MECI: A Multilingual Dataset for Event Causality Identification

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

Abstract

Event Causality Identification (ECI) is the task of detecting causal relations between events mentioned in the text. Although this task has been extensively studied for English materials, it is under-explored for many other languages. A major reason for this issue is the lack of multilingual datasets that provide consistent annotations for event causality relations in multiple non-English languages. To address this issue, we introduce a new multilingual dataset for ECI, called MECI. The dataset employs consistent annotation guidelines for five typologically different languages, i.e., English, Danish, Spanish, Turkish, and Urdu. Our dataset thus enable a new research direction on cross-lingual transfer learning for ECI. Our extensive experiments demonstrate high quality for MECI that can provide ample research challenges and directions for future research. We will publicly release MECI to promote research on multilingual ECI. The dataset is available at https://github.com/nlp-uoregon/meci-dataset.

1 Introduction

Event Causality Identification (ECI) is an important Information Extraction (IE) task that aims to identify causal relations between event mentions in text. For example, in the sentence *"After **inspection** of his computer , officers **found** that he was interested..."*, a ECI system should detect a causal relation between two events *"**inspection**"* $\xrightarrow{cause}$ *"**found**"*. ECI can provide valuable information for various applications such as event timeline construction (Shahaf and Guestrin, 2010), question-answering (Oh et al., 2016), future event forecasting (Hashimoto, 2019), and machine reading comprehension (Berant et al., 2014).

Due to its applications, ECI has been extensively studied in the natural language processing community over the past decade. The vast majority of methods for ECI involve feature engineering models (Do et al., 2011; Hu and Walker, 2017; Hashimoto, 2019; Ning et al., 2018; Gao et al., 2019) and recent deep learning architectures (Kadowaki et al., 2019; Zuo et al., 2021b; Liu et al., 2021; Zuo et al., 2021a; Man et al., 2022a). As such, the creation of large annotated datasets, e.g., EventStoryLine (Caselli and Vossen, 2017), has been critical to the development of ECI study. However, existing datasets for ECI only annotate causal relations between event mentions in data of a single language, i.e., mainly for English (Caselli and Vossen, 2017; Cybulska and Vossen, 2014; O'Gorman et al., 2016). On the one hand, this leaves many other languages unexplored for ECI, posing an important question about the generalization ability of existing methods to other languages. For instance, Spanish, Danish, and Turkish are not covered in those separated datasets for ECI. Moreover, the current single-language datasets for ECI tend to employ different annotation guidelines that prevent their combination into a larger corpus and cross-lingual transfer learning research to train and evaluate models in different languages. In all, the annotation discrepancy and limited language coverage hinder the research and development of the ECI in various dimensions, necessitating a new dataset with broader coverage for ECI.

To address this issue, this paper introduces a Multilingual Event Causality Identification (MECI) dataset to standardize and foster future research in multilingual ECI. Particularly, we present a large-scale ECI dataset for five languages, i.e., English, Danish, Spanish, Turkish, and Urdu[1] that are annotated with the same annotation guideline to enable cross-lingual transfer learning evaluation for the first time. As such, four languages, i.e., Danish, Spanish, Turkish, and Urdu, are not explored in any of the existing datasets for ECI. To facilitate open access to the dataset, we obtain the texts from

[1]We will maintain the dataset and include more languages along the way.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2346–2356
October 12–17, 2022.

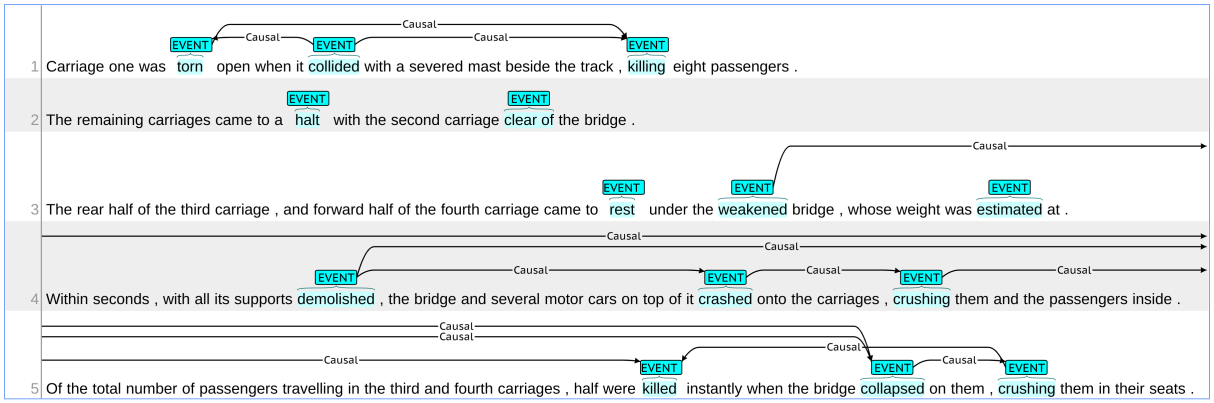

Figure 1: Our annotation interface for event causality identification.

