MCECR: A Novel Dataset for Multilingual Cross-Document Event Coreference Resolution

Amir Pouran Ben Veyseh[1], **Viet Dac Lai**[1], **Chien Van Nguyen**[1],
Franck Dernoncourt[2], **Thien Huu Nguyen**[1]
[1]Dept. of Computer Science, University of Oregon, OR, USA
[2]Adobe Research, USA
{apouranb@cs,vietl@cs,chienn,thienn}@uoregon.edu
franck.dernoncourt@adobe.com

Abstract

Event coreference resolution (ECR) is a critical task in information extraction of natural language processing, aiming to identify and link event mentions across multiple documents. Despite recent progress, existing datasets for ECR primarily focus on within-document event coreference and English text, lacking cross-document ECR datasets for multiple languages beyond English. To address this issue, this work presents the first multiligual dataset for cross-document ECR, called MCECR (Multilingual Cross-Document Event Coreference Resolution), that manually annotates a diverse collection of documents for event mentions and coreference in five languages, i.e., English, Spanish, Hindi, Turkish, and Ukrainian. Using sampled articles from Wikinews over various topics as the seeds, our dataset fetches related news articles from the Google search engine to increase the number of non-singleton event clusters. In total, we annotate 5,802 news articles, providing a substantial and varied dataset for multilingual ECR in both within-document and cross-document scenarios. Extensive analysis of the proposed dataset reveals the challenging nature of multilingual event coreference resolution tasks, promoting MCECR as a strong benchmark dataset for future research in this area.

1 Introduction

Event coreference resolution (ECR) is a fundamental task in information extraction that aims to identify mentions of events referring to the same real-world event. An event mention refers to a word or phrase that indicates the occurrence of an event, known as event trigger. The goal of an ECR system is to accurately identify all triggers that pertain to the same event. For instance, consider the following example: "*The leaders of 8 countries with the greatest industry gathered in Paris to discuss global environmental challenges*" and "*The convention of industrialized countries in Paris to combat global warming will be impactful.*" In these sentences, the event mentions "*gathered*" and "*convention*" both refer to a conference related to climate change. Recognizing such coreferences between event mentions within a document or across multiple documents is crucial for achieving a comprehensive understanding of events. This knowledge can be leveraged in various downstream applications, including question-answering, summarization, information retrieval, and knowledge base population.

Numerous methods have been proposed for event coreference resolution, encompassing a range of approaches. Early research explored feature-based models, incorporating traditional feature engineering techniques (Chen et al., 2009a; Yang et al., 2015; Lu and Ng, 2018). Additionally, graph-based models (Chen and Ji, 2009), Integer Linear Programming (Choubey and Huang, 2018), Markov Logic Networks (Lu et al., 2016), and Multi-tasking (Lu and Ng, 2021) approaches have been employed. More recently, there has been a surge of interest in leveraging large language models for event coreference resolution (Nguyen et al., 2016; Tran et al., 2021). For instance, the authors in (Xu et al., 2022) propose an ECR approach that encodes both sentence-level and document-level context using Longformer.

However, most existing work on event coreference resolution has focused on within-document scenarios, limiting its scope. While some prior studies have investigated cross-document event coreference resolution (CDECR) (Barhom et al., 2019; Cattan et al., 2021; Eirew et al., 2022a; Hsu and Horwood, 2022), comprehensive exploration of this task remains limited due to the scarcity of available resources. Moreover, multilingual cross-document event coreference resolution represents an even less-explored setting. The lack of large-scale manually annotated datasets poses a significant challenge in studying ECR in multilin-

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3869–3880

gual cross-document scenarios. Existing datasets for cross-document event coreference resolution are primarily available in English (Cybulska and Vossen, 2014; Vossen et al., 2018), providing limited annotated samples. To address these limitations, this work introduces a novel dataset, namely the Multilingual Cross-Document Event Coreference Resolution (MCECR) dataset, specifically designed to facilitate research in this domain.

To collect a diverse set of articles for annotation for event coreference, we employ news articles published in Wikinews, an online news source covering various domains. Specifically, news articles in the domains of *Politics, Crime, Health, Technology, Sports, Economy* and *Dissasters* are collected from language-specific dumps of Wikinews for five languages: *English, Spanish, Ukrainian, Turkish*, and *Hindi*. These articles serve as seed nodes to collect more related news articles from Google searches. The news articles from Google searches are then filtered based on some criteria to enhance the likelihood of encountering coreferring event mentions. To this end, the collected corpus consists of 1,456 news topics in 5,802 articles.

