CEO: Corpus-based Open-Domain Event Ontology Induction

Anonymous ACL submission

Abstract

Existing event-centric NLP models often only apply to the pre-defined ontology, which significantly restricts their generalization capabilities. This paper presents *CEO*, a novel Corpus-based Event Ontology induction model to relax the restriction imposed by pre-defined event ontologies. Without direct supervision, *CEO* leverages distant supervision from available summary datasets to detect corpus-wise salient events and exploits external event knowledge to force events within a short distance to have close embeddings. Experiments on three popular event datasets show that the schema induced by *CEO* has better coverage and higher accuracy than previous methods. Moreover, *CEO* is the first event ontology induction model that can induce a hierarchical event ontology with meaningful names on eleven open-domain corpora, making the induced schema more trustworthy and easier to be further curated. We anonymously release our dataset, codes, and induced ontology [1].

1 Introduction

Extracting and understanding real-world events described in the text are crucial information extraction tasks that lay the foundations for downstream NLP applications (Chen et al., 2021; Zhang et al., 2020; Fung et al., 2021). However, existing event-related studies are mostly restricted by the pre-defined ontology (Zhang et al., 2022; Guzman-Nateras et al., 2022). Even for the zero-shot setting, models still need a pre-defined ontology for inference (Huang and Ji, 2020; Edwards and Ji, 2022).

To address this limitation, the previous work (Shen et al., 2021) proposed the *event type induction* task, which automatically induces event ontology from documents. However, previous work only covers verbal events while ignoring the nominal ones. Moreover, it can only induce the flat ontology, which is not enough to cover the rich hierarchical ontology structure defined by humans. Last but not least, the induced ontology only contains type ids, making it hard to be verified and curated by users.

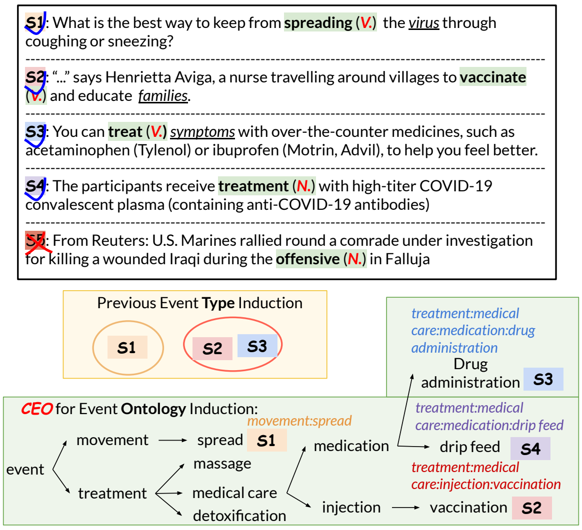


Figure 1: Instances from Covid-19 corpus with event type induced by previous work and ontology induced by *CEO*. The non-salient event *treatment*in *S4* is disregarded while others are preserved. Event **type** induction only identifies events triggered by verbs (*S1*, *S2*, *S3*) but not nouns (*S4*), and arranges events into simple clusters. *CEO* recognizes both verb- and noun-triggered events, induces tree-structure ontology and provides concrete names.

This paper introduces a new Corpus-based open-domain Event Ontology induction strategy (*CEO*). As demonstrated in Figure 1, *CEO* covers both verbal and nominal events and leverages external summarization datasets to detect salient events better. On top of that, *CEO* is also capable of inducing hierarchical event ontology with the help of a word sense ontology tree defined in WordNet (Fellbaum, 2010). To enhance the faithfulness of induced ontology and facilitate future

[1]https://sites.google.com/view/ceoeventontology

curation, *CEO* generates a meaningful name for each induced event type in the induced ontology.

In the proposed *CEO* strategy, we make two key technical contributions to better learn from open-domain events. The first technical contribution is corpus-wise salient event detection with distant supervision from available summary datasets. Following the assumption that summaries written by humans are likely to include events about the main content (Liu et al., 2018; Jindal et al., 2020), we consider events mentioned both in summary and body text as salient while those only mentioned in the body text as non-salient. To obtain corpus-wise key events, we fine-tune a Longformer-based model (Beltagy et al., 2020) to classify whether the identified events are salient or not given rich context.

The second contribution is exploiting external event knowledge for hierarchical open-domain event ontology inference. Specifically, we leverage the word sense ontology (i.e., the hypernym/hyponym relationships) trees in WordNet (Fellbaum, 2010) to improve event representations. We propose to train an autoencoder model (Domingos, 2015) to compress the original event representations in the latent space, where information is preserved by minimizing the reconstruction error. We further utilize a triplet loss (Balntas et al., 2016) to regularize the compressed embeddings, so that event pairs with senses in a short distance in the WordNet ontology tree are much closer (i.e., anchor and positive events) compared with those far away from each other (i.e., anchor and negative events). After training event data from both WordNet and the studied corpus with ontology supervision from the former, events with close compressed embeddings in the latter are expected to have short distances in the ontology tree.

In summary, we propose an effective strategy, *CEO*, to extract and understand corpus-based open-domain events. Experiments on three popular event datasets show that the proposed *CEO* could consistently induce accurate and broad-coverage event ontology without direct supervision. Moreover, to the best of our knowledge, *CEO* is the best model that could induce a hierarchical event ontology with meaningful names. We also perform event ontology induction on 11 open-domain news corpus such as *abortion, LGBT* and demonstrate the broad application of *CEO*.

2 Related Work

Event Extraction Given a set of pre-defined types and annotated samples, event extraction is typically cast as a multi-class classification task, where event types and argument roles are predicted into one of target types (Lin et al., 2020). Recently, semantic meanings of event and argument types have gained much attention to capture correlations between event mentions and types (Wang et al., 2022; Hsu et al., 2022).

Semi- and Un-supervised Event Type Induction To classify constantly emerging events of new types without annotations in an existing domain, semi-supervised learning approaches such as Vector Quantized Variational Autoencoder (Huang and Ji, 2020) and contrastive learning (Edwards and Ji, 2022; Zhang et al., 2022) have been introduced. ETypeClus (Shen et al., 2021) proposed to perform event type induction under the unsupervised setting, where neither annotations nor event types are used. Different from unutterable event clusters induced by ETypeClus, *CEO* infers underlying event type ontology including interpretable type for each mention in diverse granularities.

3 Problem Definition

Since the majority of events are triggered by **verbal** and **nominal** predicates along with relevant arguments, we denote an event mention by $<$*subject*, *predicate*, *object*$>$. For each corpus, event mentions highly relevant to its topic are considered as **salient** and constitute the extraction targets. To understand semantic relations between events, we aim at inducing a hierarchical event type **ontology** with a tree structure, where leaf nodes represent single event mentions while internal nodes are subclusters of events.

Task Definition. Given a corpus of N sentences $\mathcal{C} = \{S_1, \ldots, S_N\}$, *event ontology induction* 1) firstly extracts salient event mentions, e.g., m_{ij} for j-th event in S_i, 2) then identifies event ontology that well demonstrates correlations among all covered event types, 3) lastly infers event type names withing human readable formats from coarse-to-fine granularity.

4 CEO

In Fig. 2, we show the overview of the proposed *CEO* that extracts (*Step* 1 in §4.1) and represents salient events (*Step* 2 in §4.2) with informative

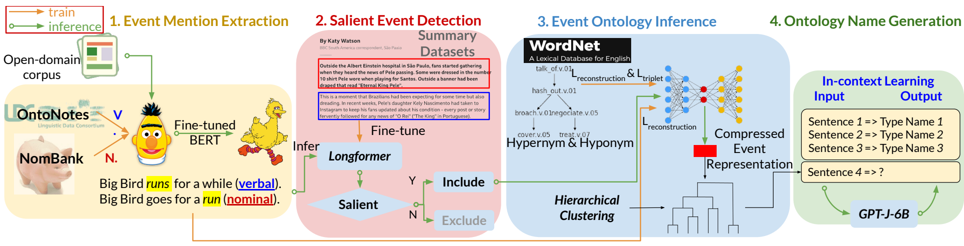


Figure 2: Framework of the proposed *CEO*. *Step 1*: extract events triggered by nouns or verbs; *Step 2*: preserve salient events with distant supervision from summaries; *Step 3*: improve event representations for hierarchical clustering with external event knowledge from WordNet; *Step 4*: generate event type names with in-context learning.

embeddings for ontology structure induction (*Step* 3 in §4.3) and name generation (*Step* 4 in §4.4).

