









Multilingual Epidemic Event Extraction

Stephen Mutuvi^{1,2}^(✉), Emanuela Boros¹, Antoine Doucet¹,
Gaël Lejeune³, Adam Jatowt⁴, and Moses Odeo²

¹ L3i, University of La Rochelle, 17000 La Rochelle, France
{[steve.mutuvi](mailto:steve.mutuvi@univ-lr.fr), [emanuela.boros](mailto:emanuela.boros@univ-lr.fr), [antoine.doucet](mailto:antoine.doucet@univ-lr.fr)}@univ-lr.fr

² Multimedia University of Kenya, Nairobi, Kenya
{[smutuvi](mailto:smutuvi@mmu.ac.ke), [modeo](mailto:modeo@mmu.ac.ke)}@mmu.ac.ke

³ STIH Lab, Sorbonne University, 75006 Paris, France
gael.lejeune@paris-sorbonne.fr

⁴ University of Innsbruck, 6020 Innsbruck, Austria
adam.jatowt@uibk.ac.at

Abstract. In this paper, we focus on epidemic event extraction in multilingual and low-resource settings. The task of extracting epidemic events is defined as the detection of disease names and locations in a document. We experiment with a multilingual dataset comprising news articles from the medical domain with diverse morphological structures (Chinese, English, French, Greek, Polish, and Russian). We investigate various Transformer-based models, also adopting a two-stage strategy, first finding the documents that contain events and then performing event extraction. Our results show that error propagation to the downstream task was higher than expected. We also perform an in-depth analysis of the results, concluding that different entity characteristics can influence the performance. Moreover, we perform several preliminary experiments for the low-resourced languages present in the dataset using the mean teacher semi-supervised technique. Our findings show the potential of pre-trained language models benefiting from the incorporation of unannotated data in the training process.

Keywords: Epidemiological surveillance · Multilingualism · Semi-supervised learning

1 Introduction

The ability to detect disease outbreaks early enough is critical in the deployment of measures to limit their spread and it directly impacts the work of health authorities and epidemiologists throughout the world. While disease surveillance has in the past been a critical component in epidemiology, conventional surveillance methods are limited in terms of both promptness and coverage, while at

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia). It has also been supported by the French Embassy in Kenya and the French Foreign Ministry.

the same time requiring labor-intensive human input. Often, they rely on information and data from past disease outbreaks, which more often than not, is insufficient to train robust models for extraction of epidemic events.

Epidemic event extraction from archival texts, such as digitized news reports, has also been applied for constructing datasets and libraries dedicated to tracking and understanding epidemic spreads in the past. Such libraries leverage the technology advantage of digital libraries by storing, processing, and disseminating data about infectious disease outbreaks. The work presented by Casey et al. [5] is an example of such an initiative aiming at analyzing outbreak records of the third plague pandemic in the period 1894 to 1952 in order to digitally map epidemiological concepts and themes related to the pandemic. Although the authors used semi-automatic approaches in their work, the discovery of documents related to epidemic outbreaks was done manually and the entity extraction was largely performed through manual annotation or the use of gazetteers, which have their own limitations. Other works devoted to the studies of past epidemics (e.g., the analysis of bubonic plague outbreak in Glasgow (1900) [10]) fully rely on manual efforts for data collection and preprocessing. We believe that automatic approaches to epidemic information extraction could also enhance this kind of scientific study.

The field of research focusing on data-driven disease surveillance, which has been shown to complement traditional surveillance methods, remains active [1, 7]. This is majorly motivated by the increase in the number of online data sources such as online news text [14]. Online news data contains critical information about emerging health threats such as what happened, where and when it happened, and to whom it happened [35]. When processed into a structured and more meaningful form, the information can foster early detection of disease outbreaks, a critical aspect of epidemic surveillance. News reports on epidemics often originate from different parts of the world and events are likely to be reported in other languages than English. Hence, efficient multilingual approaches are necessary for effective epidemic surveillance [4, 27].

Moreover, the large amounts of continuously generated unstructured data, for instance, in the ongoing COVID-19 epidemic, are often challenging and difficult to process by humans without leveraging computational techniques. With the advancements in natural language processing (NLP) techniques, processing such data and applying data-driven methods for epidemic surveillance has become feasible [2, 30, 40]. Although promising, the scarcity of available annotated corpora for data-driven epidemic surveillance is a major hindrance. Obtaining large-scale human annotations is a time-consuming and labor-intensive task. The challenge is more pronounced when dealing with neural network-based methods [18], where massive amounts of labeled data play a critical role in reducing generalization error.

Another specific challenge for the extraction of epidemic events from news text is class imbalance [19]. The imbalance exists between the disease and location entities, which when paired characterize an epidemic event. The large difference in the number of instances from different classes can negatively impact the

performance of the extraction models. Another challenge relates to data sparsity where some languages in the multilingual setup have few annotated data [15], barely sufficient to train models that achieve satisfactory performance.

In this study, we use a multilingual dataset comprising news articles from the medical domain with diverse morphological structures (Chinese, English, French, Greek, Polish, and Russian). In this dataset, an epidemic event is characterized by the references to a disease name and the reported locations that are relevant to the disease outbreak. We evaluate a specialized baseline system and experiment with the most recent Transformer-based sequence labeling architectures. Additionally, error propagation from the classification task that affects the event extraction task is also evaluated since the event extraction task is a multi-step task, comprising various sub-tasks [13, 22, 30]. The classification task filters the epidemic-related documents from the large collection of online news articles, prior to the event extraction phase. We also perform a detailed analysis of various attributes (sentence length, token frequency, entity consistency among others) of the data and their impact on the performance of the systems.

