A Multilingual Corpus for Event Coreference Resolution for Social Sciences

Anonymous ACL submission

Abstract

We propose a dataset for event coreference resolution, which is based on random samples drawn from multiple sources, languages, and countries. Early scholarship on event information collection has not quantified the contribution of event coreference resolution. We prepared and analyzed a representative multilingual corpus and measured the performance and contribution of the state-of-the-art event coreference resolution approaches. We found that almost half of the event mentions in documents co-occur with other event mentions and this makes it inevitable to obtain erroneous or partial event information. We showed that event coreference resolution could help improving this situation. Our contribution sheds light on a challenge that has been overlooked or hard to study to date. Future event information collection studies can be designed based on the results we present in this report.

1 Introduction

Event databases are of great utility in research projects in various fields of social sciences. Social actions of groups and individuals, contentious or cooperative interactions between states and societies, and among various social groups all manifest themselves as events. Thus, event data are crucial in understanding a wide variety of social and political phenomena such as modes of political participation, patterns of migration, and social and political conflict. As any type of data that serves as a source of scientific variables, completeness and reliability of event data have direct bearing on the rigor of these studies. Indeed, since many sociological, political scientific, or economic analyses that rely on event databases also inform policy, it is arguable that quality of research has indirect bearing on the well-being of citizens in some manner. This makes maximizing the quality of event databases even a worthier goal.

Social scientists have long been working on creating automated event databases. Conflict and Mediation Event Observations (CAMEO) (Gerner et al., 2002), Integrated Data for Events Analysis (IDEA) (Bond et al., 2003), and PLOVER[1] have been the main proposals of event characterizations in social sciences. Semi-automatic (Nardulli et al., 2015) and automated approaches (Leetaru and Schrodt, 2013; Boschee et al., 2013; Schrodt et al., 2014; Sönmez et al., 2016) have been developed by adopting these formalisms.

At the same time, the NLP community has achieved some consensus on the treatment of events both in terms of task definition and appropriate techniques for their detection (Pustejovsky et al., 2005; Doddington et al., 2004; Song et al., 2015; Getman et al., 2018). However, in order to be useful for social scientists, these formalisms, related language resources, and the automated systems that realize them need to be adjusted or extended in relation to certain cases. For instance the details of the event descriptions and sampling of the documents in the datasets that demonstrate application of these formalisms should reflect the richness and nuances of the events as they are reported in various social and political contexts, dialects, and languages. Moreover, the sampling of the documents to be annotated plays a critical role in determining and prioritizing linguistic characteristics that the automated approaches should handle.

The results yielded by approaches of both communities to date are either not of sufficient quality, require tremendous effort to be replicated with both in- and out-of- distribution data, are immeasurable in terms of quality as there is not any gold standard list of events, or is not comparable to each other (Wang et al., 2016; Ward et al., 2013; Ettinger et al., 2017; Plank, 2016; Demarest and Langer, 2018).

Any new project for creating an event database in this line still finds itself making design decisions such as using only the heading sentences in a news article (Johnson et al., 2016) or not con-

[1] https://github.com/openeventdata/PLOVER, accessed on October 10, 2021.

sidering event coreference information (Boschee et al., 2013; Tanev et al., 2008) without being able to quantify the effect of these decisions on quality of the output. Weischedel and Boschee (2018) assume that event coreference information may not be necessary for forecast model creation because the number of mentions in the news may already be a useful surrogate for some forecasting models. However, the same opinion piece was concluded by acknowledging the value of the event-event relation information. Therefore the effect of incorporating event-event information on use cases in social sciences domain still remains an open issue.

The event coreference, which is in-document in the scope of our study, identification is the least studied phenomenon by both NLP and social scientists. There are still many unknowns, which are either overlooked or ignored, about this phenomenon (Lu and Ng, 2021a). More information in this respect will enable the creation of precise and complete event databases by decreasing the amount of duplication and partiality of event information them (Zavarella et al., 2020). The following are only the first set of questions that should be responded in order to proceed in quantifying event coreference and improve our methodology for event information collection. What is the number of events in a news report in average? How is the information about an event is spread in a document? How information about multiple events co-occurs in a report? What is the prevalence of the expressions that refer to multiple events? How frequently sentences contain information related to multiple events? Does occurrence of event coreference differ across languages? What is the ratio of the documents and events that can benefit from event coreference resolution in a random sample? How do state-of-the-art text processing tools perform on the event coreference task? This report provides answers to majority of these questions by providing a new event coreference corpus that is created by exploiting news articles drawn from various contexts randomly and using a recall-optimized active learning approach. We also demonstrate the performance of various baseline and state-of-the-art approaches to tackle the event coreference resolution task utilizing this corpus.