To annotate the collected corpus, we hire native-speaker annotators for each language. The annotation is performed in two phases: (1) Event Detection and (2) Coreference Resolution. In the first phase, the annotators read the news articles and identify event triggers. As such, we focus on open domain events to broaden the scope of our dataset. Next, the annotators are provided with a pair of event triggers and they need to decide whether the provided event mentions refer to the same real-world event or not. Since there might be a large number of event triggers in the articles of each news topic, we first employ a pre-trained model to identify the challenging event trigger pairs, which will be annotated by the human annotators to improve efficiency. In contrast, the non-challenging event mention pairs (determined by the confidence of the predictions from the pre-trained model) are automatically labeled and sampled for manual verification with the annotators. Using this approach, out of 54,791 events, 4,266 non-singleton event clusters are detected, leading to a high-quality dataset for multilingual ECR. Compared to previous benchmark dataset for this task, i.e. ECB+ (Cybulska and Vossen, 2014), which contains only 722 non-singleton event clusters and supports only English, our proposed dataset provides more annotated samples in a diverse set of languages, serving as a strong benchmark for multilingual cross-document event coreference resolution.

In order to show the potentials of the proposed dataset for cross-document event coreference resolution in different languages, we conduct extensive experiments using the state-of-the-art ECR models. In particular, transformer-based models, joint clustering (Hsu and Horwood, 2022), and hierarchical models (Xu et al., 2022) are employed to study the challenging nature of our MCECR dataset. Additionally, we conduct experiments on the cross-lingual transfer learning settings for cross-document ECR for the first time in the literature, revealing the unique challenges of this problem for future research. Overall, our experiments demonstrate much room for further research on multilingual ECR with our dataset as the foundation to boost progress in this area. *We will publicly release our dataset for future research.*

2 Data Collection

Given the limitations of existing resources for multilingual ECR, our objective is to construct a dataset covering a diverse set of languages and provide coreference links in the cross-document setting. To achieve this, we start by collecting event-rich articles from Wikinews[1], which is an online source for publishing news. We select this source as it provides news articles in different languages and covers a variety of categories. In particular, five languages are chosen, i.e., English, Spanish, Turkish, Ukrainian, and Hindi, and the news articles from their Wikinews dumps are sampled for further processing and annotation. These languages represent different language families, and thus could better exhibit the challenges of cross-lingual ECR. Moreover, we select the following categories for article selection to promote data diversity: (1) Politics and Conflicts, (2) Crime and Law, (3) Health, (4) Science and Technology, (5) Sports, (6) Disasters and Accidents, and (7) Economy and Business. Note that we use language-specific categories to select the news articles for each language.

The selected news articles from Wikinews provide a valuable source of event-rich documents. However, in order to construct a richer dataset with many cross-document event coreference links, it is necessary to collect other articles that have more coreferring event mentions to the events mentioned in the Wikinews articles. To this end, we employ

[1] https://www.wikinews.org

	English	Spanish	Turkish	Ukrainian	Hindi
#News Topics	603	179	50	295	327
#Documents	2,600	694	193	1,123	1,192
Avg Doc. per Topic	4.31	3.88	3.86	3.81	3.65
Avg Doc. Length	557.53	471.89	362.67	340.21	675.91

Table 1: Statistics of MCECR. The document lengths are presented in terms of the numbers of words. Each news topic corresponds to a selected Wikipedia article.

	English	Spanish	Turkish	Ukrainian	Hindi
#Event Mentions	22,445	1,662	8,431	24,229	7,089
Avg. Mention per Doc.	8.84	4.93	45.28	21.86	6.07
#Event Chains	20,040	401	8,004	20,731	5,615
#Non-Singleton Chains	1,280	352	268	1,620	746
Avg. Chain Length	1.12	4.14	1.05	1.17	1.26
#Within-Doc. Co-ref.	482	973	128	410	889
#Cross-Doc. Co-ref.	2,229	61	339	4,054	591

Table 2: Statistics of the annotations in MCECR. Chain Length is computed via the number of event mentions in each chain.

Google search results to obtain such articles. In particular, for each Wikinews article, we retrieve news articles from other sources using the Google search engine. The titles of the Wikinews articles are employed as the queries in our searches as they provide short summaries for the main events in the articles. For each query, the first 50 search results are selected. These search results are further filtered to only keep the documents that have higher chance to contain referring event mentions to those in each Wikinews article. Our filtering criteria is based on the temporal correspondence heuristics (Zhang and Weld, 2013), which suggests the tendency to discuss the same events in different sources at the same or similar times (i.e., close to the occurring time of the events). As such, we only retain the searched articles whose publish dates are within 7 days of the publish dates of the original Wikinews articles to improve the precision to obtain the same events. Note that for each Wikinews article, the search results are limited to being in the same language as the article itself. Using this approach, on average, for each Wikinews article (called a news topic), 2.9 related articles are selected from Google search results. Finally, the textual content of all Wikinews articles and selected articles from Google search results are extracted to be annotated by native speakers in each language. Table 1 shows the statistics of the collected corpus.