Thus, considering the aforementioned challenges, our contributions are the following:

- We establish new performance scores after the evaluation of several pre-trained and fine-tuned Transformer-based models on the multilingual data and by comparing with a specialized multilingual news surveillance system;
- We perform a generalized, fine-grained analysis of our models with regards to the results on the multilingual epidemic dataset. This enables us to more comprehensively assess the proposed models, highlighting the strengths and weaknesses of each model;
- We show that semi-supervised learning is beneficial to the task of epidemic event extraction in low-resource settings by simulating different few-shot learning scenarios and applying self-training.

The remainder of this paper is organized as follows. Section 2 describes the related work. Section 3 presents the multilingual dataset utilized in our study. In Sect. 4, we discuss our experimental methodology and empirical results. Finally, Sect. 5 concludes this paper and provides suggestions for future research.

2 Related Work

Several works tackled the detection of events related to epidemic diseases. Some approaches include external resources and features at a sub-word representation level. For example, the Data Analysis for Information Extraction in any Language (DAnIEL) system was proposed as a multilingual news surveillance system that leverages repetition and saliency (salient zones in the structure of a news article), properties that are common in news writing [26]. By avoiding the usage of language-specific NLP toolkits (e.g., part-of-speech taggers, dependency parsers) and by focusing on the general structure of the journalistic writing

genre [21], the system is able to detect key event information from news articles in multilingual corpora. We consider it as a baseline multilingual model.

Models based on neural network architectures and which take advantage of the word embeddings representations have been used in monitoring social media content for health events [25]. Word embeddings capture semantic properties of words, and thus the authors use them to compute the distances between relevant concepts for completing the task of flu event detection from text. Another type of approach is based on long short-term memory (LSTM) [47] models that approach the epidemic detection task from the perspective of classification of tweets to extract influenza-related information.

However, these approaches, especially the recent deep learning methods such as the Transformer-based model [44], remain largely unexplored in the context of epidemic surveillance using multilingual online news text. Transformer language models can learn powerful textual representations, thus they have been effective across a wide variety of NLP downstream tasks [3, 9, 23, 46].

Despite the models requiring a large amount of data to train, annotated resources are generally scarce, especially in digital humanities [34, 39]. Having sufficient data is essential for the performance of event extraction models since it can help reduce overfitting and improve model robustness [12]. To address the challenges associated with scarcity of large-scale labeled data, various methods have been proposed [6, 12, 15].

Among them is semi-supervised learning, where data is only partially labeled. The semi-supervised approaches permit harnessing unlabeled data by incorporating the data into the training process [43, 50]. One type of semi-supervised learning method is self-training, which has been successfully applied to text classification [42], part-of-speech (POS) tagging [48] and named entity recognition (NER) [24, 37]. Semi-supervised learning methods can utilize a teacher-student method where the teacher is trained on labeled data that generates pseudo-labels for the unlabeled data, and the pseudo-labeled examples are iteratively combined with the clean labels by the student [49]. These previous attempts in addressing the problem of limited labeled data have focused on resource-rich languages such as English [6, 12, 15]. In this study, we increase coverage to other languages, and most importantly languages with limited available training data.

3 Dataset

Due to the lack of dedicated datasets for epidemic event extraction from multilingual news articles, we adapt a freely available epidemiological dataset¹, referred to as DANIEL [26]. The corpus was built specifically for the DANIEL system [26, 28], containing articles in six different languages: English, French, Greek, Russian, Chinese, and Polish. However, the dataset is originally annotated at the document level. We annotate the dataset to token-level annotations [31], a

¹ The DANIEL dataset is available at <https://daniel.greyc.fr/public/index.php?a=corpus>.

Table 1. Statistical description of the DANIEL partitions. DIS and LOC stand for the number of disease and location mentions, respectively.

| | Partition | Documents | Sentences | Tokens | Entities | DIS | LOC |
|---------|-----------|-----------|-----------|-----------|----------|-------|-------|
| French | Train | 2,185 | 62,748 | 1,786,077 | 2,677 | 1,438 | 1,239 |
| | Dev | 273 | 7,625 | 231,165 | 337 | 206 | 131 |
| | Test | 273 | 7,408 | 214,418 | 300 | 177 | 123 |
| | Total | 2,731 | 77,781 | 2,231,660 | 3,314 | 1,821 | 1,493 |
| English | Train | 379 | 7,312 | 204,919 | 524 | 319 | 205 |
| | Dev | 48 | 857 | 24,990 | 5 | 3 | 2 |
| | Test | 47 | 921 | 25,290 | 34 | 27 | 7 |
| | Total | 474 | 9,090 | 255,199 | 563 | 349 | 214 |
| Greek | Train | 312 | 4,947 | 151,959 | 259 | 144 | 115 |
| | Dev | 39 | 924 | 23,980 | 15 | 10 | 5 |
| | Test | 39 | 531 | 15,951 | 26 | 12 | 14 |
| | Total | 390 | 6,402 | 191,890 | 300 | 166 | 134 |
| Chinese | Train | 354 | 6,309 | 193,453 | 67 | 57 | 10 |
| | Dev | 44 | 838 | 26,720 | 16 | 14 | 2 |
| | Test | 44 | 624 | 19,767 | 7 | 5 | 2 |
| | Total | 442 | 7,771 | 239,940 | 90 | 76 | 14 |
| Russian | Train | 341 | 5,250 | 112,714 | 258 | 170 | 88 |
| | Dev | 43 | 618 | 14,168 | 30 | 27 | 3 |
| | Test | 42 | 547 | 11,514 | 39 | 27 | 12 |
| | Total | 426 | 6,415 | 138,396 | 327 | 224 | 103 |
| Polish | Train | 281 | 7,288 | 126,696 | 498 | 352 | 146 |
| | Dev | 35 | 954 | 17,165 | 73 | 40 | 33 |
| | Test | 36 | 998 | 17,026 | 67 | 52 | 15 |
| | Total | 352 | 9,240 | 160,887 | 638 | 444 | 194 |

common format utilized in research for the event extraction task. The token-level dataset is made freely and publicly available².