Wikipedia for annotation in all examined languages. To make it consistent with prior research and benefit from the well-designed annotation guidelines of previous datasets, we inherit the event schema from the ACE 2005 dataset (Walker et al., 2006), and the causal event relation guideline from EventStoryLine (Caselli and Vossen, 2017) (with both explicit and implicit causal relations) during the annotation process. In total, our MECI dataset involves 46K events and 11K relations that are substantially larger than those in existing ECI datasets. Figure 1 illustrates our annotation interface in this work.

In addition, we evaluate the proposed MECI dataset using the state-of-the-art models for ECI. We investigate the challenges of MECI over all examined languages through the monolingual setting where the models are trained and evaluated in the same language. The experiments show that the performance of existing ECI models, even with large pre-trained language models (PLMs), is far from satisfactory; models for non-English languages generally perform poorer than their English counterparts. We also observe the importance of choosing language-specific or multilingual PLMs for ECI models as their effectiveness varies for different languages. Moreover, we evaluate the models in the zero-shot cross-lingual setting, where the models are trained on English data and tested on the data of the other languages. The experiment suggests transferability of ECI knowledge between English and Urdu while showing a significant performance drop in other language pairs. These results can serve as baselines for future studies on cross-lingual transfer learning for ECI. Finally, we report the analysis and challenges of the MECI dataset to provide insights for future ECI research. We will publicly release MECI to promote future studies in multilingual ECI.

2 Data Annotation

2.1 Annotation Scheme

Our goal is to annotate causal relations between event mentions in text. To this end, we define the annotation scheme for event mentions following the guidelines for the ACE 2005 dataset (Walker et al., 2006) for events, while the annotation guidelines for event causality relations are obtained from those for the EventStoryLine dataset (Caselli and Vossen, 2017). This allows us to inherit the well-designed documentation in such benchmark datasets and achieve consistency with prior research for ECI.

In particular, based on the ACE 2005 annotation guideline, an event in our dataset is either (1) an occurrence involving some participants, or (2) something that happens, or (3) a change of state. Event mentions/triggers are words/phrases in text that clearly evoke some event. As we are mainly interested in event causality relations, we only annotate event mention spans and do not include event types. To accommodate different languages, we allow event mentions/triggers to span multiple words in the sentences.

Next, for event causality relations, our annotation guideline follows the EventStoryLine dataset. In particular, a causal relation represents a directional relation between two events in which an event (CAUSE) causes another event (EFFECT) to happen or hold. This definition covers standard causal relations: cause, enablement, and prevention (Caselli and Vossen, 2017). In addition, similar to EventStoryLine, our dataset covers both explicit and implicit causality. Note that this is an extension from most prior annotation schema,

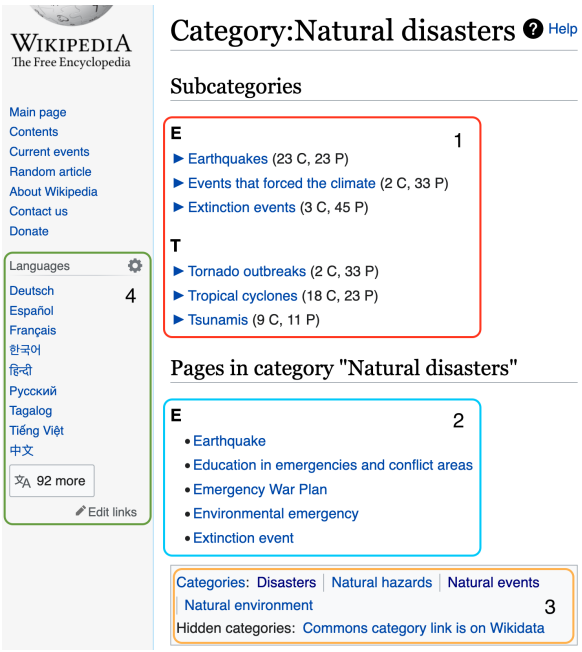

Figure 2: The Wikipedia category page of *Natural disasters* with its child categories (box 1), associated pages (2), parent categories (box 3), and interlink to the same category in other languages (box 4).

