MEE: A Novel Multilingual Event Extraction Dataset

Amir Pouran Ben Veyseh[1], **Javid Ebrahimi**[1],
Franck Dernoncourt[2], **and Thien Huu Nguyen**[1]
[1]Department of Computer Science, University of Oregon, Eugene, OR, USA
[2]Adobe Research, Seattle, WA, USA
{apouranb, thien}@cs.uoregon.edu; jebivid@gmail.com
franck.dernoncourt@adobe.com

Abstract

Event Extraction (EE) is one of the fundamental tasks in Information Extraction (IE) that aims to recognize event mentions and their arguments (i.e., participants) from text. Due to its importance, extensive methods and resources have been developed for Event Extraction. However, one limitation of current research for EE involves the under-exploration for non-English languages in which the lack of high-quality multilingual EE datasets for model training and evaluation has been the main hindrance. To address this limitation, we propose a novel Multilingual Event Extraction dataset (MEE) that provides annotation for more than 50K event mentions in 8 typologically different languages. MEE comprehensively annotates data for entity mentions, event triggers and event arguments. We conduct extensive experiments on the proposed dataset to reveal challenges and opportunities for multilingual EE.

1 Introduction

Event Extraction (EE) is one of the major tasks of Information Extraction (IE) for text. In a complete EE pipeline, three major goals should be pursued: (1) Entity Mention Detection (EMD): to recognize mentions of real world entities; (2) Event Detection (ED): to identify event mentions/triggers and their types. An event trigger is a word or phrase that most clearly refers to the occurrence of an event; and (3) Event Argument Extraction (EAE): to find participants/arguments of an event mentioned in text. A participant is an entity mention that has an specific role in a given event mention. For instance, in the sentence "*The soldiers were hit by the forces.*", there are two entity mentions "*soldiers*" and "*forces*" of types *PERSON* and *ORGANIZATION* and an event trigger "*hit*" of type *ATTACK*. Also, the two event mentions "*soldiers*" and "*forces*" play the argument roles of *Victim* and *Attacker* (respectively) in the *ATTACK* event. An EE system could be employed in other downstream applications such as Question Answering, Knowledge Base Population and Text Summarization to assist extracting information about events in text.

Multiple methods have been proposed for Event Extraction. Early work has employed feature-based models (Ahn, 2006; Ji and Grishman, 2008; Liao and Grishman, 2010; Hong et al., 2011; Li et al., 2013; Yang and Mitchell, 2016) while later methods have explored deep learning to present state-of-the-art performance for Event Extraction (Nguyen and Grishman, 2015; Chen et al., 2015; Nguyen et al., 2016; Sha et al., 2018; Wang et al., 2019; Lai et al., 2020; Veyseh et al., 2020; Lin et al., 2020; Nguyen et al., 2021a; Liu et al., 2022). However, despite all advancements on event extraction in recent years, a major limitation of current EE research is to overly focus on a few popular languages, thus failing to adequately reveal challenges and generalization of models in many other languages of the world. As such, a critical barrier for studying EE over multiple languages is the lack of high quality datasets that fully annotate data for many other languages for EE. For instance, the most popular dataset for EE, i.e., ACE 2005 (Walker et al., 2006), only provide annotations for three languages English, Chinese and Arabic while TAC KBP datasets (Mitamura et al., 2016, 2017) only supports English, Chinese and Spanish. The TempEval-2 dataset (Verhagen et al., 2010) involves 6 languages; however, it does not offer event argument annotation. Even worse, recently created datasets, e.g., MAVEN (Wang et al., 2020), RAMS (Ebner et al., 2020), and WikiEvents (Li et al., 2021), are only annotated for English. In all, such language and task limitations prevents research to comprehensively develop and evaluate EE methods over different languages and multilingual settings. Moreover, the limited size of these datasets, i.e. less than 11K and 27K in ACE 2005 and TempEval-2 respectively, hinders training of

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9603 - 9613

data-hungry deep learning models. Finally, we note that important multilingual datasets for EE, e.g., ACE 2005 and TAC KBP, are not publicly available, which further restricts research on this domain.

To address these limitations, in this work, we propose a large-scale Multilingual Event Extraction (MEE) dataset that covers 8 typologically different languages from multiple language families, including English, Spanish, Portuguese, Polish, Turkish, Hindi, Korean, and Japanese. As such, Portuguese, Polish, Turkish, Hindi, and Japanese are not explored in the popular multilingual datasets for EE, i.e., ACE 2005 and TAC KBP. Importantly, to enable public data sharing and diversity the data, we employ Wikipedia articles for the 8 languages in diverse topics (i.e., Economy, Politics, Technology, Crime, Nature and Military) for EE annotation.