3 Annotation

This section discusses the details of our annotation for the collected corpus with event triggers and event coreference links. In this work, we recruit native speakers to annotate the documents in each language. In particular, three annotators per language are hired from the freelancer website Upwork[2]. The hired freelancers are required to be native speakers of the target languages and fluent in English for training and communication. They also need to have experience in data annotation and pass an examination test to verify their ability on event and coreference annotation. Annotators are paid a fixed hourly rate of $15 per hour, which is significantly higher than the minimum wage per hour in their countries. To annotate the corpus we employ two phases: (1) Event Detection: to annotate event triggers, and (2) Event Coreference Resolution: to identify the event mentions that refer to the same real-world events in our collection of news documents.

3.1 Event Detection

In the first phase of annotation, we focus on identifying open-domain event mentions (Sims et al., 2019), aiming to facilitate the annotation process and extend the applicability of the final dataset for different domains. Concretely, annotators are instructed to find all event mentions in text regardless of their event types, i.e., only the spans of event triggers are marked for the annotation. We follow prior work (Sims et al., 2019; Pouran Ben Veyseh et al., 2022) on event detection to define event triggers and annotation guideline in our dataset. Specifically, we limit an event trigger to a word or continuous phrase that most clearly refers to the occurrence of an incident that results in a change of status of real-world entities. Note that due to the nature of some languages, event triggers might consist of multiple words. For example, in Turkish the phrase "*mahkum etmek*", which translates to "*convicted*", should be annotated as an event trigger.

3.2 Event Coreference Resolution

In the next step, we identify the event mentions that refer to the same real-world events for both within-document and cross-document scenarios. Considering each Wikipedia article and its corresponding Google-returned articles as a topic, we follow the annotation guideline in prior work (Cybulska and Vossen, 2014) to only annotate event coreference links between event mentions in the documents of the same topic. To this end, a comprehensive annotation requires all possible pairs of event men-

[2] www.upwork.com

Langauge	Krippendorff's alpha	B-Cubed F-1
English	0.82	0.95
Spanish	0.79	0.89
Turkish	0.81	0.94
Ukrainian	0.81	0.92
Hindi	0.80	0.93

Table 3: Inter-annotator agreement in event detection reported in Krippendorff's alpha and in event coreference resolution reported in B-Cubed F-1

tions in a topic to be labeled by the annotators. However, due to the high number of event mentions in the documents of a topic, the number of pairs will be prohibitively large for annotation. To address this issue, we employ a combination of automatic and manual labeling techniques. Concretely, event pairs whose coreference decisions can be confidently predicted by a pre-trained coreference models will be removed from the pool of event pairs; the remaining pairs will be used for human annotation. As such, our pre-annotation model is based on the multilingual pre-trained language model XLMR (Conneau et al., 2020) while the ECB+ dataset (Cybulska and Vossen, 2014) is leveraged for training data. Given two event mentions e_1 and e_2, our model sends the concatenation of their hosting sentences S_1 and S_2 into the XLMR model to obtain representation vectors for the words, i.e., $[h_1, \ldots, h_n] = XLRM(S_1 \ldots S_2)$. The representations of the two event mentions, i.e., h_{e_1} and h_{e_2}, are then sent to a two-layer feed-forward network classifier to predict the coreference between e_1 and e_2. After training, the model is used to make predictions for all possible pairs of event mentions in the same topics of our dataset. Pairs with a model prediction confidence over 95% are automatically labeled and removed from our human annotation pool. As such, on average, 65% of event pairs are automatically annotated for each language in our dataset. Finally, to perform event coreference annotation for the remaining pairs (called unlabeled pairs), we provide the annotators with the entire context of the hosting documents of the two input event mentions to facilitate the process.

3.3 Annotation Quality and Statistics

To perform event trigger annotation, we first sample 20% of collected articles for each language that will be co-annotated by the three language-specific annotators. Afterwards, the labels that are provided by at least two annotators for each word are selected. For the labels with which all three annotators disagree, the annotators are requested to discuss and resolve the conflicts, leading to a final version of event triggers in our dataset. Finally, the remaining 80% articles for each language will be distributed to the three annotators for separate annotation to accommodate our budget. Next, for event coreference annotation, the unlabeled event mention pairs in a sample of 20% of the topics are used for co-annotation by three annotators for each language while the unlabeled pairs in the remaining 80% will be divided and annotated separately among annotators. We use majority voting to resolve any conflict between annotators during the coreference co-annotation step. The conflict examples are also presented to the annotators to reach an agreement before conducting separate annotation on 80% of the topics.

