SPEED++: A Multilingual Event Extraction Framework for Epidemic Prediction and Preparedness

Tanmay Parekh **Jeffrey Kwan** **Jiarui Yu** **Sparsh Johri** **Hyosang Ahn**
Sreya Muppalla **Kai-Wei Chang** **Wei Wang** **Nanyun Peng**
Computer Science Department, University of California, Los Angeles
{tparekh, weiwang, violetpeng, kwchang}@cs.ucla.edu

Abstract

Social media is often the first place where communities discuss the latest societal trends. Prior works have utilized this platform to extract epidemic-related information (e.g. infections, preventive measures) to provide early warnings for epidemic prediction. However, these works only focused on English posts, while epidemics can occur anywhere in the world, and early discussions are often in the local, non-English languages. In this work, we introduce the first multilingual Event Extraction (EE) framework SPEED++ for extracting epidemic event information for a wide range of diseases and languages. To this end, we extend a previous epidemic ontology with 20 argument roles; and curate our multilingual EE dataset SPEED++ comprising 5.1K tweets in four languages for four diseases. Annotating data in every language is infeasible; thus we develop zero-shot cross-lingual cross-disease models (i.e., training only on English COVID data) utilizing multilingual pre-training and show their efficacy in extracting epidemic-related events for 65 diverse languages across different diseases. Experiments demonstrate that our framework can provide epidemic warnings for COVID-19 in its earliest stages in Dec 2019 (3 weeks before global discussions) from Chinese Weibo posts without any training in Chinese. Furthermore, we exploit our framework's argument extraction capabilities to aggregate community epidemic discussions like symptoms and cure measures, aiding misinformation detection and public attention monitoring. Overall, we lay a strong foundation for multilingual epidemic preparedness.

1 Introduction

Timely epidemic-related information is vital for policymakers to issue warnings and implement control measures (Collier et al., 2008). Social media being timely, publicly accessible, widely used, and high in volume (Heymann et al., 2001; Lamb et al., 2013; Lybarger et al., 2021) acts as a crucial information source. Previous works (Parekh et al., 2024; Zong et al., 2022) have explored utilizing Event Extraction (EE) (Sundheim, 1992; Doddington et al., 2004) to extract epidemic events from social media posts for epidemic prediction. However, these works have focused only on English; while epidemics can originate anywhere worldwide and be discussed in various regional languages.

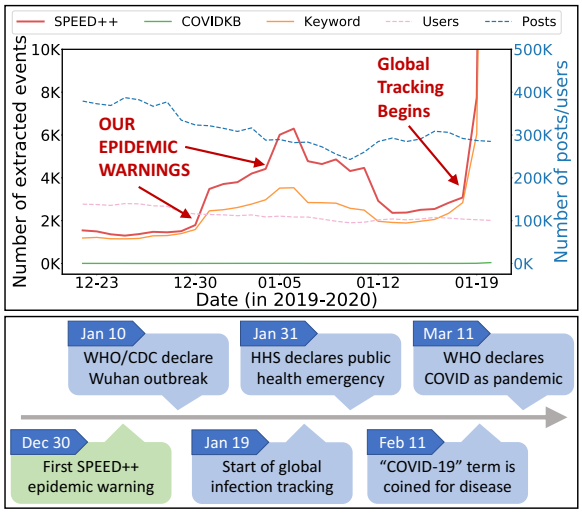

Figure 1: Zero-shot multilingual epidemic prediction in Chinese for COVID-19 pandemic. (Top) Number of epidemic events extracted in Dec-Jan 2020. Arrows indicate SPEED++ epidemic warnings. (Bottom) SPEED++ warning with respect to the general timeline of major moments of the COVID-19 pandemic.

In our work, we introduce **SPEED++** (**S**ocial **P**latform based **E**pidemic **E**vent **D**etection + Arguments + Multilinguality), the first multilingual EE framework designed for epidemic preparedness. We advance English-based SPEED (Parekh et al., 2024) multilingually by developing and benchmarking zero-shot cross-lingual models capable of extracting epidemic information across many languages. While SPEED primarily identifies basic epidemic events, we develop enhanced mod-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12936–12965
November 12-16, 2024

els capable of extracting detailed event-specific information (e.g., *symptoms*, *control measures*) by incorporating Event Argument Extraction (EAE). To integrate EAE, we enrich the SPEED ontology with event-specific roles relevant to social media to create a rich EE ontology comprising 7 event types (e.g. *infect*, *cure*, *prevent*, etc.) and 20 argument roles (e.g. *disease*, *symptoms*, *time*, *means*, etc.). Apart from English, we also annotated three other languages - Spanish, Hindi, and Japanese - to benchmark multilingual EE models. Leveraging the enriched ontology and expert annotations, we develop our SPEED++ dataset comprising 5.1K tweets and 4.6K event mentions across four different diseases (COVID-19, Monkeypox, Zika, and Dengue) in four languages.

Using SPEED++, we develop our zero-shot cross-lingual models by empowering the state-of-the-art EE models like TagPrime (Hsu et al., 2023a) with multilingual pre-training and augmented training using pseudo-generated multilingual data from CLaP (Parekh et al., 2023a). These models are trained realistically on limited English COVID-specific data. Benchmarking on SPEED++ reveals how our trained models outperform various baselines by an average of 15-16% F1 points for unseen diseases across four different languages.

To demonstrate the utility of our multilingual EE SPEED++ framework, we apply it to two epidemic-related applications. First, we utilize the framework's multilingual capabilities for epidemic prediction by aggregating epidemic events across different languages. By incorporating tweet locations, we construct a global epidemic severity meter capable of providing epidemic warnings in 65 languages spanning 117 countries. Applying our framework for COVID-19 to Chinese Weibo posts, we successfully detected early epidemic warnings by Dec 30, 2019 (Figure 1) - three weeks before the global infection tracking even began. This multilingual epidemic prediction capability can significantly enhance our global preparedness for future epidemics.

As another application, we repurpose our framework as an information aggregation system for community discussions about epidemics such as *symptoms, cure measures*, etc. Leveraging the EAE capability of our framework, we meticulously extract these event-specific details from millions of tweets across diseases and languages. Similar arguments are then agglomeratively clustered to generate an aggregated ranked bulletin. We demonstrate that this bulletin can aid misinformation detection (e.g., *cow urine* as a cure for *COVID-19*) and public attention shift monitoring (e.g., *rashes* as symptoms for *Monkeypox*). Such an automated disease-agnostic multilingual aggregation system can significantly alleviate human effort while providing insights into public epidemic opinions.

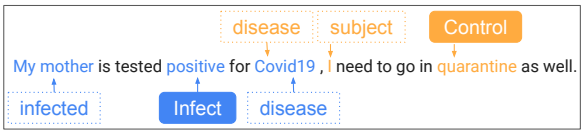

Figure 2: Illustration of Event Extraction for epidemic-related events *Infect* and *Control*. Corresponding arguments and their roles are marked in dotted boxes - that are absent in the SPEED (Parekh et al., 2024) dataset.

In conclusion, our work presents a three-fold contribution. First, we create the first multilingual Event Extraction dataset for epidemic prediction SPEED++ encompassing four diseases and four languages. Second, leveraging SPEED++, we develop models proficient in extracting epidemic-related data across a wide set of diseases and languages. Lastly, we demonstrate the robust utility of our framework through two epidemic-centric applications, facilitating multilingual epidemic prediction and the aggregation of epidemic information.

2 Background

Epidemic prediction is a classic epidemiological task that provides early warnings for future epidemics of any infectious disease (Signorini et al., 2011). Previous works (Lejeune et al., 2015; Lybarger et al., 2021) have utilized keyword-based and simple classification-based methods for extracting epidemic mentions (detailed in § 6). SPEED (Parekh et al., 2024) was the first to explore Event Extraction (EE) for extracting epidemic-based events in English. In our work, we utilize Event Extraction but focus multilingually on a broader range of languages. The extracted events are aggregated over time and abnormal influxes are reported as early epidemic warnings. To our best knowledge, we are the first to develop a multilingual Event Extraction framework for epidemic prediction.

Task Definition We adhere to the ACE 2005 guidelines (Doddington et al., 2004) to define an **event** as an occurrence or change of state associated with a specific **event type**. An **event mention** is the sentence that describes the event, and it includes an **event trigger**, the word or phrase that most clearly indicates the event. Event Extraction comprises

two subtasks: Event Detection and Event Argument Extraction. **Event Detection (ED)** involves identifying these event triggers in sentences and classifying them into predefined event types, while **Event Argument Extraction (EAE)** extracts arguments and assigns them event-specific roles. Figure 2 shows an illustration for two event mentions for the events *infect* and *control*.

3 Dataset Creation

We focus on social media, specifically Twitter as our main document source for studying four diseases - COVID-19, Monkeypox, Zika, and Dengue. SPEED (Parekh et al., 2024) focused only on Event Detection (ED) for English. Since ED identifies events but does not provide any epidemic-related information, we improve SPEED by additionally incorporating Event Argument Extraction (EAE) to develop a complete EE dataset SPEED++. Furthermore, we extend to three other languages used on social media - Spanish, Hindi, and Japanese - to enhance the multilingual capability of our framework. We detail the data creation process below while Figure 3 provides a high-level overview.

3.1 Ontology Creation

Event ontologies comprises event types and corresponding event-specific roles. For our ontology, we derive the event types from SPEED and augment them with event-specific roles in our work. We follow ACE guidelines (Doddington et al., 2004) for role definitions while also including a few non-entity roles based on GENEVA (Parekh et al., 2023b).

We initially drafted event-specific roles through a crowdsourced survey with 100 participants. Through manual inspection, we extract the frequently mentioned roles in the responses. These are augmented with more typical roles like *Time* and *Place*. We further expand the ontology with epidemic-specific roles (e.g. *Effectiveness of a cure*, *Duration of a symptom*) from a fine-grained COVID ontology ExcavatorCovid (Min et al., 2021a). Finally, the roles are renamed to reflect corresponding events (e.g. the person who gets infected in the *infect* event is named *Infected*). This multi-perspective role curation approach enhances the diversity and coverage of our ontology.

Filtering and Validation To ensure the relevance of our ontology for social media, we analyzed the event roles based on their frequency on Twitter. Specifically, we sampled 50 tweets from the SPEED dataset and annotated them with our event roles. Based on this analysis, we filtered out event roles with too few occurrences, such as *Origin* (the source of the disease) and *Manner* (how a person was infected). Additionally, we merged roles that were too similar (e.g. *Impact* and *Effectiveness* are merged). Finally, we validate our ontology with two public health experts (epidemiologists from the Dept. of Public Health). The final ontology of event types and roles is presented in Table 1 with definitions and examples in Appendix § A.

Event Type	Argument Roles
Infect	infected, disease, place, time, value, information-source
Spread	population, disease, place, time, value, information-source, trend
Symptom	person, symptom, disease, place, time, duration, information-source
Prevent	agent, disease, means, information-source, target, effectiveness
Control	authority, disease, means, place, time, information-source, subject, effectiveness
Cure	cured, disease, means, place, time, value, facility, information-source, effectiveness, duration
Death	dead, disease, place, time, value, information-source, trend

Table 1: Event Ontology for SPEED++ comprising 7 event types and 20 argument roles.

3.2 Data Processing

We utilize Twitter as the social media platform and focus on four diseases - COVID-19, Monkeypox (MPox), Zika, and Dengue. To maintain a similar distribution, we follow the data processing process from SPEED (Parekh et al., 2024). For English, we directly utilize the base data provided by SPEED which dated tweets from May 15 to May 31, 2020. For other languages, we extract tweets in the same date range as SPEED utilizing Twitter COVID-19 Endpoint as the COVID-19 base dataset. We utilize dumps from Dias (2020) as the Zika+Dengue base dataset. For tweet preprocessing, we follow Pota et al. (2021): (1) anonymizing personal information, (2) normalizing retweets and URLs, and (3) removing emojis and segmenting hashtags.

Event-based Filtering To reduce annotation costs, we utilize SPEED's event filtering technique.

