



**HAL**  
open science

# Epidemic Event Extraction in Multilingual and Low-resource Settings

Stephen Mutuvi

► **To cite this version:**

Stephen Mutuvi. Epidemic Event Extraction in Multilingual and Low-resource Settings. Document and Text Processing. Université de La Rochelle, 2022. English. NNT: 2022LAROS044 . tel-03978917v2

**HAL Id: tel-03978917**

**<https://theses.hal.science/tel-03978917v2>**

Submitted on 12 Jul 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# LA ROCHELLE UNIVERSITÉ

## *École doctorale Euclide*

Laboratoire Informatique, Image, Interaction (L3i)

**THÈSE** présentée par :

**Stephen Mutuvi**

soutenance le : **21 November 2022**

pour obtenir le grade de : **Docteur de La Rochelle Université**

Discipline : **Informatique**

## **Epidemic Event Extraction in Multilingual and Low-resource Settings**

<b>Rapportrice</b>	<b>Aurélie NÉVÉOL</b>	Directrice de Recherche, LISN, CNRS
<b>Rapporteur</b>	<b>Mathieu ROCHE</b>	Directeur de Recherche, TETIS, CIRAD
<b>Examineurs</b>	<b>Juan-Manuel TORRES-MORENO</b>	Maître de Conférences HDR, Avignon Université
	<b>Jakub PISKORSKI</b>	Research Associate, Polish Academy of Sciences
	<b>Anne VILNAT</b>	Professeure des Universités, LISN, Université Paris-Saclay
	<b>Gaël LEJEUNE</b>	Maître de Conférences, STIH, Sorbonne Université
<b>Directeur</b>	<b>Antoine DOUCET</b>	Professeur des Universités, L3i, La Rochelle Université
<b>Co-directeur</b>	<b>Moses ODEO</b>	Professeur des Universités, Multimedia University of Kenya
<b>Invités</b>	<b>Emanuela BOROŞ</b>	Post-doctorante, L3i, La Rochelle Université
<b>(Co-encadrants)</b>	<b>Adam JATOWT</b>	Professeur des Universités, University of Innsbruck, Austria



# Abstract

Epidemic event extraction aims to extract incidents of public health importance from text, such as disease outbreaks. Delayed detection and response to widespread outbreaks of infectious diseases could result in significant health, economic, and social impact. The task of extracting epidemic events from text involves finding documents containing events of interest and extracting epidemiological entities, including the disease name and location of the reported outbreak.

While event extraction has been extensively researched for high-resource languages such as English, existing systems for epidemic event extraction are sub-optimal for low-resource, multilingual settings due to limited training data. In general, only a few corpora exist for training and evaluating multilingual epidemic event extraction models. In addition to data scarcity, inherent disparities between high-resource and low-resource languages in multilingual settings result in data imbalance, which poses challenges for extraction systems. Due to these peculiarities, only a few multilingual epidemic event extraction approaches exist, while approaches based on pre-trained language models remain unexplored. This research focused on addressing these challenges, which limit the extraction of epidemic events from multilingual news texts.

First, after presenting in detail an extensive state of the art for both event extraction, in general, and epidemic event extraction, which constitutes the basis of this thesis, we tackle the data scarcity challenge. Thus, due to the lack of dedicated annotated datasets suitable for epidemic event extraction tasks, we transform an existing document-level dataset into a token-level annotated dataset suitable for supervised sequence learning. Native speakers of the respective languages were recruited to provide the annotations. The produced multilingual dataset comprises news articles from diverse languages, including low-resource languages with limited training data resources.

Second, we formulate the event extraction task as a sequence labeling task and utilize the token-level annotated dataset to train supervised machine and deep learning models for epidemic event extraction. The results show that pre-trained language models, which have demonstrated outstanding performance in a variety of information extraction tasks, produced the best overall performance across all the evaluated languages. The proposed supervised learning approaches provide strong baselines for further epidemic event extraction research. Furthermore, we ex-

amine the impact of news document structure on model performance, taking into account the inverted pyramid style commonly used in journalistic writing. We observed that the lead paragraphs of news articles, considered to contain the most fundamental information, had the greatest influence on performance.

Third, we propose a domain adaptation technique by including epidemiological entities in the vocabulary of Transformer-based pre-trained models. Domain adaptation through vocabulary expansion was found to be a viable approach for addressing the lack of annotated training data for epidemic event extraction in multilingual and low-resource settings. In particular, we observed that including epidemiological entities (disease names and locations) into the vocabulary of the tokenizers, which were not initially present in the generic datasets used to train the tokenizers, improved tokenization quality and ultimately contributed to model performance improvement.