We present related work in Section 2. Next, the protest event definition, the methodology we applied to create the corpus, and the corpus characteristics are provided in the Sections 3, 5, and 6. The Section 4 describes the conditions that lead us to consider events as the same or separate events. Our effort for tackling event coreference resolution and the results are demonstrated in the Sections 7 and 8. Finally, the Section 9 conclude this report.

2 Related work

The event coreference resolution task was first introduced in the scope of MUC 6 (Grishman and Sundheim, 1996) and MUC 7 (Chinchor, 1998) as a template filling task. Although it was not an explicitly specified task, identifying whether events are coreferent or not was a key component in this task, as it directly affects the number of templates to be filled. Automatic Content Extraction (ACE 2005) dataset (Doddington et al., 2004), ECB (Bejan and Harabagiu, 2008) and its extended version ECB+ (Cybulska and Vossen, 2014), the data released at the relatively recent evaluation campaign Knowledge Base Population (KBP) track at Text Analysis Conference (TAC) (Getman et al., 2018), OntoNotes (Pradhan et al., 2007), and Rich ERE (Song et al., 2015) are the main datasets that contain explicit annotations for event coreference. Although, many event types are covered in these datasets, the coverage is generic in terms of event types and the focus is on linguistic aspects of event manifestations. The analysis of the nuances and context dependent characteristics such as the prevalence in a random sample of news of protest events is not in the scope of these datasets

Majority of the event coreference corpora consists of documents in English. A few of the available datasets are mainly in English and incorporate data in other languages such as Chinese (Doddington et al., 2004; Getman et al., 2018), Catalan (Recasens et al., 2012), and Spanish (Huang et al., 2016; Getman et al., 2018) as well. [2]

The task event coreference resolution has not been in the scope of the studies of the social scientists that work on automated event data collection. The few protest event corpora proposed by Sönmez et al. (2016) and Makarov et al. (2016) do not include event coreference information. Although it is about protest events, work by Huang et al. (2016) focus only on temporal status of the events, which can be past, on-going, and future.

We propose the first multilingual corpus for protest event coreference resolution. The other unique features of the corpus are being based on random sampling and active learning and containing news articles that report a single event using a single trigger as well. These features enable us

[2] A detailed survey of the event coreference datasets is reported by Lu and Ng (2018).

to understand manifestation of events in a representative text collection, improve the methodology for protest event information collection by highlighting the importance of the event coreference in real world event information collection studies, and development and evaluation of event information collection systems.

3 Protest Event Definition

We define a protest as "a collective public action by a non-governmental actor who expresses criticism or dissent and articulates a societal or political demand" (Rucht et al., 1999) (p. 68), and instances or episodes of social conflict, which are based on grievances or aspirations to change the social and political order. Protest events cover any politically motivated collective action which falls outside the official mechanisms of political participation associated with formal government institutions of the country in which the said action takes place. This broad event definition is developed and fleshed out on two levels. First we identify three abstract categories of collective action, namely, political mobilizations, social protests, and group confrontations, in order to define the broad range of events that we focus on. Next, five specific categories of CP events are identified as concrete manifestations of these three modes of collective action. Demonstrations (rallies, marches, sit-ins, slogan shouting, gatherings etc.), industrial actions (strikes, slowdowns, picket lines, gheraos etc.), group clashes (fights, clashes, lynching etc.), armed militancy (attacks, bombings, assassinations etc.) and electoral politics events (election rallies) are the concrete types of events our event ontology encompasses.

We define criteria to which the news stories that report protest events must conform in order to be classified as protest news articles. The criteria are the necessity of civilian actors, and the existence of concrete or implied time and place information which ascertains that the event(s) the report mentions has definitely taken place. Only reports that mention events that took place in the past, or are taking place at the time of writing are labeled as protest news articles. The references to the future (i.e. planned, threatened, announced or expected) events are not labeled as protest, with the exception of threats of or attempts at violent actions.[3] The comparison of our definition with ACE event ontology (Doddington et al., 2004) is provided in Appendix A.

Events are annotated for their semantic types as well. The event types are

Demonstration A demonstration is a form of political action in which a demand or grievance is raised outside the given institutionalised forms of political participation in a country.

Industrial action Industrial actions are events that take place within workplaces or involve the production process in the protest.

Group clashes Group clashes are confrontations that stems from politicized conflicts (e.g. identity or economic interest based or ideological conflicts) between social groups

Armed militacy Politically motivated violent actions that fall within our event definition are included in this category.

Electoral politics These events are rallies, marches or any similar mass mobilizations that are organized within the scope of election campaigns of political parties or leaders.

Other Any CPE which does not fit in one of the categories listed above is marked with this tag.