4.1 Event Mention Extraction

We take advantage of event trigger-annotated datasets, OntoNotes (Pradhan et al., 2013) and NomBank (Meyers et al., 2004), for verb- and noun-triggered event information extraction, respectively. Concretely, we adopt a two-stage process for event information extraction: *1) event trigger detection:* we follow the practice in (Shen et al., 2021) to extract verbal tokens identified by the dependency parser as the verbal event trigger; since nouns play much more diverse roles in sentences besides predicates, we cast the nominal predicate detection as a binary classification task and fine-tune the BERT (Devlin et al., 2019) model to identify nouns labeled as event triggers in NomBank[1]. *2) joint training for event-relevant information learning:* with the identified event triggers, we follow the work for semantic role labeling (Shi and Lin, 2019; Lee et al., 2021), where the vanilla BERT model is connected with two linear layers, one for argument classification and the other for predicate sense disambiguation. The extracted event information from *CEO*, including event trigger tokens, their semantic senses, and accompanying argument tokens, comprehensively describes different perspectives of events.

4.2 Salient Event Detection

Aimed at only extracting events salient to the given corpus, prior work (Shen et al., 2021) adopted the TF-IDF idea and defined the event salience by comparing the frequency of trigger words in the studied corpus against a general-domain corpus. We argue that such a rough criterion disregards contextual information of event triggers and is prone to cause massive *false negatives*.[2] Instead, we detect salient events based on the semantic and contextual information of predicates. As shown in Tab. 1, we propose to leverage distant supervision from summarization datasets, following the assumption that an event is considered salient if a summary written by a human tends to include it (Liu et al., 2018; Jindal et al., 2020). To consider a wide window of context, we fine-tune the Longformer (Beltagy et al., 2020) model to perform binary classification: given contexts and trigger words, predict the events as salient if they appear in summary as well. For open-domain event salience inference, we provide the event sentence with context and obtain its corresponding salience score.

4.3 Event Ontology Inference

With all kinds of event-centric information for salient events, we can infer the corpus-level event ontology by incorporating the learned informative event embeddings into a wide range of off-the-shelf hierarchical clustering models (discussed in §5.3.1). For individual event mentions, we average over the following embeddings as the final comprehensive event representations: *1) contextualized embeddings* for tokens at positions predicted as the predicate, subject, and object; *2) event sentence embeddings* represented by Sentence-BERT (Reimers and Gurevych, 2019a); *3) predicate sense embeddings* composed of definition sentence representations from Sentence-BERT and contextualized token embeddings for predicate positions from example sentences.

Although there is no extra knowledge about the actual event ontology of the studied open-domain corpus, we find that the explicit hypernym/hyponym relationships among the verb

[1] NomBank is an open-domain dataset with broad coverage that considers nouns in Wall Street Journal Corpus of the Penn Treebank (Garofolo et al., 1993).

[2] For instance, the surface pattern of a trigger word could be rarely observed, but its semantic relevance to the corpus theme might be very high.

Title: Metro Briefing | New York : Brooklyn : Charter Review Meeting Disrupted .

Summary: First public hearing of *Charter* Revision *Commission* is disrupted by protesters Daniel Cantor and Arron Schildkrout, who oppose New York City Mayor Michael R Bloomberg's plan to institute nonpartisan *elections* (S)

Body Text: The first public hearing of Mayor Michael R. Bloomberg's *Charter* Revision *Commission* was disrupted last night by protesters, and two men were *arrested*. Opponents of the mayor's plan to *establish* nonpartisan *elections* burst into the Fire Department's headquarters in Brooklyn, where the hearing was held, and *chanted*, " *Change* the mayor, not the *charter*. " Two men, Daniel Cantor, 47, of Brooklyn, and Arron Schildkrout, 22, of Watertown, Mass., were *arrested* and *charged* with ...

Table 1: Instance sampled from NYT Corpus. Event triggers in the body text are marked in *italic*. Events concurrently mentioned in summary and body text are deemed salient and in *red*, while others are non-salient in *blue*.

synsets in WordNet (Fellbaum, 2010) can provide concrete guidance for the hierarchical event ontology[1]. To further improve event embeddings, we exploit the event ontology in WordNet by augmenting the standard autoencoder with an additional contrastive loss. We first assume that events within a short distance from each other in the ontology tree should be semantically similar and close in the latent space of the autoencoder (see Appx. §A.3 for distance computation and Fig. 5 for visualization). We then utilize the following loss function to augment the reconstruction loss for optimizing the autoencoder parameters[2]: $L_{\text{triplet}}(i,p,n) = \max\{d(\mathbf{e_i},\mathbf{e_p}) - d(\mathbf{e_i},\mathbf{e_n}) + \text{margin}, 0\}$, where i, p and n are anchor, positive, and negative events, $\mathbf{e_i}$, $\mathbf{e_p}$ and $\mathbf{e_n}$ are their representations in the latent space, d denotes the Euclidean distance. Compressed vectors in the latent space are adopted for ontology inference.

4.4 Ontology Name Generation

From the bottom leaf layer to the top root node in the learned ontology tree, diverse event instances are clustered according to different levels of similarities. Motivated by the in-context learning capacity of pre-trained language models, we randomly sample event instances from other available event datasets as demonstrations (see an in-context learning example in Tab. 11). For internal node name generation, the token probability distribution of event type names is averaged over all included events and the most likely is selected.

5 Experiments

In this section, we firstly introduce the utilized event datasets (§5.1) and then quantitatively evaluate the ontology (§5.3.1) and name (§5.3.2) introduction quality of *CEO*. Then we evaluate the effectiveness of different techniques incorporated in *CEO* (§5.4) via the ablation study. Lastly, we apply *CEO* to perform ontology induction on eleven open-domain corpora (§5.5) to demonstrate its effectiveness in real applications.

[1] The latest WordNet contains 13,650 verb synsets.

[2] As demonstrated in Fig. 2 and Fig. 5, to avoid distribution shift, events predicted from the studied corpus is also used for reconstruction loss besides those annotated in WordNet, but only the latter is available hence used for triplet loss.

Dataset	#Docs	#Event Mentions	#Event Types (Ontology)	%Predicates Noun/Verb
ACE 2005	599	5,349	33 (2 levels)	43.73/46.34
MAVEN	4,480	118,732	168 (4 levels)	28.60/64.23
RAMS	3,993	9,124	139 (3 levels)	39.99/55.45

Table 2: Statistics of studied event datasets show nouns are as important as verbs in expressing events.

5.1 Datasets

We summarize statistics of utilized event datasets in Tab. 2 and visualize their corresponding ontologies in Fig. 6. **ACE2005** (Doddington et al., 2004) is the widely used English event dataset with its event schema organized by a 2-level hierarchy: five types of general events, each with 1∼13 subtypes included. **MAVEN** (Wang et al., 2020) is a massive general domain event detection dataset with its event types manually derived from the linguistic resource FrameNet (Baker et al., 1998) following a 4-layer tree-structure. **RAMS** (Ebner et al., 2020) employs a three-level hierarchical event ontology with all types annotated according to a manually constructed mapping.

5.2 Implementation Details

For event mention extraction (§4.1), BERT is finetuned for event extraction model on OntoNotes for verbal predicates and Nombank for nominal predicates. For salient event detection (§4.2), we label events as salient if they also appear in summary; for New York Times, both events in summary and body text are annotated. For event ontology inference (§4.3), the encoder layers are [896, 768, 640, 512], while the decoder layers are the reverse for the Autoencoder; the learning rate is 0.005 and

Methods	ACE2005		MAVEN		RAMS	
	Purity ↑	Cost ↓ ($\times 10^9$)	Purity ↑	Cost ↓ ($\times 10^{12}$)	Purity ↑	Cost ↓ ($\times 10^9$)
hkmeans	.519	**1.00**	.356	**4.75**	.143	6.79
birch	.242	1.49	.129	6.88	.057	8.00
perch	.370	1.01	.361	4.78	.154	6.84
ghhc	.189	1.54	.027	7.22	.019	10.3
HypHC	.302	**1.00**	.027	4.81	.040	**6.75**
ward linkage	**.556**	**1.00**	**.457**	**4.75**	**.220**	6.78

Table 3: Performance of our ward linkage and other hierarchical clustering methods evaluated by dendrogram purity and Dasgupta cost. Inferred hierarchical clusters with higher purity (↑) and lower cost (↓) are more aligned with the ground-truth event ontologies.

training epochs are 100.