Finally, we evaluate self-training for epidemic event extraction and observe that the approach performs marginally better than models trained using supervised learning. This demonstrates the critical role that unlabeled data can play in achieving competitive performance while minimizing data annotation requirements. Further analysis revealed that noisy text has a negative impact on model performance and that eliminating the noise was critical for improving the performance of epidemic event extraction. Thus, we proposed a topic modeling approach for noise filtering that, contrary to our expectations, resulted in a slight decline in self-training performance, which we attributed to the preceding selection process.

## Résumé

L'extraction d'événements épidémiologiques vise à extraire d'un texte des incidents ayant une importance pour la santé publique, tels que les épidémies. Un retard dans la détection et la réponse à des épidémies de maladies infectieuses pourrait avoir un impact sanitaire, économique et social important. L'extraction d'événements épidémiologiques à partir de textes implique la recherche de documents contenant des événements d'intérêt et l'extraction d'entités épidémiologiques, notamment le nom de la maladie et le lieu de l'épidémie signalée.

Alors que l'extraction d'événements a fait l'objet de recherches approfondies pour des langues à fortes ressources comme l'anglais, les systèmes existants d'extraction d'événements épidémiologiques n'offrent pas de résultats optimaux dans les contextes multilingues peu dotés en ressources en raison de la taille limitée des données d'apprentissage dans ces langues. En général, seuls quelques corpus existent pour l'entraînement et l'évaluation des modèles multilingues d'extraction d'événements épidémiologiques. Outre la rareté des données, les disparités inhérentes entre les langues à ressources élevées et les langues à faibles ressources dans les environnements multilingues entraînent un déséquilibre des données, ce qui pose des problèmes aux systèmes d'extraction automatique. En raison de ces particularités, il n'existe que quelques approches multilingues d'extraction d'événements épidémiologiques, tandis que les approches basées sur des modèles de langue pré-entraînés restent peu explorées. Dans cette recherche, nous nous sommes donc attachés à relever les défis, qui limitent l'extraction d'événements épidémiologiques à partir de textes multilingues.

Tout d'abord, après avoir présenté un état de l'art pour l'extraction d'événements, en général, et pour l'extraction d'événements épidémiologiques en particulier, qui constitue la base de cette thèse, nous nous attaquons au défi de la rareté des données. Ainsi, en raison du manque d'ensembles de données annotées dédiées aux tâches d'extraction d'événements épidémiologiques, nous transformons un ensemble de données existantes au niveau des documents en un ensemble de données annotées au niveau des *tokens*, adapté à l'apprentissage supervisé à partir de séquences. Des locuteurs natifs des différentes langues traitées ont été recrutés pour fournir les annotations. Le jeu de données multilingue ainsi produit comprend des articles de presse dans diverses langues, en particulier des langues peu dotées (c'est-à-dire des langues pour lesquelles

la quantité de données d’entraînement sera plus limitée que pour d’autres).

Ensuite, nous avons proposé de traiter la tâche d’extraction d’événements comme une tâche d’étiquetage de séquences et utilisé l’ensemble de données annotées au niveau des *tokens* pour entraîner des modèles supervisés d’apprentissage automatique et d’apprentissage profond. Les résultats montrent que les modèles de langue pré-entraînés, qui ont démontré des performances exceptionnelles dans une variété de tâches d’extraction d’informations, ont produit la meilleure performance globale dans toutes les langues évaluées. Les approches d’apprentissage supervisé proposées constituent des bases solides pour de futures recherches sur l’extraction d’événements épidémiologiques. En outre, nous examinons l’impact de la structure des documents sur les performances du modèle, en tenant compte du style dit de “pyramide inversée” couramment utilisé dans les écrits journalistiques dans un nombre très important de langues. Nous avons observé que les paragraphes de tête des articles de presse, correspondant à la notion journalistique de “chapeau” et considérés comme contenant les informations fondamentales, avaient la plus grande influence sur les performances des systèmes automatiques.

Troisièmement, nous proposons une technique d’adaptation au domaine en incluant des entités épidémiologiques dans le vocabulaire des modèles pré-entraînés basés sur les Transformer. L’adaptation du domaine par l’expansion du vocabulaire s’est avérée être une approche viable pour remédier au manque de données d’entraînement annotées pour l’extraction d’événements épidémiologiques dans des contextes multilingues et peu dotées. En particulier, nous avons observé que l’inclusion d’entités épidémiologiques (noms de maladies et lieux) dans le vocabulaire des tokenizers, entités absentes des jeux de données génériques utilisés pour l’entraînement les tokéniseurs, a amélioré la qualité de la tokenisation et a finalement contribué à l’amélioration des performances des modèles.

