The final dataset is the result of HPLT substitution and the application of the last two filters on the HPLT source documents.

B Details of UCDP-AEC Fields

On domain openness UCDP-AEC is designed for a supervised prediction task where each instance consists of a source text and a publication date, denoted $\mathcal{X}$, and the corresponding output is a structured sequence of 14 argument roles $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{14}$. Traditionally, when $\mathcal{Y}$ is a finite set, the problem is said to be closed domain, while potentially infinite $\mathcal{Y}$ is part of open-domain problems. In practice, open-domain roles may still be partially constrained because of real-world structures. For example, location values on the country level are often limited to internationally recognised countries.

The UCDP-AEC dataset contains fields with varying degrees of domain openness. To quantify the openness of each field in UCDP-AEC, we introduced three metrics in Section 4.3, report on each metric for each field in Table 2. We here describe these metrics in more detail.

We measure average instances per unique value, the *value density*, by computing the ratio $\frac{|\mathcal{D}|}{|\mathcal{Y}|}$, where $|\mathcal{D}|$ denotes the number of instances with a value for a given role, and $|\mathcal{Y}|$ the number of unique values observed. Higher values indicate a closed domain, with a small number of unique values that are frequently observed, whereas lower values indicate a more open domain with a more diverse set of values that are less frequently observed.

We measure the *unique value overlap* between the train and test set with the ratio $\frac{|\mathcal{Y}_{\text{test}} \cap \mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|}$. This metric describes the percentage of test values observed during training, where a low overlap indicates a more open domain. A high overlap indicates that most of the values in the test set are also present in the train set.

Finally, for the *instance-level overlap*, we measure the proportion of instances in the test set with values, that has been observed in the train set with $\frac{|\mathcal{D}_{\text{test}} \cap \mathcal{D}_{\text{train}}|}{|\mathcal{D}_{\text{test}}|}$. Here, we take into account the frequency of the value.

In UCDP-AEC, the country-level location, and the death counts and estimates could be considered closed domain, as the range of possible values is limited and very repetitive.

The two actor roles, Side A and Side B, can be considered semi open domains, where many known actors will appear in the test set, but new actors will also occur. Both roles has a moderate value reuse, with about 26–27 instances per value. A similar pattern can be found for Subregion.

The more fine-grained locations show open domain characteristics. Region and Point are placenames often specific to single events, with many values in test that are not present in train. The temporal arguments have a 0% overlap because they are used for split selection (by definition, train and test do not overlap temporally).

On death estimates The fields "Deaths Low" and "Deaths High" are used when there is uncertainty with respect to the total number of fatalities. For these estimates, the detailed attribution of casualties to specific sides (Side A, Side B, Civilians, Unknown) is not specified. When summed together, those four deaths fields give the UCDP best estimate for the total number of people killed as part of an event. The fields "Deaths Low" and "Deaths High" are used when there is uncertainty with respect to the total number of fatalities. For these estimates, the detailed attribution of casualties to specific sides (Side A, Side B, Civilians, Unknown) is not specified. In any case we always have low $\leq$ best $\leq$ high and $1 \leq$ high.

C Sample Example

See Table 4. Note that we deliberately chose this article because of its short length for illustration purposes. Consisting of only 491 characters, this sample is not representative in terms of length, as the average article in UCDP-AEC contains 2116 characters.

D Detailed Results

See Table 5. Note that the accuracy for side A is given for the gold side A. Since actors are permutation invariants, this can correspond to a prediction for side B by the model. Side A and B are always different in the gold annotation, as such there is no

Input	
publication date	2022-01-06
source article	The lifeless body of a woman was discovered Wednesday morning on her farm in Mbengwi, Momo Division, Northwest region of Cameroon. Reports say there was a shootout between security forces and suspected separatist fighters around Womsei, a quarter behind the Mbon Market on Tuesday evening. Her body was only discovered Wednesday morning with the basket she was using to harvest okra still strapped on her back. Her body has been preserved at the Mbengwi mortuary pending burial arrangements.