Our dataset comprehensively annotates each document in a language for all the three sub-tasks EMD, ED, and EAE. To be consistent with prior EE research, we inherit the type anthologies for such tasks from the ACE 2005 dataset that provides well-designed guidelines and examples for the types. In particular, we include 7 entity types, 8 event types and 16 event sub-types, along with 23 argument roles in MEE to facilitate EE annotation over multiple languages. Overall, our dataset involves more than 415K entity mentions, 50K event triggers, and 38K arguments, which are much larger than previous multilingual EE datasets to better support model training and evaluation with deep learning.

Due to shared information schema over all the languages, our MEE dataset enables cross-lingual transfer learning evaluation of MEE models where training and test data comes from different languages. To this end, we conduct comprehensive experiments for both monolingual and cross-lingual learning settings to provide insights for language-specific challenges and cross-lingual generalization of EE methods. By examining both pipeline and joint inference models for EE, our experiments show that the proposed dataset present unique challenges with less satisfactory performance of existing EE models, especially for cross-lingual settings, thus calling for more research efforts for multilingual EE in the future.

2 Data Annotation

We follow the entity/event type definition and annotation guidelines from the popular ACE 2005 dataset to benefit from its well-designed documentation and be consistent with prior EE research. As such, entity mentions refer to mentions of real-world entities in text that can be expressed via names, nominals, and pronouns. Entity Mention Extraction (EMD) is more general than Named Entity Recognition that only concerns names of entities. In addition, an event is defined as an incident whose occurrence changes the state of real world entities. An event mention is the part of input text that refers to an event that consists of two components: (1) Event Trigger: the words that most clearly refer to the occurrence of the event. It is noteworthy that we allow an event trigger to span multiple words to accommodate trigger annotation for multiple languages. For instance, in the Turkish phrase "*tayin etmek*", both words are necessary to indicate an event trigger of type "*Appoint*"; and (2) Event Arguments: the entity mentions that are involved in the event with some roles.

Based on the ACE 2005 dataset, our dataset annotates entity mentions for 7 entity types: **PERSON** (human entities), **ORGANIZATION** (corporations, agencies, and other groups of people), **GPE** (geographical regions defined by political and/or social groups), **LOCATION** (geographical entities such as landmasses or bodies of water), **FACILITY** (buildings and other permanent man-made structures), **VEHICLE** (physical devices primarily designed to move an object from one location to another), and **WEAPON** (physical devices primarily used as instruments for physically harming). For event types, to avoid confusion and improve data quality, we prune the original ACE 2005 event types to only include the types that are not ambiguous across multiple languages. For instance, in Turkish, the event types *Sentence* and *Convict* are very similar (both can be evoked by the phrase "*Mahkum etmek*") so they are not retained in our dataset. As such, we preserve 8 event types and 16 sub-types that are distinct enough for annotation in our dataset. Finally, for event arguments, we preserve all 23 argument roles in the ACE 2005 dataset. Table 4 shows the list of event types along with their argument roles in our dataset.

2.1 Data Preparation

Our dataset MEE covers 8 different languages, i.e., English, Spanish, Portuguese, Polish, Turkish, Hindi, Korean and Japanese. These languages are selected based on their diversity in terms of

Category	English	Portuguese	Spanish	Polish	Turkish	Hindi	Japanese	Korean
Economy	1,095	112	168	315	297	189	199	250
Politics	3,202	308	772	1,270	1,233	349	232	248
Technology	2,171	189	400	712	815	295	312	249
Crimes	893	78	220	152	118	95	80	73
Nature	1,195	398	705	455	398	245	299	185
Military	4,444	415	1,003	1,575	1,619	326	378	495
Total	13,000	1,500	3,268	4,479	4,480	1,499	1,500	1,500

Table 1: Numbers of annotated segments in each Wikipedia subcategory for our 8 languages.

typology and their novelty with respect to existing multilingual EE datasets. For each language, we employ its latest dump of Wikipedia articles as raw data for annotation. To focus on event data, we select articles in the sub-categories under category *Event* in Wikipedia. In particular, the following sub-categories are considered to improve topic diversity: Economy, Politics, Technology, Crimes, Nature, and Military. Note that we start with these categories in English Wikipedia. Afterward, we follow interlinks between the categories in different languages to locate the intended categories for Wikipedia for non-English languages in MEE.