Enfin, nous avons évalué l’auto-apprentissage (*self-training*) pour l’extraction d’événements épidémiologiques et observé que l’approche était légèrement plus performante que les modèles entraînés à l’aide de l’apprentissage supervisé. Cela a démontré le rôle essentiel que les données non étiquetées peuvent jouer pour obtenir des performances compétitives tout en minimisant les exigences d’annotation des données. Une analyse plus poussée a révélé que les textes bruités ont un impact négatif sur les performances du modèle et que l’élimination du bruit était essentielle

pour améliorer les performances de l'extraction des événements épidémiologiques. Nous avons donc proposé une approche de modélisation thématique pour le filtrage du bruit qui, contrairement à nos attentes, a entraîné une légère baisse des performances de l'auto-apprentissage, que nous avons attribuée au processus de sélection précédent.



# Acknowledgements

I am incredibly grateful to my supervisors, Prof. Antoine DOUCET and Dr. Moses ODEO, for their unwavering support throughout my PhD studies. I have learned a lot from you that will be very useful in my future research work. I will be eternally grateful to you for shaping my research career.

I would like to express my gratitude to my thesis committee members for their thoughtful comments and discussions, which were extremely beneficial to my research. Aurélie NÉVÉOL, Mathieu ROCHE, Anne VILNAT, Jakub PISKORSKI, and Juan-Manuel TORRES-MORENO, I cannot thank you enough.

I owe a great deal of appreciation to Emanuela BOROŞ, Adam JATOWT, and Gaël LEJEUNE for the numerous opportunities for collaboration. Thank you for the tremendous sacrifice and support you have shown me since the beginning of my research journey. Many thanks to my colleagues at the University of La Rochelle, Multimedia University, and Leroy MWANZIA from CIAT, all of whom have greatly inspired me.

Additionally, this endeavor would not have been possible without the generous support of the French government, which funded my research through the French embassy in Kenya and Campus France. The work has also been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

I would be remiss if I did not express my gratitude to my family; to my wife, Faith, and daughters, Stephanie and Joan, your strong belief in me kept my spirit and motivation high throughout this process. You did, in fact, provide the much-needed moral support to navigate the journey. Last but not least, I would like to thank the Almighty for good health.

# Contents

<b>List of Figures</b>	<b>x</b>
<b>List of Tables</b>	<b>xiii</b>
<b>List of Acronyms</b>	<b>xvi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Challenges . . . . .	4
1.2 Objectives . . . . .	6
1.3 Contributions . . . . .	8
1.4 Structure of the Thesis . . . . .	9
1.5 Conclusions . . . . .	10
<b>2 State of the Art and Related Work</b>	<b>11</b>
2.1 Event Extraction in Natural Language Processing . . . . .	11
2.1.1 Topic-based Event Extraction . . . . .	17
2.1.2 Template-based Event Extraction . . . . .	19
2.2 Event Extraction Approaches . . . . .	20
2.2.1 Classification by Extraction Type . . . . .	20
2.2.2 Classification by Architecture Type . . . . .	24
2.2.3 Classification by Model Type . . . . .	26
2.2.4 Classification by Degree of Supervision . . . . .	31
2.2.5 Classification by Paradigm Type . . . . .	34
2.3 Epidemic Event Extraction Approaches . . . . .	35
2.4 Conclusions . . . . .	39

<b>3</b>	<b>Epidemiological Event-based Dataset</b>	<b>41</b>
3.1	DANIEL Dataset . . . . .	42
3.1.1	DANIEL Epidemiological Event Definition . . . . .	43
3.1.2	DANIEL Languages . . . . .	44
3.1.3	Extending the DANIEL Dataset . . . . .	45
3.1.4	Token-level DANIEL Data Annotation . . . . .	46
3.2	Conclusions . . . . .	49
<b>4</b>	<b>Supervised Learning for Multilingual Epidemic Event Extraction</b>	<b>51</b>
4.1	Document Classification . . . . .	53
4.1.1	Experimental Setup . . . . .	53
4.1.2	Results and Analysis . . . . .	55
4.2	Epidemic Event Extraction . . . . .	65
4.2.1	Model Selection and Evaluation . . . . .	65
4.2.2	Holistic Analysis . . . . .	66
4.2.3	Model-wise Analysis . . . . .	68
4.2.4	Attribute-wise Analysis . . . . .	69
4.3	Conclusions . . . . .	72
<b>5</b>	<b>Domain Adaptation for Low-resource Epidemic Event Extraction</b>	<b>74</b>
5.1	Epidemiological Domain Adaptation . . . . .	75
5.2	Tokenizer Analysis . . . . .	77
5.3	Results and Analysis . . . . .	82
5.4	Conclusions . . . . .	86
<b>6</b>	<b>Self-training with Topic Modeling for Event Extraction in Noisy Settings</b>	<b>87</b>
6.1	Mean Teacher Self-training . . . . .	88
6.2	Selection of Unlabeled Data using Topic Modeling . . . . .	89
6.3	Results and Analysis . . . . .	92
6.4	Evaluating the Impact of Errors on Topic Modeling . . . . .	96
6.5	Conclusions . . . . .	98