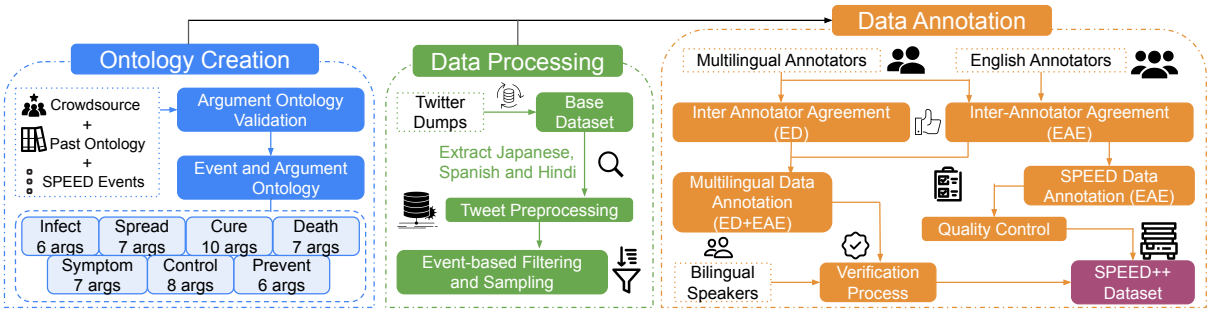

Figure 3: Overview of the data creation process. Majorly, we expand the ontology with argument roles, preprocess and filter the multilingual data, and annotate them using bilingual experts to create SPEED++.

Dataset	# Langs	# Event Types	# Arg Roles	# Sent	# EM	Avg. EM per Event	# Args	Avg. Args per Role	Domain
Genia2013	1	13	7	664	6,001	429	5,660	809	Biomedical
MLEE	1	29	14	286	6,575	227	5,958	426	Biomedical
ACE	3	33	22	29,483	5,055	153	15,328	697	News
ERE	2	38	21	17,108	7,284	192	15,584	742	News
MEE	8	16	23	31,226	50,011	3126	38,748	1685	Wikipedia
SPEED++	4	7	20	5,107	4,677	668	13,827	691	Social Media

Table 2: Data statistics for SPEED++ dataset and comparison with other standard EE datasets. Langs = languages, # = number of, Avg. = average, Sent = sentences, EM = event mentions, Args = arguments.

Specifically, each event type is associated with a seed repository of 5-10 tweets in each language. Query tweets are filtered based on their similarity to this seed repository. For procuring the multilingual event-specific seed sentences, we translate the original English seed tweets into different languages. To improve filtering efficiency, we additionally conduct keyword-based filtering for specific language-event pairs (e.g. Japanese-symptom, Japanese-cure, etc.). Here, we filtered out a query tweet if it did not contain any event-specific keywords. Finally, we apply event-based sampling from SPEED to procure the final base dataset that is utilized for data annotation. Additional details are discussed in § B.

3.3 Data Annotation

We conduct two sets of annotations to create our multilingual EE dataset: (1) EAE annotations for existing SPEED English ED data and (2) ED+EAE annotations for data in Japanese, Hindi, and Spanish. For ED, annotators were tasked to identify the presence of any events in a given tweet. For EAE, annotators were further asked to identify and extract event-specific roles that were also mentioned in the tweet. We provide further details about the annotation guidelines in § C.

Annotators and Agreement To maintain high annotation quality, our annotators were selected to be a pool of seven experts who were computer science NLP students trained through multiple annotation rounds. Of these seven, we had three annotators who were bilingual speakers of English and Japanese/Hindi/Spanish respectively. These three annotators handled the multilingual ED and EAE annotations. The remaining four annotators, along with the bilingual English-Hindi annotator, focused on English EAE annotations.

To ensure good annotation agreement, we conduct two agreement studies among the annotators: (1) ED annotations for multilingual annotators and (2) EAE annotations for all annotators. Both these studies were conducted using English data (even for multilingual annotations) to ensure that agreement could be measured in a fair manner. Agreement scores were measured using Fleiss' Kappa (Fleiss, 1971). For ED agreement, two rounds of study for the 3 multilingual annotators yielded a super-strong agreement score of 0.75 (30 samples). For EAE, the agreement score for the 7 annotators after two annotation rounds was a decent 0.6 (25 samples).

Annotation Verification To mitigate single annotator bias, each datapoint in the English data is annotated by two annotators, with a third annotator resolving inconsistencies. Owing to the scarcity of

Lang	# Sent	Avg. Length	# EM	# Args
en	2,560	32.5	2,887	8,423
es	1,012	32.4	614	1,485
hi	716	30.0	627	2,344
ja	819	89.2*	549	1,575

Table 3: Data statistics for SPEED++ split by language. # = number of, Avg = average, Lang = language, Sent = sentences, Args = arguments, *character-level.

multilingual annotators, we hire three additional bilingual speakers to verify the multilingual annotations. These verification annotators were selected through a thorough qualification test to ensure high verification quality. They were requested to judge if the current annotations were reasonable. If the original annotation was deemed incorrect, they were asked to provide feedback to correct the annotations. This feedback was finally utilized by the original multilingual annotators to rectify the annotation. We provide additional details in § C.1.

3.4 Data Analysis

Comparison with other datasets SPEED++ comprises 5,106 tweets with 4,674 event mentions and 13,815 argument mentions across four diseases and four languages. We present the main statistics along with comparisons with other prominent EE datasets like ACE (Doddington et al., 2004), ERE (Song et al., 2015), Genia2013 (Li et al., 2020), MEE (Pouran Ben Veyseh et al., 2022), and MLEE (Pyysalo et al., 2012) in Table 2. We note that SPEED++ is one of the few multilingual EE datasets, notably the first in the social media domain. Overall, SPEED++ is comparable in various event and argument-related statistics with the previous standard EE datasets.

Multilingual Statistics We provide a deeper split of data statistics per language in Table 3. Owing to cheaper annotations, English has many more annotated sentences compared to other languages. This is also a design choice, as we will solely utilize English data for training zero-shot multilingual models (discussed in § 4). In terms of event and argument densities (i.e. # EM / # Sent and # Args / # Sent), we notice a broader variation across languages, with English and Hindi being denser. The average lengths (in terms of the number of words) are similar across the languages.

Argument Study We deep-dive to study the density in terms of arguments per sentence for

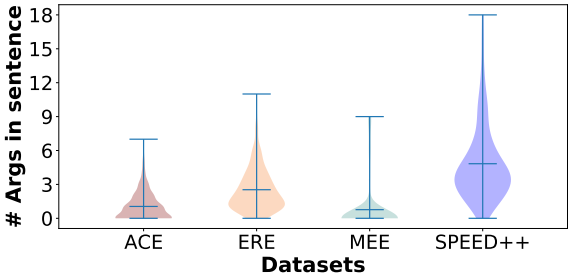

Figure 4: Distribution of the number of arguments (# Args) per sentence for SPEED++ relative to other datasets ACE, ERE, and MEE.

	Disease	Language	# Sent	# EM
Train	COVID	English	1,601	1,746
Dev	COVID	English	374	471
Test	COVID	Spanish	534	365
		Hindi	416	412
		Japanese	542	395
	Monkeypox	English	286	398
	Zika + Dengue	English	299	272
		Spanish	478	249
		Hindi	300	215
		Japanese	277	154

Table 4: Data split for epidemic event extraction. # = number of, Sent = sentences, EM = event mentions.

SPEED++ by comparing with standard EE datasets ACE and ERE and a multilingual EE dataset MEE in Figure 4. Noticeably, SPEED++ is more dense (has a higher mean argument per sentence value) and has a broader distribution with sentences up to 18 arguments as well. Furthermore, following GENEVA (Parekh et al., 2023b), we add 4 non-entity roles, which make up 20% of the total arguments. Such non-entity arguments are not present in any other multilingual datasets. Overall, the high and broad argument density and the existence of non-entity arguments render SPEED++ to be a more challenging EE dataset.

4 Zero-shot Cross-lingual Event Extraction

To validate the effectiveness of EE for epidemic events, we benchmark various EE models using SPEED++. Given the infeasibility of procuring quality data in all languages for all diseases, we benchmark in a zero-shot cross-lingual cross-disease fashion i.e. we train models only on English COVID data and evaluate on the rest. We provide the data split for our benchmarking in Table 4. For evaluation, we report F1-score for event classifi-

Model	COVID hi		COVID jp		COVID es		MPox en		Zika + Dengue en		Zika + Dengue hi		Zika + Dengue jp		Zika + Dengue es		Average	
	EC	**AC**	**EC**	**AC**	**EC**	**AC**	**EC**	**AC**	**EC**	**AC**	**EC**	**AC**	**EC**	**AC**	**EC**	**AC**	**EC**	**AC**
BASELINE MODELS																		
ACE - TagPrime	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DivED*‡	0.0	-	0.0	-	27.0	-	36.7	-	47.7	-	4.4	-	1.3	-	12.8	-	16.2	-
Keyword*	15.2	-	28.6	-	26.3	-	41.3	-	39.3	-	18.6	-	**39.7**	-	20.6	-	28.7	-
COVIDKB†	45.2	-	42.3	-	24.2	-	18.5	-	45.5	-	47.5	-	34.0	-	34.6	-	36.5	-
GPT-3.5-turbo*	35.7	14.5	36.4	15.0	43.2	16.8	46.4	24.0	56.6	33.0	45.5	20.0	29.0	11.0	39.6	15.1	41.6	18.7
TRAINED ON SPEED++ (OUR FRAMEWORK)																		
TagPrime	60.1	**39.0**	35.3	7.9	62.0	37.1	**70.2**	**45.1**	**66.7**	**48.5**	65.0	**40.7**	27.2	7.5	**49.9**	**28.2**	54.6	31.8
TagPrime + XGear	60.1	27.1	35.3	9.7	62.0	36.0	**70.2**	42.0	**66.7**	45.8	65.0	31.9	27.2	8.8	**49.9**	26.6	54.6	28.5
BERT-QA	54.7	33.9	21.2	4.0	60.6	28.1	66.1	39.0	63.0	45.5	50.8	31.1	4.6	0.8	41.9	24.1	45.4	25.8
DyGIE++	61.0	35.7	38.2	2.0	61.7	**39.1**	67.4	39.3	64.1	46.4	61.4	32.2	27.5	0.4	45.6	22.7	53.4	27.2
OneIE	**61.9**	34.3	12.0	11.4	44.5	37.3	68.8	42.6	**66.7**	47.9	61.9	38.5	12.0	5.0	44.5	25.4	46.5	30.3
TagPrime + CLaP	58.6	32.9	**48.4**	**19.1**	**62.6**	37.7	**70.2**	**45.1**	**66.7**	**48.5**	**65.2**	40.6	39.2	**18.8**	49.7	28.1	**57.6**	**33.9**

Table 5: Benchmarking EE models trained on SPEED++ for extracting event information in the cross-lingual cross-disease setting. EC = Event Classification, AC = Argument Classification, hi = Hindi, jp = Japanese, es = Spanish, and en = English. *Numbers are higher due to string matching evaluation. †Binary classification evaluation.

cation (**EC**) and argument classification (**AC**) (Ahn, 2006) measuring the classification of events and arguments respectively. We use TextEE (Huang et al., 2024) for most implementations, with specific details discussed in § D. Additional benchmarking experiments are provided in § E.

EE Models Most EE models solely focus on English and can not be directly utilized in the cross-lingual setting. To this end, we adopt the following models using multilingual pre-trained models and tokenization: (1) TagPrime (Hsu et al., 2023a), (2) EEQA (Du and Cardie, 2020), (3) DyGIE++ (Wadden et al., 2019a), (4) OneIE (Lin et al., 2020a), (5) XGear (Huang et al., 2022). Since XGear is an EAE-only model, we combine it with the TagPrime ED model. To further improve the models, we train them with pseudo-generated data using a label projection model CLaP (Parekh et al., 2023a).