We process the collected articles with the WikiExtractor tool (Attardi, 2015) to obtain clean textual data and meta-data for each article. The textual data is then split into sentences and tokenized into words by the multilingual NLP toolkit Trankit (Nguyen et al., 2021b). Afterward, to annotate the data with entity and event mentions, one approach is to directly ask annotators to read each article entirely for annotation. However, as the articles in Wikipedia might be lengthy, this approach can be overwhelming for annotators, thus hindering their attention and lowering quality of annotated data. To address this issue, we follow prior dataset creation efforts for EE, i.e., RAMS (Ebner et al., 2020), to divide the articles into segments of five consecutive sentences. Each segment will then be annotated separately for EE tasks so annotators can better capture the entire context to provide entity and event annotation. Note that similar to RAMS, we annotate all event arguments in a text segment for each event trigger, thus allowing event arguments to appear in different sentences from the event trigger (i.e., document-level EAE). Finally, to accomodate our budget, a sample of text segments is obtained for each language for annotation. The numbers of selected text segments for each category per language in our dataset are presented in Table 1.

2.2 Annotation Process

To annotate the sampled article segments, we employ the crowd-sourcing platform upwork.com that allows us to hire freelancers across the globe with different expertise. For each language in our dataset, we choose native speakers as annotator candidates. In addition, we require them to be fluent in English, have experience in related tasks (i.e., data annotation for information extraction), and have approval rate higher than 95% (i.e., provided in their profiles). The candidates are first provided with annotation guidelines and interfaces in English. Afterward, they are invited to an annotation test for entity mentions, event triggers, and arguments. Those candidates who correctly annotate all test cases are then officially hired to work on our annotation jobs. Table 3 shows the numbers of annotators who are hired to annotate data for each language in our dataset. Next, before the actual annotation process, the English annotation guideline and examples are translated to each target language by the hired annotators. Any language-specific confusions and rules for annotation is discussed and included in the translation to create a common understanding. Finally, our language experts will review the annotation guideline in each language to avoid conflicts across languages to be used for actual annotation.

Our annotation process is done in three separate steps to annotate data for three EE tasks with entity mentions, event triggers, and event arguments in this order. In particular, the annotation for a later task will be performed over the text segments that have been annotated and finalized for previous tasks (e.g., event arguments will be annotated over segments that are already provided with entity mentions and event triggers). As such, for each task, 20% of text segments for each language will be co-annotated by the annotators to measure agreement score. The remaining 80% of text seg-

Language	#Seg.	Avg. Length	#Entities	#Triggers	#Arguments	Challenging Entity Type	Challenging Trigger Type	Language Family
English	13,000	123	190,592	17,642	13,548	GPE	Personnel	Germanic
Spanish	3,268	112	48,001	6,064	802	GPE	Conflict	Italic
Portuguese	1,500	102	25,463	1,953	12,329	Location	Personnel	Italic
Polish	4,479	108	62,971	10,875	3,395	Facility	Transaction	Balto-Slavic
Turkish	4,480	117	38,469	8,390	1,416	GPE	Personnel	Turkic
Hindi	1,499	98	18,797	1,810	2,117	Facility	Conflict	Indo-Iranian
Japanese	1,500	99	19,174	2,152	3,399	Location	Personnel	Japonic
Korean	1,500	103	12,508	1,125	1,742	GPE	Personnel	Koreanic
Total (MEE)	31,226	-	415,975	50,011	38,748	-	-	-

Table 2: Statistics of the MEE dataset. #Seg. represents the numbers of annotated text segments for each language. All annotated segments consist of 5 sentences and their lengths (Avg. Length) are computed in terms of numbers of tokens. "*Challenging Type*" indicates the types where entity or event trigger annotation involves largest disagreement between annotators in each language.

Language	#Annotator	EMD	ED	EAE
English	10	0.792	0.834	0.820
Spanish	10	0.788	0.812	0.823
Portuguese	5	0.791	0.803	0.799
Polish	8	0.780	0.799	0.813
Turkish	10	0.785	0.813	0.822
Hindi	6	0.790	0.803	0.812
Japanese	5	0.793	0.789	0.780
Korean	6	0.802	0.810	0.825

Table 3: Number of annotators and agreement scores for 8 languages in MEE for Entity Mention Detection (EMD), Event Detection (ED) and Event Argument Extraction (EAE).

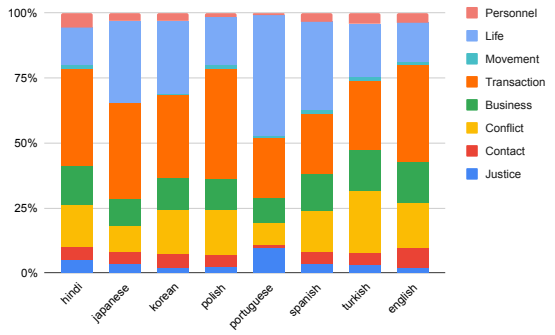

Figure 1: Distributions of event types in each language.

