Token-level Multilingual Epidemic Dataset for Event Extraction*

Stephen Mutuvi[1,2][0000−0002−3067−9806], Emanuela Boros[2][0000−0001−6299−9452], Antoine Doucet[2][0000−0001−6160−3356], Gaël Lejeune[3][0000−0002−4795−2362], Adam Jatowt[4][0000−0001−7235−0665], and Moses Odeo[1][0000−0001−5068−3450]

[1] Multimedia University, Nairobi, Kenya
[2] University of La Rochelle, La Rochelle, France
[3] Sorbonne University, Paris, France
[4] University of Innsbruck, Innsbruck, Austria

Abstract. In this paper, we present a dataset and a baseline evaluation for multilingual epidemic event extraction. We experiment with a multilingual news dataset which we annotate at the token level, a common tagging scheme utilized in event extraction systems. We approach the task of extracting epidemic events by first detecting the relevant documents from a large collection of news reports. Then, event extraction (disease names and locations) is performed on the detected relevant documents. Preliminary experiments with the entire dataset and with ground-truth relevant documents showed promising results, while also establishing a stronger baseline for epidemiological event extraction.

Keywords: Epidemiological Surveillance · Multilingualism · Sequence Labeling.

1 Introduction

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 the same time requiring labor-intensive human input. Recently, approaches that complement the traditional surveillance methods with data-driven approaches which rely on internet-based data sources such as online news articles have been advanced [1, 3]. With the progress in natural language processing (NLP), processing and analyzing news data for epidemic surveillance has become feasible. Although this research is promising, the scarcity of available annotated multilingual corpora for data-driven epidemic surveillance is a major hindrance.

Online news data contains critical information about emerging health threats such as what happened, where it happened, when, and to whom it happened [11].

* 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.

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 [2].

Several works have tackled the detection of events related to epidemic diseases. For example, the Data Analysis for Information Extraction in any Language (DAnIEL) was proposed as a multilingual dataset and a news surveillance system that leverages repetition and saliency (salient zones in the structure of a news article), properties that are common in news writing [9]. Models based on neural network architectures which take advantage of the word embeddings representations have been used in monitoring social media content for health events [8]. Other methods were based on long short-term memory networks (LSTMs) [12] that approached the epidemic detection task from the perspective of classification of documents (in this case, tweets) to extract influenza-related information.

In this study, we formulate the problem of extracting the disease names and locations in the text as a sequence labeling task. We use the DAnIEL multilingual dataset (Chinese, English, French, Greek, Polish, and Russian) comprising news articles from the medical domain with diverse morphological structures. We establish a baseline performance using a specialized baseline system and experiment with the most recent neural sequence labeling architectures.

2 Dataset

Due to the lack of dedicated datasets for epidemic event extraction from multilingual news articles, we adapt a freely available epidemiological dataset[1], called DAnIEL [9]. The dataset consists of news articles in six different languages, namely French, Polish, English, Chinese, Greek, and Russian. In this dataset, an epidemiological event is represented by a disease name and the location of the reported event.

However, the DAnIEL dataset is annotated at document level, which differentiates it from typical datasets (token or word level annotations) utilized in research for the event extraction task (i.e., ACE 2005[2], TAC KBP 2014-2015[3]). A document is either reporting an event of interest (a disease-place pair appears in a relevant document) or not (an irrelevant document).

An example of a relevant document is contained in the following sentence: *Ten tuberculosis patients in India described as having an untreatable form of the lung disease may be quarantined to thwart possible spread, a health official said [. . .].*

[1] The dataset is available at https://daniel.greyc.fr/public/index.php?a=corpus.
[2] https://catalog.ldc.upenn.edu/LDC2006T06
[3] https://catalog.ldc.upenn.edu/LDC2020T13

In this case, the document is annotated with *Tuberculosis* as the disease name, and *India* as the location.

We begin by performing sentence segmentation, thus obtaining the individual sentences from the text corpus. The data is then annotated using the Doccano annotation tool[4], a collaborative annotation tool that provides annotation features for various tasks, among them sequence labelling task. The annotation guidelines required the annotators to identify and mark the spans for the key entities from the text. The occurrence of an epidemic event is characterized by mentions of disease name and the location of the disease outbreak, labeled DIS and LOC, respectively. Three native speakers annotators were recruited for each language.