Baseline Models We consider the following baselines: (1) ACE - TagPrime, a TagPrime model trained on the multilingual EE dataset ACE (Doddington et al., 2004) and transferred to SPEED++. (2) DivED (Cai et al., 2024), a Llama2-7B model fine-tuned on a diverse range of event definitions, (3) COVIDKB (Zong et al., 2022), an epidemiological work using a BERT classification model. Since the original output classes are different, we train it to simply classify tweets as epidemic-related or not. (4) Keyword baseline inspired from an epidemiological work (Lejeune et al., 2015) curates a set of keywords for each event. (5) GPT-3.5-turbo (Brown et al., 2020), a Large Language Model (LLM) baseline using seven in-context examples.

Results We present our per-disease per-language results in Table 5. We note that most of the baseline models do not perform well for our task, as was noted also in Parekh et al. (2024). The GPT-based LLM baseline performs better in English but exhibits poor performance across other languages. On the other hand, we observe stronger performance by the **supervised baselines trained on our SPEED++ dataset with TagPrime providing the best overall average performance**. We also note how CLaP further improves performance by 2-3 F1 points in the cross-lingual setting, especially for character-based language Japanese.

5 Applications

To validate its practical utility for epidemic preparedness, we demonstrate our framework's use in two downstream applications: Global Epidemic Prediction and Epidemic Information Aggregation. For this, we train a multilingual TagPrime model on the entire SPEED++ dataset. Further details about these applications are provided below.

5.1 Global Epidemic Prediction

To showcase the robust multilingual utility of our framework, we highlight its extensive language coverage and provide an in-depth analysis of COVID-19 predictions from Chinese data.

‡English-based Llama performs poorly multilingually.

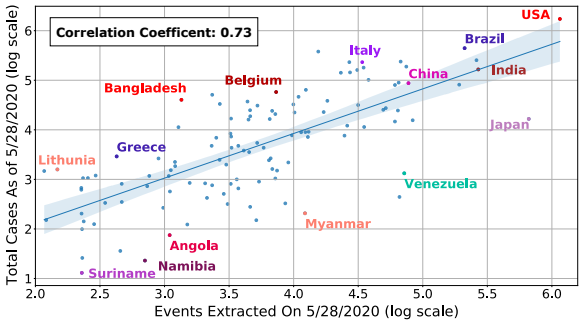

Figure 5: Number of extracted events plotted against the number of reported cases for each country. Both of them are in log scale.

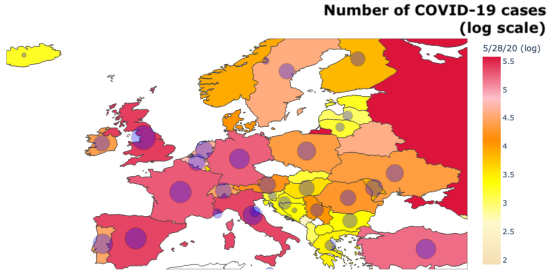

Figure 6: Geographical distribution of the number of reported COVID-19 cases as of May 28, 2020 in Europe. Red depicts more spread, and yellow/white indicates less spread. The blue dots indicate the events extracted by our model and its size depicts the number of epidemic events for the specific country (log scale).

Global Epidemic Monitoring Validating our framework for each language is resource-intensive and infeasible. Instead, we perform a preliminary study of the broad language coverage of our framework by demonstrating its capability to detect COVID-related events across 65 languages. We analyze tweets across all languages from a single day (May 28, 2020) and extract epidemic events using our framework. Utilizing user locations, we map the tweets from these languages to 117 countries. For reference, we plot the extracted events for each country with the actual number of reported COVID-19 cases[1] in Figure 5. Countries with significant COVID-19 spread appear in the top-right, while some outliers are also shown in the figure. Our framework achieves a healthy correlation of 0.73 with the actual reported cases, indicating strong performance across a broad range of languages.

We further extend these plots geographically in Figure 6. Each country is color-coded on the number of COVID cases, with lighter shades indicating fewer cases while darker shades indicate a massive spread. The extracted number of events from the country-mapped tweets by our framework are plotted as translucent circles. Bigger dots indicate more events extracted for the given country. In this plot, we observe extracted events and COVID-19 spread more in Western European countries like the United Kingdom, France, Spain, Italy, and Germany, while lesser events spread in Eastern European countries. We provide additional details along with a world map geographical plot in § F.

[1] https://www.worldometers.info/coronavirus/

Chinese Posts	Translation
武汉华南海鲜市场出现 [infect]多个不明原因肺炎病例，请同道们提高警惕，早期发现，早期隔离 [prevent]	Multiple cases of pneumonia of unknown origin have appeared [infect] in Wuhan's Huanan Seafood Market. Please be more vigilant, detect and isolate [prevent] them as early as possible.
近日，武汉进入流感高发期 [spread]，多家医院感冒 [symptom]发烧的患者数量猛增。	Recently, Wuhan has entered a period of high influenza incidence [spread], and the number of patients with colds [symptom] and fevers in many hospitals has increased sharply.

Table 6: Sample Weibo posts in Chinese with their translations identified by SPEED++ framework as epidemic-related from late December 2019. Event types and their trigger words are marked in blue.

COVID-19 Epidemic Prediction using Chinese As a case study, we examine the earliest stages of the COVID-19 pandemic, analyzing Chinese social media posts from Dec 16, 2019, to Jan 21, 2020, using Weibo data from Hu et al. (2020). Using our trained TagPrime model, we infer on Chinese in a *zero-shot fashion* (i.e. without prior training on Chinese). We aggregate the 7-day rolling average of our extracted event mentions across time and report any sharp increases as epidemic warnings, as illustrated in Figure 1. Since case reporting had not begun, actual COVID-19 case numbers are unavailable for this period. Instead, we also plot the total number of Weibo posts and active users (Guo et al., 2021). Additionally, we compare trends with baselines from COVIDKB (Zong et al., 2022) and a keyword-based approach (Lejeune et al., 2015).

Figure 1 demonstrates how our SPEED++ framework provides epidemic warnings three weeks before the global tracking of infection cases began. While the keyword-based method also provides some signals, they are relatively weaker. Furthermore, Table 5 shows that the keyword baseline performs worse for morphologically richer languages,

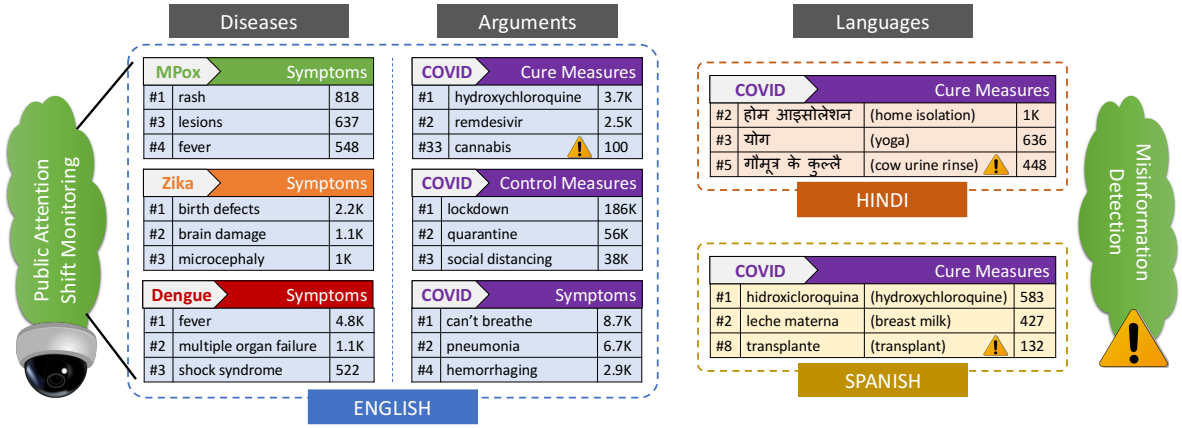

Figure 7: Information Assimilation Bulletin as extracted by our SPEED++ framework and agglutinatively clustered. The first column represents different diseases, the second column represents different argument roles, and the third column represents the different languages. We also highlight the utility of these bulletin for two applications of Public Attention Shift Monitoring and Misinformation Detection.

making it less robust. Additionally, the number of posts and active users do not provide any epidemic-related signals. For further validation, we present sample event mentions extracted by our framework in Table 6. In the bottom timeline[2] in Figure 1, we demonstrate the efficacy of our framework as it provided epidemic warnings 6 weeks before the "COVID-19" term was coined and used in social media. Overall, we show how **our framework can provide early epidemic warnings multilingually without relying on any target language data, making it suitable for global deployment.**

5.2 Epidemic Information Aggregation

Our framework possesses a strong EAE capability to extract detailed information about epidemic events such as *symptoms*, *preventive measures*, *cure measures*, etc. Aggregating such information from millions of social media posts can provide insights into public opinions regarding various epidemic aspects. To this end, we develop an information aggregation system for community epidemic discussions. Specifically, we use our framework to extract arguments for various event roles, project them into a representative space, and merge similar arguments using agglomerative clustering. The final arguments and their counts for different diseases and languages, extracted from 6M tweets, are presented as a bulletin in Figure 7. Further details and a complete table are reported in § G.

Our framework effectively extracts various arguments for COVID-19 in English (middle column of Figure 7), including cure measures such as *hydroxychloroquine* and *remdesivir*, control measures like *lockdown* and *quarantine*, and symptoms such as *pneumonia*. Additionally, this capability extends to other diseases, such as Monkeypox, Zika, and Dengue (left column), and across languages, including Hindi and Spanish (right column). This condensed information is crucial for Public Attention Shift Monitoring, aiding policymakers in devising better control measures (Liu and Fu, 2022). We demonstrate this in the form of extracted symptoms such as *rashes* and *lesions* for Monkeypox and *fever* and *shock syndrome* for Dengue (left column). Simultaneously, our framework can assist in **Misinformation Detection** (Mendes et al., 2023). Shown as caution signs in the figure, we highlight various potential COVID-19 cure misinformation such as *cow urine rinse*, *cannabis*, and *transplants* extracted across languages by our framework. As evidence, we also provide some of the actual tweets in Hindi and Spanish flagged by our framework for mentioning these terms in Table 7. We provide example tweets comprising these arguments as extracted by our framework in § G.

6 Related Works

Event Extraction Datasets Event Extraction (EE) aims at detecting events (Event Detection) and extracting details about specific roles associated with each event (Event Argument Extraction) from natural text. Unlike document parsing (Tong et al., 2022; Suvarna et al., 2024), we utilize EE only at the sentence/tweet level in our work. Over-

[2] https://www.cdc.gov/museum/timeline/covid19.html

Tweet	Translation
HINDI - COVID-19 - CURE MEASURES	
कोरोना के खिलाफ एक योद्धा बन के योग से कोरोना को हराना है।	Become a warrior against Corona and defeat it through yoga
संबित कोरोना पॉजिटिव, हालत नाजुक आत्मनिर्भर अभियान के तहत गौमूत्र के कुल्ले करवाकर उपचार किया जा रहा है।	Sambit is corona positive, condition is critical, and is being treated by gargling with cow urine under the self-reliant campaign.
संबित जी के कोरोना प्रभावित होने की खबर मिली। जानकर दुःख हुआ, लेकिन वे अस्पताल में स्वास्थ्य लाभ ले रहे हैं, और अब बहुत बेहतर है	Received the news of Sambit ji being affected by Corona! Sorry to hear, but he's recovering in the hospital, and is much better now.
SPANISH - COVID-19 - CURE MEASURES	
Científicos rusos sugieren que una proteína presente en la leche materna puede ser clave en la lucha contra el covid-19 (url)	Russian scientists suggest that a protein present in breast milk may be key in the fight against covid-19 (url)
Este transplante de pulmones a un paciente con COVID-19 es una operación realizada hasta ahora sólo en China y que por primera vez se lleva a cabo en Europa	This lung transplant to a patient with COVID-19 is an operation carried out so far only in China and is being carried out for the first time in Europe.
el mundo acumula más evidencia de la efectividad de la ivermectina para el tratamiento en casa de pacientes con estadios leves de #(COVID)-19	the world accumulates more evidence of the effectiveness of ivermectin for the home treatment of patients with mild stages of #(COVID)-19

Table 7: Illustration of some actual tweets in Hindi and Spanish mentioning various cure measures related to COVID-19. The terms related to *cure measures* extracted by our framework are highlighted in red.

all, EE is a well-studied task with earliest works dating back to MUC (Sundheim, 1992; Grishman and Sundheim, 1996), ACE (Doddington et al., 2004) and ERE (Song et al., 2015), but there have been various newer diverse datasets like MAVEN (Wang et al., 2020), WikiEvents (Li et al., 2021), FewEvent (Deng et al., 2019), DocEE (Tong et al., 2022), and GENEVA (Parekh et al., 2023b). While most of these datasets are only in English, some datasets like ACE (Doddington et al., 2004), ERE (Song et al., 2015), and MEE (Pouran Ben Veyseh et al., 2022) provide EE data in ten languages for general-purpose events in the news and Wikipedia domains. SPEED (Parekh et al., 2024) introduces data for epidemic-based events in social media but is limited to Event Detection and focuses on English. Overall, SPEED++ extends SPEED to four languages and Event Argument Extraction.