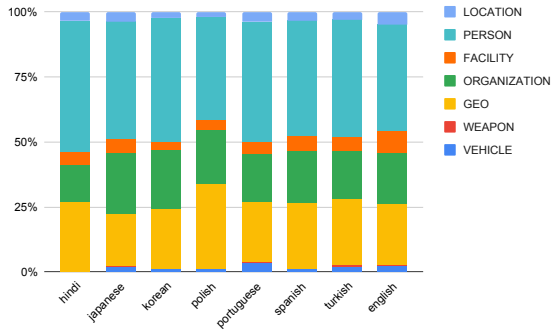

Figure 2: Distributions of entity types in each language.

ments will be distributed and annotated separately by the annotators for each language. Based on the Krippendorff's alpha (Krippendorff, 2011) with MASI distance metric (Passonneau, 2006), we report the inter-annotator agreements (IAA) for each task and language in Table 3, showing high agreement scores and quality of our MEE dataset. Note that after independent annotation for each EE task, the annotators also share their annotations and communicate with each other to resolve any conflicts and finalize our data.

2.3 Data Analysis

Table 2 shows the main statistics of MEE for each language. As such, comparing to the popular multilingual ACE 2005 dataset (Walker et al., 2006) for EE, our MEE dataset provides more languages (i.e., 3 vs. 8) and much more event mentions (i.e., 11K vs. 50K). For other multilingual datasets for EE, i.e., TAC KBP (with three languages and 6.5K event mentions) (Mitamura et al., 2016, 2017) and TempEval-2 (with 6 languages and 27K event mentions) (Verhagen et al., 2010), they do not annotate entity mentions and event arguments. In contrast, our MEE dataset fully annotates texts for three EE tasks (i.e., EMD, ED, and EAE) and also with more languages and event mentions. This clearly demonstrates the advantages of our dataset over existing multilingual datasets for EE.

In addition, from the table, we find that the languages in our dataset exhibits diverse densities for entity mentions, event triggers, and arguments in texts. In particular, while the average number of entities in a text segment in Portuguese is 16.9, this number is only 8.3 in Korean. For event density, in Polish, there are 2.4 event mentions per article segment on average while the average number in Korean is only 0.75. Similarly for event arguments, the average number of arguments per event is 6.1 in Portuguese and only 0.75 in English. Fur-

ID	Event	Arguments
1	Life_Be-Born	Person, Time, Place
2	Life_Marry	Person, Time, Place
3	Life_Divorce	Person, Time, Place
4	Life_Injure	Agent, Victim, Instrument, Time, Place
5	Life_Die	Agent, Victim, Instrument, Time, Place
6	Movement_Transport	Agent, Artifact, Vehicle, Price, Origin, Destination, Time
7	Transaction_Transfer-Ownership	Buyer, Seller, Beneficiary, Price, Artifact, Time, Place
8	Transaction_Transfer-Money	Giver, Recipient, Beneficiary, Money, Time, Place
9	Business_Start-Organization	Agent, Organization, Time, Place
10	Conflict_Attack	Attacker, Target, Instrument, Time, Place
11	Conflict_Attack	Entity, Time, Place
12	Contact_Meet	Entity, Time, Place
13	Contact_Phone-Write	Entity, Time
14	Personnel_Start-Position	Person, Entity, Position, Time, Place
15	Personnel_End-Position	Person, Entity, Position, Time, Place
16	Justice_Arrest-Jail	Person, Agent, Crime, Time, Place

Table 4: Event types and argument roles for each type in MEE. The types and roles are inherited from the event extraction annotation guideline in the ACE 2005 dataset (Walker et al., 2006).

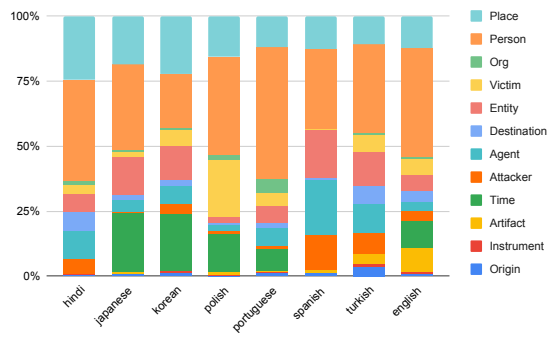

Figure 3: Distributions of most argument roles for each language.

ther, Table 2 highlights the divergences between languages regarding challenging entity and event types. Specifically, we employ the disagreement rates (i.e., number of disagreements divided by frequency of mentions) between annotators for each entity and event types. Those types that have highest disagreement rates are selected as challenging entity or event types. Finally, Figures 2, 1, and 3 present the distributions of entity types, event types, and argument roles (respectively) for each language in our dataset, which further demonstrate the differences between languages in MEE. In all, such differences over various dimensions will cause significant challenges for EE models to adapt to new languages (e.g., for cross-lingual transfer learning), thus presenting ample opportunities for multilingual EE research with our dataset.