To assess the quality of our annotations, we evaluate the agreements among three annotators over the co-annotated data for each language. For the event detection phase, as the task is modeled via sequence labeling, we report the Krippendorff's alpha (Krippendorff, 2011) with MASI distance metric (Passonneau, 2006) for the inter-annotator agreement score of each language in our dataset. For the event coreference resolution phase, following prior work (Wang et al., 2022), we report the average of B-Cubed F-1 score (Bagga and Baldwin, 1998) over every pair of event chains detected by the annotators over the same topics. Table 3 shows the results. The table suggests that there is a high agreement between annotators across different languages for both tasks. Finally, to evaluate the quality of the automatically labeled event pairs for coreference, we sample 10% of such pairs and assign them to one annotator for manual verification for each language. Our evaluation shows an accuracy of at least 97% for the automatically labeled event pairs across all the languages to further highlight the quality of our dataset.

Table 2 shows the main statistics of the final annotated corpus. Note that in this table, each event chain corresponds to a fully connected component in a coreference graph where event mentions in a topic serve as the nodes and coreference links are used for the edges.

3.4 Annotation Challenges

ECR annotation is a challenging task involving challenges for both event trigger identification and event coreference resolution. Performing the task for multiple languages further complicates this pro-

	Synonym				XLMR				Hierarchical				Joint			
	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC
English	44.9	89.5	89.4	55.9	33.7	98.3	98.0	61.7	42.6	96.4	97.8	58.8	45.4	98.0	95.3	57.9
Spanish	15.2	38.5	18.7	19.1	85.0	86.7	71.2	42.3	76.2	75.1	62.2	41.2	70.1	75.3	59.8	35.8
Turkish	18.9	88.2	89.0	52.0	32.0	98.5	97.6	62.5	40.8	97.3	97.0	65.2	39.8	95.2	98.2	60.8
Ukrainian	65.1	76.3	76.7	45.6	72.8	98.9	98.3	75.6	73.1	96.2	97.2	74.4	70.6	90.2	97.9	71.3
Hindi	55.1	92.5	88.6	67.7	37.5	89.7	86.0	59.0	35.1	87.2	86.0	57.2	36.2	89.1	85.3	60.2

Table 4: Performance (F1 score) of the models on the test set of each language in Within-Doc settings. The models are trained on the training set of the corresponding language.

	Synonym				XLMR				Hierarchical				Joint			
	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC
English	65.4	87.9	87.3	63.3	59.7	94.8	93.6	65.8	61.7	95.1	94.1	67.2	57.9	93.2	94.2	64.7
Spanish	5.0	24.2	2.8	12.2	13.5	41.4	20.9	20.6	12.6	40.2	19.8	18.5	12.0	38.0	17.8	19.2
Turkish	72.0	79.6	79.3	62.7	75.4	98.4	97.8	75.6	78.9	98.1	97.1	75.2	76.5	97.9	98.0	75.3
Ukrainian	44.1	89.6	89.3	57.6	70.0	92.3	91.5	68.4	69.3	91.2	90.5	67.3	70.1	91.8	90.7	70.6
Hindi	41.3	81.1	73.7	65.0	23.5	85.8	80.6	54.5	21.8	85.0	79.5	53.1	24.0	83.6	79.2	54.0

Table 5: Performance (F1 score) of the models on the test set of each language in Cross-Doc settings. The models are trained on the training set of the corresponding language.

cess. In the following, we summarize the major barriers encountered during the annotation of MCECR and and our adopted solutions:

Event Significance: MCECR is annotated in open domains where any event type is supposed to be annotated. However, it can result in confusions on how significant an event should be annotated. For instance, in the sentence "*The police officer opened the door of his car to inspect the accident.*", while "*inspect*" and "*accident*" are important event mentions to annotate, annotators might disagree on whether or not we should mark "*opened*" due to its trifle. Based on the discussions, we decide to only label significant event mentions to improve the data quality.

Conflicting Coreference: Our coreference annotation presents one pair of event mentions at a time for annotators to simplify the required task. The coreference annotation r for a pair of event mentions (e_1, e_2) in two different event chains C_1 and C_2 ($e_1 \in C_1$ and $e_2 \in C_2$) will thus induce the same relation r for other pairs of event mentions between C_1 and C_2. As such, a conflicting situation might occur if the annotators later vote for a different relation from r for another event mention pair (e_1', e_2') ($e_1' \in C_1$, $e_2' \in C_2$), causing confusion for our dataset. To resolve these situations, we require the annotators to discuss the conflicts and update the annotations as the final step to generate our dataset.

Lack of Background: To identify the event mentions that refer to the same real-world events, in some cases, the context of the text itself is not sufficient. Specifically, the annotators may require information about the events that are not directly presented in the articles. In these cases, there might be conflicting annotations for event coreference relations between annotators due their different background for events. To address this issue, we require the annotators to limit the coreference annotation only to those that are most confidently apparent in the context of the presented documents.