5.3 Evaluations of Event Ontology Induction

In this section, we evaluate induced event ontologies from two perspectives: mention clustering accuracy and cluster name preciseness.

5.3.1 Hierarchical Clustering

Metrics We evaluate the quality of inferred hierarchical clusters using the widely-adopted *dendrogram purity* (Heller and Ghahramani, 2005), and the more recent *Dasgupta cost* (Dasgupta, 2016). Higher purity and lower cost indicate more accurate clustering. We leave their concrete formulae in Appx. §A.1.

Baselines We perform comprehensive evaluations on discrete optimization methods from two classes: top-down divisive –*Hierarchical Kmeans* and *Birch* (Zhang et al., 1997), and bottom-up agglomerative –*Ward Linkage* (Ward Jr, 1963) and *Perch* (Kobren et al., 2017). Furthermore, we consider recent gradient-based continuous optimization methods which benefit from stochastic optimization: *gHHC* (Monath et al., 2019) and *HypHC* (Chami et al., 2020).

Results As shown in Tab. 3, we adopt ***ward linkage*** algorithm, which achieves the best performance for ontology induction evaluated by both purity and cost consistently. On MAVEN and RAMS with more complicated event ontologies, the enlarged performance gap is observed between continuous optimization methods and discrete ones. We speculate that hundreds of clusters and input dimensions make it challenging for the continuous approach to outperform discrete methods based on heuristics, which is in contrast to observations reported on small-scale datasets (Monath et al., 2019; Chami et al., 2020).

We further demonstrate the alignment of inferred event ontology with coarsest event type annotations for ACE 2005 in Fig. 3 and the other two datasets in Fig. 7. We observe that events of identical coarse-grained types are clustered together compared with those annotated by different labels. In Fig. 3, the most popular *conflict* events cluster in the left branches while the less popular *justice* events gather in the middle branches.

5.3.2 Name Generation

Metrics We treat the ground-truth coarse-to-fine label names, $E_r = \{e_r^i | 1 \leq i \leq n_r\}$ of n_r levels, as an ordered reference. We compare E_r with the generated type names, which are composed of node names from root to leaf in the ontology tree, $E_p = \{e_p^j | 1 \leq j \leq n_p\}$ of n_p levels. We utilize the following metrics: *1) Sim dist* is self-defined to consider both semantic similarity and granularity difference between each pair of reference e_r^i and generated name e_p^j (see Appx. §A.1 for the formula); *2) Rouge-L*: type names from coarse to fine granularities are combined into a single sentence and Rouge-L score (Lin, 2004) is used to compare the generated against the reference sentence. *3) BERTScore* (Zhang et al., 2019): similar to Rouge-L, the similarity F1 score is computed for token pairs in the generated and reference sentence.

Baselines With clustered events predicted by *CEO*, we utilize either statistical strategies – *Most frequent* and *tf-idf*, or off-the-shelf language models – *RoBERTa-large* (Liu et al., 2019) and *GPT-J-6B* (Wang and Komatsuzaki, 2021), to generate cluster names. Keywords extracted by *textrank* (Mihalcea and Tarau, 2004), *topicrank* (Bougouin et al., 2013) or *KeyBERT* (Grootendorst, 2020) are also utilized as cluster names. Besides, we introduce the *wordnet synset* strategy that adopts the least common ancestor hypernym of event triggers (Fellbaum, 2010). We describe more methodology details in Appx. §A.2.

Results We evaluate the qualities of our in-context learning ***GPT-J-6B*** and other name generation strategies and show results in Tab. 4. The language model *GPT-J-6B* achieves the best performance evaluated by three metrics on all studied datasets. Compared with other statistical methods, keyword extraction strategies can hardly extract

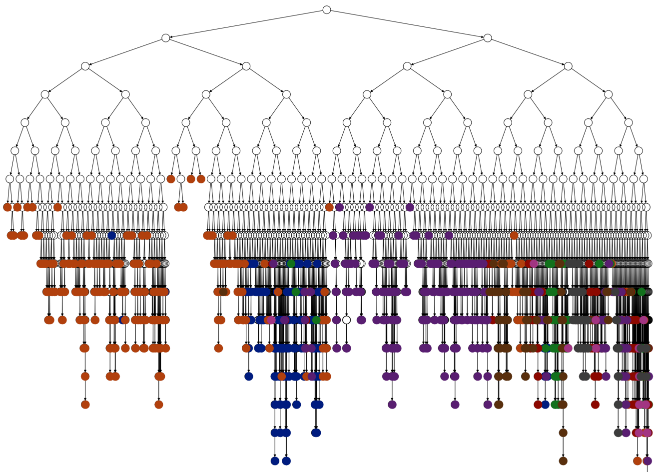


Figure 3: Event ontology induced by ward linkage on ACE2005. Each leaf node represents one event mention and is colored by its actual coarsest event type: *Life*, *Personnel*, *Justice*, *Conflict*, *Transaction*, *Movement*, *Contact*, *Business*. The ontology hierarchies of the other two datasets are visualized in Fig. 7.

Method	ACE2005			MAVEN			RAMS		
	Sim dist ↑	rougeL ↑	BERTScore ↑	Sim dist ↑	rougeL ↑	BERTScore ↑	Sim dist ↑	rougeL ↑	BERTScore ↑
most frequent	.508	.167	.869	.466	.043	.836	.448	.041	.849
tf-idf	.505	.184	.869	.464	.041	.835	.447	.038	.849
topicrank	.437	.024	.824	.380	0.0	.721	.413	.006	.817
textrank	.418	.035	.813	.376	0.0	.724	.399	.016	.811
keybert	.462	.072	.838	.427	0.0	.795	.425	.014	.830
WordNet	.438	.055	.827	.418	.006	.814	.411	.003	.825
RoBERTa-large	.510	.191	.871	.462	.041	.838	.440	.027	.842
GPT-J-6B	**.513**	**.210**	**.880**	**.466**	**.051**	**.840**	**.466**	**.086**	**.851**

Table 4: Evaluation of type names from our GPT-J-6B and other generation methods for event ontologies. For all metrics, higher scores indicate higher similarity of generated names to the annotated hierarchical event labels.

Preference	ACE2005	MAVEN	RAMS
GPT-J-6B better	**.75**	**.58**	**.59**
2nd best better	.21	.30	.22
Same	.04	.12	.19

Table 5: Human preferences on event names generated by GPT-J-6B and 2nd best strategy for each dataset.

salient event triggers from thousands of tokens. Overall, deep language models perform much better than statistical ones.

Human Evaluations For each event dataset, we randomly sample 100 instances and ask annotators to compare type names from *GPT-J-6B* and the 2nd best strategy in Tab. 4. As demonstrated in Tab. 5, event names generated by *GPT-J-6B* are consistently preferred across three datasets.

Case Study We randomly sample three event instances and demonstrate their type names generated from different strategies in Tab. 6. For easy instances such as *T1* and *T2*, we observe that statistical strategies are able to produce type names as accurately as pre-trained LMs. However, for the challenging instance *T3*, most generation strategies mistakenly provide descriptions semantically opposite to *robs*, e.g., *lend* and *borrow* from *WordNet Sysnet*. Only *GPT-j-6B* successfully captures the critical meaning of the event: *attack* and *steal*.

5.4 Ablation Studies

In this section, we showcase the effectiveness of different techniques introduced in *CEO*.