i.e., Causal-TimeBank (Mirza and Tonelli, 2014), RED (O'Gorman et al., 2016), BECauSe (Dunietz et al., 2017), that have only considered explicit relations covering the three causal concepts: *cause, enable,* and *prevent* through a verb-based lexicalization (Wolff, 2007). In our view, causality is a tool for humans to understand the world, and its existence is independent of the actual language for presentation (Neeleman et al., 2012). Hence, event causality relations might be established without explicit ground in the text. In other words, there are implicit causal relations between events that are not covered by the above lexicalization (Caselli and Vossen, 2017; Webber et al., 2019). To capture this important type of event causality relations, our annotation guideline is extended to cover implicit relations which require background knowledge, e.g., common-sense, domain-specific knowledge, for successful identification. Finally, similar to prior datasets, we annotate both intra- and inter-sentential causal relations between two events (Mirza and Tonelli, 2014; Caselli and Vossen, 2017).

2.2 Data Collection & Preparation

The documents for our MECI dataset are collected from Wikipedia for five topologically different languages, i.e., English, Danish, Spanish, Turkish, and Urdu. In particular, we focus on 5 topics: aviation accidents, railway accidents, natural disasters, conflicts, and economic crisis, to expect a high yield of events and event causality relations. Wikipedia organizes articles into a hierarchical graph of categories. A category is a group of articles sharing a topic that might be further split into finer subcategories as shown in Figure 2. Furthermore, the hierarchical category systems in Wikipedia for different languages are interconnected through interlinks between identical categories. Therefore, by exploiting the category systems and language interlinks, we are able to obtain Wikipedia articles of the same topics across many languages.

Given the list of five categories for the examined languages, we crawl all the articles associated with their category descendants (i.e., subcategories, subsubcategories) in the hierarchy up to the depth of 6. After this step, we obtain at least 1,000 articles per category for each language. The obtained articles are cleaned by removing format elements (i.e., lists, images, URLs, and markups) to retain only textual data. Afterward, the articles are split into sentences and tokenized into words by Trankit (Nguyen et al., 2021), a multilingual text processing tool with state-of-the-art performance. The detailed list of subcategory URLs will be included in the final dataset package.

Given an article, a direct method for data annotation for ECI is to ask the annotators to label all the event mention spans and event mention pairs with causal relations. However, as the number of event mention pairs in a document grows quadratically with respect to the number of event mentions, a long Wikipedia article can easily overwhelm the annotators, thus affecting the quality of the annotated data. To address the issue, we split the Wikipedia articles into smaller chunks that span five consecutive sentences for separate annotation, following prior practices (Mostafazadeh et al., 2016; Ebner et al., 2020). These chunks are called documents in our dataset. In this way, the annotators only need to consider a shorter context at a time to enhance the attention and quality of annotated data.

2.3 Human Annotation

To annotate the obtained documents, we hire annotators from `upwork.com`, a crow-sourcing platform with freelancers from all around the globe. We only consider candidates that are (1) native to the target language, (2) fluent in English, and

Language	Event	Relation
Danish	0.68	0.58
English	0.92	0.80
Spanish	0.84	0.66
Turkish	0.69	0.61
Urdu	0.65	0.75

Table 1: Kappa scores for the MECI dataset.

(3) highly approved among the Upwork employers. We can access this information from the annotators' profiles on the platform. The candidates are then given annotation guidelines and a test for performing both event annotation and event causality relation extraction tasks. The top two candidates are hired for each language. We use BRAT annotation tool for our annotation (Stenetorp et al., 2012) and illustrated in Figure 1.

Our annotation consists of two tasks, i.e., event mention annotation and event causal relation annotation. For each language, we annotate event causality relations over the outputs from event mention annotation (i.e., after event mention annotation has been completed and finalized for all documents). Given a sample of selected documents for a language, for each task, the two annotators for that language independently annotate event mentions/event causal relations for the documents. Afterward, the annotation conflicts will be presented to the annotators for further discussion and revision to produce the final version of annotated documents for the current task. This will help to ensure high agreement and consistency for our dataset.

2.4 Data Analysis

Table 1 presents our Kappa scores for annotation agreements of event mentions and event causality relations over different languages. Note that these scores are computed by comparing the independent annotations of the annotators over the documents before engaging in discussion to resolve conflicts. As can be seen, the scores are very close to either substantial or almost perfect agreement for all the tasks and languages, thus demonstrating the high quality of our created MECI dataset. We also find that non-English languages tend to have lower annotation agreement scores for both event mention and causality relation extraction tasks, thus highlighting the challenges of ECI for non-English languages and showing the importance of additional research for multilingual ECI.