Typically in event extraction, this dataset is characterized by class imbalance. Only around 10% of the documents are relevant to epidemic events, which is very sparse. The number of documents in each language is rather balanced, except for French, having about five times more documents compared to the rest of the languages. More statistics on the corpus can be found in Table 1.

In this dataset, a document generally talks about an epidemiological event and the task of extracting the event comprises the detection of all the occurrences of a disease name and the locations of the reported event, as shown in Fig. 1. The document talks about the ending of a *polio* outbreak in *India*, more exactly

² The token-level annotated dataset is available at <https://bit.ly/3kUQcXD>.

Today marks one year since the last case of **polio** was recorded in **India** when the virus paralysed an 18-month-old girl in **Howrah**, near **Kolkata**. If pending test results return absent of the virus in coming weeks, India will be removed from the list of endemic **polio** countries. But **India** still remains at serious risk of fresh outbreaks if the virus is brought back into the country from overseas, and **polio** experts say the country's massive immunisation regimen must be maintained.

Fig. 1. Excerpt from an English article in the DANIEL dataset that was published on January 13th, 2012 at <http://www.smh.com.au/national/health/polio-is-one-nation-closer-to-being-wiped-out-20120112-1pxho.html>.

in *Howrah* and *Kolkata*. An event extraction system should detect all the *polio* event mentions, along with the aforementioned locations.

4 Experiments

Our experiments are performed in two setups:

1. Supervised learning experiments:
 - Our first experiments focus on the *epidemic event extraction* utilizing the entire dataset.
 - Next, like most approaches for text-based disease surveillance [22], we follow a two-step process by first applying *document classification* into either relevant (documents that contain event mentions) or irrelevant (documents without event mentions) and then performing the *epidemic event extraction* task through the detection and extraction of the disease names and locations from these documents.
2. Semi-supervised learning experiments:
 - For these experiments, we simulate several few-shot scenarios for the low-resourced languages in our dataset, and we apply semi-supervised training with the mean teacher method in order to assess the ability of the models to alleviate the challenge posed by the lack of annotated data.

Models. We evaluate the pre-trained model BERT (Bidirectional Encoder Representations from Transformers) proposed by [11] for token sequential classification³. We decided to use BERT not only because it is easy to fine-tune, but it has also proved to be one of the most performing technologies in multiple NLP tasks [9, 11, 38]. Due to the multilingual characteristic of the dataset, we use the multilingual BERT pre-trained language models and fine-tune them on our epidemic-specific labeled data. We will refer to these models as BERT-multilingual-cased⁴

³ For this model, we used the parameters recommended in [11].

⁴ <https://huggingface.co/bert-base-multilingual-cased>. This model was pre-trained on the top 104 languages having the largest Wikipedia edition using a masked language modeling (MLM) objective.

and BERT-multilingual-uncased⁵. We also experiment with the XLM-RoBERTa-base model [8] that has shown significant performance gains for a wide range of cross-lingual transfer tasks. We consider this model appropriate for our task due to the multilingual nature of our dataset⁶.

Evaluation. The epidemic event extraction evaluation is performed in a coarse-grained manner, with the entity as the reference unit [29]. We compute precision (P), recall (R), and F1-measure (F1) at the micro-level (error types are considered over all documents).

4.1 Supervised Learning Experiments

We chose DANIEL [26] as a baseline model for epidemic event extraction. This is an unsupervised method that consists of a complete pipeline that first detects the relevant documents and then extracts the event triggers. The system considers text as a sequence of strings and does not depend on language-specific grammar analysis, hence can easily be adapted to a variety of languages. This is an important attribute of epidemic extraction systems for online news text, as the text is often heterogeneous in nature. Figure 2 presents the full procedure for the supervised learning experiments.

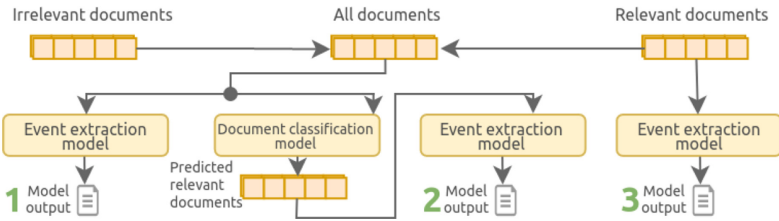


Fig. 2. Illustration of the types of experiments carried out: (1) using all data instances (*relevant and irrelevant documents*), (2) testing on the *predicted relevant documents* provided by the document classification step, (3) using only the *ground-truth relevant documents*.

For document classification, we chose the fine-tuned BERT-multilingual-uncased [11, 30] whose performance on text classification is a F1 of 86.25%. The performance in F1 with regards to the relevant documents per language is 28.57% (Russian), 87.10% (French), 50% (English), 100% (Polish), and 50% (Greek). One drawback of this method is the fact the none of the Chinese relevant documents was found by the classification model, and thus, none of the events will be further detected.

⁵ <https://huggingface.co/bert-base-multilingual-uncased>. This model was pre-trained on the top 102 languages having the largest Wikipedia editions using a masked language modeling (MLM) objective.

⁶ XLM-RoBERTa-base was trained on 2.5 TB of newly created clean CommonCrawl data in 100 languages.

Holistic Analysis. We now present the results of the evaluated models, namely the DAnIEL system and the Transformer-based models. We first observe in Table 2 that all the models significantly outperform our baseline, DAnIEL. As it can be seen in Table 2, under *relevant and irrelevant documents* (1), when the models are trained on the entire dataset (i.e., the relevant and irrelevant documents), the BERT-multilingual-uncased model recorded the highest scores, with a very small margin when compared to the other two fine-tuned models, the cased BERT and XLM-RoBERTa-base.