Benefits of Event Embedding We first show the capability of *CEO* for *covering more actual event mentions* in Tab. 7: 1) the transformer model jointly trained for predicate/argument identification and sense disambiguation improves the recall of **verbal** mentions by around 10% compared with those identified by POS tagging in ETypeClus; 2) with an additional model trained on NomBank for nominal predicates detection, *CEO* can capture the majority of **nominal events** and lead to an

Dataset	Event Instances and Names
ACE2005	*T1:* Peterson Trial Scott Peterson has been found guilty of ***murdering*** his wife Laci and their unborn son, and he now faces the death penalty. **Gold types**: life:die Most Frequent: kill:die:murder TF-IDF: kill:die:murder WordNet Synset: killing:die:murder RoBERTa-large: kill:die:murder GPT-j-6B: death:murder
MAVEN	*T2:* The robbers attempted to ***flee*** the scene, Phillips on foot and Matasareanu in their getaway vehicle while continuing to exchange fire with the officers. **Gold types**: Action:Motion:Self_motion:Escaping Most Frequent: attack:meet:send:move:fly:transport:carry TF-IDF: become:destroy:receive:occupy:evacuate:flee WordNet Synset: range:destroy:pit:inflict:seize:flee RoBERTa-large: hold:destroy:receive:occupy:evacuate:flee GPT-j-6B: attack:transport:escape
RAMS	*T3:* Corruption in oil production - one of the world's richest industries and one that touches us all through our reliance on petrol - fuels inequality, ***robs*** people of their basic needs and causes social unrest in some of the world's poorest countries **Gold types**: conflict:attack Most Frequent: urge:donate:lend:borrow:rob TF-IDF: urge:donate:lend:borrow:rob WordNet Synset: rede:donate:borrow:rob RoBERTa-large: urge:donate:end:rob GPT-j-6B: attack:transfer:steal

Table 6: Generated names for instances sampled from three event datasets. We mark the predicted ***predicates***, while type names are separated by ":" and arranged from coarse to fine.

Predicate		ACE2005	MAVEN	RAMS
Nominal	ETypeClus	-	-	-
	CEO	**.630**	**.612**	**.600**
Verbal	ETypeClus	.713	.770	.764
	CEO	**.808**	**.880**	**.876**
Combined	ETypeClus	.396	.544	.471
	CEO	**.729**	**.801**	**.770**

Table 7: Event extraction performance comparison between *CEO* and EtypeClus. Recall numbers are recorded to fulfill the goal of extracting as many events as possible. False positives are tolerable since they could be filtered in salient event detection.

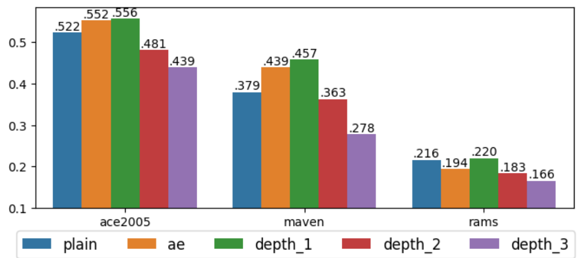


Figure 4: Impact of different utilization methods of external WordNet knowledge on hierarchical clustering (*purity* by *linage ward*). When both reconstruction and contrastive loss are employed, we also show the influence of the distance threshold. Dasgupta costs are omitted for statistically insignificant value variances.

overall 30% more events coverage.

Furthermore, we perform flat event clustering with representations learned by *CEO* and ETypeClus[1]. On the set of common salient events detected by both approaches[2], we follow prior work (Shen et al., 2021) by investigating five clustering algorithms: *kmeans*, Spherical KMeans (*sp-Kmeans*), Agglomerative Clustering(*AggClus*), *JCSC* (Huang et al., 2016) and *EtypeClus* (Shen et al., 2021), and evaluate with three metrics: *ARI* (Hubert and Arabie, 1985), *BCubed-F1* (Bagga and Baldwin, 1998) and *NMI*. We find that results from different metrics are positively related, hence demonstrating performance evaluated by ARI in Tab. 8 and leaving the other two in Tab. 12. In Tab. 8, we observe significant performance gain when the embeddings learned by *CEO* are utilized compared with ETypeClus. We also find that the impact of different event embeddings is less obvious on RAMS, where event types are annotated considering contexts rather than single sentences.

Benefits of Distant Supervision from Summary Datasets We first fine-tune Longformer (Beltagy et al., 2020) on three widely-adopted summary datasets for salient event detection: New York Times corpus (Sandhaus, 2008), CNN/Daily Mail (See et al., 2017) and Multi-News (Fabbri et al., 2019)[3]. We list salient event detection performance compared with existing approaches on summary datasets in Tab. 13. In Tab. 9, we show benefits of distant supervision on studied corpora: the model trained on any of the summary datasets is able to capture more salient events compared with ETypeClus, covering all event types. We utilize salient events detected by the model trained on NYT for ontology and type name generation[4].

[1]ETypeClus represents events by concatenating predicates and objects, which are not instance-specific but contextual vectors averaged over all occurrences. Conversely, we exclusively represent each event with its respective context considered.

[2]We find that salient events identified by EtypClus are always covered by *CEO*. We therefore directly use salient events identified by ETypeClus. The very few events missed by *CEO* can still be represented with sentence embeddings.

[3]For NYT corpus, the events in body texts and their salience labels are provided by (Liu et al., 2018). For DailyMail and Multi-News, we extract events triggered by either verbal or nominal predicates with *CEO* and automatically annotate them as salient if they also appear in the summary.

[4]Multiple sources of distant supervision might be helpful for more accurate salient event extraction and we leave this for future work.

Dataset	spkmeans		kmeans		aggclus		jcsc		EtypeClus	
	EtypeClus	*CEO*	EtypeClus	*CEO*	EtypeClus	*CEO*	EtypeClus	*CEO*	EtypeClus	*CEO*
ACE2005	.215	.350	.205	.422	.157	.413	.397	.525	.452	.433
MAVEN	.226	.317	.199	.280	.117	.367	.314	.308	.326	.404
RAMS	.197	.246	.189	.202	.186	.208	.204	.214	.240	.206

Table 8: Flat clustering performance (ARI) of different algorithms given events represented by EtypeClus and *CEO*. Higher scores indicate better performance. Contextualized event embeddings improved by external event knowledge in *CEO* help most algorithms achieve much higher ARI than those from EtypeClus. Results evaluated by BCubed-F1 and NMI are similar in Tab. 12.

Event	Method	ACE2005	MAVEN	RAMS
Mention F1 ↑	ETypeClus	.132	.401	.202
	CEO-NY	**.207**	.419	**.213**
	CEO-DM	.161	**.524**	.199
	CEO-MN	.141	.480	.166
Type Coverage ↑	ETypeClus	.848	.970	.885
	CEO-NY	**1.0**	**1.0**	**1.0**
	CEO-DM	.909	**1.0**	**1.0**
	CEO-MN	.909	**1.0**	**1.0**

Table 9: Performance of event mention detection and type coverage with distant supervision from New York Times (NY), Daily Mail (DM), and Multi-News (MN).

Topic	Event Instances & Generated Names
Abortion	*S1*: Women have to have two in-person doctor appointments prior to receiving an *abortion* and must undergo a state-mandated ultrasound. *GPT-J-6B*: abortion
	S2: ...none would have said "because he will make sure to appoint justices to the Supreme Court who, given the chance, will *overturn* Roe." *GPT-J-6B*: abortion:cause:decision:change
	S3: By a vote of 5-to-4, the court's most conservative members *upheld*, for now, a Texas law that, in effect, bans abortions after about six weeks. *GPT-J-6B*: abortion:cause:restrict:app:decision:pass:protect
LGBT	*S4*: ...and the First Amendment that the ADF used in the Supreme Court to argue that Phillips shouldn't be required to bake a cake for a same-sex *wedding*. *GPT-J-6B*: make:marriage:wedding
	S5: The First Amendment Defense Act, as written, would do exactly what Jeb Bush *believes* – and much more. *GPT-J-6B*: make:change:be:create:think:belief
	S6: ..., 35 percent chose "strongly disapprove," showing passion is higher among those opposed to marriage *equality*. *GPT-J-6B*: make:change:election:cause:equality

Table 10: Identified events and type names generated by *GPT-J-6B* for instances sampled from two topics. Refer to Tab. 14 and Tab. 15 for the other 9 topics.

Benefits of External Knowledge on Ontology Inference In Fig. 4, we verify the utility of the external hierarchical event relationship for open-domain ontology induction by comparing performance among *1)* ***plain***: original embeddings without leveraging external knowledge; *2)* ***ae***: fine-tuned embeddings only with the reconstruction loss; *3)* ***depth_1/2/3***: rich embeddings with both reconstruction and contrastive loss. We therefore have the following observations: 1) simply treating event mentions in WordNet as additional instances with the reconstruction loss can hardly guarantee performance gain; 2) selecting event mentions with direct hypernym-hyponym relations (depth_1) as anchors and positives are effective enough to surpass the performance when no external knowledge is utilized.