4 Event Separation for Coreference Annotation

If an event is referred with multiple words in a sentence, these mentions are marked as coreferent. This is the case in *Ex1* and *Ex2* in Table 1. Coreferent event mentions may occur across sentences as well, e.g. *Ex3*. The news articles may report more than one event and pieces of information about one event might not be applicable to the other event. In this case, we need to distinguish different events within the article and link the arguments to the correct event mentions. This is referred to as event separation and is subject to a number of rules to ensure coherence in annotation. Note that in separating events we need to think of the news text rather than the actual reality that the text recounts. That is to say, we are more interested in the separate event references in the text than whether the said events are actually separate from each other in real life. As will be clearer in the examples demonstrated in Table 1, sometimes it is not possible to know or show for certain whether separate event references correspond to separate real life events. For instance, there are two separate events in *Ex4*.

[3] Although planned events and protest threats could have a role in our analysis (Huang et al., 2016), they are neither relevant in the protest reporting context nor their prevalence, which is below 0.5% of a random sample according to our observations, allow their automated analysis.

BJP workers' demonstration is the first event and the attack at the train station is the second event (in the order in which they appear in the document).

The separation of event references is based on difference in at least one of the following:

Time Events that occur at different times are separated from each other. The time difference necessary for separation is 24 hours. Events that continue throughout the same day are not separated even if they are reported to occur at different times of the day.

Location Events which are reported to take place in different locations are separated as different events. Locations can be event places or facilities. An event that has started at some place and continued at another, e.g. a march that started at a location and proceeded at somewhere else, is not separated. However, if an event is happening simultaneously at multiple locations or at multiple locations at different times but not in continuum are separated such that every location reference count as a separate event. Demonstrations in Bangalore and Mysore are annotated as belonging to separate events in *Ex5*, although they share the event trigger *demonstration*.

Participant or organizer Events which are carried out by different participants or organizers with separate goals and motivations are separated. This separation takes place even in cases where different protests occur at the same time and location. The separation is based on event motivations or goals but since motivation info is not something that we annotate and might at times be elusive, we distinguish events based on participants and organizers. The most frequent cases which exemplify this situation are that of counter-protests where two groups of participants or organizers demonstrating against each other and/or with conflicting agendas. Note that in cases where there are multiple types of participants and/or organizers that protest together, the event will not be separated.

Semantic event category Events which occur at the same time, place and facility, and organized and participated by the same participants but have a different semantic category are separated as different events. In other words, as a result of this, the triggers of each event in a document that is separated by its respective event number will have only one semantic category tag. Although this case is rare, it can be encountered when rallies, marches or other types of demonstrations occur during industrial strikes.

Event information can be spread over a document and occur in a sentence that does not contain the respective event mention. This event information is not annotated. This is to say only event information that co-occur with the related event mention in a sentence is annotated. Moreover, there might be event triggers (types or mentions) that are plural such as *Ex5*, i.e. refer to more than one event that are separated. We have a unique procedure for separating these events. In a nutshell, if an article contains a plural event reference, such as "protests" which refers to e.g. two different events, each of which are reported in the article, that article will have three separate event numbers. This is because, the event reference "protests" is counted as an event reference on its own.[4]

5 Methodology for Corpus Creation

A corpus that has the capacity to support creation of automated systems for event information collection in the wild must be representative of the event type occurrence in real life (Halterman et al., 2021). Therefore, our corpus is based on a randomly sampled news articles from online archives of local news sources from India, China, South Africa, Argentine, and Brazil. English data was collected from *The Hindu*, *South China Morning Post (SCMP)*, *New Indian Express*, *Indian Express*, *Guardian*, and *African News Agency* journals. The news articles in Portuguese were retrieved from *Folha* and *Estadao*. Finally, the Spanish documents were gathered from *Clarin*, *Pagina12*, and *La Nacion*. The news archives mainly cover the period between 2000 and 2019. Although the majority of the documents are from random samples, we facilitated a high recall active sampling to extend the random samples in cases they do not contain sufficient number of positive samples for modelling protest events.

The annotation starts with labelling the articles as containing a protest event or not. Next, the same

[4]The reason for this is that, a plural event mention might have different arguments from the singular events that it designates. For instance, in the sentence "The plaza was the scene of protests for the last two weeks" the reference "protests" has the time argument "last two weeks". The references to events that make up these "protests" will have their corresponding and distinct time arguments elsewhere in the article, as in, "last week", and "the week before last week".

Ex1: The students organized a (**e_1: protest**) by (**e_1: marching**) against the payment seat decision.
Ex2: Commenting on the (**e_1: strike**) which was flagged off on Monday, the union secretary stated "(**e_1: it**) will continue as long as our demands are not met.
Ex3: CPI(M) stages (**e_1: protest**) rally in Bhavnagar. The Bhavnagar unit of communist party of India CPI(m) on Friday staged a (**e_1: demonstration**) opposite the local post office here.
Ex4: At noon, BJP workers (**e_1: gathered**) in the square and shouted slogans, condemning the failure of the Union Government in delivering justice to the victims of last year's terror (**e_2: attack**) at the train station where armed militants killed 25 people.
Ex5: Karnataka State Government Employees Association organized (**$e_{1,2}$:demonstrations**) in Bangalore and Mysore yesterday, urging the government not to go ahead with the new retirement scheme.