Table 2. Evaluation results for the detection of disease names and locations on all languages and all data instances (relevant and irrelevant documents).

| Models | P | R | F1 |
|---------------------------------------|--------------|--------------|--------------|
| DAnIEL Baseline | 38.97 | 47.32 | 42.74 |
| Relevant and irrelevant documents (1) | | | |
| BERT-multilingual-cased | 80.66 | 79.72 | 80.19 |
| BERT-multilingual-uncased | 82.25 | 79.77 | 80.99 |
| XLM-RoBERTa-base | 82.41 | 76.81 | 79.52 |
| Predicted relevant documents (2) | | | |
| BERT-multilingual-cased | 52.13 | 89.43 | 65.87 |
| BERT-multilingual-uncased | 53.66 | 92.28 | 67.86 |
| XLM-RoBERTa-base | 53.10 | 90.65 | 66.97 |
| Ground-truth relevant documents (3) | | | |
| BERT-multilingual-cased | 85.40 | 90.95 | 88.08 |
| BERT-multilingual-uncased | 87.16 | 89.79 | 88.46 |
| XLM-RoBERTa-base | 88.53 | 89.56 | 89.04 |

In Table 2, under *ground-truth relevant documents* (3), when evaluating the ground-truth relevant examples only, the task is obviously easier, particularly in terms of precision, while, when we test on the predicted relevant documents in Table 2, under *predicted relevant documents* (2), the amount of errors that are being propagated to the event extraction step is extremely high, reducing all the F1 scores by over 20% points for all models. Since there is a considerable reduction in the number of relevant instances after the classification step, this step alters the ratio between the relevant instances and the retrieved instances. Thus, not only in F1 but also a significant drop in precision is observed across all the models, when compared with the ground-truth results. The drop in precision is due to a number of relevant documents being discarded by the classifier.

Since our best results were not obtained after applying document classification for the relevant article detection, we consider that our best models are those applied on the initial dataset comprised of relevant and irrelevant documents. Thus, we continue by presenting the performance of these models for each language in the dataset.

Table 3. Evaluation scores (F1%) of the analyzed models for the predicted relevant documents per language, found by the classification model. The Chinese language was not included in the table because the classification model did not detect any relevant Chinese document.

| Model | French | English | Greek | Chinese | Russian | Polish |
|---------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| BERT-multilingual-cased | 83.60 | 65.52 | 75.00 | 80.00 | 63.64 | 82.35 |
| BERT-multilingual-uncased | 84.17 | 80.70 | 73.47 | 50.00 | 60.27 | 84.62 |
| XLNet-RoBERTa-base | 84.67 | 52.00 | 72.73 | 66.67 | 61.11 | 81.90 |

As shown in Table 3, BERT-multilingual-uncased obtained the highest scores for three out of the four low-resource languages, while BERT-multilingual-cased was more fitted for Polish. The reason for the higher results in the case of the low-resourced Greek, Chinese, and Russian languages could be motivated by considering the experiments performed in the paper that describes the XLM-RoBERTa model [8]. The authors concluded that, initially, when training on a relatively small amount of languages (between 7 and 10), XLM-RoBERTa is able to take advantage of positive transfer which improves performance, especially on low resource languages. On a larger number of languages, the *curse of multilinguality* [8] degrades the performance across all languages due to a trade-off between high-resource and low-resource languages. As pointed out by Conneau et al. [8], adding more capacity to the model can alleviate this *curse of multilinguality*, and thus the results for low-resource languages could be improved when trained together.

Model-Wise Analysis. As demonstrated in the results, different models perform differently on different datasets. Thus, we move beyond the holistic score assessment (entity F1-score) and compare the strengths and weaknesses of the models at a fine-grained level.

We analyzed the individual performance of the models and the intersections of their predicted outputs by visualizing them in several UpSet plots⁷. As seen in Fig. 3(a), there are approximately 70 positive instances that none of the systems was able to find. The highest intersection, approximately 340 instances, represents the true positives found by the three systems. BERT-multilingual-cased was able to find a higher number of unique true positive instances, instances not detected by the other models.

BERT-multilingual-uncased had the highest number of true positive instances cumulatively, the second-highest number of unique true positives, and the lowest number of false positive instances. This reveals the ability of the BERT-multilingual-uncased model to find the relevant instances in the dataset and to correctly predict a large proportion of the relevant data points, thus the high recall and precision, and overall F1 performance.

⁷ <https://jku-vds-lab.at/tools/upset/>.

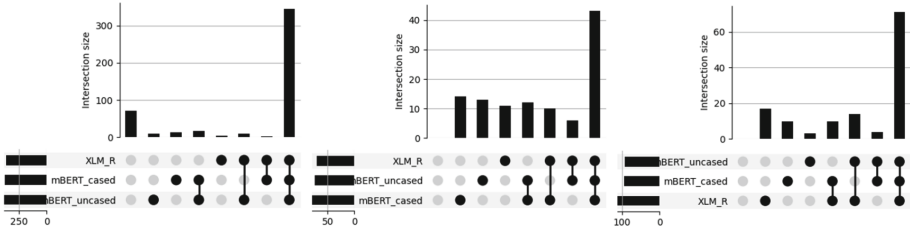


Fig. 3. Intersection of models predictions. The figures represent (from left) the true positive, false positive and false negative intersection sizes. The x-axis is interpreted as follows; from left to right, the first bar represents the number of instances that no system was able to find, the next three bars show the instances found by the respective individual models, the next three denote instances found by a pair of systems, while the last bar (the highest intersection) represents instances jointly found by all systems.

The overall performance is, generally, affected by the equally higher number of false positive and false negative results, as presented in Figs. 3(b, c). XLM-RoBERTa-base recorded the highest false negative rate and the lowest number of true positive instances, which explains the low recall and F1 scores.

Attribute-Wise Analysis. We chose to utilize an evaluation framework for interpretable evaluation for the named entity recognition (NER) task [16] that proposes a fine-grained analysis of entity attributes and their impact on the overall performance of the information extraction systems⁸.