5.5 Open-domain Event Ontology Inference

We collect articles over eleven topics from Allsides, including the long-term popular topic *elections* and recently heated debate over *abortion* and *gun control rights*. We consider articles tagged with the same topic as an open domain and show their statistics in Fig. 8. For events sampled from *abortion* and *LGBT* corpus, we display the generated type names in Tab. 10, which are highly correlated with their respective topics. The finer granularity of names, the more details about events as well as their contexts are reflected. For instance, the event type of the trigger *overturn* (*S2*) is firstly named with the general token *abortion*, then finer token *cause* and *decision*, and lastly the most precise token *change*. We also observe some less appropriate generation, especially among the general type names, such as *make* and *change* for event *believes* (*S5*) and *equality* (*S6*). We attribute the less accurate coarse types to the single root restriction for the induced event ontology and leave multi-root ontology induction for future investigation.

6 Conclusion

To understand events expressed in open domains free from the restriction of pre-defined ontologies, we propose a new Corpus-based open-domain Event Ontology induction strategy *CEO* to automatically induce hierarchical event ontology structure and provide interpretable type names for further curation. On three event datasets, we find it can capture salient events more accurately, induce ontology structures aligning well with ground truth and generate appropriate coarse-to-fine type names. We also show the broad application of *CEO* on open domains from Allsides.

Limitations

An important caveat to this work is the assumption that all event types in the studied open-domain corpus could be covered by a single tree-structured schema. However, sometimes events in a corpus could be quite different and we can hardly categorize them with a single coarse type as the root node of the ontology tree. Meanwhile, we restrict the induced event ontology in a tree structure. Although event schemas pre-defined by humans in popular event datasets follow the tree structure, it is likely other styles of ontology can better describe events and their relations in emerging corpora. As the first event ontology induction model that can induce a hierarchical event ontology with meaningful names, we advocate more efforts in exploring event ontology in the open-domain setting.

Ethical Consideration

CEO is an effective strategy for event ontology induction that leverages widely-adopted textual data and NLP models pretrained on fairly neutral corpora. To the best of our knowledge, *CEO* helps understand events from all studied datasets in this paper without raising privacy issues or increasing bias in the induced event ontology.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. TopicRank: Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan. Asian Federation of Natural Language Processing.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ines Chami, Albert Gu, Vaggos Chatziafratis, and Christopher Ré. 2020. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *Advances in Neural Information Processing Systems*, 33:15065–15076.

Muhao Chen, Hongming Zhang, Qiang Ning, Manling Li, Heng Ji, Kathleen McKeown, and Dan Roth. 2021. Event-centric natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 6–14, Online. Association for Computational Linguistics.

Sanjoy Dasgupta. 2016. A cost function for similarity-based hierarchical clustering. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 118–127.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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).

Pedro Domingos. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

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.

Carl Edwards and Heng Ji. 2022. Semi-supervised new event type induction and description via contrastive loss-enforced batch attention. *arXiv preprint arXiv:2202.05943*.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. 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 1683–1698, Online. Association for Computational Linguistics.

John Garofolo, David Graff, Doug Paul, and David Pallett. 1993. Csr-i (wsj0) complete ldc93s6a. *Web Download. Philadelphia: Linguistic Data Consortium*, 83.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Luis Guzman-Nateras, Minh Van Nguyen, and Thien Nguyen. 2022. Cross-lingual event detection via optimized adversarial training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5588–5599, Seattle, United States. Association for Computational Linguistics.

Katherine A Heller and Zoubin Ghahramani. 2005. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, pages 297–304.

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.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R. Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–268, Berlin, Germany. Association for Computational Linguistics.

Lifu Huang and Heng Ji. 2020. Semi-supervised new event type induction and event detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 718–724, Online. Association for Computational Linguistics.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Disha Jindal, Daniel Deutsch, and Dan Roth. 2020. Is killed more significant than fled? a contextual model for salient event detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 114–124, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ari Kobren, Nicholas Monath, Akshay Krishnamurthy, and Andrew McCallum. 2017. A hierarchical algorithm for extreme clustering. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 255–264.

Celine Lee, Anjana Tiha, Deng Yuqian, and Tissot Hegler. 2021. English semantic role labeling (srl) demo. `https://github.com/CogComp/SRL-English`.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengzhong Liu, Chenyan Xiong, Teruko Mitamura, and Eduard Hovy. 2018. Automatic event salience identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1226–1236, Brussels, Belgium. Association for Computational Linguistics.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 24–31, Boston, Massachusetts, USA. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Nicholas Monath, Manzil Zaheer, Daniel Silva, Andrew McCallum, and Amr Ahmed. 2019. Gradient-based hierarchical clustering using continuous representations of trees in hyperbolic space. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 714–722.

Benjamin Moseley and Joshua Wang. 2017. Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. *Advances in neural information processing systems*, 30.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Jiaming Shen, Yunyi Zhang, Heng Ji, and Jiawei Han. 2021. Corpus-based open-domain event type induction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5427–5440, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Ben Wang and Aran Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

Dingkang Wang and Yusu Wang. 2018. An improved cost function for hierarchical cluster trees. *arXiv preprint arXiv:1812.02715*.

Sijia Wang, Mo Yu, Shiyu Chang, Lichao Sun, and Lifu Huang. 2022. Query and extract: Refining event extraction as type-oriented binary decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Senhui Zhang, Tao Ji, Wendi Ji, and Xiaoling Wang. 2022. Zero-shot event detection based on ordered contrastive learning and prompt-based prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2572–2580, Seattle, United States. Association for Computational Linguistics.

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1997. Birch: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, 1(2):141–182.

Tianran Zhang, Muhao Chen, and Alex AT Bui. 2020. Diagnostic prediction with sequence-of-sets representation learning for clinical events. In *International Conference on Artificial Intelligence in Medicine*, pages 348–358. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 Evaluation Metrics

Hierarchical Clustering As discussed in §5.3.1, we leverage the following two metrics to compare the induced event ontologies with the ground truth:

- *Dendrogram Purity* (Heller and Ghahramani, 2005): Given the dataset X, the k-th ground-truth flat cluster $\mathcal{C}_k^*$ and the inferred tree structure $\mathcal{T}$, dendrogram purity is the average purity of the least common ancestors of pairs of points belonging to the same ground truth cluster:

$$P(\mathcal{T}) = \frac{1}{|\mathcal{P}^*|} \sum_{k=1}^{K} \sum_{x_i, x_j \in \mathcal{C}_k^*} \mathrm{pur}\Big(\underbrace{\mathrm{lvs}\big(\mathrm{lca}(x_i, x_j)\big)}_{\text{inferred } \mathcal{T}}, \mathcal{C}_k^*\Big),$$

  where $|\mathcal{P}^*|$ represents the number of data point pairs in the same ground-truth cluster, $\mathrm{lca}(x_i, x_j)$ gives the least common ancestor of x_i and x_j in the inferred tree $\mathcal{T}$, $\mathrm{lvs}(n)$ gives a set of leaf node descendants of node n, while $\mathrm{pur}(\cdot, \cdot)$ measures the fraction of data points under its first cluster (i.e., the inferred cluster) that are members of the second (i.e., the ground-truth cluster).

- *Dasgupta's Cost* (Dasgupta, 2016): Good trees acknowledged by Dasgupta cost should cluster data such that similar data points have least common ancestors much further from the root than that of dissimilar data points:

$$C(\mathcal{T}) = \sum_{x_i, x_j \in X} \omega_{i,j} \big|\mathrm{lvs}\big(\mathrm{lca}(x_i, x_j)\big)\big|,$$

  where $\omega_{i,j}$ measures pairwise similarity. In summary, inferred trees with higher purity and lower cost achieve more accurate hierarchical event clustering.

Name Generation *Sim dist* is self-defined to consider both semantic similarity and granularity difference between each pair of reference e_r^i and generated name e_p^j:

$$sim\_dist = 1/(n_r \cdot n_p) \sum_{i,j} \underbrace{\big(1 - |i/n_r - j/n_p|\big)}_{\text{granularity difference}} \cdot \underbrace{\Big(\cos\big(emb(e_r^i), emb(e_p^j)\big) + 1\Big)/2}_{\text{semantic similarity}},$$

where emb is phrase representation from SBERT (Reimers and Gurevych, 2019b).