Multilingual Epidemiological Information Extraction Early epidemiological works (Lindberg et al., 1993; Rector et al., 1996; Stearns et al., 2001) largely focused on defining extensive ontologies for usage by biomedical experts. BioCaster (Collier et al., 2008) and PULS (Du et al., 2011) explored utilizing rule-based methods for the news domain. Early information extraction systems tackled predicting influenza trends from social media (Signorini et al., 2011; Lamb et al., 2013; Paul et al., 2014). More recently, IDO (Babcock et al., 2021) and DO (Schriml et al., 2022) are two extensive ontologies for human diseases. CIDO (He et al., 2020), ExcavatorCovid (Min et al., 2021b), CACT (Lybarger et al., 2021) and COVIDKB (Zong et al., 2022; Mendes et al., 2023) were developed specifically focused on COVID-19 events. While most of these works are English-focused, some other works (Lejeune et al., 2015; Mutuvi et al., 2020; Sahnoun and Lejeune, 2021) support multilingual systems by using keyword-based and simple classification methods. Overall, most of these systems are English-centric, disease-specific, not suitable for use for social media and utilize very rudimentary models. In our work, utilizing disease-agnostic annotations and powerful multilingual models, we develop models that can detect events for any disease mentioned in any language.

7 Conclusion and Future Work

In our work, we pioneer the creation of the first multilingual Event Extraction (EE) framework for application in epidemic prediction and preparedness. To this end, we create a multilingual EE benchmarking dataset SPEED++ comprising 5K tweets spanning four languages and four diseases. To realistically deploy our models, we develop zero-shot cross-lingual cross-disease models and demonstrate their capability to extract events for 65 languages spanning 117 countries. We prove the effectiveness of our model by providing early epidemic warnings for COVID-19 from Chinese Weibo posts in a zero-shot manner. We also show evidence that our framework can be utilized as an information aggregation system aiding in misinformation detection and public attention monitoring. In conclusion, we provide a strong utility of multilingual EE for global epidemic preparedness.

Acknowledgements

We express our gratitude to Anh Mac, Syed Shahriar, Di Wu, Po-Nien Kung, Rohan Wadhawan, and Haw-Shiuan Chang for their valuable time, reviews of our work, and constructive feedback. We thank the anonymous reviewers and the area editors for their feedback. This work was supported by NSF 2200274, 2106859, 2312501, DARPA HR00112290103/HR0011260656, NIH U54HG012517, U24DK097771, as well as Optum Labs. We thank them for their support.

Limitations

We benchmark our framework for four languages, but it is possible that it would be poor in performance for many others. Owing to the lack of annotated data, it is difficult to conduct a holistic multilingual evaluation of our framework. Our experiments on global epidemic prediction and information aggregation have been done on a single day of social media posts. Furthermore, owing to the expensive cost of procuring massive social media data, it is infeasible run our framework across languages for a longer duration of time. Finally, our major experiments are based on four diseases. We would like to expand this further, but owing to budget constraints, we restrict ourselves to four diseases only in this work.

Ethical Considerations

Our framework extracts signals from social media wide range of languages to provide epidemic information. However, the internet and access and usage of social media are disparate across the globe - leading to biased representation. This aspect should be considered when utilizing our framework for inferring for low-resource languages or under-represented communities.

Since our work utilizes actual tweets, there could be some private information that could not be completely anonymized in our pre-processing. These tweets may also possess stark emotional, racial, and political viewpoints and biases. Our work doesn't focus on bias mitigation and our models may possess such a bias. Due consideration should be taken when using our data or models.

Finally, despite our best efforts, our framework is far from being ideal and usable in real-life scenarios. Our framework can output false positives frequently. The goal of our work is to provide a strong prototype and encourage research in this direction. Usage of our models/framework for practical use cases should be appropriately considered.

Note: We utilize ChatGPT in writing the paper better and correctly grammatical mistakes.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Shane Babcock, John Beverley, Lindsay G. Cowell, and Barry Smith. 2021. The infectious disease ontology in the age of COVID-19. *J. Biomed. Semant.*, 12(1):13.

Hritik Bansal, Po-Nien Kung, P. Jeffrey Brantingham, Kai-Wei Chang, and Nanyun Peng. 2024. Genearl: A training-free generative framework for multimodal event argument role labeling. *CoRR*, abs/2404.04763.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Zefan Cai, Po-Nien Kung, Ashima Suvarna, Mingyu Derek Ma, Hritik Bansal, Baobao Chang, P. Jeffrey Brantingham, Wei Wang, and Nanyun Peng. 2024. Improving event definition following for zero-shot event detection. *CoRR*, abs/2403.02586.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Hung Quoc Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinform.*, 24(24):2940–2941.

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. *Preprint*, arXiv:1911.02116.

Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2019. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. *CoRR*, abs/1910.11621.

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.

Guilherme Dias. 2020. Tweets dataset on Zika virus.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Mian Du, Peter von Etter, Mikhail Kopotev, Mikhail Novikov, Natalia Tarbeeva, and Roman Yangarber. 2011. Building support tools for russian-language information extraction. In *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 380–387. Springer.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018a. Code-switched language models using dual RNNs and same-source pretraining. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3083, Brussels, Belgium. Association for Computational Linguistics.

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018b. Dual language models for code switched speech recognition. In *19th Annual Conference of the International Speech Communication Association, Interspeech 2018, Hyderabad, India, September 2-6, 2018*, pages 2598–2602. ISCA.

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

Shuhui Guo, Fan Fang, Tao Zhou, Wei Zhang, Qiang Guo, Rui bi Zeng, Xiaohong Chen, Jianguo Liu, and Xin Lu. 2021. Improving google flu trends for covid-19 estimates using weibo posts. *Data Science and Management*, 3:13 – 21.

Yongqun He, Hong Yu, Edison Ong, Yang Wang, Yingtong Liu, Anthony Huffman, Hsin-Hui Huang, John Beverley, Asiyah Yu Lin, William D. Duncan, Sivaram Arabandi, Jiangan Xie, Junguk Hur, Xiaolin Yang, Luonan Chen, Gilbert S. Omenn, Brian D. Athey, and Barry Smith. 2020. CIDO: the community-based coronavirus infectious disease ontology. In *Proceedings of the 11th International Conference on Biomedical Ontologies (ICBO) joint with the 10th Workshop on Ontologies and Data in Life Sciences (ODLS) and part of the Bolzano Summer of Knowledge (BoSK 2020), Virtual conference hosted in Bolzano, Italy, September 17, 2020*, volume 2807 of *CEUR Workshop Proceedings*, pages 1–10. CEUR-WS.org.

David L Heymann, Guénaël R Rodier, et al. 2001. Hot spots in a wired world: Who surveillance of emerging and re-emerging infectious diseases. *The Lancet infectious diseases*, 1(5):345–353.

I Hsu, Xiao Guo, Premkumar Natarajan, Nanyun Peng, et al. 2021. Discourse-level relation extraction via graph pooling. *arXiv preprint arXiv:2101.00124*.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

I-Hung Hsu, Kuan-Hao Huang, Shuning Zhang, Wenxin Cheng, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2023a. TAGPRIME: A unified framework for relational structure extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12917–12932, Toronto, Canada. Association for Computational Linguistics.

I-Hung Hsu, Zhiyu Xie, Kuan-Hao Huang, Prem Natarajan, and Nanyun Peng. 2023b. AMPERE: AMR-aware prefix for generation-based event argument extraction model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10976–10993, Toronto, Canada. Association for Computational Linguistics.

I-Hung Hsu, Zihan Xue, Nilay Pochhi, Sahil Bansal, Prem Natarajan, Jayanth Srinivasa, and Nanyun Peng. 2024. Argument-aware approach to event linking. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12769–12781, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yong Hu, Heyan Huang, Anfan Chen, and Xian-Ling Mao. 2020. Weibo-COV: A large-scale COVID-19 social media dataset from Weibo. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2)*

at EMNLP 2020, Online. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 4633–4646, Dublin, Ireland. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. Textee: Benchmark, reevaluation, reflections, and future challenges in event extraction. *Preprint*, arXiv:2311.09562.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, Atlanta, Georgia. Association for Computational Linguistics.

Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. 2015. Multilingual event extraction for epidemic detection. *Artif. Intell. Medicine*, 65(2):131–143.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.

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*, pages 894–908, Online. Association for Computational Linguistics.

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

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

D. A. Lindberg, B. L. Humphreys, and A. T. McCray. 1993. The unified medical language system. *Methods of information in medicine*, 32(4):281—-291.

Lu Liu and Yifei Fu. 2022. Study on the mechanism of public attention to a major event: The outbreak of covid-19 in china. *Sustainable Cities and Society*, 81:103811.

Kevin Lybarger, Mari Ostendorf, Matthew Thompson, and Meliha Yetisgen. 2021. Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. *J. Biomed. Informatics*, 117:103761.

Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15817–15835, Toronto, Canada. Association for Computational Linguistics.

Bonan Min, Benjamin Rozonoyer, Haoling Qiu, Alexander Zamanian, and Jessica MacBride. 2021a. Excavatorcovid: Extracting events and relations from text corpora for temporal and causal analysis for covid-19. *Preprint*, arXiv:2105.01819.

Bonan Min, Benjamin Rozonoyer, Haoling Qiu, Alexander Zamanian, Nianwen Xue, and Jessica MacBride. 2021b. ExcavatorCovid: Extracting events and relations from text corpora for temporal and causal analysis for COVID-19. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 63–71, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Mutuvi, Antoine Doucet, Gaël Lejeune, and Moses Odeo. 2020. A dataset for multi-lingual epidemiological event extraction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4139–4144, Marseille, France. European Language Resources Association.

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding linguistic accommodation in code-switched human-machine dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 565–577, Online. Association for Computational Linguistics.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023a. Contextual label projection for cross-lingual structure extraction. *CoRR*, abs/2309.08943.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023b. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686, Toronto, Canada. Association for Computational Linguistics.

Tanmay Parekh, Anh Mac, Jiarui Yu, Yuxuan Dong, Syed Shahriar, Bonnie Liu, Eric Yang, Kuan-Hao

Huang, Wei Wang, Nanyun Peng, and Kai-Wei Chang. 2024. Event detection from social media for epidemic prediction. *CoRR*, abs/2404.01679.

Michael J Paul, Mark Dredze, and David Broniatowski. 2014. Twitter improves influenza forecasting. *PLoS currents*, 6.

Marco Pota, Mirko Ventura, Hamido Fujita, and Massimo Esposito. 2021. Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. *Expert Syst. Appl.*, 181:115119.

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

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Junichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinform.*, 28(18):575–581.

Alan L Rector, Jeremy E Rogers, and Pam Pole. 1996. The galen high level ontology. In *Medical Informatics Europe'96*, pages 174–178. IOS Press.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sihem Sahnoun and Gaël Lejeune. 2021. Multilingual epidemic event extraction : From simple classification methods to open information extraction (OIE) and ontology. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1227–1233, Held Online. INCOMA Ltd.