Table 1: Event coreference examples i) *Ex1*, *Ex2*, and *Ex3* contain event triggers that express the same event, ii) The triggers in *Ex4* are about separate events, and iii) The trigger in *Ex5* denotes events that take place in Bangalore and Mysore.

procedure is applied on the sentences of the documents that are ensured to have protest information by applying adjudication, spotcheck, and error correction. Both at the document and sentence levels, at least one event trigger must occur in the instance to qualify for the positive label. The positively labeled sentences are annotated at token level for event triggers, arguments such as time, place, and event actors, and semantic category of these event triggers. Finally, the event triggers are connected to each other in case they are about the same event. Document and sentence level labelling is applied on an online tool we have developed in-house. The event sentence grouping and token level annotations are performed utilizing FLAT.[5]. Annotators always see complete documents and any annotations that are agreed upon from previous level(s).

We pay particular attention to the quality of the annotations. Detailed annotation manuals were prepared and updated as they are tested against the data. Each annotation on an instance at any level is performed by two graduate students who are studying social or political science and trained on the annotation methodology. Moreover, they were trained about the socio-political context of the country the news articles to be annotated. Therefore, if a news article reports on an event that had not occurred in the target country, this article is only labelled at document and sentence levels. But it is not included in the event coreference dataset. The English text from India, China, and South Africa was annotated by a team of annotators whose native language is Turkish and living in Turkey. The annotations on Spanish and Portuguese text from Argentine and Brazil respectively was prepared by a team of annotators whose native language is Portuguese and live in Brazil.

Disagreements between annotators are adjudicated by the annotation supervisor, who is a political scientist and responsible for maintaining annotation manuals for each annotation task, such as document labelling, sentence labelling, and token level event annotation. The annotation supervisor performs a spotcheck to around 10% of the agreements. Finally, for each task semi-automated quality checks were performed by using the adjudicated data for both training and testing a machine learning model. The disagreements between the predictions and annotations were analyzed by the annotation supervisor. The quality enhancement efforts has enabled us to update around 10% of all of the annotations.

6 Corpus Characteristics

The corpus consists of documents in English (EN), Portuguese (PR), and Spanish (SE), which are represented with 896, 97, and 106 documents respectively. The inter-annotator agreement (IAA) was measured using Krippendorf's alpha (Krippendorff et al., 2016) for the document, sentence, and token level annotations. Table 2 provides the average IAA scores in the rows *Document*, *Sentence*, and *Token* for each language. The columns *Time*, *Trigger*, *Place*, *Facility*, *Participant*, *Organizer*, and *Target* break down the average *Token* scores. The IAA for event coreference annotation was measured by comparing labels of the annotators with the adjudicated annotations using scorch - a Python implementation of CoNLL-2012 average score for the test data (Pradhan et al., 2014). [6] The scores for EN, PR, and ES are 88.58, 89.72, and 68.64.

[5] https://github.com/proycon/flat, accessed on October 10, 2021.

[6] https://github.com/LoicGrobol/scorch, accessed on October 28, 2021.

	English	Portuguese	Spanish
Document	.75	.82	.83
Sentence	.65	.72	.79
Token	.39	.48	.39
Time	.59	.52	.53
Trigger	.38	.44	.45
Place	.41	.47	.49
Facility	.34	.42	.32
Participant	.36	.51	.39
Organizer	.45	.67	.26
Target	.25	.41	.25
Native	Turkish	Portuguese	Portuguese

Table 2: The inter-annotator agreement for document, sentence, and token levels in terms of Krippendorff's alpha. Token level scores are provided for the trigger and its arguments as well. Finally, the row *Native* provides the native language of the annotation teams.

	EN	PR	SE
#docs	896	97	105
#docs1e	532	60	68
#sents	13,584	1,397	2,669
#esents	3,559	352	449
#sents1e	3,255	320	404
#events	1,651	180	183
#events1sent	763	86	82

Table 3: The number of documents (#docs), sentences (#sents), and events (#events) in English (EN), Portuguese (PR), and Spanish (SE). Documents and sentences that contain information about one event (#docs1e and #sents1e) and events mentioned only in one sentence (#events1sent) show the prevalence of event coreference.

	EN	PR	SE
#train	628	67	74
#validation	134	15	16
#test	134	15	16
Positive ratio	.58	.59	.53

Table 4: The number of documents in the train, validation, and test splits for English (EN), Portuguese (PR), and Spanish (ES). The ratio of the documents that contain events is provided in the row *Positive ratio*.