4 Experiments

To facilitate the evaluation and development of multilingual ECR models, this section studies how typical ECR models perform on our proposed dataset MCECR. In the literature, ECR is often modeled as a binary classification task (Hsu and Horwood, 2022; Ravi et al., 2023). Given two event mentions (in the same or different documents) along with their context, typical ECR models first concatenate the context for the two mentions to form a single input text $W = w_1, w_2, \ldots, w_n$, where $1 \leq s_1 \leq e_1 < s_2 \leq e_2 \leq n$ are the start and end indexes for the spans of the first and second event mentions in W. The models then aims to classify W into 1 or 0 to indicate whether the two event mentions corefer to each other or not.

In the inference time, once the labels for all pairs of event triggers in a document (for within-document ECR) or across documents in a topic (for cross-document ECR) are predicted, the event chains will be constructed for coreference evaluation. To evaluate the performance of the models, following prior work (Wang et al., 2022), we report the coreference evaluation metrics MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$ (Luo, 2005), and BLANC (Recasens and Hovy, 2011). We use the implementations provided

by (Wang et al., 2022) to compute these metrics.

To prepare the proposed MCECR dataset for model evaluation, we divide the topics (i.e., Wikipedia articles) for each language into train/dev/test sets with a ratio of 80/10/10. For each set, we create two versions: (1) Within-Doc: to use only coreference links for event mentions in the same documents to create data, and (2) Cross-Doc: to use all coreference links for event mentions in the documents of the same topic to form event chains and data. Note that in the Cross-Doc setting, two event triggers might be in the same document or in different documents of the same topic. For the experiments, to address documents with long context, for both settings, we limit the context of each event mention to three surrounding sentences, i.e., the sentence that contains the event trigger plus the previous and next sentences.

4.1 Models

We evaluate the following ECR models in the experiments:

Synonym: In this baseline, the semantic similarity of the event trigger words are used to determine the coreference relations. To generalize this approach to all languages, we employ the contextualized representations of words obtained by running the multilingual pre-trained model XLMR over the input text W, i.e., $h_1, \ldots, h_n = XLMR(w_1, \ldots, w_n)$. Next, the representations for the event mentions are computed using the averages of their word representations, i.e., in spans (s_1, e_1) and (s_2, e_2). Finally, the cosine similarity between the event mention representations is computed to capture their semantic similarity score. The model's prediction will be positive if the similarity score is higher than a threshold α and negative otherwise in this approach.

Fine-Tuned XLMR: This baseline fine-tunes the multilingual pre-trained language models XLMR (Conneau et al., 2020) for ECR. The input text W is first encoded using the XLMR model: $h_{[CLS]}, h_1, \ldots, h_n = XLMR([CLS], w_1, \ldots, w_n)$. Next, the averages of word representations in the text spans of event mentions are used for their representations h_{t_1} and h_{t_2}. Finally, the concatenation of $h_{[CLS]}$, h_{t_1} and h_{t_2} is sent to a two-layer feed-forward network with softmax in the end to obtain a probability distribution for coreference prediction.

Hierarchical: This baseline (Xu et al., 2022) represents the context of the event mentions in three different levels, i.e., sentence, document, and topic. For the sentence level, the event triggers are masked with $[MASK]$ in their hosting sentences that will be encoded by a transformer-based language model. In the original model, the $[MASK]$ representations are then used to predict event types for the mentions. However, as event types are not available in our dataset, we instead predict the actual event triggers in this model. For the document level, the entire document of each event mention is encoded by Longformer (Beltagy et al., 2020) for English and by the long version of XLMR for non-English text[3]. For the topic-level representation, a Variational AutoEncoder (VAE) model is employed to infer the topic of each event mention based on the words in its context. The representations for the event mentions in all three levels are then combined to perform coreference prediction for the event mention pair.

Joint Clustering: In this baseline (Hsu and Horwood, 2022), instead of pair-wise prediction, the context for each event mention, i.e., its hosting sentence and the first two sentences from the beginning of its hosting document, is encoded by an XLMR-based encoder. The representations of the event mentions are then sent to an agglomerative clustering model to form the clusters (i.e., chains) of event mentions. Note that the encoder is trained with the Siamese network architecture so the contexts of the event mentions with coreference relations are represented more closely.

We tune the hyper-parameters for the fine-tuned XLMR and Synonym models using the MUC scores over development data of the English datasets. For the fine-tuned XLMR model, we choose $1e$-3 for the learning rate, 16 for batch size, and 200 for the dimensionality of the feed-forward networks. The prediction threshold α for Synonym is set to $\alpha = 0.9$. The hyper-parameters for the other models, i.e., Hierarchical (Xu et al., 2022) and Joint (Hsu and Horwood, 2022), are inherited from their original papers.