Lynn M. Schriml, James B. Munro, Mike Schor, Dustin Olley, Carrie McCracken, Victor Felix, J. Allen Baron, Rebecca C. Jackson, Susan M. Bello, Cynthia Bearer, Richard Lichenstein, Katharine Bisordi, Nicole Campion, Michelle G. Giglio, and Carol Greene. 2022. The human disease ontology 2022 update. *Nucleic Acids Res.*, 50(D1):1255–1261.

Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. 2001. SNOMED clinical terms: overview of the development process and project status. In *AMIA 2001, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 3-7, 2001*. AMIA.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Ashima Suvarna, Xiao Liu, Tanmay Parekh, Kai-Wei Chang, and Nanyun Peng. 2024. QUDSELECT: selective decoding for questions under discussion parsing. *CoRR*, abs/2408.01046.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019a. Entity, relation, and event extraction with contextualized span representations. 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 5784–5789, Hong Kong, China. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019b. Entity, relation, and event extraction with contextualized span representations. *ArXiv*, abs/1909.03546.

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

Di Wu, Xiaoxian Shen, and Kai-Wei Chang. 2024. Metakp: On-demand keyphrase generation. *CoRR*, abs/2407.00191.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2022. Extracting a knowledge base of COVID-19 events from social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3810–3823, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Ontology Creation: Role Definitions

We provide our complete event ontology, including argument definitions along with corresponding examples in Table 24-30. We underline the arguments corresponding to each role in the examples. We note that our ontology can be further utilized for other tasks as well, like relation extraction (Hsu et al., 2021) and event linking (Hsu et al., 2024).

B Dataset Filtering and Sampling

While there are works which focus on Event Extraction from multimodal tweets (Bansal et al., 2024), we restrict our work to text-based tweets only. We associate each event with 5-10 seed tweets inspired by SPEED (Parekh et al., 2024). Utilizing embedding-space similarity of query tweets and our seed tweet repository, we filter out tweets related to epidemic events. For multilingual languages, we translate the English seed tweets into individual languages. We modify and further correct the translations with the help of human experts. We provide some seed tweets per language per event in Table 31. We utilize the sentence-transformer model (Reimers and Gurevych, 2019) for embedding the tweets.

Furthermore, we utilize the event-similarity to uniformly sample tweets based on events. More specifically, we over-sample tweets from frequent events and under-sample for the non-frequent ones. Such uniform sampling has proved elemental to more robust model training, as noted in Parekh et al. (2023b).

C Annotation Guidelines and Details

We conduct two sets of annotations in our work and describe both in more detail here. First, we conduct ED annotations for multilingual data in Japanese, Hindi, and Spanish. We refer to ACE (Doddington et al., 2004) and SPEED (Parekh et al., 2024) to chalk our guidelines. We provide instructions and examples to the annotators in English while they are expected to annotate data in the respective target languages. We provide the exact annotation guidelines in Figure 10.

Next, we conduct EAE annotations for all the languages. Inspired by ACE (Doddington et al., 2004) and GENEVA (Parekh et al., 2023b), we design our guidelines with special instructions. We present the instructions in Figure 11 along with simple argument definitions in Figure 12.

Language	Inconsistent rate	Verification acceptance rate
Hindi	27.46%	31.48%
Japanese	17.52%	81.73%
Spanish	19.01%	66.66%

Table 8: Inconsistencies identified (as percentage) and verifications accepted (as percentage) for the multilingual verifications. The inconsistent rate is the percentage of annotations with which the bilingual speakers did not agree, while the verification rate is the percentage of suggestions from the bilingual speakers that are accepted by our multilingual annotators.

C.1 Multilingual Data Verification

Due to the scarcity of multilingual annotators, we adopted a verification process different from the English data verification. The entire verification procedure can be divided into four phases: qualification task, inter-annotator agreement (IAA) study, verification task, and correction task.

We opt to choose bilingual speakers of English and Hindi/Japanese/Spanish as the verifiers. Not that we do not consider code-switching (Garg et al., 2018a,b) in our work, but bilingualism helps to ensure that the instructions are well understood by the verifiers. To ensure verification quality, each bilingual speaker must pass a qualification test before entering the verification process. They are provided with a guideline explaining their primary task introducing the essence ED/EAE annotation (Figure 13), argument definitions (Figure 12) along with two pairs of positive and negative examples (Figure 14) - all in English. Although not directly tasked with ED/EAE annotation, they must understand the standards of ED/EAE annotation to fairly judge the correctness of a given annotated example. After reading the instructions, the bilingual speaker must correctly answer at least 4 out of 5 test questions to pass the qualification test. Selected by our multilingual annotators, these test questions are in Japanese, Hindi, and Spanish, respectively, and are in the same format as the verification questions. Failing the qualification task indicates an insufficient understanding of the verification task, thereby disqualifying the bilingual speaker from proceeding further. We select one verifier for each language after filtering from this round. Each of the verifiers was paid $150 in total for 6 hours of their service at the rate of $25/hr in line with other works (Parekh et al., 2020).

Next, we asked the three qualified bilingual speakers to verify 40 English examples as part of an inter-annotator agreement (IAA) study to ensure an adequate agreement rate among them. These IAA examples are in the same format as the actual verification questions. They reached a final IAA score of 0.6 on the 40 English samples.

Next, the qualified bilingual speakers participate in the final verification process. Along with the tweet text, they are shown all the events and arguments identified by our annotators. If they agree with the current annotation, no action is needed; otherwise, they should check the "incorrect" box and provide their reasons (as shown in Figure 14).

Following the verification process, our multilingual annotators addressed the correction task: reviewing the comments and deciding on the final annotation. We provide statistics about the total corrections suggested and accepted by the multilingual annotators for each language in Table 8.

D Benchmarking Model: Implementation Details

We use the EE benchmarking tool TextEE (Huang et al., 2024) to conduct the benchmarking experiment of the models. We present details about each ED, EAE, and end-to-end model that we benchmark, along with the extensive set of hyperparameters and other implementation details.

D.1 TagPrime

TagPrime (Hsu et al., 2023a) is a sequence tagging model with a word priming technique to convey more task-specific information. We run our experiments on the ED and EAE tasks of TagPrime on an NVIDIA RTX A6000 machine with support for 8 GPUs. The models are fine-tuned on XLM-RoBERTa-Large (Conneau et al., 2020). We train this model separately for ED and EAE. The major hyperparameters are listed in Table 9 for the ED model and Table 10 for the EAE model.

Pre-trained LM	XLM-RoBERTa-Large
Training Batch Size	16
Eval Batch Size	4
Learning Rate	0.001
Weight Decay	0.001
Gradient Clipping	5
Training Epochs	10
Warmup Epochs	5
Max Sequence Length	250
Linear Layer Dropout	0.2

Table 9: Hyperparameter details for TagPrime ED model.

Pre-trained LM	XLM-RoBERTa-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	0.001
Weight Decay	0.001
Gradient Clipping	5
Training Epochs	90
Warmup Epochs	5
Max Sequence Length	250
Linear Layer Dropout	0.2

Table 10: Hyperparameter details for TagPrime EAE model.

D.2 XGear

XGear (Huang et al., 2022) is a language-agnostic model that models EAE as a generation task. This model is similar to other generative models like DEGREE (Hsu et al., 2022) and AMPERE (Hsu et al., 2023b) but focuses on zero-shot cross-lingual transfer. We run our experiments on the EAE tasks of XGear on an NVIDIA RTX A6000 machine with support for 8 GPUs. The model is fine-tuned on mT5-Large (Xue et al., 2021). The major hyperparameters for this model are listed in Table 11. To evaluate its end-to-end performance, we complement it with the TagPrime ED model.

Pre-trained LM	mT5-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	0.00001
Weight Decay	0.00001
Gradient Clipping	5
Training Epochs	90
Warmup Epochs	5
Max Sequence Length	400

Table 11: Hyperparameter details for XGear EAE model.

D.3 BERT-QA

BERT-QA (Du and Cardie, 2020) is a classification model utilizing label semantics via transforming the EE task into a question-answer task. We run our experiments on the EAE tasks of BERT-QA with support for 8 GPUs. The model is fine-tuned on XLM-RoBERTa-Large (Conneau et al., 2020). The major hyperparameters for this model are listed in Table 12 for the ED model and Table 13 for the EAE model.

Pre-trained LM	XLM-RoBERTa-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	0.001
Weight Decay	0.001
Gradient Clipping	5
Training Epochs	30
Warmup Epochs	5
Max Sequence Length	250
Linear Layer Dropout	0.2

Table 12: Hyperparameter details for BERT-QA ED model.

Pre-trained LM	XLM-RoBERTa-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	0.00001
Weight Decay	0.00001
Gradient Clipping	5
Training Epochs	90
Warmup Epochs	5
Max Sequence Length	400
Linear Layer Dropout	0.2

Table 13: Hyperparameter details for BERT-QA EAE model.

D.4 DyGIE++

DyGIE++ (Wadden et al., 2019b) is an end-to-end model that simultaneously leverages span graph propagation for EE, entity recognition, and relation extraction tasks. We run our experiments on the end-to-end tasks of DyGIE++ on an NVIDIA RTX A6000 machine with support for 8 GPUs. The model is fine-tuned on XLM-RoBERTa-Large (Conneau et al., 2020). The major hyperparameters are listed in Table 14.

Pre-trained LM	XLM-RoBERTa-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	0.001
Weight Decay	0.001
Gradient Clipping	5
Training Epochs	60
Warmup Epochs	5
Max Sequence Length	250
Linear Layer Dropout	0.4

Table 14: Hyperparameter details for DyGIE++ end-to-end model.

D.5 OneIE

OneIE (Lin et al., 2020b) is an end-to-end model that extracts a globally optimal information network from input sentences to capture interactions among entities, relations, and events. We run our experiments on the end-to-end tasks of OneIE on an NVIDIA RTX A6000 machine with support for 8 GPUs. The models are fine-tuned on XLM-RoBERTa-Large (Conneau et al., 2020). The major hyperparameters are listed in Table 15.

Pre-trained LM	XLM-RoBERTa-Large
Training Batch Size	6
Eval Batch Size	10
Learning Rate	0.001
Weight Decay	0.001
Gradient Clipping	5
Training Epochs	60
Warmup Epochs	5
Max Sequence Length	250
Linear Layer Dropout	0.4

Table 15: Hyperparameter details for OneIE end-to-end model.

D.6 CLaP

CLaP (Parekh et al., 2023a) is a multilingual data-augmentation technique for structured prediction tasks utilizing constrained machine translation for label projection. Specifically, we translate the English SPEED++ into other languages using CLaP. We utilize five in-context examples for each language - Hindi, Japanese, and Spanish - with the original CLaP prompt using the Llama2-13B (Touvron et al., 2023) model. We apply post-processing from SPEED++ to reduce the distribution difference between SPEED++ and the generated multilingual data. We train a separate model for each language as it provided better results than joint training on all language data.

D.7 DivED

DivED (Cai et al., 2024) is trained with the LLaMA-2-7B (Touvron et al., 2023) models on DivED and GENEVA (Parekh et al., 2023b) dataset for zero-shot event detection. They utilize 200 + 90 event types from DivED and Geneva datasets respectively. Training is done using ten event definitions, ten samples, and ten negative samples per sample for each event type while incorporating the ontology information and three hard-negative samples. We utilize their available trained model for our experiments.