3 Experiments

This section evaluates the state-of-the-art models for Event Extraction to reveal challenges in our new dataset MEE. To this end, the annotated article segments for each language in MEE are randomly split into training/development/test portions with the ratios of 80/10/10. Here, to prevent any information leakage, we ensure that different segments of an article (if any) are only assigned to one portion of the data split for each language. We examine EE models in two different settings: (1) monolingual learning where training and test data of models comes from the same language; (2) cross-lingual transfer learning where models are trained on training data of one language (i.e., the source language), but evaluated directly on test data of the other languages (i.e., the target languages).

Models: We evaluate two typical approaches for EE models with pipeline and joint inference in this work. First, for the pipeline approach, a model is trained separately for each of the three tasks in EE, i.e., entity mention detection (EMD), event detection (ED), and event argument extraction (EAE). Here, the EMD and ED tasks are modeled as sequence labeling problems, aiming to predict BIO tag sequences for each input sentence to capture spans and types of entity and event mentions. As such, motivated by previous work (Wang et al.,

Language	Pipeline			OneIE			FourIE		
	Entity	Event	Argument	Entity	Event	Argument	Entity	Event	Argument
English	70.32	70.58	61.14	62.18	70.09	62.94	69.72	72.19	65.89
Spanish	70.39	66.19	60.16	70.27	65.00	60.31	71.89	67.49	62.19
Portuguese	75.13	71.33	69.15	73.19	70.13	71.27	74.98	72.99	70.17
Polish	69.27	59.12	60.09	60.09	59.44	60.14	68.23	60.98	61.32
Turkish	71.88	66.09	56.19	71.98	61.27	58.72	72.33	65.13	59.80
Hindi	66.22	57.77	57.78	61.72	58.18	59.44	65.23	59.88	60.82
Japanese	68.19	67.89	68.19	71.40	65.01	63.17	70.88	66.88	70.19
Korean	57.17	61.26	67.87	55.87	61.10	65.41	58.18	60.09	69.23
Avg.	68.57	65.03	62.57	65.84	63.78	62.68	68.93	65.70	64.95

Table 5: Performance (F1 scores) of models in the monolingual setting using mBERT on MEE.

Language	Pipeline			OneIE			FourIE		
	Entity	Event	Argument	Entity	Event	Argument	Entity	Event	Argument
English	70.22	71.28	66.34	70.39	70.29	68.68	71.19	73.14	68.23
Spanish	70.33	64.32	61.12	70.18	62.46	62.23	72.87	65.90	63.11
Portuguese	70.39	71.88	71.75	72.16	69.43	70.33	73.98	70.43	72.23
Polish	69.14	60.45	61.23	72.22	63.77	60.15	70.25	62.87	62.84
Turkish	76.13	67.18	55.78	74.45	65.31	57.40	75.19	67.29	58.23
Hindi	65.14	59.34	58.22	61.72	58.18	59.44	66.69	61.99	62.19
Japanese	71.34	67.77	69.19	68.20	62.89	70.90	72.82	65.27	73.55
Korean	59.13	62.34	69.70	59.99	60.55	66.89	60.24	61.18	70.09
Avg.	68.98	65.57	64.17	68.84	64.36	64.57	70.40	66.01	66.31

Table 6: Performance (F1 scores) of models in the monolingual setting using XLM-RoBERTa on MEE.

2020), our EMD and ED models leverage a pre-trained transformer-based language model to encode the input text. The representation for each token in input text (obtained via average of hidden vectors of word-pieces in the last transformer layer) is then sent into a feed-forward network to compute a tag distribution for the token for training and decoding. For EAE, the task is formulated as a text classification problem in which the input consists of an input text and two word indices for the positions of an event trigger and an entity mention of interest. The goal is to predict the argument role that the entity mention plays for the event. To this end, we also use a pre-trained language model to obtain representations for the tokens in input text. Next, the representations for the event trigger and entity mention words are concatenated and sent to a feed-forward network to predict argument role. Note that the EAE model employs golden entity mentions and event triggers during the training process while the outputs from the EMD and ED models are fed into the EAE model in the test time.