A.2 Baselines

Hierarchical Clustering

- *Hierarchical Kmeans*: it splits data into two clusters at each iteration using Kmeans [1].
- *Birch* (Zhang et al., 1997): it adopts a dynamically growing tree structure with points inserted greedily using the node statistics and split operation invoked when the branching factor is exceeded.
- *Ward Linkage* (Ward Jr, 1963): the algorithm uses the Ward variance minimization algorithm to calculate the distance between the newly formed cluster and other clusters in the forest
- *Perch* (Kobren et al., 2017): it incrementally builds a tree structure by inserting points as a sibling of their nearest neighbor and performs local tree re-arrangements.
- *gHHC* (Monath et al., 2019): it represents uncertainty over tree structures with vectors in the Poincaré ball and optimizes hyperbolic embeddings of internal nodes using an objective related to Dasgupta's cost (Dasgupta, 2016; Wang and Wang, 2018).
- *HypHC* (Chami et al., 2020): it derives a continuous relaxation of Dasgupta's discrete objective (Dasgupta, 2016) by introducing a continuous analog for the notion of the lowest common ancestor.

Name Generation

- *Most frequent*: the token that appears most in the event triggers are extracted as the cluster name.
- *tf-idf*: following (Shen et al., 2021), we obtain more popular trigger tokens in the studied corpus with regard to their frequency in general corpora.
- *textrank* (Mihalcea and Tarau, 2004), *topicrank* (Bougouin et al., 2013) and *KeyBERT* (Grootendorst, 2020): we cast the cluster name generation as the keyword extraction task, hence the above three strategies are utilized to extract keywords given sentences from the same cluster.

[1] We use Bisecting K-Means as the direct analog of hierarchical KMeans (Moseley and Wang, 2017).

- *wordnet synset*: since WordNet (Fellbaum, 2010) describes the relatedness of word synsets in the hypernym-hyponym format, we introduce the *wordnet synset* strategy where the cluster is named after the least common ancestor hypernym of event triggers.

- *RoBERTa* (Liu et al., 2019): given the context of even triggers, the masked language model *RoBERTa-large* is employed to obtain token probabilities of the trigger position and the token with the highest probability over all instances is adopted as the cluster name.

- *GPT-J* (Wang and Komatsuzaki, 2021): motivated by the in-context learning capabilities of generative language models (Brown et al., 2020), we provide the sentence, the trigger phrase as well as the finest label name of instances sampled from other corpora as the demonstration and acquire the label distribution of testing instances from *GPT-J-6B* [1].

A.3 Autoencoder Design to Improve Event Embeddings

As introduced in §4.3, an autoencoder optimized by reconstruction and triplet loss exploits external event knowledge from WordNet. To extract anchor synsets and their corresponding positive and negative ones, we first define the distance between different synsets in the ontology tree. Considering the synset *treat.v.01* in the partial ontology demonstrated in Fig. 5 as an anchor event: its distance to the first-level hypernym *interact.v.01* is 1 and the second-level hypernym *act.v.01* is 2; furthermore, its distance to the loosely related synset *hash_out.v.01* is 5. Suppose the threshold distance to distinguish positive from negative events is 2, then we treat *interact.v.01* and *act.v.01* as positive event mentions while *hash_out.v.01* as the negative.

Template		Demonstration
Input	sentence:	*Do you think Arafat's death will help or hurt the Israeli-Palestinian peace process?*
	predicate:	*death*
Output	event type:	*Die*

Table 11: Example input-output pair for event type name generation. To retrieve the event type of a test instance, several demonstrations with input and output are randomly sampled and the token with the maximum probability from the PLM is adopted as the type name.

[1] In the unsupervised setting, we use examples from other datasets to provide the finest label name required in the demonstrations. Similar to RoBERTa, the output token with the highest probability across instances in the same cluster is adopted as the label name.

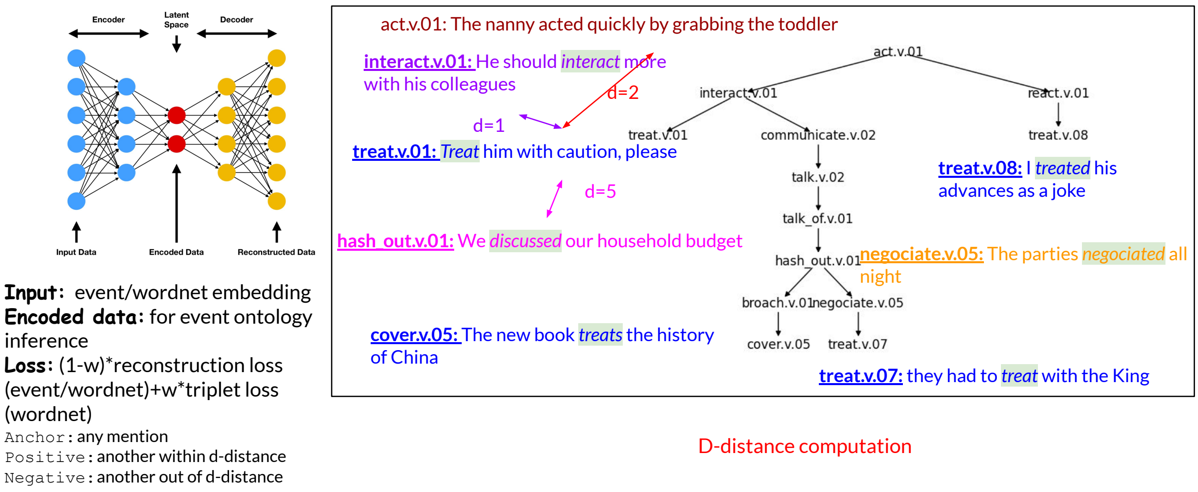


Figure 5: The proposed autoencoder model to improve event embeddings by leveraging external knowledge. The typical autoencoder architecture is optimized with the weighted sum of reconstruction loss and contrastive triplet margin loss (left). The event mention triplet in the form of <anchor, positive, negative> is selected based on the d-distance, which is calculated according to the pre-defined ontology of WordNet (right).

Dataset	spkmeans		kmeans		aggclus		jcsc		EtypeClus	
	EtypeClus	CEO	EtypeClus	CEO	EtypeClus	CEO	EtypeClus	CEO	EtypeClus	CEO
					BCubed_f1					
ACE2005	.378	.500	.398	.536	.351	.527	.533	**.576**	.510	.388
MAVEN	.241	.390	.226	.370	.162	**.421**	.358	.366	.295	**.395**
RAMS	.310	**.371**	.302	.359	.306	.380	.380	**.385**	.351	.364
					NMI					
ACE2005	.524	.629	.537	.631	.481	.628	.626	**.651**	.609	.437
MAVEN	.522	.676	.503	.663	.428	**.695**	.636	.626	.567	.688
RAMS	.665	.701	.662	.688	.663	**.706**	.697	.685	.702	.697

Table 12: Flat clustering performance of different algorithms given events represented by EtypeClus and our \CEO. Higher scores indicate better clustering performance for both metrics.

Dataset	Method	P@1	P@5	P@10	R@1	R@5	R@10	AUC
NYT	KCE (Liu et al., 2018)	.618	.523	0.444	.116	.395	.580	.803
	CEE-IEA (Jindal et al., 2020)	.654	.542	.449	.131	.420	.596	-
	CEO	**.741**	**.604**	**.488**	**.173**	**.493**	**.662**	**.874**
DailyMail	*CEO*	.438	.309	.316	.169	.491	.639	.753
Multi-News	Longformer	.512	.365	.267	.169	.475	.626	.769

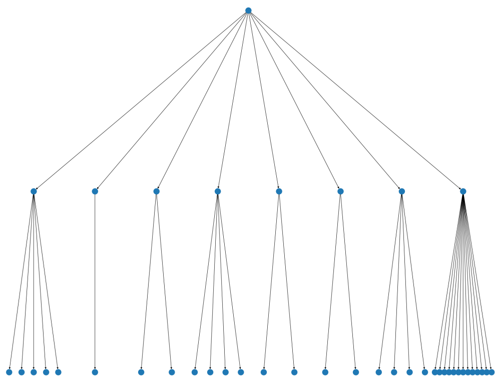
Table 13: Salient Event Detection Performance on the test set of three datasets. The proposed *CEO* fine-tunes the Longformer model to process long documents for contextualized embedding learning. It outperforms baselines with the performance reported in their papers: KCE is a kernel-based approach to learning from different statistical features, while CEE-IEA leverages token-level embeddings of all constituents from the document encoded using BERT.