D.8 COVIDKB

COVIDKB (Zong et al., 2022) is a simple BERT-classification model trained on a multi-label classification objective on the COVIDKB Twitter corpus. Since our ontology differs from their model, we

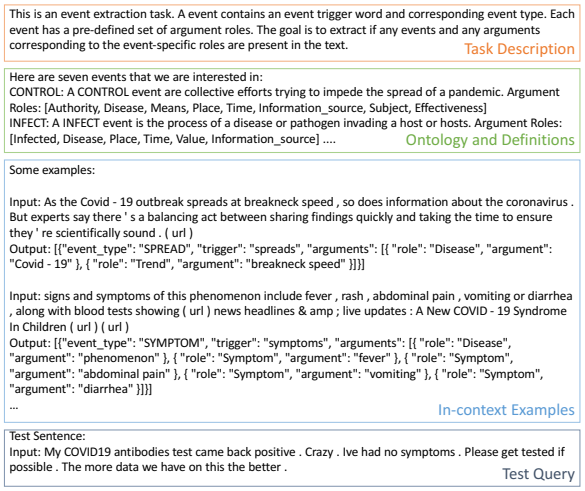

Figure 8: Illustration of the prompt used for GPT-3.5 model. It includes a task description, followed by ontology details of event types and their definitions. Next, we show some in-context examples for each event type and, finally, provide the test sentence.

train it as a binary classification model. We run our experiments on the end-to-end tasks of this model on an NVIDIA RTX A6000 machine with 8 GPUs. The model is fine-tuned on the multilingual BERT (Devlin et al., 2019). The major hyperparameters are listed in Table 16.

Pre-trained LM	mBERT
Training Batch Size	64
Learning Rate	0.00002
Training Epochs	4
Max Sequence Length	128
Number of Classes	2

Table 16: Hyperparameter details for COVIDKB binary classification model.

D.9 Keyword

This model curates a list of keywords specific to each event and predicts a trigger for a particular event if it matches one of the curated event keywords. We utilize the base set of keywords from SPEED (Parekh et al., 2024) for English. We translate these English event-specific keywords for other languages. Recent works have developed advanced keyword extraction techniques (Wu et al., 2024). However, we couldn't explore them in the scope of our work.

D.10 GPT-3

We use the GPT-3.5 turbo model as the base GPT model. We illustrate our final prompt template in Figure 8. It majorly comprises a task definition, ontology details, 1 example for each event type along with corresponding arguments, and the final test query. We conducted a looser evaluation for GPT and only matched if the predicted trigger text matched the gold trigger text.

Model	COVID			MPox	Zika + Dengue				Avg
	hi	jp	es	en	en	hi	jp	es	
BASELINE MODELS									
ACE - TagPrime	0	0	0	0	0	0	0	0	0
DivED*	0	0	16	25	32	1	0	7	10
Keyword*	8	11	10	14	12	12	20	8	12
GPT-3.5-turbo*	13	14	20	35	45	12	12	14	21
TRAINED ON SPEED++ (OUR FRAMEWORK)									
TagPrime	50	0	35	62	**59**	**56**	0	26	36
TagPrime + XGear	50	0	35	62	**59**	**56**	0	26	36
BERT-QA	46	0	31	60	57	43	0	22	32
DyGIE++	**51**	0	**36**	62	58	54	0	21	35
OneIE	50	0	35	**63**	58	55	0	23	36
TagPrime + CLaP	45	**27**	**36**	62	**59**	**56**	**27**	**28**	**42**

Table 17: Benchmarking EE models trained on SPEED++ for extracting event information in the cross-lingual cross-disease setting. The evaluation used is Trigger Classification (TC). Here, hi = Hindi, jp = Japanese, es = Spanish, and en = English. *Numbers are higher compared to others as evaluation is done using string matching.

Model	COVID			MPox	Zika + Dengue				Avg
	hi	jp	es	en	en	hi	jp	es	
TagPrime	**55**	**21**	**51**	**58**	**67**	**61**	**25**	**50**	**49**
XGear	29	17	49	54	63	44	19	47	40
BERT-QA	**55**	17	46	53	64	59	7	49	44

Table 18: Benchmarking EAE models trained on SPEED++ for extracting event information in the cross-lingual cross-disease setting. The evaluation used is Argument Classification (AC). Here, hi = Hindi, jp = Japanese, es = Spanish, and en = English.

E Additional Epidemic Event Extraction Experiments

In addition to the performance evaluation of the event and argument classification of SPEED++, we benchmark its event *trigger* classification (**TC**) performance using the same benchmarking settings. More simply, TC is a stricter evaluation metric that computes the F1 score of the *(trigger, event type)* pair. We present our results in Table 17. We continue to observe the strongest overall performance by the **supervised baselines trained on our SPEED++ dataset with TagPrime**. Models trained without CLaP perform poorly with a zero

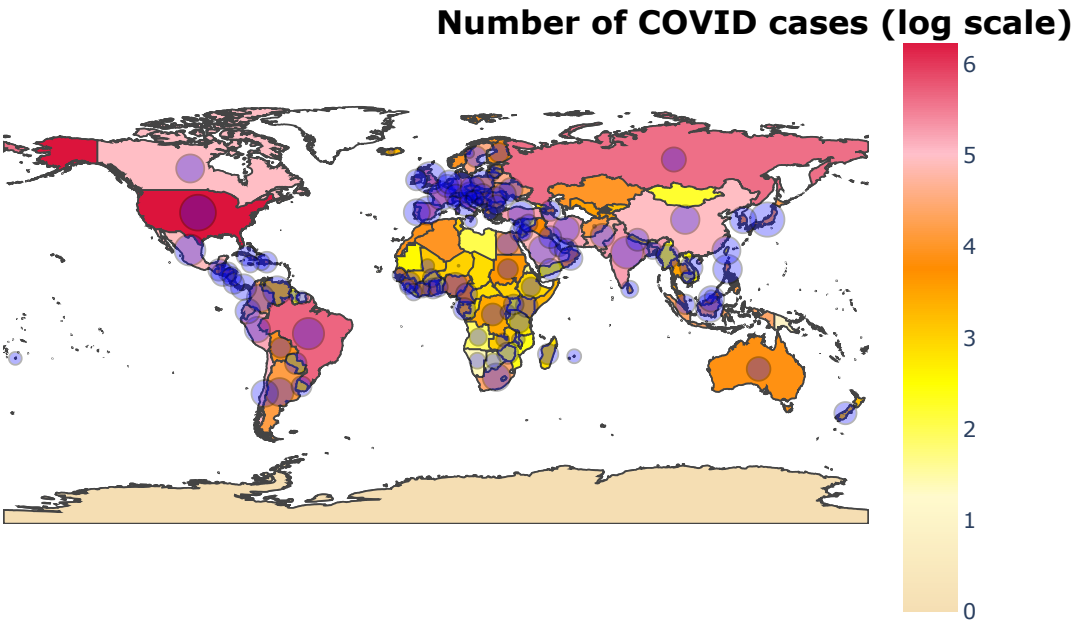

Figure 9: Geographical distribution for the whole world of the number of reported COVID-19 cases as of May 28, 2020. The blue dots indicate the events extracted by our model and the size indicates the number of epidemic events for the specific country (log scale).

F1 score for Japanese. This can be attributed to tokenization difference as Japanese is treated as a character-level language. Since models are trained on English component of SPEED++, it has a strong prior for trigger words to be a single token. But for Japanese, each character is treated as a token and it possesses multi-token triggers. Due to this mismatch, we note that the SPEED++-trained models provide zero scores. This is improved when additional training is done using augmented data from CLaP.

EAE Benchmarking We also benchmark our pure EAE models trained with SPEED++ and present our results in Table 18. These models are provided with the gold event annotations and required to predict all possible arguments corresponding to the gold events. As evident from Table 18, TagPrime model preforms the best across the different families of models across the different languages and diseases.

F Global Epidemic Prediction: Additional Details

To validate the breadth and coverage of our multilingual framework, we utilize it to detect COVID-19 pandemic-related events from social media. Specifically, we focus on all tweets from a single day (chosen at random) - May 28, 2020. We utilize Twitter's Language Identification for sorting the tweets into different languages, resulting in a total of 65 langauges. Next, we map each tweet to a specific location and country, as explained below

Location mapping We utilize the user's location to map each tweet to a specific country. For tweets without location specified, we pool them into a set of unspecified tweets. We estimate the country distribution of tweets per language using the tweets with specified locations. Utilizing this distribution, we extrapolate the locations to the unspecified tweets to approximate the actual location distribution of the tweets. Mapping tweets from the 65 languages results in a country distribution over 117 countries worldwide.

Geographical plotting We consider the 117 countries mapped for these plots. Majorly, we pan these countries out on a map and color them based on the number of COVID-19 cases reported[3] until May 28, 2020. Lighter shades indicate fewer cases, while darker shades indicate massive spread in those countries. We utilize our framework to extract the number of events from the country-mapped tweets and plot them as transculent circles. Bigger dots indicate more events extracted for the given country. We show this geographically for the whole world in Figure 9 and just for Europe in Figure 6.

[3] https://www.worldometers.info/coronavirus

Rank	Clustered Argument	Count
ENGLISH - MONKEYPOX - SYMPTOMS		
1	rash	818
2	sick	746
3	lesions	637
4	fever	548
5	side effect	484
6	itching	441
7	rashes	412
8	cough	175
ENGLISH - ZIKA - SYMPTOMS		
1	birth defects	2.2K
2	brain damage	1.1K
3	microcephaly	990
4	health problems	723
5	nerve disorder	705
6	congenital syndrome	391
7	nerve damage	382
8	damages placenta	196
ENGLISH - DENGUE - SYMPTOMS		
1	fever	4.8K
2	multiple organ failure	1.1K
3	shock syndrome	522
4	symptoms	326
5	high fever	292
6	severe disease	256
7	disease	250
8	rashes	200

Table 19: Aggregated information about symptoms for Monkeypox, Zika, and Dengue from English tweets using our SPEED++ framework.

Rank	Clustered Argument	Count
ENGLISH - COVID-19 - SYMPTOMS		
1	can't breathe	8.8K
2	pneumonia	6.7K
3	sick	4.2K
4	hemorrhaging	2.9K
5	prevents me from staying home	2.7K
6	cough	2.1K
7	symptoms	1.9K
8	critically ill	1.2K
ENGLISH - COVID-19 - CURE MEASURES		
1	hydroxychloroquine	3.7K
2	remdesivir	2.5K
3	drug	2.1K
4	treatment	1.7K
5	hcq	1.3K
6	vaccine	485
7	zinc	448
8	lockdown	425
ENGLISH - COVID-19 - CONTROL MEASURES		
1	lockdown	187K
2	quarantine	56K
3	social distancing	38K
4	deny entry	28K
5	response	21K
6	title 32 orders	15K
7	masks	15K
8	executive order	10K

Table 20: Aggregated information about various arguments for COVID-19 from English tweets using our SPEED++ framework.

From the world map, we note how many of the red countries (United States of America, India, Brazil) have large dots associated with them - indicating more events found for countries where the spread of the disease was high. Similarly, countries where the spread was lesser correspondingly have smaller dots - indicating lesser epidemic events found for these countries. We observe a large cluster for Europe and thus plot it separately, shown in Figure 6 and discussed in § 5.1.

G Epidemic Information Aggregation: Additional Details

In § 5.2, we discussed how we utilize the EAE capability of our trained TagPrime model for creating an information aggregation bulletin. Here we specify more details about this process. First, we utilize our EE framework to extract all possible arguments for the event-specific roles. Since many similar arguments can be extracted, we merge them together by clustering. To this end, we project the arguments into a higher-dimensional embedding space using a Sentence Transformer[4] (Reimers and Gurevych, 2019) encoding model. Next, we utilize a hierarchical agglomerative clustering (HAC) algorithm to merge similar arguments. We implement the clustering using sklearn[5] utilizing euclidean distance as the distance metric and a threshold of 1 as the stopping criteria. After generating the clusters, we rank the clusters by the occurrence count of all arguments in the cluster and label them based on the most frequent argument.

We report the top-ranked clustered arguments for several event roles for COVID-19 from English tweets in Table 20. We report similar tables for different diseases from English tweets in Table 19 and COVID-19 from multilingual tweets in Table 21. Despite some irrelevant extractions owing to model inaccuracies, most of these top clustered arguments are relevant and reflect the language and disease-specific properties quite accurately.