We conduct an attribute-wise analysis that compares how different attributes affect performance on the DANIEL dataset, (e.g., how entity or sentence length correlates with performance). The entity attributes considered are entity length (eLen), sentence length (sLen), entity frequency (eFreq), token frequency (tFreq), out-of-vocabulary density (oDen), entity density (eDen) and label consistency. The label consistency describes the degree of label agreement of an entity on the training set. We consider both entity and token label consistencies, denoted as eCon and tCon. eCon represents the number of entities in a sentence. To perform the attribute-wise analysis, bucketing is applied, a process that breaks down the performance into different categories [16, 17, 33].

The process involves partitioning the attribute values into $m = 4$ discrete parts, whose intervals were obtained by dividing the test entities equally, with in some cases the interval method being customized depending on the individual characteristics of each attribute [16]. For example, in the entity length (eLen), entities in the test set with lengths of $\{1, 2, 3\}$ and >4 are partitioned into four buckets corresponding to the lengths. Once the buckets are generated, we calculate the F1 score with respect to the entities of each bucket.

The results in Table 4 illustrate that for our dataset the performance of all models varies considerably and it is highly correlated with oDen, eCon, tCon,

⁸ The code [16] is available here: <https://github.com/neulab/InterpretEval..>

Table 4. Attribute-wise F1 scores (%) per bucket for the following entity attributes: entity length (eLen), sentence length (sLen), entity frequency (eFreq), token frequency (tFreq), out of vocabulary density (oDen), entity density (eDen), entity consistency (eCon) and token consistency (tCon).

| Model | F1 | Bucket | F1 | | | | | | | |
|---------------------------|-------|--------|-------------|--------------|--------------|--------------|-------------|-------------|-------------|--------------|
| | | | eDen | oDen | eCon | tCon | tFreq | sLen | eFreq | eLen |
| BERT-multilingual-cased | 80.19 | 1 | 84.15 | 86.00 | 59.11 | 18.18 | 74.76 | 74.16 | 76.78 | 81.12 |
| | | 2 | 84.15 | 83.54 | 85.62 | 84.07 | 86.59 | 77.35 | 90.47 | 79.24 |
| | | 3 | 88.03 | 70.32 | 100 | 87.94 | 83.52 | 85.58 | 85.50 | 92.30 |
| | | 4 | 89.88 | 53.33 | 100 | 96.15 | 84.26 | 88.23 | 81.96 | 0 |
| Standard deviation | | | 2.48 | 12.97 | 16.69 | 31.14 | 4.48 | 5.76 | 4.99 | 36.80 |
| BERT-multilingual-uncased | 80.99 | 1 | 86.13 | 84.09 | 59.75 | 31.25 | 77.72 | 72.50 | 77.51 | 80.61 |
| | | 2 | 87.12 | 84.61 | 84.60 | 81.22 | 85.17 | 78.67 | 86.40 | 76.36 |
| | | 3 | 88.67 | 68.88 | 100 | 87.35 | 83.01 | 81.19 | 82.60 | 83.33 |
| | | 4 | 82.75 | 55.88 | 100 | 94.33 | 83.00 | 90.36 | 81.30 | 0 |
| Standard deviation | | | 2.17 | 11.90 | 16.45 | 24.85 | 2.74 | 6.42 | 3.17 | 34.77 |
| XLM-RoBERTa-base | 79.52 | 1 | 81.24 | 84.28 | 53.06 | 100 | 72.27 | 72.80 | 76.40 | 79.73 |
| | | 2 | 84.57 | 80.00 | 85.15 | 81.05 | 84.04 | 74.99 | 86.88 | 76.00 |
| | | 3 | 85.43 | 63.52 | 87.50 | 87.60 | 84.04 | 81.73 | 81.20 | 92.30 |
| | | 4 | 87.35 | 57.57 | 100 | 87.50 | 81.25 | 89.77 | 81.60 | 0 |
| Standard deviation | | | 2.20 | 11.10 | 17.32 | 6.86 | 4.83 | 6.61 | 3.70 | 36.30 |

and eLen. This proves that the prediction difficulty of an event mention is influenced by label consistency, entity length, out-of-vocabulary density, and sentence length. Regarding the entity length, the third bucket had fewer entities among the first three buckets and the highest F1 score among the four buckets, an indication that a majority of entities were correctly predicted. A very small number of entities had a length of size 4 or more, and at the same time, those entities were poorly predicted by the evaluated models (F1 of zero).

Moreover, the standard deviation values observed for BERT-multilingual-uncased are the lowest when compared with the other two models across the majority of the attributes (except for tCon, oDen, and sLen), which can be an indication that this model is not only the best performing, but it is also the most stable, thus being particularly robust.

4.2 Semi-supervised Learning Experiments

Due to the limited availability of annotated datasets in epidemic event extraction, we employ the self-training semi-supervised learning technique in order to analyze whether our dataset and models could benefit from having relevant unannotated documents. We then experiment with the mean teacher (MT) training method, a semi-supervised learning method where a target-generating teacher model is updated using the exponential moving average weights of the student model [41]. As such, the approach can handle a wide range of noisy input such as digitized documents, which often are susceptible to optical character recognition (OCR) errors [32, 36, 45].

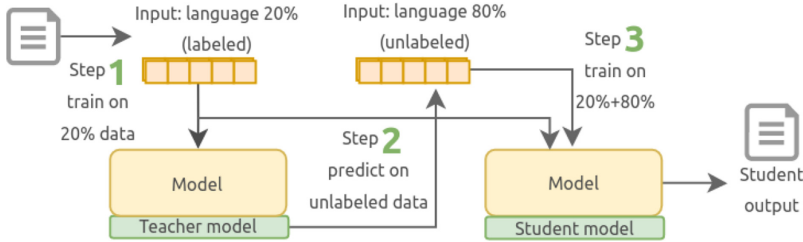


Fig. 4. The self-training process in the 20% for training and 80% unannotated data few-shot setting.