The IAA score for some of the token level annotations are relatively low. This can be speculated to be caused by the native language of the annotators, which is provided in the *Native* column in 2. The quality assurance steps that are 100% double annotation, adjudication of all disagreements, spotcheck of the 10% of the annotations agreed on, and semi-automated annotation error correction ensure the low IAA scores not to affect the utilization of the corpus.

Table 3 demonstrates the number of documents, sentences, and event mentions in the rows *#docs*, *#sents*, and *#events* for English (EN), Portuguese (PR), and Spanish (SE) respectively. Moreover, the Table provides information on the amount of event information that could be identified precisely under the assumptions 1) a document contain information about a single event, 2) a sentence contain information about a single event, and 3) information about an event is reported in a single sentence. The first assumption could capture the information presented in *#docs1e* which shows it holds for 532 (59.38%), 60 (61.86%), and 68 (64.15%) documents. The average number of events in a news articles that reports a protest event is two. The second allow 3,255, 320, and 404 out of 3,559, 352, and 449 sentences to be processed based on this assumption respectively. Around 10% of the sentences contain mentions of multiple separate events, which is around 15% of the total event information. The third is valid only for 763 (46%), 86 (47.77%), and 82 (44.80%) of the events. Although the documents that contain information about a single event are more than the ones that contain event information about multiple events, more than half of the event information occur in documents that contain information about multiple events.

Last but not least, the event mentions that refer to more than one event is around 9% across all languages.

We have created train, validation, and test splits that has the ratio .70, .15, and .15 respectively in order to facilitate experimentation, benchmarking, and reproducibility. The splits are presented in Table 4. The ratio, which is provided in the row *Positive ratio*, of the documents that contain events is more or less the same across splits in a language.

7 Event Coreference Resolution Methodology

We evaluated performance of a state-of-the-art monolingual and multilingual transformer models in an architecture proposed by Yu et al. (2020), which is illustrated in Figure 1, on the corpus. Moreover, we have calculated a dummy baseline score on the validation and test data. The baseline predicts all events as being in the same cluster in a document, i.e., maximum cluster prediction (MaxC). This baseline is the reflection of assuming a document contains information about a single event.

In addition to use the standard threshold, which is .50 for predicting coreference relation, we optimized it by evaluating all values starting from .01 until .99 by increasing the threshold by .01 as a threshold on the validation set for each language.

Neither the models nor the baseline fully utilize the event information that occurs in event mentions that refer to multiple events and sentences that contain event mentions about more than one event. The event label that occurs more than other event labels assigned to an event mention is the final label of the event mention. In case the occurrence frequency of the assigned event labels are the same, the one that occurs first is used.

The sentences that contain more than 512 tokens are ignored if all event mentions are not in the first 512 tokens. This was the case in only nine sentence pairs in Spanish training data. [7]

8 Results

The transformer models utilized are SpanBERT (Lu and Ng, 2021b)[8], RoBERTa (Liu et al., 2019)[9], and mBERT (Devlin et al., 2019)[10]. The training data is set as English and validation and test data is the respective subsets in each language.[11]

Table 5 demonstrates the performance of MaxC, SpanBERT, and RoBERTa on the validation and test sets. The multilingual modeling is achieved using mBERT. All scores are generated using a single random seed, which is 44, and measured utilizing scorch for the scores in terms of F1, MUC, B^3, $CEAF_e$, Blanc, and CoNLL 2012. The CoNLL 2012 score is used for comparing the systems as it is the average of MUC, B^3, and $CEAF_e$ as each of the three metrics represents a different aspects of the performance (Pradhan et al., 2012)

Although, RoBERTa has obtained the best CoNLL 2012 score, which is 82.82, on the English test set, the results of SpanBERT are comparable. The threshold optimization does not help any of these two models. The performance of $mBERT_{EN,EN}$ that is trained and validated on the respective splits of the English data is slightly higher than SpanBERT and RoBERTa. The mBERT models that are trained on English data and validated and tested on respective splits of Portuguese and Spanish data is reported in the rows $mBERT_{EN,PR}$ and $mBERT_{EN,ES}$ respectively. $mBERT_{EN,PR}$ outperforms the baseline by obtaining 81.76 CoNLL 2012 score. However, threshold optimization on validation set does not improve performance on test data. Finally, the performance of $mBERT_{EN,ES}$ remain below the baseline even after threshold optimization.

9 Conclusion

We have explored the prevalence of event coreference in a random sample of news articles collected from multiple sources, languages, and countries. We have found that the news articles contain two events in average and state-of-the-art transformer models can improve determination of separate events in most of the evaluation scenarios.

We aim at tackling multilingual event coreference resolution by first testing and improving the work reported by Phung et al. (2021) Awasthy et al. (2021), and Tan et al. (2021) on our dataset.