<b>7</b>	<b>Conclusions and Future Work</b>	<b>99</b>
7.1	Conclusions . . . . .	99
7.2	Future Work . . . . .	101
	<b>Publications</b>	<b>103</b>
	<b>Appendices</b>	<b>105</b>
<b>A</b>	<b>Overview of Neural Network Approaches for Information Extraction</b>	<b>107</b>
A.1	Neural Networks for Sequential Data Modeling . . . . .	107
A.1.1	Convolutional Neural Networks . . . . .	108
A.1.2	Recurrent Neural Networks . . . . .	109
A.1.3	Transformer-based Neural Networks . . . . .	113
A.2	Pre-Trained Language Models . . . . .	116
A.3	Bidirectional Encoder Representations from Transformers (BERT) . . . . .	117
A.3.1	Tokenization in Transformer-based Models . . . . .	119
A.4	Conclusions . . . . .	120
	<b>Bibliography</b>	<b>121</b>

# List of Figures

1.1	Timeline (in weeks) of COVID-19 outbreak. <b>C</b> denotes the number of cases. . . . .	3
2.1	An illustration of event extraction. The sentence has two events: <b>Die</b> and <b>Attack</b> . The <b>Die</b> event is triggered by “died” and its argument roles are <b>Place</b> , <b>Victim</b> and <b>Instrument</b> . The <b>Attack</b> event is triggered by “fire” with <b>Place</b> , <b>Target</b> and <b>Instrument</b> as the argument roles. Source: <a href="#">Sha et al. (2018)</a> . . . . .	14
2.2	A template of extracted event trigger and argument information for the <b>Die</b> event, from the example sentence presented in Figure 2.1 using closed-domain event extraction. . . . .	19
2.3	Pipeline-based Event Extraction. Source: <a href="#">Li et al. (2021)</a> . . . . .	24
2.4	Joint-based Event Extraction. Source: <a href="#">Li et al. (2021)</a> . . . . .	25
2.5	Sequence Labeling-based Event Extraction. Source: <a href="#">Li et al. (2021)</a> . . . . .	34
3.1	Excerpt from an English article in the DANIEL dataset that was published on January 13th, 2012 at <a href="http://www.smh.com.au/national/health/polio-is-one-nation-closer-to-being-wiped-out-20120112-1pxho.html">http://www.smh.com.au/national/health/polio-is-one-nation-closer-to-being-wiped-out-20120112-1pxho.html</a> . . . . .	44
4.1	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. . . . .	52
4.2	Impact of data size on performance of the best performing model: fine-tuned BERT (multilingual-uncased). . . . .	62

4.3	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. . . . .	70
4.4	Attribute-wise analysis of performance . . . . .	72
5.1	Tokenizer fertility score per language and model type. . . . .	80
5.2	Proportion of continued words per language and model type. . . . .	81
5.3	The ratios of unknown words per language and model type. . . . .	82
5.4	Relationship between tokenization quality and F1 performance. The languages considered were English, Greek, Polish, and Russian. The assessment of tokenizer quality was based on the real values of continued entities, continued words, OOVs, and fertility. . . . .	83
6.1	An illustration of the mean-teacher self-training approach using 20% labeled and 80% unlabeled data few-shot setting. . . . .	90
6.2	Performance (F1-scores) comparison of mBERT (Multilingual BERT) and XLM-R (XLM-RoBERTa) models per data split. Baseline denotes a model trained using supervised learning, self-training (MT) is mean teacher self-training, self-training (BERTopic) and self-training (LDA) represent a scenario where topic modeling techniques (BERTopic and LDA) were employed for noise filtering before self-training. . . . .	92
6.3	Performance (F1-scores) of mBERT (Multilingual BERT) and XLM-R (XLM-RoBERTa) models per language. The results show how performance scores change across the data splits for the baseline, mean teacher self-training, self-training with BERTopic, and self-training with LDA. The dotted line denotes the performance attained by the model when trained on the entire training dataset.	93

A.1	Convolutional Neural Networks for sentence classification. Source: <a href="#">Zhang and Wallace (2015)</a> . . . . .	110
A.2	Recurrent Neural Network Architecture <sup>1</sup> . $U$ and $V$ are the weights of the hidden layer and output layer respectively, while $W$ represents the transition weights of the hidden state. $\mathbf{x}_t$ and $\mathbf{O}_t$ are the input vector and output result at time $t$ , respectively. . . . .	111
A.3	Long Short-term Memory. Source: <a href="https://www.deeplearningbook.org/contents/rnn.html">https://www.deeplearningbook.org/contents/rnn.html</a> . . . . .	112
A.4	The Transformer architecture. Source: <a href="#">Vaswani et al. (2017)</a> . . . . .	114