Second, for the joint inference approach, EE models simultaneously predicts entity mentions, event triggers, and arguments in end-to-end fashion to avoid error propagation and leverage inter-dependencies between tasks. To this end, we evaluate two state-of-the-art (SOTA) joint EE models, OneIE (Lin et al., 2020) and FourIE (Nguyen et al., 2021a), in this work due to their language-agnostic nature. Both OneIE and FourIE utilize pre-trained language models to represent input texts and capture cross-task dependencies for joint inference. Note that these models are original designed to include the relation extraction task between entities. To adapt them to EE, we obtain their implementations from the original papers and remove the relation extraction components. Finally, for performance measure, we report the performance (F1 scores) of EE models over three tasks EMD (Entity), ED (Event), and EAE (Argument) using the same correctness criteria as in prior work (Lin et al., 2020) (i.e., requiring correct prediction for both offsets and types of entity mentions, event triggers, and argument roles).

Hyper-parameters: To facilitate evaluation with multiple languages, we leverage the multilingual pre-trained language models (PLMs) mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) (base versions) to encode texts for EE models. For the pipeline approach, we fine-tune the hyper-parameters for the EMD, ED, and EAE mod-

Language	XLM-RoBERTa			mBERT		
	Entity	Event	Argument	Entity	Event	Argument
English	69.72	72.19	65.89	71.19	73.14	68.23
Spanish	61.96	59.70	52.23	60.72	60.06	50.77
Portuguese	59.98	54.80	52.23	56.17	52.98	50.28
Polish	52.89	51.78	52.44	53.44	50.29	53.56
Turkish	60.13	53.32	52.19	59.19	52.76	53.10
Hindi	56.32	59.76	57.17	55.39	58.44	55.65
Japanese	41.13	44.95	40.13	42.43	43.76	41.18
Korean	45.78	42.99	43.04	44.78	40.22	41.14

Table 7: Cross-lingual performance (F1 scores) of **FourIE** when it is trained on English training data and evaluated on test data of other languages in MEE.

els over development data for English and apply the selected values for all experiments for consistency. In particular, our hyper-parameters for the pipeline model include: 2 hidden layers with 250 hidden units in each layer for the feed-forward networks, 8 for mini-batch size, and $1e$-2 for learning rate with the Adam optimizer. For the joint IE models, we utilize the same hyper-parameters suggested in the original papers, i.e., OneIE (Lin et al., 2020) and FourIE (Nguyen et al., 2021a).

Results: The results for monolingual experiments over different languages in MEE are presented in Tables 5 and 6 (i.e., with mBERT and XLM-RoBERTa encoders respectively). There are several observations from the tables. First, the models' performance on individual languages and on average for all three tasks EMD, ED, and EAE is still far from being perfect (i.e., all average performance is less than 69%), thus indicating considerable challenges in our multilingual EE dataset for future research. In addition, comparing the current state-of-the-art joint IE model (i.e., FourIE) with the pipeline method, we find that FourIE is better than the pipeline model on average, especially for the EAE task with significant performance gap. As such, we attribute this to the ability of joint models to mitigate error propagation to EAE from EMD and ED to boost the performance. Due to its best average performance, FourIE will be leveraged in our next experiments. Finally, we find that XLM-RoBERTa generally has better performance than mBERT (i.e., on average) for EE models. Future research can thus focus on XLM-RoBERTa to develop better EE models for multilingual settings.

Cross-lingual Evaluation: To further understand the cross-lingual generalization challenges in MEE, Table 7 reports the performance of FourIE in the cross-lingual transfer learning settings where the model is trained on English training data (source language) and tested on test data of the other languages in MEE. As can be seen, compared to performance on English test set, FourIE suffers from significant performance drops over different tasks and multilingual encoders when it is evaluated on other languages. It thus demonstrates inherent challenges of cross-lingual generalization for complete EE models that can be further studied with MEE. In addition, the performance loss due to cross-lingual testing varies across different target languages (e.g., 10.88% loss for Spanish vs. 33.42% loss for Japanese in EAE task). These variations can be attributed to different levels of divergence between languages (e.g., sentence structures and morphology) that hinder cross-lingual knowledge transfer for EE.

Language	Entity	Event	Argument
English (Devlin et al., 2019)	70.21	73.18	66.19
Spanish (Cañete et al., 2020)	67.29	65.14	60.13
Portuguese (Souza et al., 2020)	70.21	68.88	67.13
Polish (Kłeczek, 2021)	65.78	61.23	59.14
Turkish (MDZ, 2021)	67.34	64.19	58.72

Table 8: Test data performance (F1) of **FourIE** in monolingual learning using available language-specific BERT models on MEE. The citations indicate the sources of the language-specific models.

Language	Entity	Event	Argument
English (Liu et al., 2019)	70.32	72.28	69.19
Spanish (MMG, 2021)	70.23	61.34	60.28
Polish (CLARIN-PL, 2021)	68.12	60.89	60.34
Hindi (Parmar, 2021)	64.91	59.09	60.38
Japanese (Wongso, 2021)	69.72	60.45	71.45

Table 9: Test data performance (F1) of **FourIE** in monolingual learning using available language-specific RoBERTa models on MEE. The citations indicate the sources of the language-specific models.