Table 5. The four few-shot scenarios with a comparison between their increasing number of training sentences and the amounts of DIS and LOC per scenario and per language.

| Model | Greek | | | Chinese | | | Russian | | | Polish | | |
|-------|-------|-----|-----|---------|-----|-----|---------|-----|-----|--------|-----|-----|
| | Sent | DIS | LOC | Sent | DIS | LOC | Sent | DIS | LOC | Sent | DIS | LOC |
| 20% | 854 | 35 | 19 | 1,280 | – | – | 1,115 | 44 | 22 | 1,486 | 56 | 28 |
| 30% | 1,315 | 53 | 32 | 1,904 | 17 | 0 | 1,617 | 62 | 35 | 2,501 | 64 | 29 |
| 40% | 1,801 | 69 | 46 | 2,427 | 22 | 1 | 2,031 | 72 | 40 | 3,078 | 83 | 47 |
| 50% | 2,228 | 87 | 58 | 3,230 | 22 | 1 | 2,488 | 95 | 50 | 3,697 | 112 | 56 |
| 100% | 4,947 | 144 | 115 | 6,309 | 57 | 10 | 5,250 | 170 | 88 | 7,288 | 352 | 146 |

First, we consider the documents in four low-resource languages from the DANIEL dataset: Greek, Chinese, Russian, and Polish. These languages are around 80% less represented in this dataset when compared to French. For these experiments, we simulate several few-shot scenarios by implementing a strategy in which we split our training data into annotated and unannotated sets, starting from 20%, and increasing iteratively by 10% points until 50%. Thus, we obtain four few-shot learning scenarios as detailed in Table 5. For example, in the 20% scenario, 20% of the data is considered annotated and 80% unannotated. The process of self-training, as presented in Fig. 4, has the following steps:

- Step 1: Each of our models will be, at first, the teacher, trained and fine-tuned on the event extraction task using a cross-entropy loss and a small percentage of the DANIEL dataset (i.e., we keep 20% for training). The rest of the data is considered unlabeled (i.e., the rest of 80%).
- Step 2: This data (80%) is annotated using the teacher model generating in this manner the pseudo labels which are added to the annotated percentage of data (20%) to form the final dataset.
- Step 3: Next, each of the models will be the student, trained and fine-tuned on this dataset using KL-divergence consistency cost function.

Table 6. The results for the low-resourced languages from DANIEL when all data for all languages is trained together, and when the languages are trained separately.

| Model | Greek | Chinese | Russian | Polish |
|---|--------------|--------------|--------------|--------------|
| Data instances trained on all languages | | | | |
| BERT-multilingual-cased | 73.47 | 50.00 | 60.27 | 84.62 |
| BERT-multilingual-uncased | 75.00 | 80.00 | 63.64 | 82.35 |
| XLNet-RoBERTa-base | 72.73 | 66.67 | 61.11 | 81.90 |
| Data instances trained per language | | | | |
| BERT-multilingual-cased | 80.77 | 85.71 | 60.00 | 86.67 |
| BERT-multilingual-uncased | 81.56 | 80.00 | 64.00 | 86.79 |
| XLNet-RoBERTa-base | 80.77 | 80.00 | 63.16 | 86.54 |

Table 7. The results for the low-resourced languages in DANIEL in the four few-shot scenarios (F1%).

| Language | 100% | Baseline | | | | Self-training (MT) | | | |
|---------------------------|-------|----------|-------|-------|-------|--------------------|--------------|--------------|--------------|
| | | 20% | 30% | 40% | 50% | 20% | 30% | 40% | 50% |
| BERT-multilingual-cased | | | | | | | | | |
| Greek | 80.77 | 58.54 | 72.73 | 77.55 | 83.33 | 55.56 | 77.19 | 60.47 | 77.27 |
| Chinese | 80.00 | 0.0 | 80.00 | 80.00 | 66.67 | 0 | 80.00 | 72.73 | 80.00 |
| Russian | 63.16 | 41.79 | 50.67 | 54.84 | 53.73 | 34.15 | 47.06 | 53.85 | 40.82 |
| Polish | 86.54 | 74.51 | 74.23 | 78.10 | 78.50 | 75.25 | 73.47 | 80.39 | 76.47 |
| BERT-multilingual-uncased | | | | | | | | | |
| Greek | 81.56 | 51.43 | 72.73 | 72.73 | 80.77 | 45.71 | 72.73 | 70.83 | 78.26 |
| Chinese | 80.00 | 0 | 80.00 | 80.00 | 80.00 | 0 | 80.00 | 80.00 | 80.00 |
| Russian | 64.00 | 36.11 | 48.00 | 52.78 | 52.17 | 31.37 | 53.38 | 53.52 | 50.00 |
| Polish | 86.79 | 72.55 | 73.47 | 77.36 | 78.90 | 75.25 | 69.31 | 75.23 | 76.19 |
| XLNet-RoBERTa-base | | | | | | | | | |
| Greek | 76.36 | 50.00 | 74.51 | 72.00 | 75.47 | 53.33 | 75.00 | 72.34 | 77.55 |
| Chinese | 85.71 | 0.0 | 66.67 | 72.73 | 66.67 | 0 | 66.67 | 72.73 | 66.67 |
| Russian | 60.00 | 32.50 | 47.62 | 53.66 | 53.16 | 34.38 | 49.28 | 55.74 | 53.52 |
| Polish | 86.67 | 65.55 | 70.59 | 75.00 | 75.23 | 67.83 | 73.27 | 77.67 | 79.61 |

Holistic Analysis. In Table 6, we compare the results obtained when the languages were all trained and tested together and when the languages were trained separately. One can notice that higher scores were obtained in the second case, showing the positive impact of fine-tuning one model per language. This could also be explained by the *curse of multilinguality* [8] that degrades the performance across all languages due to a trade-off between high-resource and low-resource languages when the languages are trained together. Meanwhile, the

advantages of training them separately considerably increase the performance for each of the languages.