4.2 Results

Monolingual: We first evaluate the ECR models on our dataset in the monolingual settings where the models are trained and tested over data of the same language. Tables 4 and 5 show model performance for the Within-Doc and Cross-Doc settings. From these tables, it is clear that all of the existing

[3] https://huggingface.co/markussagen/xlm-roberta-longformer-base-4096

	XLMR				Hierarchical				Joint			
Target	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC
Spanish	68.2	72.3	48.9	30.0	60.2	70.2	50.6	31.2	66.5	71.2	41.3	27.3
Turkish	29.8	90.6	89.3	57.2	27.3	88.0	86.4	55.3	28.2	87.3	85.1	56.3
Ukrainian	65.4	87.3	89.4	66.4	61.1	85.2	87.9	62.4	63.3	85.4	82.0	62.5
Hindi	29.3	75.3	71.3	50.7	30.2	71.8	70.0	52.0	30.3	74.1	70.9	50.4

Table 6: Performance (F1 score) of the models on the test set of each language in Within-Doc setting. The models are trained on the training set of the corresponding language.

	XLMR				Hierarchical				Joint			
Target	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC	MUC	B^3	$CEAF_e$	BLANC
Spanish	10.2	32.3	18.7	15.3	9.9	31.4	17.6	15.0	9.0	31.8	17.7	14.8
Turkish	68.2	90.8	89.3	71.8	65.1	87.9	88.3	70.2	65.4	85.4	86.1	71.9
Ukrainian	63.0	83.3	89.7	60.2	60.3	80.9	88.2	59.7	60.4	83.0	88.2	57.2
Hindi	16.7	80.9	71.2	50.3	17.0	72.2	71.0	49.3	15.1	71.9	70.1	46.2

Table 7: Performance (F1 score) of the models on the test set of each language in Cross-Doc setting. The models are trained on English and tested on the target language.

models still underperform a perfect model. Specifically, for the Within-Doc setting, the average F1 score using the MUC metric across all models and languages is 50.52%. The same number for the Cross-Doc setting is 48.47%. These numbers show that the proposed MCECR dataset is challenging and further research is necessary to improve the performance of the ECR models on this benchmark. Another observation from the tables is that there is a considerable difference between the performance of the models across different languages. For example, the average F1 score using the BLANC metric across all models for each language in the Cross-Doc setting, ranges from 18.21% in Spanish to 73.12% in Turkish. Such differences of model performance over different languages corroborate the importance of further exploration for the challenges of ECR in multilingual settings.

In addition, the tables suggests that the Cross-Doc setting for ECR is more challenging than Within-Doc ECR. In particular, the average F1 score across all metrics, languages, and models in the Within-Doc setting is 70.36% while number for the Cross-Doc setting is only 64.84%. Finally, comparing the performance of the models, we observe the best performance is generally achieved by XLMR. The higher performance of the simple model XLMR compared to previous state-of-the-art baselines (i.e., Hierarchical and Joint) indicates that the existing architectures are more tailored to English and cannot perform well in other languages. Also, compared to the simple baseline Synonym, the better performance of fine-tuned XLMR-based model highlights the importance of using effective encoders for ECR.

Cross-Lingual: In order to shed more light on the operation of the existing methods for ECR, we examine the models in the cross-lingual transfer learning setting. Tables 6 and 7 show the performance of the models for Within-Doc and Cross-Doc ECR when they are trained in English training data and directly evaluated on test data of other languages. Here, as training data is not needed for the Synonym baseline, we do not report the performance of this model in the tables. The first observation from these tables is that the performance of the models significantly drops when tested in the cross-lingual setting compared to the monolingual setting. Concretely, the average F1 score of all models across all metrics and languages decreases by 7.44% in the Within-Doc setting and 8.59% in the Cross-Doc setting. The significant performance loss in cross-lingual transfer learning indicates the differences in ECR patterns across different languages, calling for more research to address the challenges of cross-lingual learning for ECR. Moreover, from the tables, it is obvious that performance losses in different languages are not the same. Specifically, the performance loss of all models in the Within-Doc setting across all metrics ranges from 2.33% in Spanish to 12.33% in Hindi. This number for the Cross-Doc setting ranges from 2.32% in Spanish to 8.28% in Turkish. These variances reveal the necessity to explore the challenges of ECR for specific languages. For example, the lower performance loss in Spanish compared to Hidi in the cross-lingual evaluation can be attributed to the higher similarity between Spanish (the target language) and English (the source language). Finally, considering performance of the models in

the cross-lingual evaluation, we observe the same pattern as in the monolingual setting. In particular, XLMR tends to outperform the other baselines over different target languages, which confirms the effectiveness of this model for multilingual and cross-lingual learning in ECR.

5 Related Work

Event coreference resolution (ECR) is one of the important tasks for any event understanding and information extraction systems. To this end, there has been a considerable body of prior work for this problem. We study the prior work for ECI in two dimensions, i.e., datasets and models:

Datasets: ECR data has already been provided in some of the existing event extraction datasets. Specifically, ACE 2005 (Walker et al., 2006), MUC (Grishman and Sundheim, 1996), TAC KBP (Ellis et al., 2015, 2016; Getman et al., 2017), OntoNotes (Pradhan et al., 2007), and MAVEN (Wang et al., 2022) are the popular event datasets with manually-annotated event coreference information. However, such datasets are mainly developed for English and some popular languages, i.e., Spanish, Arabic, or Chinese. Also, each previous ECR dataset on its own only supports at most 3 languages (e.g., ACE 2005 and TAC KBP). As such, these datasets cannot extensively evaluate models in less popular languages to better support multilingual research for ECR. Most importantly, all of these datasets are only annotated for within-document event coreference that hinders model development for cross-document ECR with multiple languages.