References

Parul Awasthy, Jian Ni, Ken Barker, and Radu Florian. 2021. IBM MNLP IE at CASE 2021 task 1: Multigranular and multilingual event detection on protest news. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 138–146, Online. Association for Computational Linguistics.

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

Doug Bond, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles Lewis Taylor. 2003. Integrated data for events analysis (IDEA): An event typology for automated events data development. *Journal of Peace Research*, 40(6):733–745.

[7] The models we have created can be found on https://www.dropbox.com/sh/7j2j3f06kbn5ziv/AACVvvoFe5HH52PSKWTLph2Oa?dl=0

[8] https://huggingface.co/SpanBERT/spanbert-base-cased, accessed on November 15, 2021.

[9] https://huggingface.co/roberta-base, accessed on November 15, 2021.

[10] https://huggingface.co/bert-base-multilingual-uncased, accessed on November 15, 2021.

[11] Although they are not used to train the models for Portuguese and Spanish, the splits are provided for all languages as we believe these splits are critical for benchmarking purposes.

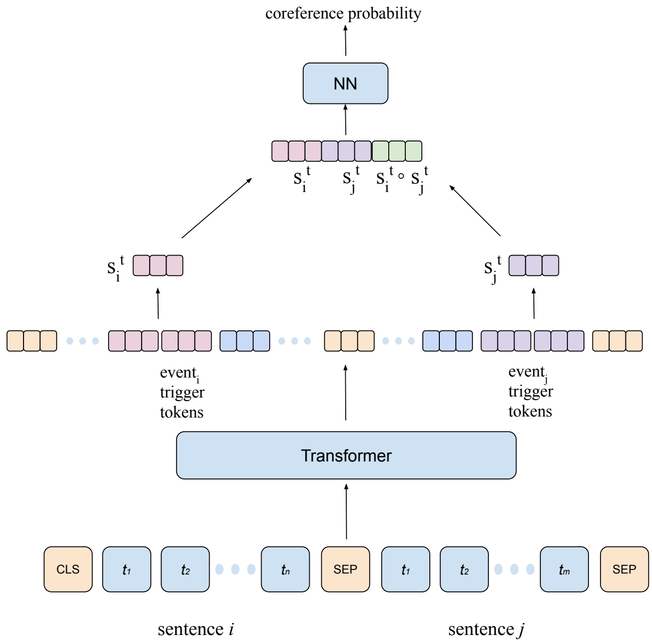

Figure 1: The architecture that was proposed by Yu et al. (2020). The sentence pairs are fed to the transformer model to get token embeddings. To obtain the final trigger vector for a given event mention, the point-wise average of tokens, which are part of the trigger span, of the sentence is calculated, to have fixed size event representations. These tokens might come from different words or as subtokens of a single word. Lastly, the trigger vectors are concatenated with their point-wise multiplication to compose the final representation of trigger pairs in sentences *i* and *j*. The final representation is fed into a two-layer multi layer perceptron (MLP) that yields the probability of being coreferent for a given trigger pair.

	thres	Validation						Test					
		F1	MUC	B^3	$CEAF_e$	Blanc	CoNLL	F1	MUC	B^3	$CEAF_e$	Blanc	CoNLL
$MaxC_{EN}$	-	73.48	90.64	82.64	60.57	86.62	77.95	72.80	91.75	82.76	62.79	86.41	79.10
SpanBERT	.50	79.94	89.86	83.75	68.52	86.13	80.71	80.06	91.11	84.20	71.42	85.84	82.24
	.53	79.71	90.00	83.93	69.43	86.07	81.12	79.95	90.90	84.13	71.49	85.96	82.18
RoBERTa	.50	80.83	91.07	84.12	65.99	87.17	80.39	81.33	93.04	85.21	70.20	88.13	**82.82**
	.54	81.00	91.28	84.15	66.44	86.98	80.62	81.52	92.94	85.03	69.87	87.99	82.61
$mBERT_{EN,EN}$	.50	77.73	90.51	83.27	65.19	85.92	79.66	80.38	92.14	84.63	70.05	86.91	82.27
	.87	76.32	89.89	82.95	66.62	85.06	79.82	79.60	91.44	84.74	71.44	85.85	82.54
$MaxC_{PR}$	-	74.27	93.80	85.38	72.57	86.17	83.92	72.07	89.07	79.24	58.21	84.95	75.51
$mBERT_{EN,PR}$	.50	78.69	94.64	86.35	75.88	86.36	85.62	77.23	92.03	84.59	68.66	87.04	81.76
	.56	78.93	94.64	86.35	75.88	86.36	85.62	77.23	92.03	84.59	68.66	87.04	**81.76**
$MaxC_{ES}$	-	68.92	89.41	74.93	45.37	82.63	69.89	66.86	91.78	79.39	58.85	81.95	**76.67**
$mBERT_{EN,ES}$	.50	73.55	90.47	77.41	50.91	83.72	72.93	67.38	90.27	77.50	54.81	79.78	74.20
	.97	73.86	89.99	78.74	52.23	80.78	73.65	64.44	88.73	76.32	54.13	78.63	73.06

Table 5: Baseline and transformer model performances for event coreference resolution on our corpus. *MaxC* is the baseline calculated by assuming all event mentions in a document refer to the same event. SpanBERT and RoBERTa are trained and tested using respective splits of the English data. mBERT is trained using the English training data and validated and tested on the target language, which is Portuguese for $mBERT_{EN,PR}$ and Spanish for $mBERT_{EN,ES}$. The *thres* column is the probability threshold for determining whether two event mentions are coreferent.