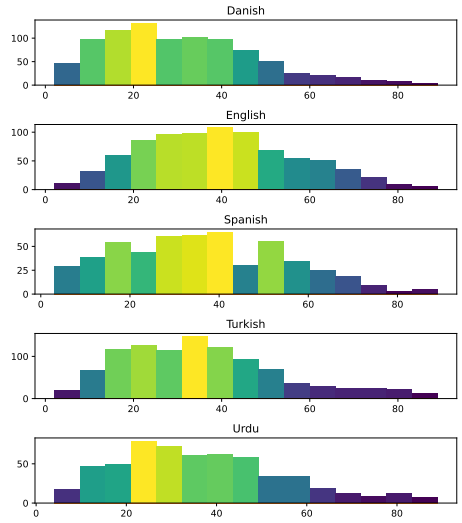

Figure 3: Distributions of distances between two event mentions with causal relations in MECI. Distances are measured via the number of words.

In addition, Table 2 show other statistics for our MECI dataset. Across five languages, each document contains an average of 13.0 event triggers, which account for 2.6 event triggers per sentence. This reveals a challenge of MECI for ECI models that might need to handle the ambiguity due to the overlap of the context of event mention pairs in both sentence and document levels. Furthermore, each document contains approximately 3.1 relations on average; however, there is a discrepancy in event causality relation density in documents among languages. In particular, English and Turkish represent a much denser level of event causality relations per document than other languages, especially Spanish and Urdu. As such, the divergences in the density of event causality relations (and event mentions) pose another robustness challenge for ECI models that should be able to bridge the gaps and transfer event causal knowledge across languages.

Finally, Figure 3 presents the distributions of distances between two event mentions with causal relations for five examined languages in MECI (the distances are counted via the number of words in between). There are several observations from the figure. First, for all the languages, a majority of event mentions are 10 to 50 words away from each other in the documents. This suggests diverse levels of context information between event mentions that an ECI model needs to capture to perform well for the languages in MECI. Second, there are clear divergences between the distance distributions of

causal event mention pairs over languages. For instance, the distances between event mentions for Danish and Urdu seem to be more distributed in the shorter ranges than those of English and Spanish. Such distribution differences require ECI models to introduce robust mechanisms to induce language-transferable representations for diverse causal contexts in cross-lingual learning for ECI.

Dataset Comparison: Table 2 also compares our MECI dataset with previous public datasets for ECI. Note that we focus on the datasets that explicitly consider causal relations between event mentions/triggers to make them comparable. It is clear from the table that our MECI dataset has a much larger scale with more event mentions, causal relations, and languages than all previous datasets for ECI. This will enable the training of larger models and a more comprehensive evaluation for ECI.

2.5 Challenges

Unlike most prior ECI datasets, our MECI dataset includes implicit causal relations, which allow causal relations to be derived from various implicit reasoning sources such as common-sense knowledge. This section illustrates some types of implicit reasoning for causal relations between events discovered in our dataset.

Implicit inference of causal cues: In the following example, considering two event mentions: *"derailed"* and *"running into"*, there is no triggering verb-based expression to signal the causal relationship between the two events. However, with the presence of the trailing comma between the two event mentions, our annotators can easily realize that the *"derail"* event is the cause of the *"running into"* event. As such, the annotators might have implicitly inferred the reduced relative clause "*which makes the train*" (presented in the brackets) between the two event mentions to make the causal decision. To this end, a model will also need to recognize such implicit reasoning cues based on the context to successfully perform ECI.

> *The Granville rail disaster ... when a crowded commuter train* ***derailed****, [~~which makes the train~~]* ***running into*** *the supports of a road bridge that ...*

Implicit transitivity: Consider three event mentions *"trouble"*, *"bail out"*, and *"killed"* in the following example. The ground text explicitly expresses the causal relation *"bail out"* $\xrightarrow{cause}$ *"killed"* via the adverb *"consequently"*. However, there is no clear signal of the causality between *"trouble"* and *"bail out"*, which requires common-sense knowledge to successfully recognize for the causal order of such events, i.e., *"trouble"* $\xrightarrow{cause}$ *"bail out"*. This increases the difficulty for identifying the causality *"trouble"* $\xrightarrow{cause}$ *"killed"*, which might entail transitivity reasoning between implicit and/or explicit causal relations, i.e., *"trouble"* $\xrightarrow{cause}$ *"bail out"* and *"bail out"* $\xrightarrow{cause}$ *"killed"*.