[4] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

[5] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Rank	Argument	Translation	Count
HINDI - COVID-19 - CURE MEASURES			
1	इलाज	treatment	1.1K
2	होम आइसोलेशन	home isolation	1K
3	योग	yoga	636
4	स्वास्थ्य लाभ	recover	500
5	गौमूत्र के कुल्ले	cow urine rinse	448
6	आपके आशीर्वाद	your blessings	240
7	डिस्चार्ज	discharge	126
8	दवाओं	medicines	120
SPANISH - COVID-19 - CURE MEASURES			
1	hidroxicloroquina	hydroxychloroquine	583
2	leche materna	breastmilk	427
3	medicamentos	medicines	252
4	tratamientos	treatments	226
5	red integrada covid	covid integrated network	214
6	ivermectina	ivermectin	157
7	remdesivir	remdesivir	152
8	transplante	transplant	132

Table 21: Aggregated information about various arguments for Hindi and Spanish for COVID-19 using our SPEED++ framework.

Example tweets We also provide qualitative example tweets mentioning some of these arguments to prove the efficacy of our EAE framework. Table 22 presents various English tweets for COVID-19 related mentions. Table 23 presents various English tweets for other diseases of Monkeypox, Zika, and Dengue with their mentions. Table 7 presents various Hindi and Spanish tweets for COVID-19-related mentions. Through this table, we see the diverse set of tweets and how our framework can extract these arguments across them.

Tweet
ENGLISH - COVID-19 - SYMPTOMS
My mum has pneumonia and it might be because of corona, praying for her man
Autopsies of African Americans who died of #(COVID) in New Orleans reveal hemorrhaging ...
Apply for a test if you have symptoms of #(coronavirus): a high temperature, a new continuous cough, loss or change to your sense of smell or taste
ENGLISH - COVID-19 - CURE MEASURES
Trump reveals he's taking hydroxychloroquine in effort to prevent and cure coronavirus symptoms
In this new Covid-19 audio interview, editors discuss newly published studies of remdesivir that highlight its potential and its problems
Hydroxychloroquine combined with zinc has shown effective in treating covid-19
ENGLISH - COVID-19 - CONTROL MEASURES
We should not return to school, we should not undo any other aspect of the lockdown until the test, trace and isolation policy is fully in place
Boris Johnson says from Monday, up to six people will be allowed to meet outside subject to social distancing rules in England
Masks work. Everyone has to wear a mask when in any business ...

Table 22: Illustration of actual tweets in English mentioning various symptoms, cure measures, and control measures related to COVID-19. The terms extracted by our system are highlighted in red.

An Event is defined as something happens in a sentence. In this task, we are trying to identity whether one or more of the following events exist in a given string: *infect, spread, symptom,prevent,control, cure, and death*. And if an event exist, what is the major **triggering word** that mostly manifest its occurrence.

Event	Definition
infect	The process of a disease/pathogen invading host(s).
spread	The process of a disease spreading/pervailing massively at a large scale.
symptom	Individuals displaying physiological features indicating the abnormality of organisms.
prevent	Individuals trying to prevent the infection of a disease.
control	Collective efforts trying to impede the spread of a pandemic.
cure	Stopping infection and relieving individuals from infections/symptoms.
death	End of life of individuals due to infectious disease.

If there exist any explicit negation of an Event, we say that Event does NOT exist and do not mark it.

Important Notes:

There can be sentences without any events. No need to annotate anything for such sentences.

A trigger word can be linked to one or more events. Choose all possible events in such cases.

Multiple events can be presented in a given sentence. Mark all such events.

The same event can occur multiple times (at different parts) in the same sentence. Mark all occurrences of the event.

You will be able to submit the HIT at the last sentence once you finish annotating all the sentences.

Select "flag" event if you see multiple triggering words or any other tricking situations that needs revisiting, but do not abuse this function.

Figure 10: Guidelines for ED annotations for the SPEED++ dataset.

Rule of thumb:

1. Marking arguments acrossing sentences is allowed (if necessary), but try to localize the arguments as close to the trigger as possible in the first place.
2. Entity arguments are usually short words or concise phrases, while non entity arguments, such as effectiveness, may be longer.
3. Don't be restricted to identifiying the most synthesized word as we did in ED; Instead. capture as many meaningful arguments as possible.
4. If there are multiple candidates for a single argument role, instead of marking "A, B, and C" as one argument, map "A", "B" and "C" to that argument.
5. If a tweet contains multiple events, try to localize the arguments instead of sharing them in the first place.

Specific emphasis:

1. Do not mark too implicit mentions.
eg. "XXX institution of America advices that people should … during the pandemic." --> "XXX institution of America" is the authority of the control event but we don't additionally mark "America" as place argument.
2. information-source need not to be authoritative but must be meaningful.
eg. Do not mark "**I** think ..." or "**I** feel ..." as information-source; "**My brother** tells ..." or "**YY News** states..." are both valid information-source.
3. Many entity arguments need not to be numerical.
eg. value arguments can be "**80%** of the population","**a lot of** people","**58** students"...
eg. time arguents can be "**yesterday**","until **May 22rd**","**before the end of next week**"...
eg. symptoms arguments can be "**cough**", "**really uncomfortable**","**sickness**"...

Figure 11: Guidelines for EAE annotation for the SPEED++ dataset.

Here are the events and corresponding arguments we are interested in for each event. The question following helps you to understand their meaning.			
infect	**DEFINITION**: The process of a disease/pathogen invading host (s).	spread	**DEFINITION**: The process of a disease spreading/pervailing massively at a large scale.
Infected	who are infected?	**Population**	among who?
Disease	what infects the infected?	**Disease**	What is spreading?
Place	where it happened?	**Place**	where it happened?
Time	when it happened?	**Time**	when it happened?
Value	how many are infected?	**Value**	how many are infected?
Information-source	How do the sentence know about the infections?	**Information-source**	How do the sentence know about the spread?
		Trend	How is it going?
symptom	**DEFINITION**: Individuals displaying physiological features indicating the abnormality of organisms.	death	**DEFINITION**: End of life of individuals due to infectious disease.
Person	who has symptoms?	**Dead**	who dies?
Symptom	what symptoms?	**Disease**	from what?
Disease	what may cause it?	**Place**	where it happened?
Place	where it happened?	**Time**	when it happened?
Time	when it happened?	**Value**	how many are dead?
Duration	How long symptoms lasts?	**Information-source**	How do the sentence know about the death?
Information-source	How do the sentence know about this?	**Trend**	How is the death count changing?
prevent	**DEFINITION**: Individuals trying to prevent the infection of a disease.	control	**DEFINITION**: Collective efforts trying to impede the spread of a pandemic.
Agent	who is taking action to prevent?	**Authority**	who is imposing control measures?
Disease	what is prevented against?	**Disease**	what is controled against?
Means	How to prevent?	**Means**	How to control?
Target	who is protected?	**Place**	where it happened?
Information-source	How do the sentence know about the prevention?	**Time**	when it happened?
Effectiveness	Is the means working?	**Information-source**	How do the sentence know about the control?
		Subject	who should impletement the measures?
		Effectiveness	is the means working?
cure	**DEFINITION**: Stopping infection and relieving individuals from infections/symptoms.		
Cured	who recovers?		
Disease	recovered from what?		
Means	by what means?		
Place	where it happened?		
Time	when it happened?		
Value	how many are cured?		
Facility	who helps to cure?		
Information-source	How do the sentence know about the recovery?		
Effectiveness	is the means working?		
Duration	how long it takes to recover?		

Figure 12: Argument definitions provided as part of the EAE annotation process.

An Event is defined as something happens in a sentence. Each event may has several arguments, which contain more specific information we are interested in. In this task, we are trying to identify whether one or more of the following events exist in a given string: *infect, spread,symptom,prevent,control, cure*, and *death* as well as some of their arguments.

If an event exists, there will be a major **triggering word** that mostly manifests its occurrence.
We then search for arguments related to this event from the sentence (which may or may not exist).

For each sentence, **we have already marked the Event (if any) and its triggering word, as well as their arguments (if any)**. Your task is to look through these annotations and check "incorrect" if you disagree with them and write down your opinion.

Generally, do the three things for each sentence you receive:
1. Look through each event and their arguments, if you agree with everything, move on to the next question.
2. If you think any annotations are incorrect, clicked the "incorrect" box and write down your opinion in the box below.
3. If you think current event should not exist or other events should exist in this sentence, clicked the "incorrect" box and write down your opinion in the box below.

While using your best knowledge to judge does each event and each argument make sense, here are some rules specific to this task:

1. If there is explicit negation of an event, that event should not exist.
In "This vaccine cannot cure the illness", event cure does not exist.

2. Arguments are not required to be rigorous or specific.
"Recently", "at home", "a lot of" are valid time, place, value arguments respectively.

3. One sentence can have multiple events. All annotations will be presented to you at once.
In "I wear mask to **protect** myself from flu but I still **got** it", both EVENT prevent and Event infect exist.

4. Each event is not required to have all the arguments; arguments not present are left blank.

5. Multiple phrases may share the same argument role.
In "Amy and Bob are **infected** by the virus", "Amy" and "Bob" are both marked as ARGUMENT [infected] of EVENT [infect].

Notes:
1. Your answers (both checkings and writings) are automatically saved. If you navigated to previous pages and do not see your previous answers, they are already saved and you DON'T need to retype.
2. Once you finish all five examples, please scroll down and click "Submission".

Here are the events we are looking for and marking in the sentences:

Events	Definition
infect	The process of a disease/pathogen invading host(s).
spread	The process of a disease spreading/ prevailing massively at a large scale.
symptom	Individuals displaying physiological features indicating the abnormality of organisms.
prevent	Individuals trying to prevent the infection of a disease.
control	Collective efforts trying to impede the spread of a pandemic.
cure	Stopping infection and relieving individuals from infections/symptoms.
death	End of life of individuals due to infectious disease.

Figure 13: Intructions provided for the multilingual verification task.

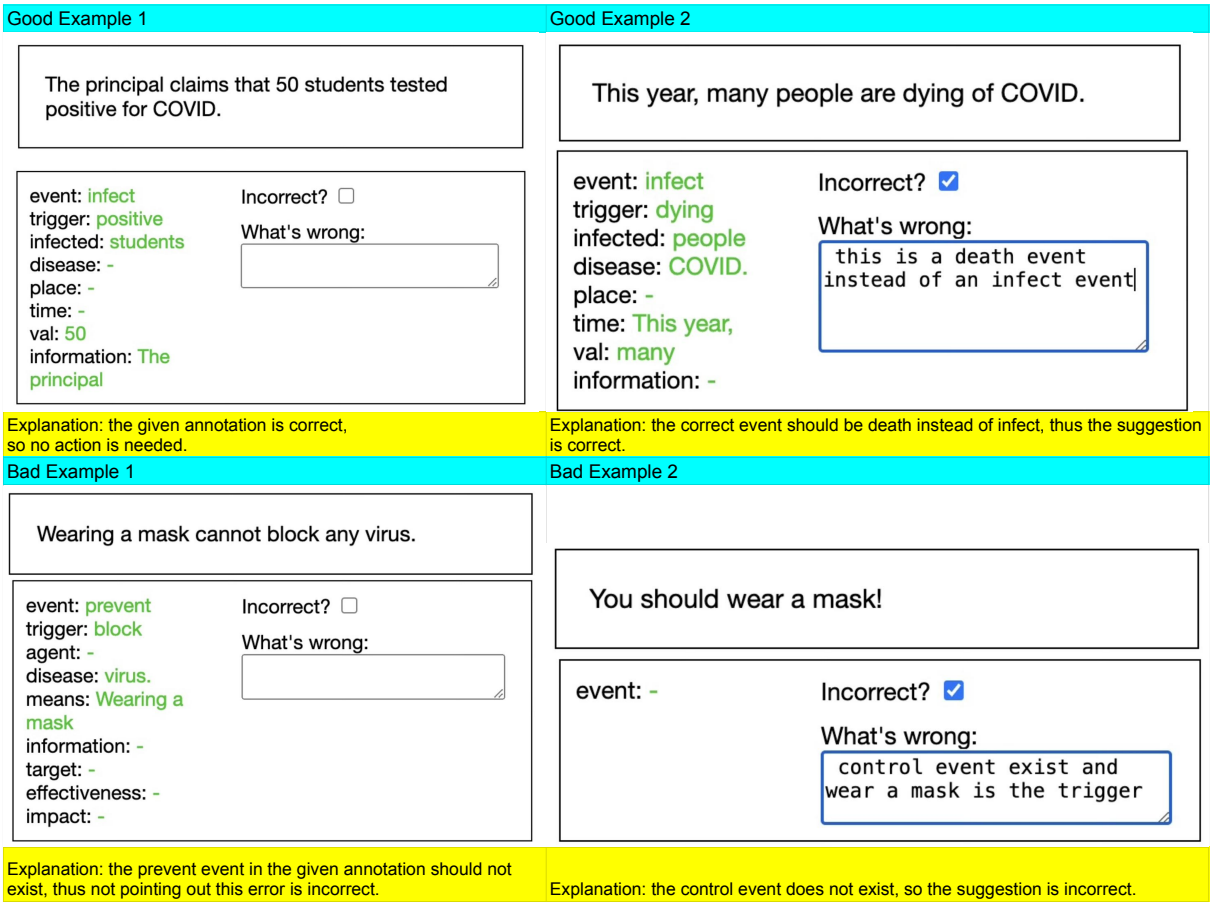

Figure 14: Illustrations provided for the multilingual verification task.