Table 7 presents the four few-shot scenarios, the F1 score when the models are trained on the entire language data, the F1 scores for the baselines (the models trained in a supervised manner on the few samples), and with the self-training using the mean teacher method. For the latter, we fine-tune all our models on between 800–3000 sentences of training data for each language (as shown in Table 5) and use it as a teacher model. Larger improvements in performance were noticed in the case of the XLM-roBERTa-base model, where self-training leads to 2.29% average gains on Greek (from 67.99% to 69.55%), 4.19% on Polish (from 71.59% to 74.59%), and on Russian, 3.21% (from 46.73% to 48.23% while remaining unchanged for Chinese).

In the majority of the cases and for all the models, the performance improvements can also be due to the fact that, because of the few-shot scenarios that are created from our initial dataset, the simulated unannotated data remains in-domain with the labeled data. It was proven that using biomedical papers for a downstream named entity recognition (NER) biomedical task considerably improves the performance of NER compared to using unannotated news articles [20]. Meanwhile, for the cases where we observed a decrease in the performance after self-training, it would mean that the teacher model was not that strong, leading to noisier annotations compared to the full or baseline dataset setup.

5 Conclusions

In this study, we evaluated supervised and semi-supervised learning methods for multilingual epidemic event extraction. First, with supervised learning, we observe low precision values when training and testing on all data instances and predict relevant documents. This is not surprising since the number of negative examples, with potential false positives, rises up to around 90%.

While the task of document classification, prior to event extraction, was expected to result in performance gains, our results reveal a significant drop in performance. This can be attributed to error propagation to the downstream task. Further, the fine-grained error analysis provides a comprehensive assessment and better understanding of the models. This facilitates the identification of the strengths of a model and aspects that can be enhanced to improve the performance of the model.

Regarding the semi-supervised experiments, we show that the mean teacher self-training technique can potentially improve the model results, by utilizing the fairly readily available unannotated data. As such, the self-training method can be beneficial to low-resource languages by alleviating the problems associated with the scarcity of labeled data.

In future work, we propose to focus on the integration of real unannotated data to improve our overall performance scores on the low-resourced languages. Also, since directly applying self-training on pseudo labels results in gradual drifts due to label noises, we propose to study in future work a judgment model

to help select sentences with high-quality pseudo labels that the model predicted with high confidence. Further, we intend to explore the semi-supervised method under different noise levels and types to determine the robustness of our models to noise.

References

1. Aiello, A.E., Renson, A., Zivich, P.N.: Social media-and internet-based disease surveillance for public health. *Ann. Rev. Public Health* **41**, 101–118 (2020)
2. Bernardo, T.M., Rajic, A., Young, I., Robiadek, K., Pham, M.T., Funk, J.A.: Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *J. Med. Internet Res.* **15**(7), e147 (2013)
3. Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y.: COMET: commonsense transformers for automatic knowledge graph construction. arXiv preprint [arXiv:1906.05317](https://arxiv.org/abs/1906.05317) (2019)
4. Brixstel, R., Lejeune, G., Doucet, A., Lucas, N.: Any language early detection of epidemic diseases from web news streams. In: 2013 IEEE International Conference on Healthcare Informatics, pp. 159–168. IEEE (2013)
5. Casey, A., et al.: Plague dot text: text mining and annotation of outbreak reports of the Third Plague Pandemic (1894–1952). *J. Data Min. Digit. Humanit. HistoInf.* (2021). <https://jdmdh.episciences.org/7105>
6. Chen, S., Pei, Y., Ke, Z., Silamu, W.: Low-resource named entity recognition via the pre-training model. *Symmetry* **13**(5), 786 (2021)
7. Choi, J., Cho, Y., Shim, E., Woo, H.: Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health* **16**(1), 1–10 (2016)
8. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, 5–10 July 2020, pp. 8440–8451. Association for Computational Linguistics (2020). <https://www.aclweb.org/anthology/2020.acl-main.747/>
9. Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 32, pp. 7059–7069. Curran Associates, Inc. (2019). <http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf>
10. Dean, K., Krauer, F., Schmid, B.: Epidemiology of a bubonic plague outbreak in Glasgow, Scotland in 1900. *R. Soc. Open Sci.* **6**, 181695 (2019). <https://doi.org/10.1098/rsos.181695>
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, June 2019. <https://doi.org/10.18653/v1/N19-1423>
12. Ding, B., et al.: DAGA: data augmentation with a generation approach for low-resource tagging tasks. arXiv preprint [arXiv:2011.01549](https://arxiv.org/abs/2011.01549) (2020)
13. Doan, S., Ngo, Q.H., Kawazoe, A., Collier, N.: Global health monitor-a web-based system for detecting and mapping infectious diseases. In: Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II (2008)