> *... when his Spitfire developed engine* ***trouble*** *between the islands of Skiathos and Skópelos over the Aegean Sea . He attempted to* ***bail out*** *of the aircraft, but his altitude was too low for his parachute to open, and he was consequently* ***killed****.*

3 Experiments

We randomly split the documents for each language in MECI into three separate parts with a ratio of 3/1/1 to serve as training, development, and test data respectively for experiments. To study the challenges of ECI presented in MECI, we evaluate the performance of the state-of-the-art models for ECI on this dataset. Each model will be comprehensively evaluated in the monolingual learning (i.e., trained and tested on data of the same language) and multilingual learning (i.e., trained and tested on the data of different language) settings with MECI.

3.1 ECI Models

We explore the following representative models for ECI in the literature:

PLM: This model is inherited from the BERT baseline in (Tran Phu and Nguyen, 2021). Given an input document D, this model concatenates the words from all sentences and sends it into a pre-trained language model, e.g., BERT (Devlin et al., 2019), to obtain representation vectors for each word-piece using the hidden vectors in the last transformer layer. Afterward, given the spans A and B for two event mentions e_A and e_B of interest in D, we compute the representations $\mathbf{r}_A$, $\mathbf{r}_B$ for the two event mentions by averaging the representation vectors of the word pieces within the corresponding spans A and B. Finally, we form an overall representation vector $\mathbf{r}_{A \to B} = [\mathbf{r}_A, \mathbf{r}_B, \mathbf{r}_A - \mathbf{r}_B, \mathbf{r}_A * \mathbf{r}_B]$ ($*$ is the element-wise multiplication operation) for ECI.

Dataset	Lang	#Documents	#Relations	#Events	Relation Type
Causal-TimeBank (Mirza et al., 2014)		100	318	11,000	Explicit
RED (O'Gorman et al., 2016)		95	*4,969	8,731	Explicit
BECauSE-2.0 (Dunietz et al., 2017)	English	118	1,803	-	Explicit
CaTeRS (Mostafazadeh et al., 2016)		320	488	2,708	Explicit, Implicit
EventStoryline (Caselli and Vossen, 2017)		258	5,519	7,275	Explicit, Implicit
MECI	Danish	519	1,377	6,909	Explicit, Implicit
	English	438	2,050	8,732	
	Spanish	746	1,312	11,839	
	Turkish	1,357	5,337	14,179	
	Urdu	531	979	4,975	
MECI (total)	Various	3591	11,055	46,634	Explicit, Implicit

Table 2: Comparison of public ECI datasets. #Relations indicates the number of causal relations in the datasets. * designates the numbers that include other event-event relations, i.e., temporal and hierarchical relations.

	Model	MECI English			EventStoryLine		
		P	R	F	P	R	F
BERT	PLM	35.6	44.9	39.7	27.3	35.3	30.8
	RichGCN	**48.1**	**69.5**	**56.8**	**42.6**	**51.3**	**46.6**

Table 3: Performance of models on MECI (English) and EventStoryLine datasets.

This vector will be fed into a feed-forward network with a sigmoid function in the end to predict the causal relationship between e_A and e_B in D.

RichGCN (Tran Phu and Nguyen, 2021): Similar to **PLM**, **RichGCN** employs a PLM to encode the entire input document and compute an overall representation vector $\mathbf{r}_{A \to B}$ for identifying the causal relationship between two given event mentions. To enhance representation learning, **RichGCN** also introduce several interaction graphs (with words and event mentions in the input document as the nodes) to capture relevant context information/interactions for the causal relationship between two event mentions. In particular, to adapt **RichGCN** to MECI with multiple languages, we implement four interaction graphs to represent an input document: (1) *Sentence Boundary Graph* where words or event mentions within each sentence in the document are connected to each other; (2) *Event Mention Span Graph* where words within each event mention span are connected to the event mention; (3) *Syntax-based Graph* where words within each sentence are connected to each other following the dependency tree structure of the sentence; and (4) *Semantic-based Graph* where words across the document are connected to each other; the weights for the connections are measured via the similarity between the word representations (computed from PLM). In **RichGCN**, each interaction graph is represented by an adjacency matrix. A final graph V to capture relevant connections for the two event mentions is formed by learning a linear combination of the adjacency matrices of the four graphs. Finally, the graph V is then sent into a Graph Convolutional Network (GCN) (Kipf and Welling, 2017; Nguyen and Grishman, 2018) to compute a richer representation for the two event mentions with more relevant context to perform ECI.