Language-Specific Encoders: To study the effectiveness of pre-trained language models as text encoders for EE models, we compare the performance of FourIE when the multilingual encoders mBERT or XLM-RoBERTa are replaced with comparable language-specific encoders (i.e., BERT-based models for mBERT and RoBERTa-based models for XLM-RoBERTa). Using publicly available pre-trained language models for our languages in MEE, Tables 8 and 9 show the monolingual performance over test data of the languages for BERT-based and RoBERTa-based models (respectively). Comparing corresponding performance in Tables 5, 6, 8 and 9, it is clear that language-specific language models all under-perform their multilingual counterparts over different EE tasks and languages, thus

Language	Trained on English			Trained on Polish		
	Entity	Event	Argument	Entity	Event	Argument
Portuguese	53.22	50.79	40.21	54.17	51.70	42.33
Spanish	50.76	43.72	41.16	51.72	45.22	45.81
Turkish	54.44	50.12	55.71	53.99	50.78	56.15
Hindi	51.44	52.78	45.27	52.21	53.00	47.24
Japanese	36.16	41.23	38.13	37.17	40.13	39.28
Korean	43.72	40.08	37.29	42.08	39.78	37.10

Table 10: Cross-lingual performance (F1) of **FourIE** with XLM-RoBERTa encoder when it is trained on English or Polish training data, and tested on test data of the other languages in MEE. We use 3,500 random annotated segments from the training sets of English and Polish to train the model.

suggesting the benefits of multilingual data to train language model encoders to boost EE performance over different languages.

Source Language Impact: Finally, to study the impact of the source language for cross-lingual transfer learning for EE, we compare the performance of FourIE when either English or another comparable language is used as the source language to provide training data to train the model. In particular, we choose Polish as a comparable language for English as it has the same sentence structure (i.e., both languages have Subject-Verb-Object order) and entails similar density and type distributions for entity/event mentions as English. Table 10 shows the performance of the models when they are tested over test data of the other 6 languages in MEE. Here, to make it comparable, we use the same number of annotated segments (i.e., 3,500) sampled from training data of English and Polish to train the FourIE model. Interestingly, we find that Polish can lead to better performance for FourIE than English over a majority of task and target language pairs (i.e., over 4 languages for EMD and ED, and 5 languages for EAE). A possible explanation for this issue comes from richer event patterns that Polish might introduce to produce allow better cross-lingual generalization for EE than those for English. As such, this superior performance of Polish challenges the common practice of using English as the source language in cross-lingual transfer learning studies for EE and NLP. Future research can explore this direction to better understand the differences between languages to best select a source language to optimize performance over a target language for EE.

4 Related Works

Due to its importance, various datasets have been recently developed for EE in different domains, including CySecED (Man et al., 2020) (for cybersecurity domain), LitBank (for literacy) (Sims et al., 2019), MAVEN (Wang et al., 2020), RAMS (Ebner et al., 2020), and WikiEvents (Li et al., 2021) (for Wikipedia texts). However, these datasets are only annotated for English texts. There exist several multilingual datasets for EE, ACE (Walker et al., 2006), TAC KBP (Mitamura et al., 2016, 2017), and TempEval-2 (Verhagen et al., 2010); however, such datasets only provide annotation for a handful of popular languages with limited number of event mentions and might not fully support all EE tasks (e.g., missing EAE in TAC KBP and TempEval-2).

Regarding model development, existing EE methods can be categorized into feature-based (Ahn, 2006; Ji and Grishman, 2008; Liao and Grishman, 2010; Hong et al., 2011; Li et al., 2013; Yang and Mitchell, 2016) or deep learning (Chen et al., 2015; Nguyen et al., 2016; Sha et al., 2018; Wang et al., 2019; Lin et al., 2020; Veyseh et al., 2021a,b; Liu et al., 2022; Veyseh and Nguyen, 2022; Nguyen et al., 2022) methods. While most prior EE methods have been designed for one popular language, there have been growing interests in multilingual and cross-lingual learning for EE in recent work, featuring multilingual PLMs (i.e., mBERT and XLMR) as the key component for representation learning (Chen and Ji, 2009; M'hamdi et al., 2019; Ahmad et al., 2021; Nguyen et al., 2021c; Huang et al., 2022; Guzman-Nateras et al., 2022). However, as such works only rely on existing multilingual EE datasets, their evaluation is limited to a few popular languages and fails to evaluate the generalization over many other languages.