# List of Tables

3.1	Dataset statistics for the extended dataset used for the document classification task.	45
3.2	The number of documents (percentage of relevant documents) per dataset split. The dataset consists of both relevant and irrelevant documents and was used for the document classification task.	47
3.3	Inter-annotator agreement score (Kappa coefficient) per language for the token-level corpus	48
3.4	Statistics of the token-level dataset. The terms DIS and LOC represent the number of disease and location mentions, respectively.	49
3.5	Number of tokens and sentences for the relevant documents per language.	50
4.1	Evaluation scores of the analyzed models for the relevant documents for all languages. The models evaluated, using the dataset configurations presented in Table 3.2, were Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB). Additionally, pre-trained BERT base-multilingual models were evaluated, both the <sup>†</sup> fine-tuned and models trained on BERT features.	57
4.2	F1-scores of the analyzed models for the relevant documents per language. The models evaluated, using the dataset configurations presented in Table 3.2, were Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB). Additionally, pre-trained BERT base-multilingual models were evaluated, both the <sup>†</sup> fine-tuned and models trained on BERT features.	59
4.3	Performance based on the sections of the documents using fine-tuned multilingual BERT (uncased) and VGCN with BERT. All positions of text have a limit of 512 tokens.	61



4.4	Evaluation scores of the BERT (multilingual-uncased) <sup>†</sup> model fine-tuned on the relevant documents in a zero-shot transfer learning setting. . . . .	63
4.5	Evaluation results for the detection of disease names and locations on all languages and all data instances (relevant and irrelevant documents). . . . .	68
4.6	Evaluation scores (F1) of the analyzed models for the predicted relevant documents per language, found by the classification model. . . . .	68
4.7	Attribute-wise F1-measures (%) 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). . . . .	71
5.1	Statistics of the DANIEL dataset. DIS = disease, LOC = location. . . . .	77
5.2	Language family, vocabulary size of individual language-specific models, and the size of vocabulary shared with mBERT and XLMR. . . . .	78
5.3	Model performance per language and model type. . . . .	83
5.4	Comparison between the default pre-trained model and the extended model (E-Model). The E-model is obtained by enriching the tokenizer vocabulary with domain-specific words (entities) that the tokenizer splits into subwords. . . . .	84
5.5	Number of entities added to the vocabulary per language and model. DIS (%) and LOC (%) denote the percentage of unseen disease and location entities, respectively, which were not found in the tokenizer vocabulary. . . . .	85
6.1	The studied scenarios, with an increasing number of training sentences and the number of disease names (DIS) and locations (LOC) entities per scenario and language. . . . .	91
6.2	The results of joint self-training on low-resource languages (F1%) and zero-shot learning with English as the source language. Baseline represents a model finetuned on labeled data while self-training (MT) represents a model trained using the mean teacher self-training method on both labeled and unlabeled data instances. . . . .	95

6.3 Topic Stability and Coherence performance of LDA and NMF of OCR generated text. . . . .	97
---	----

# List of Acronyms

**NLP** Natural language processing

**NER** Named entity recognition

**RE** Relation extraction

**EE** Event extraction

**NN** Neural network

**ACE** Automatic Content Extraction

**PLMs** Pre-trained language models

**DAnIEL** Data Analysis for Information Extraction in any Language

**ProMED** Program for Monitoring Emerging Diseases

**GPHIN** Global Public Health Intelligence Network

**EBS** Event-based surveillance

**IBS** Indicator-based surveillance

**PHCs** Primary healthcare centers

**MUC** Message Understanding Conference

**ACE** Automatic Content Extraction

**EI** Epidemic intelligence

**PULS** Pattern-based Understanding and Learning System

**OCR** Optical Character Recognition

**LR** Logistic Regression

**RF** Random Forest

**SVM** Support Vector Machine

**NB** Naive Bayes

**ED** Event detection

**BERT** Bidirectional Encoder Representations from Transformers

**LSTM** Long Short-Term Memory

**BiLSTM** Bidirectional LSTM

**CNNs** Convolutional neural networks

**RNNs** Recurrent neural networks

**FFN** Feed-forward network

**ODEE** Open-domain event extraction

**LDA** Latent Dirichlet Allocation

**BPE** Byte-pair encoding

# CHAPTER 1

---

## Introduction

---

The prevention and control of infectious diseases remain a public health priority globally ([Bloom and Cadarette, 2019](#)). Diseases account for 11.78% of all deaths globally ([Roser and Ritchie, 2021](#)) and for 46.8% of deaths in low-resource countries ([World Health Organization, 2022](#)). An estimated 60% of the known infectious diseases and up to 75% of newly emerging ones are zoonotic diseases, also known as zoonoses ([Otte and Pica-Ciamarra, 2021](#); [UNEP and ILRI, 2020](#); [Woolhouse and Gowtage-Sequeria, 2005](#)). The zoonoses, which are transmitted from animals to humans, such as avian influenza, brucellosis, tuberculosis, coronavirus, and nipah virus ([Ochani et al., 2019](#)), are estimated to be responsible for 2.5 billion cases of human illness and 2.7 million human deaths worldwide each year. While infectious disease mortality rates declined in the years leading up to 2019, the trajectory shifted due to the COVID-19 pandemic, with over 6.2 million COVID-19-related deaths reported as of April 2022 ([World Health Organization, 2022](#)).