Elizabeth Boschee, Premkumar Natarajan, and Ralph Weischedel. 2013. Automatic Extraction of Events from Open Source Text for Predictive Forecasting. In V.S. Subrahmanian, editor, *Handbook of Computational Approaches to Counterterrorism*, pages 51–67. Springer New York, New York, NY.

Nancy A. Chinchor. 1998. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Agata Cybulska and Piek Vossen. 2014. Guidelines for ecb+ annotation of events and their coreference. In *Technical Report*. Technical Report NWR-2014-1, VU University Amsterdam.

Leila Demarest and Arnim Langer. 2018. The study of violence and social unrest in Africa: A comparative analysis of three conflict event datasets. *African Affairs*, 117(467):310–325.

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).

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10. Association for Computational Linguistics.

Deborah J Gerner, Philip A Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. 2002. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association, New Orleans*.

Jeremy Getman, Joe Ellis, Stephanie Strassel, Zhiyi Song, and Jennifer Tracey. 2018. Laying the groundwork for knowledge base population: Nine years of linguistic resources for TAC KBP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Andrew Halterman, Katherine Keith, Sheikh Sarwar, and Brendan O'Connor. 2021. Corpus-level evaluation for event QA: The IndiaPoliceEvents corpus covering the 2002 Gujarat violence. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4240–4253, Online. Association for Computational Linguistics.

Ruihong Huang, Ignacio Cases, Dan Jurafsky, Cleo Condoravdi, and Ellen Riloff. 2016. Distinguishing past, on-going, and future events: The eventstatus corpus. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 44–54.

Erik W Johnson, Jonathan P Schreiner, and Jon Agnone. 2016. *The Effect of New York Times Event Coding Techniques on Social Movement Analyses of Protest Data*, volume 40, pages 263–291. Emerald Group Publishing Limited.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality and quantity*, 50(6):2347–2364.

Kalev Leetaru and Philip A Schrodt. 2013. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

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. *CoRR*, abs/1907.11692.

Jing Lu and Vincent Ng. 2018. Event coreference resolution: A survey of two decades of research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5479–5486. International Joint Conferences on Artificial Intelligence Organization.

Jing Lu and Vincent Ng. 2021a. Conundrums in event coreference resolution: Making sense of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jing Lu and Vincent Ng. 2021b. Span-based event coreference resolution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13489–13497.

Peter Makarov, Jasmine Lorenzini, and Hanspeter Kriesi. 2016. Constructing an annotated corpus for protest event mining. In *Proceedings of the First*

Workshop on NLP and Computational Social Science, pages 102–107, Austin, Texas. Association for Computational Linguistics.

Peter F. Nardulli, Scott L. Althaus, and Matthew Hayes. 2015. A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data. *Sociological Methodology*, 45(1):148–183.

Duy Phung, Hieu Minh Tran, Minh Van Nguyen, and Thien Huu Nguyen. 2021. Learning cross-lingual representations for event coreference resolution with multi-view alignment and optimal transport. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 62–73, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *CoRR*, abs/1608.07836.

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

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, page 1–40, USA. Association for Computational Linguistics.

Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in ontonotes. In *International Conference on Semantic Computing (ICSC 2007)*, pages 446–453.

James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Saurí. 2005. Temporal and event information in natural language text. *Language resources and evaluation*, 39(2-3):123–164.

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

Dieter Rucht, Ruud Koopmans, Friedhelm Niedhardt, Mark R Beissinger, Louis J Crishock, Grzegorz Ekiert, Olivier Fillieule, Pierre Gentile, Peter Hocke, Jan Kubik, et al. 1999. Acts of dissent: new developments in the study of protest.

Philip A Schrodt, John Beieler, and Muhammed Idris. 2014. Three'sa charm?: Open event data coding with el: Diablo, Petrarch, and the open event data alliance. In *ISA Annual Convention*.

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 The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Çağıl Sönmez, Arzucan Özgür, and Erdem Yörük. 2016. Towards building a political protest database to explain changes in the welfare state. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 106–110. Association for Computational Linguistics.