5 Conclusion

We present a novel multilingual EE dataset, i.e., MEE, that covers 8 typologically different languages with more than 50K event mentions to support training of large deep learning models. MEE provides complete annotation for three EE subtasks, i.e., entity mention detection, event detection, and event argument extraction. To study the challenges in MEE, we conduct extensive analysis and experiments with different EE methods in the monolingual and cross-lingual learning settings. Our results demonstrate various challenges for EE in the multilingual settings that can be further pur-

sued with MEE. In the future, we will extend our dataset to include more languages and tasks for IE.

Limitations

In this work we present a novel large-scale multilingual dataset for Event Extraction. As it is shown in the experiments, our dataset introduces many challenges that can inform future research on multilingual Event Extraction. However, there are still some limitations in the current work that can be improved in future research. First, cross-lingual transfer learning for EE is a challenging task that requires specifically designed models and methods. However, in this work, we have mainly focused on existing state-of-the-art EE models that are originally developed for the monolingual settings. As such, future work can study cross-lingual transfer models that are specifically designed to address the gaps between languages to better understand the challenges in our multilingual EE dataset. Second, in addition to data scarcity for multilingual EE, another challenge for this problem is the lack of resources for text encoding and processing in multiple languages. In particular, pre-trained language models and text processing tools might not be available for some languages (e.g., low-resource languages) that hinder dataset creation and model development efforts. As such, our work has not explored datasets and methods for low-resource languages for EE. In addition, as shown in our experiments, a majority of existing language-specific text encoders under-perform their multilingual counterparts for EE models. However, our work has not studied methods to improve such language-specific language models for EE. Finally, although our experiments empirically challenge English as the main source language for cross-lingual learning, we have not explored why other languages might be better options for the source language in this setting. Future research can perform more comprehensive analysis to shed light on this direction.

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

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

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

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

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 International Conference on Learning Representations (ICLR) 2020*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zheng Chen and Heng Ji. 2009. Can one language bootstrap the other: A case study on event extraction. In *Proceedings of the NAACL-HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*.

CLARIN-PL CLARIN-PL. 2021. Polish roberta.

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*.

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.

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*.

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

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Darek Kłeczek. 2021. Polish bert.

Viet Dac Lai, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5405–5411, Online. Association for Computational Linguistics.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82.

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

Shasha Liao and Ralph Grishman. 2010. Filtered ranking for bootstrapping in event extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*.

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

Xiao Liu, Heyan Huang, Ge Shi, and Bo Wang. 2022. Dynamic prefix-tuning for generative template-based event extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5216–5228, Dublin, Ireland. Association for Computational Linguistics.

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

Duc Trong Hieu Man, Duc Trong Le, Amir Pouran Ben Veyseh, Thuat Nguyen, and Thien Huu Nguyen. 2020. Introducing a new dataset for event detection in cybersecurity texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Digital Library team MDZ. 2021. Turkish bert.

Meryem M'hamdi, Marjorie Freedman, and Jonathan May. 2019. Contextualized cross-lingual event trigger extraction with minimal resources. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2016. Overview of TAC-KBP 2016 event nugget track. In *Proceedings of the Text Analysis Conference (TAC)*.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2017. Events detection, coreference and sequencing: What's next? overview of the TAC KBP 2017 event track. In *Proceedings of the Text Analysis Conference (TAC)*.

MMG MMG. 2021. Spanish roberta.

Minh Van Nguyen, Viet Dac Lai, and Thien Huu Nguyen. 2021a. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021b. Trankit: A light-weight 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.

Minh Van Nguyen, Bonan Min, Franck Dernoncourt, and Thien Nguyen. 2022. Joint extraction of entities, relations, and events via modeling inter-instance and inter-label dependencies. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4363–4374, Seattle, United States. Association for Computational Linguistics.

Minh Van Nguyen, Tuan Ngo Nguyen, Bonan Min, and Thien Huu Nguyen. 2021c. Crosslingual transfer learning for relation and event extraction via word category and class alignments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5414–5426, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Suraj Parmar. 2021. Hindi roberta.

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)*.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

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.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021a. Unleash GPT-2 power for event detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6271–6282, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Nghia Ngo Trung, Bonan Min, and Thien Huu Nguyen. 2021b. Modeling document-level context for event detection via important context selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Amir Pouran Ben Veyseh and Thien Nguyen. 2022. Word-label alignment for event detection: A new perspective via optimal transport. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 132–138, Seattle, Washington. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Graph transformer networks with syntactic and semantic structures for event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3651–3661, Online. Association for Computational Linguistics.

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

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019. Adversarial training for weakly supervised event detection. 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 998–1008, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wilson Wongso. 2021. Japanese roberta.

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.