The annotations were then transformed into IOB (Inside, Outside, Beginning) tagging scheme. For example, each token of a disease name, based on the spans, is assigned the tags B-DIS, I-DIS, and O, marking the beginning (B-), intermediate (I-), and out-of-span markers (O). We then compute the Inter-annotator agreemen (IAA) using the Kappa coefficient introduced by Cohen [4]. The average IAA for all languages was 0.66.

Table 1. Number of relevant tokens and sentences per dataset split per language.

Split	**Sentences**	**Tokens**	**French**	**English**	**Polish**	**Chinese**	**Greek**	**Russian**
Training	6,638	201,043	156,221	13,404	11,741	4,853	7,028	7,796
Validation	861	26,022	19,427	2,321	1,453	346	819	1,656
Test	862	26,134	21,634	1,221	1,498	434	687	660

Table 1 presents the statistics for this dataset from which we can observe the particularities and challenges of this dataset. DAnIEL dataset is not only multilingual, but it is also imbalanced considering the low-resourced languages (Chinese, Greek, and Russian).

3 Experiments and Results

We first consider the specialized event extraction system, DAnIEL [9], which we consider as a strong baseline. Then, we experiment with deep learning models based on a bidirectional LSTM (BiLSTM) [7, 10] that use character and word representations[5]. Additionally, due to the multilingual characteristic of the dataset, we use the multilingual BERT pre-trained language models [6] for token sequential classification and fine-tune them on our dataset. We will refer to these models as BERT-multilingual-cased[6] and BERT-multilingual-uncased[7]. We also experiment with the XLM-RoBERTa-base model [5] that has shown significant

[4] https://github.com/doccano/doccano.
[5] The hyperparameters for both models are detailed in the papers [7, 10].
[6] https://huggingface.co/bert-base-multilingual-cased.
[7] https://huggingface.co/bert-base-multilingual-uncased.

performance gains for a wide range of cross-lingual transfer tasks. We consider this model appropriate for our task and dataset due to the multilingual nature of the data[8].

As shown in Table 2, BERT-multilingual-uncased recorded the highest F1, recall and precision scores with 80.99%, 79.77% and 82.25% respectively, on the dataset comprising both relevant and irrelevant examples. We observe in Table 2 that all the models significantly outperform our DAnIEL baseline.

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**
All data instances (relevant and irrelevant)			
BiLSTM+LSTM	79.68	70.07	74.57
BiLSTM+CNN	73.38	71	72.17
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
Only relevant documents			
BiLSTM+LSTM	**91.32**	85.38	88.25
BiLSTM+CNN	87.29	84.45	85.85
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**

When evaluating the ground-truth relevant examples only, the task is obviously easier, particularly in terms of precision. Overall, XLM-RoBERTa-base attained the best F1-measure score of 89.04%. The model with the best recall was BERT-multilingual-cased (90.95%), while the BiLSTM+LSTM model had the highest precision.

4 Conclusions

In this paper, we present a token-level dataset and a strong baseline evaluation for multilingual epidemic event extraction. The results of the preliminary experiments suggest that the approaches based on pre-trained language models performed better than other deep learning models, thus, they can be utilized as strong baselines for epidemic event extraction. As future work, a further investigation of these preliminary results could reveal the underlying reasons of the different performance values, and thus, further work will focus on a more fine-grained analysis of the methods. Moreover, we also propose to further examine the classification of relevant and irrelevant documents, in order to ascertain the level of error propagation from the document classification.

[8] XLM-RoBERTa-base was trained on 2.5TB of CommonCrawl data in 100 languages.

References

1. Aiello, A.E., Renson, A., Zivich, P.N.: Social media–and internet-based disease surveillance for public health. Annual Review of Public Health **41**, 101–118 (2020)
2. Brixtel, 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)
3. 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)
4. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 8440–8451. Association for Computational Linguistics (2020), https://www.aclweb.org/anthology/2020.acl-main.747/
6. 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, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
7. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016)
8. 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)
9. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: Multilingual event extraction for epidemic detection. Artificial intelligence in medicine **65** (07 2015). https://doi.org/10.1016/j.artmed.2015.06.005
10. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1064–1074. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1101, https://www.aclweb.org/anthology/P16-1101
11. 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. Canada Communicable Disease Report **46**(6), 186–191 (2020)
12. 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)