Regarding cross-document ECR, ECB+ (Cybulska and Vossen, 2014) has served as the major dataset to boost research progress for this problem. However, the lack of annotations in non-English documents is a critical shortcoming in ECB+ that prevents multilingual learning research for cross-document ECR. Also, it is noteworthy that ECB+ provides much less non-singleton event clusters than our dataset (i.e., 722 vs. 4,266), making ECB+ less suitable for developing data-hungry deep learning models. In contrast, MCEMR represents the first dataset that annotates both within-document and cross-document event coreference for multiple languages (i.e., beyond English). Compared to existing ECR datasets, our dataset supports the largest number of languages (i.e., five languages), covering Turkish, Ukrainian, and Hindi for the first time in ECR research. With much more non-singleton event clusters, our dataset also enables training of larger models for ECR.

In addition to manually-annotated datasets, there are some other ECR datasets that are automatically collected and annotated, including MEANTIME (Minard et al., 2016), GVC (Vossen et al., 2018), and WEC (Eirew et al., 2021)[4]. However, due to the inherent noises in the fully automatic annotation, these datasets cannot guarantee the highest quality for multiple languages to provide reliable resources for model development for ECR. As such, our MCECR dataset leverages human annotation to control and produce a higher-quality dataset in multilingual languages for both with-document and cross-document ECR to significantly facilitate future research in this area. Finally, due to the relatedness of ECI and the event and event-event relation extraction (EERE) tasks (Do et al., 2011; Man et al., 2022, 2024), we also note some recent multilingual datasets for event extraction (Veyseh et al., 2022; Pouran Ben Veyseh et al., 2022) and EERE (Lai et al., 2022b,a).

Models: The ECR task in the literature has been approached with different methods ranging from feature-based models to deep learning methods. In particular, for the feature-based approach (Ahn, 2006; Chen et al., 2009b), the typical models for ECR have employed SVMs (Lu et al., 2016), Markov Logic Network (Chen and Ng, 2016), and Integer Linear Programming (Choubey and Huang, 2018). Recently, deep learning has been used extensively to solve ECR (Barhom et al., 2019; Choubey et al., 2020; Eirew et al., 2022b). The authors in (Huang et al., 2019) employ LSTM to encode input text and model the compatibility of arguments in event clusters. More recently, the application of large language models, e.g., BERT or RoBERTa, has increased in ECR models. The authors in (Hsu and Horwood, 2022) employ a contrastive learning technique to train the RoBERTa-based model for ECR. In (Xu et al., 2022), the authors leverage BERT-based encoders to encode local and global context for events. For multilingual learning, (Phung et al., 2021) explores cross-lingual transfer learning for within-document ECR with adversarial training. However, due to the lack of necessary datasets, none of those previous work has explored multilingual cross-document ECR as we do. Finally, Finally, it is worth noting that the modeling

[4]Note that in MEANTIME the English portion is manually labeled. In WEC, the evaluation sets are also manually labeled.

approaches for ECR shares some similarities with the popular task of Relation Extraction in Information Extraction (Veyseh et al., 2020b,a).

6 Conclusion

In this work, we introduce MCECR, a multilingual event coreference resolution dataset. Compared to previous dataset, MCECR is the first ECR dataset that provides annotation for both within-document and cross-document event coreference for multiple languages. Our dataset is annotated on Wikipedia articles and related news articles obtained from Google searches in 5 different languages (i.e., English, Spanish, Turkish, Ukrainian, and Hindi). We study the challenging nature of this dataset by evaluating the performance of strong baselines in monolingual and cross-lingual settings. Our experiments reveal the necessity of further research on the proposed MCECR dataset to improve the performance of ECR models in multilingual learning.

Limitations

The proposed MCECR dataset is meant to promote future research on multilingual and cross-document event coreference resolution. Although our experiments show the difficulty of this task and the necessity for future work, we highlight the following limitations and risks involved in the proposed dataset: (1) Lack of event types: In the proposed dataset, we aim to annotate events in general domains, so no event types is presented. This could be restricting for the methods that rely on event types to identify coreference; (2) Lack of event arguments: Some prior work for ECR resorts to the consistency between event arguments to identify event chains. MCECR does not annotate event arguments, thus hindering the application of argument-based methods for ECI; (3) Noise in annotation: As mentioned in the annotation details, we employ a pre-trained ECR model to identify the easy event mention pairs for coreference and remove them from the pool of annotation. This is necessary to address the prohibitively expensive costs for comprehensively annotating every possible pair of event mentions in the dataset. This method also lead to a high-quality dataset as demonstrated in our human verification step. However, this approach is still not perfect and it might still introduce a small portion of noises/errors, which could be addressed to further improve the dataset.