Know (Liu et al., 2021): By treating the event mentions as concepts in ConceptNet (Speer et al., 2017), **Know** retrieves related concepts and relations for the two input event mentions in our ECI problem from ConceptNet. The retrieved information is then used to augment the input text. As such, **Know** also utilizes a PLM to encode the augmented text to compute prediction representation for ECI. In addition, this model employs a masking mechanism to obtain event-agnostic context from input text, serving as another source of information to be encoded by the PLM and incorporated into representation learning for our task.

3.2 Experiment Setups

In the monolingual learning settings, for each language in MECI, we train the ECI models on the training data and evaluate model performance on the test data of the same language. We explore both multilingual PLMs, i.e., mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2020), and language-specific PLMs for the languages in MECI as the encoder for the ECI models in the experiments. In particular, we utilize the following language-specific PLMs that are available for MECI languages, i.e., BERT (Devlin et al., 2019)

	Model	English			Danish			Spanish			Turkish			Urdu		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
mBERT	PLM	38.4	46.0	41.9	25.2	26.6	25.9	43.9	41.5	42.7	36.2	48.7	41.6	31.9	34.3	33.0
	Know	35.8	56.7	43.9	25.8	36.0	30.1	39.7	38.3	39.0	39.7	46.9	43.0	36.7	35.3	36.0
	RichGCN	48.4	67.1	56.2	29.7	38.0	33.4	**51.2**	52.0	51.6	50.0	59.9	54.5	40.1	50.0	44.5
XLMR	PLM	48.7	59.9	53.7	**35.9**	36.2	36.0	50.6	49.1	49.9	44.0	59.4	50.5	40.4	43.2	41.8
	Know	39.3	42.6	40.9	31.4	11.4	16.7	39.9	28.4	33.2	36.5	46.7	41.0	**41.1**	22.2	28.9
	RichGCN	**50.6**	**68.0**	**58.1**	31.9	**50.0**	**38.9**	50.7	**55.0**	**52.8**	**50.5**	**64.6**	**56.7**	37.7	**56.0**	**45.1**

Table 4: Monolingual learning performance of ECI models on MECI with mBERT and XLMR.

	Model	English			Danish			Spanish			Turkish			Urdu		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
*	PLM	35.6	44.9	39.7	23.2	23.0	23.1	42.7	44.6	43.6	40.4	56.0	46.9	20.2	33.5	25.2
	RichGCN	**48.1**	**69.5**	**56.8**	**27.1**	**35.0**	**30.6**	**59.8**	**48.2**	**53.4**	**54.7**	**62.0**	**58.1**	**31.1**	**47.9**	**37.7**

Table 5: Monolingual learning performance of ECI models on MECI with language-specific PLMs.

for English; BotXO[2] for Danish, BETO (Cañete et al., 2020) for Spanish, BERTurk (Schweter, 2020) for Turkish, and UrduHack[3] for Urdu.

The support of multiple languages with the same annotation guideline for event causality relations in MECI allows us to perform cross-lingual transfer learning evaluation for ECI models. In particular, for cross-lingual settings, ECI models are trained on the training data of one language (the source language); however, they are evaluated on test data of new target languages. In the experiments, we treat English as the source language and other languages in MECI as the target languages for cross-lingual evaluation. To facilitate the prediction over multiple languages, we leverage the multilingual PLMs mBERT and XLMR in cross-lingual experiments.

Hyper-parameters: We employ the same hyper-parameters from the original works for the ECI models: **RichGCN** (Tran Phu and Nguyen, 2021), and **Know** (Liu et al., 2021) in the experiments. The multilingual NLP toolkit Trankit (Nguyen et al., 2021) is leveraged to obtain dependency trees for sentences in multiple languages for the **RichGCN** model. Also, we utilize the multilingual version of ConceptNet (Speer et al., 2017) to retrieve augmented information for **Know**. Finally, we employ the base versions for all the multilingual and monolingual PLMs considered in this work.