14. Dórea, F.C., Revie, C.W.: Data-driven surveillance: effective collection, integration and interpretation of data to support decision-making. *Front. Vet. Sci.* **8**, 225 (2021)
15. Feng, X., Feng, X., Qin, B., Feng, Z., Liu, T.: Improving low resource named entity recognition using cross-lingual knowledge transfer. In: *IJCAI*, pp. 4071–4077 (2018)
16. Fu, J., Liu, P., Neubig, G.: Interpretable multi-dataset evaluation for named entity recognition. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6058–6069 (2020)
17. Fu, J., Liu, P., Zhang, Q.: Rethinking generalization of neural models: a named entity recognition case study. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 7732–7739 (2020)
18. Glaser, I., Sadegharmaki, S., Komboz, B., Matthes, F.: Data scarcity: Methods to improve the quality of text classification. In: *ICPRAM*, pp. 556–564 (2021)
19. Grancharova, M., Berg, H., Dalianis, H.: Improving named entity recognition and classification in class imbalanced Swedish electronic patient records through resampling. In: *Eighth Swedish Language Technology Conference (SLTC)*. Förlag Göteborgs Universitet (2020)
20. Gururangan, S., et al.: Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964* (2020)
21. Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., Gipp, B.: Giveme5W: main event retrieval from news articles by extraction of the five journalistic W questions. In: Chowdhury, G., McLeod, J., Gillet, V., Willett, P. (eds.) *iConference 2018*. LNCS, vol. 10766, pp. 356–366. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78105-1_39
22. Joshi, A., Karimi, S., Sparks, R., Paris, C., Macintyre, C.R.: Survey of text-based epidemic intelligence: a computational linguistics perspective. *ACM Comput. Surv. (CSUR)* **52**(6), 1–19 (2019)
23. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: SpanBERT: improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* **8**, 64–77 (2020)
24. Kozareva, Z., Bonev, B., Montoyo, A.: Self-training and co-training applied to Spanish named entity recognition. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) *MICAI 2005*. LNCS (LNAI), vol. 3789, pp. 770–779. Springer, Heidelberg (2005). https://doi.org/10.1007/11579427_78
25. Lampos, V., Zou, B., Cox, I.J.: Enhancing feature selection using word embeddings: the case of flu surveillance. In: *Proceedings of the 26th International Conference on World Wide Web*, pp. 695–704 (2017)
26. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: Multilingual event extraction for epidemic detection. *Artif. Intell. Med.* **65** (2015). <https://doi.org/10.1016/j.artmed.2015.06.005>
27. Lejeune, G., Brixtel, R., Lecluze, C., Doucet, A., Lucas, N.: Added-value of automatic multilingual text analysis for epidemic surveillance. In: Peek, N., Marín Morales, R., Peleg, M. (eds.) *AIME 2013*. LNCS (LNAI), vol. 7885, pp. 284–294. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38326-7_40
28. Lejeune, G., Doucet, A., Yangarber, R., Lucas, N.: Filtering news for epidemic surveillance: towards processing more languages with fewer resources. In: *Proceedings of the 4th Workshop on Cross Lingual Information Access*, pp. 3–10 (2010)
29. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., et al.: Performance measures for information extraction. In: *Proceedings of DARPA Broadcast News Workshop*, Herndon, VA, pp. 249–252 (1999)

30. Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., Odeo, M.: Multilingual epidemiological text classification: a comparative study. In: COLING, International Conference on Computational Linguistics (2020)
31. Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., Odeo, M.: Token-level multilingual epidemic dataset for event extraction. In: Berget, G., Hall, M.M., Brenn, D., Kumpulainen, S. (eds.) TPD 2021. LNCS, vol. 12866, pp. 55–59. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86324-1_6
32. Mutuvi, S., Doucet, A., Odeo, M., Jatowt, A.: Evaluating the impact of OCR errors on topic modeling. In: Dobрева, M., Hinze, A., Žumer, M. (eds.) ICADL 2018. LNCS, vol. 11279, pp. 3–14. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04257-8_1
33. Neubig, G., et al.: compare-MT: a tool for holistic comparison of language generation systems. arXiv preprint [arXiv:1903.07926](https://arxiv.org/abs/1903.07926) (2019)
34. Neudecker, C., Antonacopoulos, A.: Making Europe’s historical newspapers searchable. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 405–410. IEEE (2016)
35. Ng, V., Rees, E.E., Niu, J., Zaghool, A., Ghiasbeglou, H., Verster, A.: Application of natural language processing algorithms for extracting information from news articles in event-based surveillance. *Can. Commun. Dis. Rep.* **46**(6), 186–191 (2020)
36. Nguyen, N.K., Boros, E., Lejeune, G., Doucet, A.: Impact analysis of document digitization on event extraction. In: 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020) co-located with the 19th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2020), vol. 2735, pp. 17–28 (2020)
37. Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., Ji, H.: Cross-lingual name tagging and linking for 282 languages. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1946–1958 (2017)
38. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. Arxiv (2018)
39. Riedl, M., Padó, S.: A named entity recognition shootout for German. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 120–125 (2018)
40. Salathé, M., Freifeld, C.C., Mearu, S.R., Tomasulo, A.F., Brownstein, J.S.: Influenza a (H7N9) and the importance of digital epidemiology. *N. Engl. J. Med.* **369**(5), 401 (2013)
41. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint [arXiv:1703.01780](https://arxiv.org/abs/1703.01780) (2017)
42. Van Asch, V., Daelemans, W.: Predicting the effectiveness of self-training: application to sentiment classification. arXiv preprint [arXiv:1601.03288](https://arxiv.org/abs/1601.03288) (2016)
43. van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. *Mach. Learn.* **109**(2), 373–440 (2019). <https://doi.org/10.1007/s10994-019-05855-6>
44. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
45. Walker, D., Lund, W.B., Ringger, E.: Evaluating models of latent document semantics in the presence of OCR errors. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 240–250 (2010)
46. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. arXiv preprint [arXiv:1804.07461](https://arxiv.org/abs/1804.07461) (2018)

47. Wang, C.K., Singh, O., Tang, Z.L., Dai, H.J.: Using a recurrent neural network model for classification of tweets conveyed influenza-related information. In: Proceedings of the International Workshop on Digital Disease Detection Using Social Media 2017 (DDDSM-2017), pp. 33–38 (2017)
48. Wang, W., Huang, Z., Harper, M.: Semi-supervised learning for part-of-speech tagging of mandarin transcribed speech. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP 2007, vol. 4, pp. IV-137. IEEE (2007)
49. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189–196 (1995)
50. Zhu, X.J.: Semi-supervised learning literature survey (2005)