Topic	Event Instances & Generated Names
Economy	*S9*: Across the nation, protesters are taking to the streets and business owners are *filing* lawsuits objecting to the shutdown rules. *GPT-J-6B*: pay:create:cause:spend:give:claim:seek
	S10: A lockdown targeted to protecting the highest-risk group, people 65 and over, instead of confining all age groups would slash deaths by half but at only half the economic cost of a total *shutdown* ... *GPT-J-6B*: pay:create:cause:l:shut:prevent
	S11: A sharp *devaluation* of the ruble would mean a drop in the standard of living for the average Russian, economists and analysts said. *GPT-J-6B*: pay:create:cause:trade
	S12: But the NBER has other criteria that can constitute a recession, which is particularly applicable to the COVID-19 *crisis* given the speed of the economic downturn. *GPT-J-6B*: pay:create:cause:recession:cat:crisis
Education	*S13*: On July 28, the American Federation of Teachers, the second-largest education *union*, threatened "safety strikes" if reopening plans aren't entirely to its liking. *GPT-J-6B*: pay:education:teach:organ:organization
	S14: ...Obama said during an online commencement address to *graduates* of historically black colleges and universities (HBCUs) on Saturday. *GPT-J-6B*: pay:education:get
	S15: ...a conspiracy theory pushed by the president that accuses Obama of attempting to frame Trump for colluding with Russia to win the 2016 *election*. *GPT-J-6B*: pay:education:cause:app:vote:election
	S16: Yet ... six of them carry the *support* of more than 50 percent of committed liberals ... *GPT-J-6B*: pay:education:cause:enjoy:support
Environment	*S17*: Satellite data published by the National Institute for Space research (Inpe) shows an increase of 85% this year in *fires* across Brazil... *GPT-J-6B*: be:cause:burn
	S18: Indeed, when the scientists drew up their first *report*, in 1990, the diplomats tried so hard to water down their conclusions that the whole enterprise nearly collapsed. *GPT-J-6B*: be:cause:report:find:release
	S19: It is likely going to make the world sicker, hungrier, poorer, gloomier and way more dangerous in the next 18 years with an "unavoidable" *increase* in risks... *GPT-J-6B*: be:cause:make:change:reduce:growth:increase
	S20: Supporters of Mr. Obama's *plan*, including some Democratic-led states and environmental groups, argue it will create thousands of clean -energy jobs and help... *GPT-J-6B*: be:cause:policy:plan
Gun Control Rights	*S21*: LaPierre told Friday's audience "every NRA member is in mourning" because of the Uvalde *shooting*, which he said was the work of a "criminal monster." *GPT-J-6B*: kill:shoot
	S22: ...Houston and the gun *safety* group Moms Demand Action, held protests outside the convention center Friday. *GPT-J-6B*: kill:control:make:cause:safety
	S23: Mr. Biden also *urged* lawmakers to expand background checks for gun purchases, change liability laws to allow gun manufacturers to be sued for shootings... *GPT-J-6B*: kill:control:make:cause:protest:spend:motion:closing:request
	S24: It would raise the federal age of purchasing a rifle from 18 to 21; *restrict* ammunition magazine capacity, though existing magazines are "grandfathered" in... *GPT-J-6B*: kill:control:make:ban:restrict
Immigration	*S25*: There were *immigrants* from El Salvador, China, Honduras and countries in between. *GPT-J-6B*: cause:imigration
	S26: ...She spoke the same night President Trump in a message on Twitter said that Immigration and Customs Enforcement next week would begin *deporting* "millions" of immigrants who are living in the U.S. illegally. *GPT-J-6B*: cause:immigration:death:travel:seek:arrest:hold:removal
	S27: Democrats are likely to face questions about whether they agree with Ocasio-Cortez's comments about concentration camps and the Trump administration's *detention* centers as they return to Washington this week. *GPT-J-6B*: cause:immigration:death:travel:seek:arrest:hold
	S28: ... progressives and Democratic congressional leaders have been pressuring Biden to *end* the use of the policy that turns back families and single adults at the border. *GPT-J-6B*: cause:closing:end:process

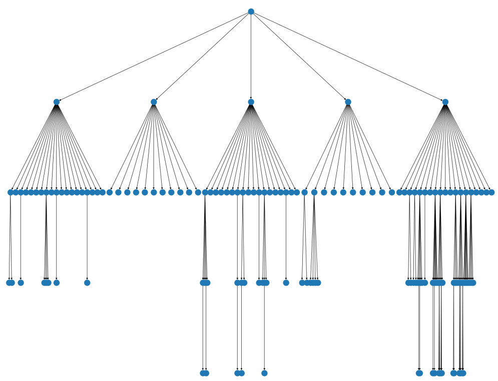
Table 14: Identified events and generated type names for instances sampled from 5 topics of Allsides.

Topic	Event Instances & Generated Names
Elections	*S29*: That's consonant with broad ***support*** for police generally. *GPT-J-6B*: election:debate:cause:support
	S30: A number of prominent figures have explicitly ***called*** for defunding or abolition of police. *GPT-J-6B*: election:win:be:think:make:call
	S31: A majority of members of the City Council of Minneapolis... ***announced*** over the weekend their plans to "begin the process of ending the Minneapolis Police Department." *GPT-J-6B*: election:debate:cause:support:end:announce:campaign
	S32: ...Democratic presidential candidate Joe Biden said Monday he ***opposes*** "defunding the police," declining to embrace a rallying cry that has gained support... *GPT-J-6B*: election:debate:cause:support:attack:contest:opposition
Race	*S33*: In San Francisco, the mob ***demolished*** statues of Ulysses S. Grant, Junipero Serra, and Francis Scott Key. *GPT-J-6B*: kill:cause:protest:crit:ban:celebr:end:destruction
	S34: Last week a mob in downtown Washington, D.C. decided to ***tear*** down a statue of a man called Albert Pike. *GPT-J-6B*: kill:be:cause:removal:destruction:t
	S35: This is a serious and highly organized political ***movement*** . *GPT-J-6B*: kill:be:cause:give:host:protest
	S36: ***Reforms*** have also been proposed under "8 Can't Wait," an initiative released in the wake of the protests by Campaign Zero, a group advocating police reform. *GPT-J-6B*: kill:cause:death:process:reform
Sports	*S37*: The United States ***beat*** the Netherlands in the 2019 Women's World Cup on Sunday 2-0, following a month-long tournament that attracted more attention to the sport... *GPT-J-6B*: protest:be:watch:give:win
	S38: After other hits including "Earned It" and "Save Your Tears,"The Weeknd concluded the 13-minute ***show*** with his smash single "Blinding Lights," a song that references... *GPT-J-6B*: protest:advertising:cause:give:meet:view:coverage:performance
	S39: But this year, many advertising insiders ***expect*** the Super Bowl spots to steer clear of the #MeToo movement opposing the sexual harassment and abuse of women... *GPT-J-6B*: protest:be:watch:give:agreement:predict
	S40: ...city councils, governors and state legislatures all too often respond by ***offering*** lucrative "inducement payments." *GPT-J-6B*: protest:be:watch:give
Technology	*S41*: Moreno accused Assange of behaving badly at the embassy, interfering with building security and attempting to ***access*** security files. *GPT-J-6B*: cause:communication:service:access
	S42: "When users ***violate*** these policies repeatedly, like our policies against hate speech and harassment or our terms prohibiting circumvention of our enforcement measures... *GPT-J-6B*: cause:ban:repe:cance:break:removal
	S43: The InfoWars broadcaster's past tweets will, however, ***remain*** viewable to others while his account is locked in a "read-only" mode. *GPT-J-6B*: cause:control:keep:be:hold
	S44: Mr Jones subsequently ***posted*** a video in which he discusses the move to a separate @Infowars feed - with about 431,000 followers - which he described as being a "sub-account". *GPT-J-6B*: cause:publish:question:post

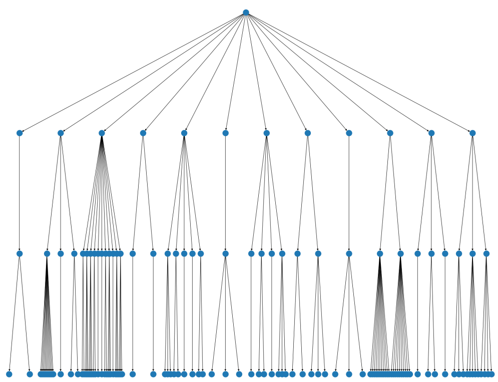
Table 15: Identified events and generated type names for instances sampled from 4 topics of Allsides.