The large-scale pandemic has demonstrated that delayed detection and response to widespread infectious disease outbreaks not only increases morbidity and mortality rates but could also

have adverse economic and social ramifications. It is worth noting that many countries around the world have made significant progress in establishing mechanisms that ensure timeliness in detecting health-related events by strengthening critical public health functions, such as surveillance. However, the improvements need to be sustained and expanded to cover the growing number of emergencies. The establishment and implementation of robust epidemiological surveillance systems are critical for the timely detection and response to disease outbreaks (Heymann, Rodier, et al., 2001; Jung et al., 2019; Njeru et al., 2020). Epidemiological surveillance describes the continuous, systematic collection, analysis, and interpretation of health-related data to extract events of public health importance (Thacker and Berkelman, 1988). The extracted epidemiological event information serves as the basis for deploying appropriate intervention measures to contain the disease outbreaks (Rutherford, 1998).

Epidemic surveillance approaches can broadly be categorized into event-based surveillance (EBS) and indicator-based surveillance (IBS), both of which comprise the epidemic intelligence (EI) framework that serves the early warning and response functions (Balajee et al., 2021; Paquet et al., 2006). Being a case-based surveillance method, IBS primarily relies on formal reports from primary healthcare centers (PHCs) to detect health-related events. While generally reliable due to the use of well-defined data sources, the collection of structured data through conventional surveillance systems covers limited Sentinel locations and their laboratory-based verification of suspected cases could result in delayed detection of the outbreaks (Chan et al., 2010; Jebara and Shimshony, 2006). In the case of the COVID-19 pandemic, for instance, it took nearly a month from the time the first local case was reported to when the pandemic was declared a public health emergency. As many as 7,800 people had contracted the virus by the time the World Health Organization issued an official declaration and a response plan to states, as shown in Figure 1.1.

The EBS, on the other hand, monitors epidemic events using informal data sources such as online news text. Digital data, which includes data in text format, is a remarkable source of real-time disease information. Among the factors driving EBS adoption is the recent rapid increase in the amount of data generated, which is attributed to the advancement and widespread adoption of Internet technology. Digital data, such as online news text, is a valuable source of real-