3.3 Results

Monolingual Performance: Table 4 shows the performance of the three ECI models on the monolingual learning settings across all the languages with the multilingual PLMs: mBERT and XLMR. Among the ECI models, we find that **RichGCN** maintains its top performance across all the languages and multilingual PLMs, thus demonstrating the effectiveness of its language-agnostic document structure to represent documents for ECI. Nonetheless, the best performance by **RichGCN** for English, Danish, Spanish, Turkish, and Urdu is 58.1, 38.9, 52.8, 56.7, and 45.1. These performance is far from being perfect, thus suggesting the challenges for ECI across languages and presenting ample research opportunities to improve the performance in the future. In addition, among the models, **Know** exhibits mixed performance with mBERT and worst performance with XLMR across languages. We attribute this phenomenon to the unstable quality of the concept retrieval with ConceptNet and context modification in **Know** that might exclude important causal context from the input texts to cause poor performance in different languages. Finally, comparing the multilingual PLMs, we find that XLMR performs significantly better than mBERT over all the languages with the **PLM** and **RichGCN** models, thus suggesting the benefits of XLMR for future ECI research.

Effects of language-specific PLMs: To better understand the effectiveness of PLMs for ECI, Table 5 reports the performance of **PLM** and **RichGCN** in the monolingual learning settings where language-specific PLMs for each language are employed as the encoder for the models. As can be seen, using the best model **RichGCN** and the best multilingual PLM XLMR as the anchors, ECI performance for English, Spanish and Turkish is very close with monolingual and multilingual PLMs (i.e., less than 2% difference in F1 scores). However, multilingual PLMs are substantially better than monolingual

[2]https://huggingface.co/Maltehb/danish-bert-botxo
[3]https://github.com/urduhack/urduhack

	Model	English → Danish			English → Spanish			English → Turkish			English → Urdu		
		P	R	F	P	R	F	P	R	F	P	R	F
mBERT	PLMd	12.4	35.4	18.4	11.4	63.3	19.3	21.5	47.6	29.6	17.0	44.2	24.6
	Know	7.8	**62.0**	13.8	7.2	**69.4**	13.0	20.4	55.5	29.9	14.2	61.5	23.0
	RichGCN	23.7	45.3	31.1	20.6	58.6	30.5	44.5	52.0	48.0	35.0	56.8	43.3
XLMR	PLM	20.1	59.2	30.1	16.0	66.4	25.8	36.1	**60.5**	45.2	25.7	**62.0**	36.3
	Know	13.3	42.1	20.3	10.4	47.3	17.1	25.8	57.6	35.7	19.3	54.5	28.5
	RichGCN	**28.5**	43.7	**34.5**	**22.7**	62.4	**33.3**	**46.4**	55.0	**50.3**	**38.6**	55.2	**45.5**

Table 6: Zero-shot cross-lingual learning performance on MECI using English as source language.

PLMs for Danish and Urdu (up to 7% difference in performance). This can be attributed to the lower resources in Danish and Urdu that hinder effective training for language-specific PLMs. With multilingual PLMs, such low-resource languages can benefit more from data in other languages to train multilingual PLMs.

Cross-lingual Performance: To investigate the transferability of ECI knowledge across languages, Table 6 presents the performance of the ECI models in the cross-lingual learning settings. Note that in these experiments English is the source languages while other languages are the targets. Among the three models, **RichGCN** is still the best performer across all target languages. However, the model's performance drops significantly for the three target languages Danish (by 4.4%), Spanish (by 19.5%), and Turkish (by 6.4%) compared to their monolingual performance with XLMR. This illustrates the challenges and necessity of further research on cross-lingual transfer learning for ECI that can now be enabled with our multilingual dataset.

Interestingly, compared to the monolingual settings, the performance on Urdu of **RichGCN** is slightly improved (by 0.4%) in the cross-lingual setting. One potential reason is due to the smallest size of the training data for Urdu in MECI that allows the larger English training data to train better models for Urdu test data. In addition, among the four target languages, we observe a wide range of cross-lingual performance from the model trained on English data, thus showing the diverse nature of data and languages in MECI for future research.

4 Related Work

As an important task in IE, ECI has attracted extensive research effort to develop effective models (Do et al., 2011; Hashimoto et al., 2014; Hidey and McKeown, 2016; Hu and Walker, 2017; Kadowaki et al., 2019; Zuo et al., 2020; Liu et al., 2021; Tran Phu and Nguyen, 2021; Man et al., 2022b). To support model development for ECI, several datasets have been introduced for this task, including PDTB (Prasad et al., 2008), Causal-TimeBank (Mirza, 2014), ECB (Cybulska and Vossen, 2014), Richer Event Description (O'Gorman et al., 2016), BeCause (Dunietz et al., 2017), and EventStoryLine (Caselli and Vossen, 2017), CaTeRS (Mostafazadeh et al., 2016). However, these previous work and datasets only focus on English data, presenting a strong demand for new research and datasets on other languages for ECI.