Tweet
ENGLISH - MONKEYPOX - SYMPTOMS
I have a pretty mild rash on my stomach. A little bit of itchy. The extremely optimistic part of my brain is like Ẅhat if it's monkey pox?
Anyone can get #(monkey pox) through close skin-to-skin contact ... Healthcare providers must be vigilant and test any patient with a suspicious lesion or sore.
Its inappropriate to say but the amount of itching Ive done from the bites makes me nervous people are gonna think Ive got monkey pox or some shit
ENGLISH - ZIKA - SYMPTOMS
More birth defects seen in (url) areas where Zika was present (url)
Zika brain damage may go undetected in pregnancy
study sheds light on how Zika causes nerve disorder (url)
ENGLISH - DENGUE - SYMPTOMS
In the evening the fever is skyrocketing & the joint pain is born-breaking & nauseating, vomiting is constant
Dengvaxia = yellow fever vaccine + live attenuated dengue virus. Multiple organ failure was already established as its key symptom
TMI but I've had rashes on my arms and legs for a couple of days now. Tried to tell Scott I have dengue fever but ...

Table 23: Illustration of actual tweets in English mentioning various symptoms related to Monkeypox, Zika, and Dengue. The terms extracted by our system are highlighted in red.

Arguments of INFECT event		
Argument Name	**Argument Definition**	**Example**
Infected	The individual(s) being infected	300 people tested positive.
Disease	The disease or virus that invaded the host	I tested positive for COVID.
Place	The place that the individual(s) are infected	5 students at school are infected.
Time	The time that the individual(s) are infected	300 people tested positive on May 15.
Value	The number of people being infected	Some people are infected.
Information-source	The source that is providing this information regarding to infection	According to CDC, if you have COVID...

Table 24: Complete definition and examples of arguments of INFECT event

Arguments of SPREAD event		
Argument Name	**Argument Definition**	**Example**
Population	The population among which the disease spreads	16000 Americans are infected.
Disease	The disease/virus/pandemic that is prevailing	Monkey-box is spreading ...
Place	The place at which the disease is spreading	the flu prevails in the U.S..
Time	The time during which the disease is spreading	the flu prevails in the U.S. in winter.
Value	The number of people being infected	16000 Americans are infected.
Information-source	The source that is providing this information regarding to the transmission of the disease	My mom says COVID is spreading again.
Trend	The possible change of a transmission of a disease with respect to past status	COVID spreads faster than we've expected.

Table 25: Complete definition and examples of arguments of SPREAD event

Arguments of SYMPTOM event		
Argument Name	**Argument Definition**	**Example**
Person	The individual(s) displaying symptoms	I'm coughing now.
Symptom	The concrete symptom(s) that are displayed	You may have severe fever and stomach-ache.
Disease	The disease(s)/virus that are potentially causing the symptoms	If you cough, that's probably COVID.
Place	The place at which the symptom(s) are displayed	Students are showing illness at school.
Time	The time during which the symptom(s) are displayed	I feel sick yesterday.
Duration	The time interval that the symptom(s) last	My fever lasts three days.
Information-source	The source that is providing this information regarding to symptoms of the disease	He said half of his class were ill.

Table 26: Complete definition and examples of arguments of SYMPTOM event

Arguments of PREVENT event		
Argument Name	**Argument Definition**	**Example**
Agent	The individual(s) attempting to avoid infections	You should wear a mask to protect yourself and others.
Disease	The disease/virus/illness being defensed against	prevent COVID infection.
Means	actions/means that may prevent infection	You should wear a mask to protect yourself and others.
Information-source	The source that is providing this information regarding to the prevention of this disease	CDC proves masks can efficiently blocks virus.
Target	The individual(s)/population to which the agent attempts to prevent the disease transmission	You should wear a mask to protect yourself and others.
Effectiveness	How effective is the means against the disease	CDC proves masks can efficiently blocks virus.

Table 27: Complete definition and examples of arguments of PREVENT event

Arguments of CONTROL event		
Argument Name	**Argument Definition**	**Example**
Authority	The authority implementing/advocating the control of a pandemic	To impede COVID transmission, Chinese government required quarantine upon arrival.
Disease	The intruding disease/virus/pandemic being defensed against	To impede COVID transmission...
Means	the enacted/advocated policies/actions that may control the pandemic	To impede COVID transmission, Chinese government required quarantine upon arrival.
Information-source	The source that is providing this information regarding to the control of this disease	CNN reports massive pandemic lockdowns in China.
Place	The individual(s)/population to which the agent attempts to prevent the disease transmission	NY will enforces mask policy from June.
Time	The individual(s)/population to which the agent attempts to prevent the disease transmission	NY will enforces mask policy from June.
Effectiveness	How effective is the means against the disease	The infection rate does not decrease since the enforcement of mask policy.
Subject	The individual(s)/population encouraged/ordered to implement the control measures	Due to the pandemic, students are required to wear masks in class.

Table 28: Complete definition and examples of arguments of CONTROL event

Arguments of CURE event		
Argument Name	**Argument Definition**	**Example**
Cured	The individuals(s) recovered/receiving the treatments	My grandma recovered from COVID yesterday.
Disease	The disease/illness that the patients get rid of	My grandma recovered from COVID yesterday.
Means	The therapy that (potentially) treat the disease	Just get rest and your fever will go away.
Information-source	The source that is providing this information regarding to the cure/recovery of this disease	CNN reports that XX company claimed to developed COVID treatment.
Place	The place at which the recovery takes place	In the U.S., 15670 people recovered and 16000 died of COVID.
Time	The time at which the recovery takes place	By May 15, 15670 Americans recovered and 16000 died of COVID.
Effectiveness	How effective is the means against the disease	The new COVID treatment is not fully effective.
Value	The number of people being cured	By May 15, 15670 Americans recovered and 16000 died of COVID.
Facility	The individual(s)/organization(s) utilizing/inventing certain means to facilitate recoveries	CNN reports that XX company claimed to developed COVID treatment.
Duration	The time interval that the treatment takes	I received the treatment for two month before full recovery.

Table 29: Complete definition and examples of arguments of CURE event

Arguments of DEATH event		
Argument Name	**Argument Definition**	**Example**
Dead	The individuals(s) who die of infectious disease	By March, 500 people died of COVID in CA.
Disease	The disease/virus/pandemic that (potentially) causes the death	By March, 500 people died of the virus in CA.
Information-source	The source that is providing this information regarding to fatality of this disease	Daily news: new death of COVID ...
Place	The place at which the death takes place	By March, 500 people died of COVID in CA.
Time	The time at which the death takes place	By March, 500 people died of COVID in CA.
Value	The number of death due to infectious disease	By March, 500 people died of COVID in CA.
Trend	The possible change of death counts caused by disease infection compared to the statistics from the past.	The COVID death toll is still increasing...

Table 30: Complete definition and examples of arguments of DEATH event

Event	English	Hindi	Spanish	Japanese
Infect	I caught the virus earlier today	मैं आज सुबह वायरस से बीमार हो गया हूं	Contraje el virus temprano hoy	今日ウイルスに感染した
	My brother tested positive for COVID-19	मेरे भाई का COVID-19 टेस्ट पॉजिटिव आया	Mi hermano dio positivo por COVID-19.	私の兄はCOVID-19陽性でした
Spread	The COVID-19 outbreak put WHO in alert that the pandemic may develop into global scale	COVID-19 के प्रकोप ने WHO को सतर्क कर दिया है कि यह महामारी वैश्विक स्तर पर विकसित हो सकती	El brote Covid-19 alertó al OMS que la pandemia puede alcanzar en escala global	COVID-19の発生により、WHOはパンデミックが世界的な規模に発展する可能性に警戒を強めている
	A new flu is sweeping across Los Angeles	लॉस एंजिल्स में एक नया फ्लू फैल रहा है	Una nueva gripe está propagando a traves de Los Ángeles	ロサンゼルスで新型インフルエンザが流行
Symptom	Many of my friends have a cold	मेरे कैइ दोस्तों को सर्दी है	Muchos de mis amigos tienen un resfriado	私の多くの友人が風邪をひいた
	I became incredibly ill after catching the virus	वायरस की चपेट में आने के बाद मैं अविश्वसनीय रूप से बीमार हो गया	Me enfermé increíblemente después de contagiarme de el virus	ウイルスに感染した後、私はすごく体調を崩した
Prevent	Medical experts encourage young kids to wash their hands	चिकित्सा विशेषज्ञ छोटे बच्चों को हाथ धोने के लिए प्रोत्साहित करते हैं	Los expertos médicos alientan a los niños a lavarse las manos	医療専門家が幼児に手洗いを奨励
	Wear a mask to protect your family from the disease	अपने परिवार को बीमारी से बचाने के लिए मास्क पहनें	Use una máscara para proteger a su familia de la enfermedad	家族を疾病から守るためにマスクを着用すること
Control	The WHO has published new guidelines in response to the rising cases of COVID-19	WHO ने COVID-19 के बढ़ते मामलों के जवाब में नए दिशानिर्देश प्रकाशित किए हैं	La OMS ha publicado nuevas pautas en respuesta a los casos crecientes de Covid-19	WHOはCOVID-19の感染者増加を受けて新しいガイドラインを発表した。
	Government officials have imposed a lockdown on certain districts	सरकारी अधिकारियों ने कुछ जिलों में लॉकडाउन लगा दिया है	Los funcionarios gubernamentales han impuesto un aislamiento a ciertos distritos	政府当局が特定の地区にロックダウンを課した
Cure	There is no magic cure for the pandemic	अभी तक कोविड का कोई प्रभावी इलाज नहीं	No existe una cura mágica para la pandemia	パンデミックに特効薬はない
	Unfortunately doctors were unable to save him from the pandemic	दुर्भाग्य से डॉक्टर उसे महामारी से बचाने में असमर्थ थे	Desfortunadamente, los médicos no pudieron salvarlo de la pandemia	残念ながら、医師たちは彼をパンデミックから救うことはできなかった。
Death	700 people killed by COVID	कोविड से 700 लोगों की मौत	700 personas matadas por Covid	COVIDによる死亡率：７００人
	The mortality rate of the pandemic has decreased as experts figure out how to treat it	महामारी की मृत्यु दर में कमी आई है क्योंकि विशेषज्ञ यह पता लगा रहे हैं कि इसका इलाज कैसे किया जाए	La tasa de mortalidad de la pandemia ha disminuido a medida porque los expertos están descubriendo cómo tratarla	パンデミックの死亡率は、専門家による治療法の解明のために減少している。

Table 31: Sample translated seed tweets for the different event types in our ontology for the different languages. Triggers are highlighted in red.