(a) ACE 2005

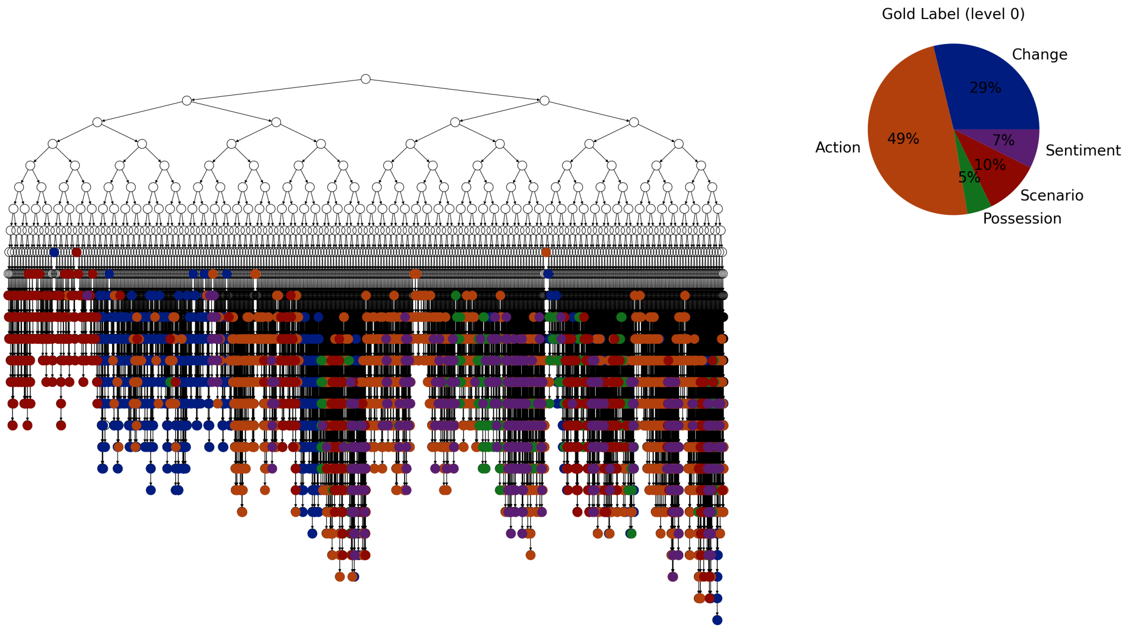


(b) MAVEN

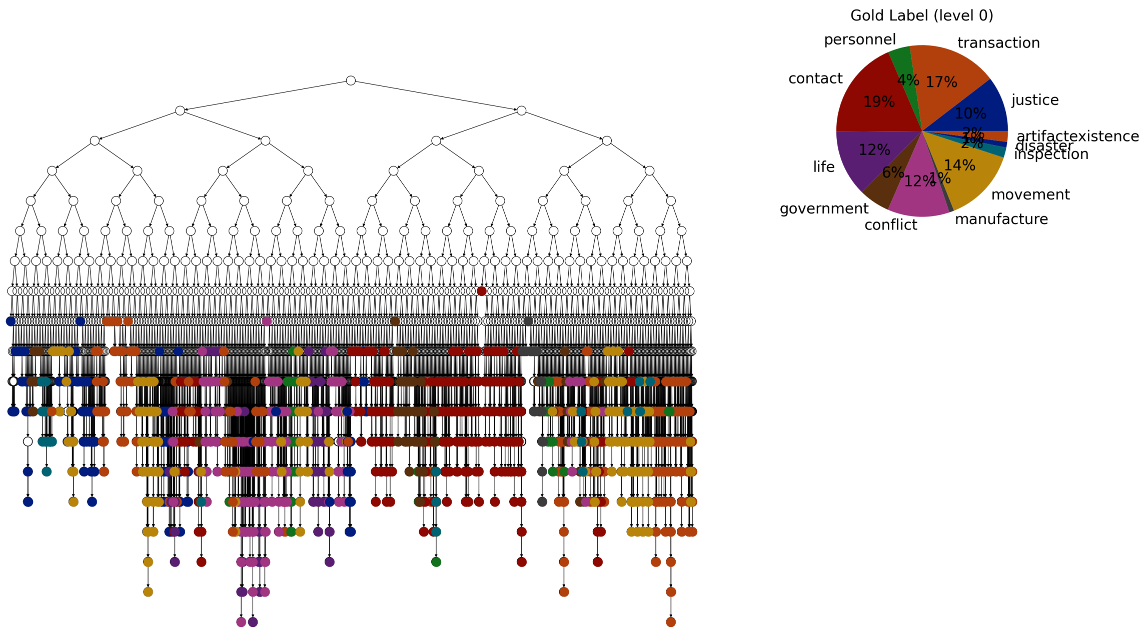


(c) RAMS

Figure 6: Event ontologies of three studied datasets.



(a) MAVEN



(b) RAMS

Figure 7: Event ontology induced by ward linkage algorithm and level-1 event type distributions on MAVEN and RAMS.

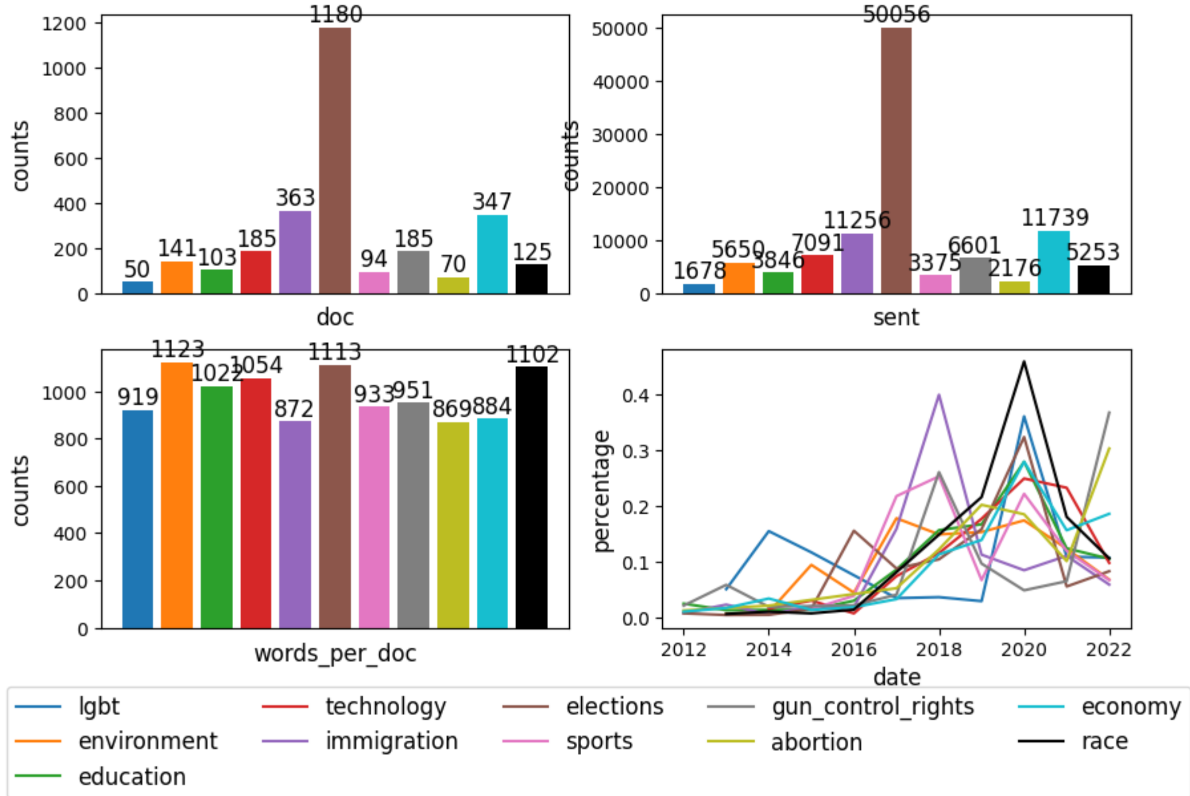


Figure 8: Data statistics of the collected articles concerning 11 topics from Allsides. We record the number of documents, sentences, words per document, and distribution of released dates.