To this end, there are a few efforts on creating causality corpora for other languages, such as German (Rehbein and Ruppenhofer, 2020), Arabic (Sadek et al., 2018) and Persian (Rahimi and Shamsfard, 2021). However, these corpus consider not only event mentions, but also entities, clauses, and sentences, thus, not directly solving ECI as we do. In addition, most existing annotation efforts for ECI focus on explicit event causality relationships. EventStoryLine (Caselli and Vossen, 2017) and CaTerRS (Mostafazadeh et al., 2016) are the only two prior datasets that also explore implicit causal relationships between events. However, they do not provide annotation for multiple languages as we do in MECI. Finally, we also note recent efforts on creating multilingual datasets for other NLP tasks, including event detection (Pouran Ben Veyseh et al., 2022), natural language understanding (e.g., slot filling) (FitzGerald et al., 2022), and acronym extraction (Veyseh et al., 2022).

5 Conclusion

We present a new dataset for event causality identification in five different languages across diverse typologies. The dataset is annotated consistently for all languages, offering a large number of event mentions/causal relations and covering four languages that have not been explored in the prior ECI resources. Our extensive experiments and analysis

reveal the quality and challenges of our dataset for ECI. In addition, our dataset enables cross-lingual transfer learning research that is not possible with current resources for ECI. In the future, we plan to extend the dataset to include more languages such as Arabic and Hindi to broaden its coverage.

Ethical Considerations

In this work we present a dataset annotated over the publicly accessible articles of `wikipedia.org`. Complying with the discussion presented by Benton et al. (2017), research with human subject information is exempted from the required full Institutional Review Board (IRB) review if the data is already available from public sources (as with Wikipedia) or if the identity of the subjects cannot be recovered.

Acknowledgement

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112 and the NSF grant CNS-1747798 to the IU-CRC Center for Big Learning. This research is also based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the Better Extraction from Text Towards Enhanced Retrieval (BETTER) Program. 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 ARO, ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

References

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

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, Online. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Guidelines for ecb+ annotation of events and their coreference.

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.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics.

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

Jack FitzGerald, Christopher Hench, and Charith Peris et al. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. 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 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics.

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from Wikipedia. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2988–2999, Hong Kong, China. Association for Computational Linguistics.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997, Baltimore, Maryland. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Zhichao Hu and Marilyn Walker. 2017. Inferring narrative causality between event pairs in films. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 342–351, Saarbrücken, Germany. Association for Computational Linguistics.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5816–5822, Hong Kong, China. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*.

Jian Liu, Yubo Chen, and Jun Zhao. 2021. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3608–3614.

Hieu Man, Nghia Trung Ngo, Linh Van Ngo, and Thien Huu Nguyen. 2022a. Selecting optimal context sentences for event-event relation extraction. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Hieu Man, Minh Nguyen, and Thien Nguyen. 2022b. Event causality identification via generation of important context words. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 323–330, Seattle, Washington. Association for Computational Linguistics.

Paramita Mirza. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17, Baltimore, Maryland, USA. Association for Computational Linguistics.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Ad Neeleman, Hans Van de Koot, et al. 2012. The linguistic expression of causation. *The theta system: Argument structure at the interface*, 20.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. 2016. A semi-supervised learning approach to why-question answering. In *Thirtieth AAAI Conference on Artificial Intelligence*.

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.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Zeinab Rahimi and Mehrnoush Shamsfard. 2021. Persian causality corpus (percause) and the causality detection benchmark. *CoRR*, abs/2106.14165.

Ines Rehbein and Josef Ruppenhofer. 2020. A new resource for German causal language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5968–5977, Marseille, France. European Language Resources Association.

Jawad Sadek, Farid Meziane, et al. 2018. Building a causation annotated corpus: the salford arabic causal bank-proclitics. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Stefan Schweter. 2020. Berturk - bert models for turkish.

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 623–632, New York, NY, USA. Association for Computing Machinery.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Nicole Meister, Seunghyun Yoon, Rajiv Jain, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Macronym: A large-scale dataset for multilingual and multi-domain acronym extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*.

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

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Phillip Wolff. 2007. Representing causation. *Journal of experimental psychology: General*, 136(1):82.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021a. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, Online. Association for Computational Linguistics.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021b. LearnDA: Learnable knowledge-guided data augmentation for event causality identification. 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 3558–3571, Online. Association for Computational Linguistics.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1544–1550, Barcelona, Spain (Online). International Committee on Computational Linguistics.