Acknowledgements

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112, the NSF grant CNS-1747798 to the IUCRC Center for Big Learning, and the NSF grant # 2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

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

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, pages 5100–5107.

Chen Chen and Vincent Ng. 2016. Joint inference over a lightly supervised information extraction pipeline: Towards event coreference resolution for resource-scarce languages. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2913–2920.

Zheng Chen and Heng Ji. 2009. Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 54–57.

Zheng Chen, Heng Ji, and Robert Haralick. 2009a. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 17–22.

Zheng Chen, Heng Ji, and Robert Haralick. 2009b. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In *Proceedings of the Workshop on Events in Emerging Text Types*.

Prafulla Kumar Choubey and Ruihong Huang. 2018. Improving event coreference resolution by modeling correlations between event coreference chains and document topic structures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 485–495.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *EMNLP*.

Alon Eirew, Avi Caciularu, and Ido Dagan. 2022a. Cross-document event coreference search: Task, dataset and modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 900–913.

Alon Eirew, Avi Caciularu, and Ido Dagan. 2022b. Cross-document event coreference search: Task, dataset and modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 900–913, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2498–2510.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M. Strassel. 2015. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *Proceedings of the 2015 Text Analysis Conference, TAC 2015, Gaithersburg, Maryland, USA, November 16-17, 2015, 2015*.

Joe Ellis, Jeremy Getman, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M. Strassel. 2016. Overview of linguistic resources for the TAC KBP 2016 evaluations: Methodologies and results. In *Proceedings of the 2016 Text Analysis Conference, TAC 2016, Gaithersburg, Maryland, USA, November 14-15, 2016*.

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie M. Strassel. 2017. Overview of linguistic resources for the TAC KBP 2017 evaluations: Methodologies and results. In *Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, USA, November 13-14, 2017*.

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*.

Benjamin Hsu and Graham Horwood. 2022. Contrastive representation learning for cross-document coreference resolution of events and entities. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3644–3655.

Yin Jou Huang, Jing Lu, Sadao Kurohashi, and Vincent Ng. 2019. Improving event coreference resolution by learning argument compatibility from unlabeled data. 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 785–795.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. In *University of Pennsylvania*.

Viet Dac Lai, Hieu Man, Linh Ngo, Franck Dernoncourt, and Thien Nguyen. 2022a. Multilingual SubEvent relation extraction: A novel dataset and structure induction method. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5559–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022b. MECI: A multilingual dataset for event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2346–2356, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jing Lu and Vincent Ng. 2018. Event coreference resolution: A survey of two decades of research. In *IJCAI*, pages 5479–5486.

Jing Lu and Vincent Ng. 2021. Constrained multi-task learning for event coreference resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4504–4514.

Jing Lu, Deepak Venugopal, Vibhav Gogate, and Vincent Ng. 2016. Joint inference for event coreference resolution. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3264–3275.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Hieu Man, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Mastering context-to-label representation transformation for event causality identification with diffusion models. In *The AAAI Conference on Artificial Intelligence (AAAI)*. Association for the Advancement of Artificial Intelligence.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11058–11066.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422.

Thien Huu Nguyen, , Adam Meyers, and Ralph Grishman. 2016. New york university 2016 system for kbp event nugget: A deep learning approach. In *TAC*.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. In *The International Conference on Language Resources and Evaluation (LREC)*.

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.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Nguyen. 2022. MINION: a large-scale and diverse dataset for multilingual event detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2286–2299, Seattle, United States. 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.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwarz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. *arXiv preprint arXiv:2302.09715*.

Marta Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural language engineering*, 17(4):485–510.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Hieu Minh Tran, Duy Phung, and Thien Huu Nguyen. 2021. Exploiting document structures and cluster consistencies for event coreference resolution. 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 4840–4850, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020a. Exploiting the syntax-model consistency for neural relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8021–8032, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020b. A joint model for definition extraction with syntactic connection and semantic consistency. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. MEE: A novel multilingual event extraction dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

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

Christopher Walker, Stephanie Strassel, Julie Medero, , and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. In *Linguistic Data Consortium*, page 57.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941.

Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2022. Improving event coreference resolution using document-level and topic-level information. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6765–6775.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent Bayesian model for event coreference resolution. *Transactions of the Association for Computational Linguistics*, pages 517–528.

Congle Zhang and Daniel S. Weld. 2013. Harvesting parallel news streams to generate paraphrases of event relations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1776–1786, Seattle, Washington, USA. Association for Computational Linguistics.