Output			
Role	**Argument**	**Role**	**Argument**
Side A	Government of Cameroon	Start date	2022-01-04
Side B	Ambazonia insurgents	End date	2022-01-04
Location country	Cameroon	Location region	North-West region
Location subregion	Momo département	Location point	Mbengwi village
Deaths side A	0	Deaths side B	0
Deaths civilian	1	Deaths unknown	0
Deaths low	1	Deaths high	1

Table 4: Example of sample from the test set (sample id: 428220).

precision–recall trade-off to be taken advantage of despite evaluation through accuracy.

E Licences

The two data sources we are using are UCDP GED released under CC-BY and HPLT v2.0 realesed under CC0. We license our dataset UCDP-AEC under CC-BY.

We release our modifications to Text2Event and DEGREE under the same licence as the original code, that is MIT for Text2Event and Apache 2.0 for DEGREE.

We release the code of the Flan-T5 and Llama baselines, the code used to create the dataset, its evaluation script and scripts computing dataset statistics under the GNU AGPL licence.

F Evaluation Details

Since in UCDP-AEC all fields always have exactly one assigned value for all events, for a given field, the recall of a model is the number of true positive divided by the number of samples in the dataset, that is the accuracy. Similarly, the precision is also equal to the accuracy if the model predicts a single value for each field – as it should since its precision would only decrease otherwise – therefore we have precision = recall = F_1 = accuracy, thus explaining why we only report accuracies instead of the usual precision–recall–F_1 triplets.

For the date and deaths categories, we investigated the use of smooth metrics such as RMSE and RMSLE. However, as errors tend to be infrequent but important in magnitude, we did not find those metrics informative enough to be of interest. The provided evaluation script does compute RMSE and other metrics we considered such as semantic MRR.

For the semantic evaluation, we investigated using different models including S-BERT following Sharif et al. (2024). However we observed little variations apart from larger models providing slightly better P@1. Given the nature of the insight provided by the semantic evaluation, we did not find this warranted the extra computational cost and relied on `bert-base-uncased` instead.

G Model Details and Computational Budget

For both Text2Event (Lu et al., 2021) and DEGREE (Hsu et al., 2022), we reused the hyperparameters of the original models, except the input token limit, which we increased to 512 in both cases.

For the Flan-T5 baseline, we used learning rates of 3×10^{-5} with a maximum of 18 epochs. 10 intermediary evaluations were performed with early stopping on validation aggregate accuracy.

For the Llama-3.2-Instruct models, we set the maximum sequence length to 2048, used up to 900 training steps, applied a weight decay of 0.01, and a learning rate of 2e-4.

All models were trained on either P100 or A100 GPUs. Including development runs, the total GPU usage was around 200 GPU hours.

Model	Actors		Date		Location				Deaths					
	Side A	Side B	Start	End	Country	Region	Subregion	Point	Side A	Side B	Civilian	Unknown	Low	High
Text2Event	75.8	64.4	35.3	41.0	37.0	11.1	5.1	3.2	81.3	78.2	56.6	67.7	21.3	15.3
DEGREE	77.8	69.3	63.0	65.9	97.8	75.8	46.0	38.4	84.7	84.6	82.6	85.6	68.8	81.6
Flan-T5 large	77.0	70.8	60.3	65.7	97.8	73.3	45.7	34.3	82.3	81.9	82.9	87.3	69.8	82.1
Llama-3.2-I-1B	78.2	63.8	68.3	72.6	96.5	68.0	43.3	31.7	80.2	78.5	73.7	82.0	65.7	78.2
Llama-3.2-I-3B	78.9	67.1	72.8	74.7	97.1	76.7	47.5	34.4	83.9	80.3	75.5	83.0	68.6	79.7

Table 5: Accuracies for each individual role.

Instruction	Generate structured event information related to a socio-political conflict in the UCDP event format, based on the content of a source UCDP document. The generated data should include the following fields: `side_a_name`, `side_b_name`, `start_date`, `end_date`, `location_root_name`, `location_adm1_name`, `location_adm2_name`, `location_where_name`, `deaths_side_a`, `deaths_side_b`, `deaths_civilian`, `deaths_unknown`, `deaths_low`, `deaths_high`. Output the event information as a valid JSON object corresponding to the armed conflict incident described in the following source UCDP document.
Input	{Source article text}
Output	{Required output in JSON format}

Table 6: Instruction prompt used for fine-tuning.

H Instruction Prompt

The prompt used for the Llama models can be found in Table 6.