Fiona Anting Tan, Sujatha Das Gollapalli, and See-Kiong Ng. 2021. NUS-IDS at CASE 2021 task 1: Improving multilingual event sentence coreference identification with linguistic information. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 105–112, Online. Association for Computational Linguistics.

Hristo Tanev, Jakub Piskorski, and Martin Atkinson. 2008. Real-time news event extraction for global crisis monitoring. In *Natural Language and Information Systems*, pages 207–218, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wei Wang, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. 2016. Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1503.

Michael D Ward, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013. Comparing gdelt and icews event data. *Event Data Analysis*, 21(1):267–297.

Ralph Weischedel and Elizabeth Boschee. 2018. What can be accomplished with the state of the art in information extraction? a personal view. *Comput. Linguist.*, 44(4):651–658.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2020. Paired representation learning for event and entity coreference. *ArXiv*, abs/2010.12808.

Vanni Zavarella, Jakub Piskorski, Camelia Ignat, Hristo Tanev, and Martin Atkinson. 2020. Mastering the media hype: Methods for deduplication of conflict events from news reports. In *Proceedings of AI4Narratives — Workshop on Artificial Intelligence for Narratives*.

A Comparison of our Protest Event Definition with ACE Event Ontology

The ATTACK and DEMONSTRATE categories in the CONFLICT heading of the ACE English Annotation Guidelines for Events coding manual (Doddington et al., 2004)[12], have commonalities with our event ontology, however, they are not applicable in the latter setting due to fundamental differences between how events are defined in the two annotation schemes. ACE annotation principles define events as any "specific occurrence involving participants. An event is something that happens" (p.5). This abstracts the actors from the definition, making event type and sub-type definitions neutral in terms of actors. Namely, ACE event type labels are employed based solely on the nature of the occurrences -"acts" in relevant types-regardless of the nature of participants. On the other hand, our event ontology focuses on CPEs, which, by their nature, involve a particular type of actor from the outset, namely, civilian, that is non-state actors. In this respect, the ATTACK event type, which is defined as any "violent physical act causing harm or damage" (p.33) in ACE event coding rules, is not applicable in CPE coding as it includes state actions, such as international wars and military actions against non-state actors. In other words, despite many event examples of the ATTACK type enumerated in ACE manual, such as "attack", "clash", "bomb", "explode", overlap with our event definition, they will be excluded from the latter when their authors are state actors due to their different, non-contentious politics nature.

The second similar event type category in ACE event annotation guidelines is the DEMONSTRATE category. It is defined as including events that occur "whenever a large number of people come together in a public area to protest or demand some sort of official action" (p.34). This definition is better aligned with the CPE ontology we define due to the fact that it designates actions of social and/or political actors that are non-state. However, this definition, in itself, is too restrictive to be applicable in terms of a broad understanding of contentious politics for two reasons. First, as it seems to limit the scope of this event type to spontaneous (that is unorganized) gatherings of people, it excludes certain actions of political and/or grassroots organizations such as political parties and NGOs. Protest actions of such organizations sometimes do not involve mass participation despite aiming at challenging authorities, raising their political agendas or issuing certain demands. Putting up posters, distributing brochures, holding press declarations in public spaces are examples of such protest events. Secondly, the requirement of mass participation in a public area leaves many protest actions such as on-line mass petitions and boycotts, which are not necessarily tied to specific locations where people actually gather, and actions of individuals or small groups such as hunger strikes and self-immolation. Due to the fundamental incompatibilities detailed above, we opted to develop a specific event ontology and annotation guidelines[13] that are different from event definitions in ACE guidelines.

B Reproducibility notes

The following libraries were utilized to conduct the experiments: python == 3.8.10, torch == 1.9.0, pytorch_lightning == 1.3.8, and transformers == 4.8.2

The following hyperparameters are optimized:

Threshold The probability of being coreferent for two event mentions are tested from .01 to .99 by incrementally increasing the threshold by .01.

Learning Rate Each model was trained using the learning rate of 5-e6 which was searched in {1-e5, 5-e5, 1-e6, 5-e6, 1-e7}.

AdamW Eps We used AdamW optimizer for our models. Eps value for our optimizer was selected as 1-e6 which was searched in {1-e6, 1-e7, 1-e8}

Hidden Unit Each model used used identical classifier heads which was a two-layer MLP. 128 was the selected hidden unit which was searched in 32, 64, 128, 256.

We have used fixed parameters for each model.

The number of epochs needed for each model to be trained is 2 to get shared baseline results. The average run-time for an epoch is 10 minutes.

All experiments were performed on the same machine with 10 Intel i9-10900X CPUs, and 2 NVIDIA RTX 2080 (8 GB) GPUs. We did not perform distributed training among the GPUs. Full

[12] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf, accessed on October 10, 2021

[13] The detailed guidelines will be provided either as supplementary material or upon acceptance of the paper.

memory of a single GPU was enough to perform each experiment.