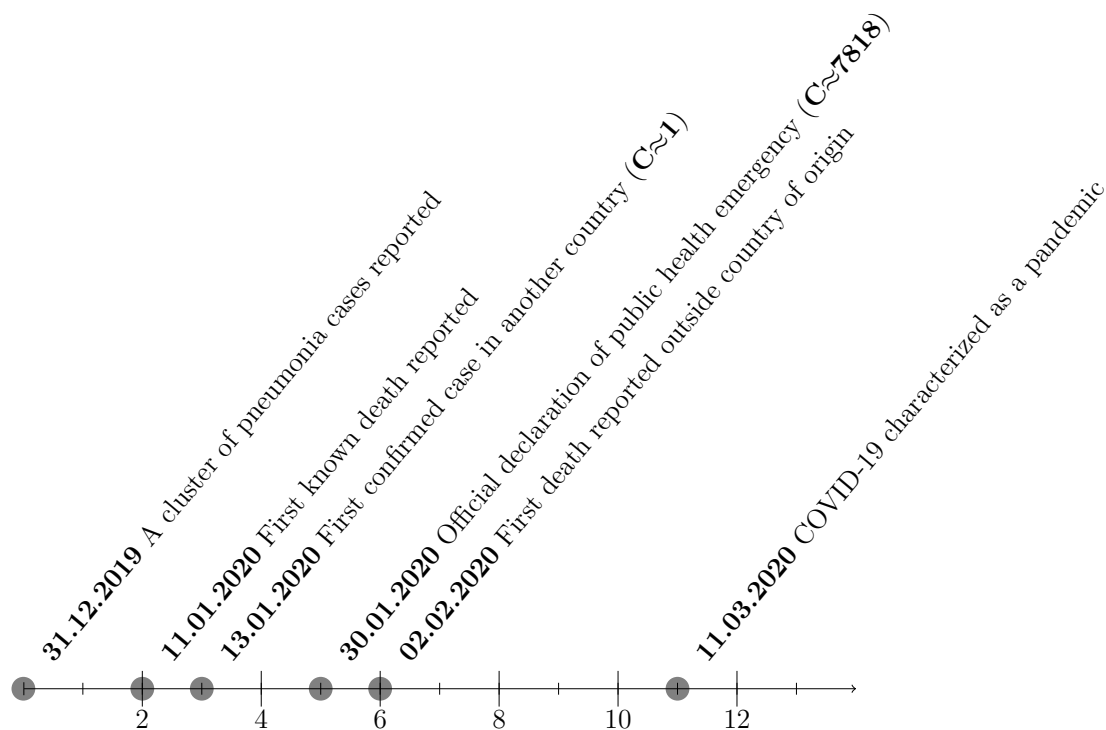


Figure 1.1: Timeline (in weeks) of COVID-19 outbreak.  $C$  denotes the number of cases.

time disease information (Paolotti et al., 2014; Valentin et al., 2020c), providing an opportunity to develop data-driven epidemic surveillance approaches. The approaches could complement traditional surveillance methods, increasing the geographic scope and timeliness of surveillance activities (Arsevska et al., 2018; Bahk et al., 2015; Mawudeku and Blench, 2005). Online news reports originate from all over the world, thus potentially providing timely and useful signals on confirmed acute public health events or potential events of concern that span a large geographical area.

Besides the increased availability of text data, advancements in natural language processing (NLP) techniques play an important role in the development of event-based disease surveillance. Using NLP techniques could bolster the extraction of events of public health importance (e.g., infectious disease outbreaks) from written natural language texts. Such text is computationally opaque and difficult for computers to immediately interpret (Busser and Moens, 2006). However, training high-performing models typically requires both high-quality and large amounts of annotated data. Obtaining the annotations is prohibitively expensive, especially in specialized domains (e.g., epidemiological surveillance), where domain experts are required to annotate the

data (Yang et al., 2019d). As a result, cutting-edge machine and deep learning-based methods remain largely underutilized for epidemic surveillance (Valentin et al., 2021). To our knowledge, approaches based on pre-trained language models (e.g., XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2018) and which have been proven to learn high-quality representations of texts (Lyu, 2022), have yet to be adequately explored in epidemic surveillance. A pre-trained language model typically comprises a deep Transformer-based neural network (Vaswani et al., 2017) trained on large-scale unlabeled corpora, which can then be fine-tuned on labeled data for downstream tasks such as event extraction (EE). While pre-training is computationally intensive, fine-tuning is relatively inexpensive and allows for significant reductions in data annotation efforts.

Despite the empirical success of the pre-trained models in NLP downstream tasks, the models still lack the knowledge needed for extracting information in specialized domains (e.g., epidemic event extraction) (Tai et al., 2020). Therefore, further research on the suitability of the pre-trained models for domain-specific multilingual contexts is required in order to improve their performance, especially for low-resource languages (Lejeune et al., 2013). This is essential for increasing the coverage and timeliness of epidemic surveillance systems since news reports about an outbreak are likely to be first published in the local language rather than in English.

In this thesis, we proposed and evaluated approaches to epidemic event extraction from multilingual corpora comprising texts from both high-resource and low-resource languages with limited annotated data resources. We annotated a token-level multilingual epidemiological dataset, explored domain adaptation through the injection of domain knowledge into pre-trained models, and investigated self-training that leverages unlabeled data along with a few labeled training instances for model training. These were found to be beneficial for epidemic event extraction involving low-resource languages.

## 1.1 Challenges

Despite the promise of data-driven disease surveillance to improve timeliness and coverage of disease surveillance, the textual data used in training the EBS systems present various peculiari-



ties. The large volume and variety of news data generated on a daily basis can be overwhelming for surveillance systems. This information overload makes it difficult for users to select information and accurately apply it for problem solving and decision making (Edmunds and Morris, 2000). In addition, the text data is often characterized by redundancy and linguistic ambiguities, where words and phrases can have multiple meanings, thus having multiple alternative interpretations. Redundancy manifests in having multiple documents that convey the same message being generated from multiple sources. Another common problem associated with textual data is noise. Because text data is user-generated and/or produced through a digitization process, it is difficult to guarantee the absence of errors (e.g., spelling mistakes or Optical Character Recognition (OCR) errors) (Srivastava et al., 2020; Subramaniam et al., 2009). The impact of these text peculiarities on model performance could be exacerbated by the unique characteristics of multilingual and low-resource settings, typified by domain-specific extraction tasks such as epidemic event extraction. Therefore, low-resource settings not only concern languages but also non-standard domains and tasks (Hedderich et al., 2020). More specifically, epidemic event extraction faces the following challenges:

1. **Data scarcity:** While the availability of sufficient and high-quality data is a critical requirement for developing high-performing data-driven systems (LeCun et al., 2015; Liu et al., 2017; Mihalcea and Chklovski, 2003; Sarker, 2021), there exist only a few corpora for training and evaluating epidemic event extraction models. The few available datasets are small in size and predominantly in English. Acquisition of additional labeled data necessitates data annotation, a process that is prohibitively expensive and time-consuming, sometimes requiring annotators with expert knowledge, particularly for specialized domains such as epidemic surveillance. Among the publicly available epidemiological corpora is the dataset used for the evaluation of the Data Analysis for Information Extraction in any Language (DAnIEL) system (Lejeune et al., 2015), an unsupervised multilingual epidemiological news surveillance system. This dataset, however, is annotated at the document level, necessitating transformation into a token-level format suitable for sequence labeling tasks such as event extraction.
2. **Limited multilingual epidemic event extraction approaches:** Because news reports

typically originate from divergent sources and languages, the relevant datasets used in epidemic event extraction from news text are inherently multilingual. Given a large number of languages across the world, developing models that can handle multiple languages and dialects remains challenging (Joshi et al., 2020; Ogueji et al., 2021). Previous research on multilingual epidemic event extraction primarily focused on pattern-based systems (Collier et al., 2008; Linge et al., 2010) or systems based on unsupervised learning (Lejeune et al., 2015). State of the art NLP approaches based on pre-trained language models remain unexplored for epidemic event extraction (Valentin et al., 2021).

3. **Inherent disparities between high-resource and low-resource languages:** Well-described languages such as English and French usually receive more attention from researchers and thus have readily available training data resources compared to their low-resource counterparts. In a multilingual setup, data imbalance among the languages poses an optimization tension between high-resource and low-resource languages (Li and Gong, 2021). As a result of assigning few parameters to the low-resource languages, the resultant multilingual model is often sub-optimal for the low-resource languages due to their limited training data size, which ultimately impacts the model performance (Chau et al., 2020; Wu and Dredze, 2020a).

## 1.2 Objectives

The goal of this thesis is to advance event extraction research in multilingual and low-resource settings, where annotated datasets are limited. We primarily focus on addressing the challenges (outlined in Section 1.1) facing the extraction of epidemic events from online news texts. The extraction task is formulated as a sequence labeling task that extracts epidemiological information, such as disease names and locations, from the text. Since in-depth token-level understanding is essential in sequence labeling tasks (Wang et al., 2019a), our first objective was to annotate an existing multilingual document-level dataset, known as the DANIEL dataset (Lejeune et al., 2015), into a token-level dataset suitable for supervised sequence learning. The multilingual dataset consists of news articles in French, English, Russian, Polish, Greek, and Chinese. In this dataset, only the French language had a significant number of documents (> 2000).

Our second objective was to use the annotated data to train supervised models for epidemic event extraction. Various machine learning and deep learning methods were evaluated and compared. This was essential in determining the most suitable supervised learning techniques for extracting epidemic events in multilingual settings, which are characteristic of most news-based surveillance systems. Pre-trained language models, previously not evaluated for epidemic event extraction, were also examined in this study.

Third, we sought to address the problem resulting from data imbalances among languages in multilingual, low-resource settings. In such settings, some languages are under-resourced in terms of the available training data, which could impact the performance of event extraction models. To deal with the imbalance and data scarcity in these settings (multilingual and low-resource), we explored domain adaptation through injection of in-domain data into the vocabulary of pre-trained models. The incorporation of epidemiological-specific terms into the vocabulary of pre-trained models was intended to improve the tokenization quality of the models, thereby improving their performance.

Finally, we investigated self-training, a semi-supervised learning approach that leverages unlabeled data along with the few available labeled examples to train models. The utilization of unlabeled data aids in addressing the issue of data scarcity while minimizing data annotation requirements. In the same context, we also explored the extent to which filtering noisy unlabeled data using topic modeling impacts the performance of self-training. In epidemic event extraction, noise is characterized by documents with epidemiological entities that are not necessarily linked to an event ([Valentin, 2020](#)). These errors present a challenge to predicting phenomena of interest from data. Models such as Bidirectional Encoder Representations from Transformers ([BERT](#)) are sensitive to noise and break down easily in the presence of errors such as spelling mistakes ([Soper et al., 2021](#); [Srivastava et al., 2020](#)). For instance, in the study by [Srivastava et al. \(2020\)](#), the performance drop in BERT was attributed to the inability of the BERT tokenizer to handle misspelled words.

More specifically, our objective is to answer the following questions:

1. How suitable is sequence labeling of token-level annotated data for epidemiological surveillance?

2. To what extent does supervised learning improve over the selected baseline system on the epidemic event extraction task in resource-constrained multilingual settings?
3. How effective is tokenizer vocabulary expansion in adapting pre-trained models to epidemic event extraction?
4. To what extent does the use of unlabeled data through self-training improve the performance of epidemic event extraction?

### 1.3 Contributions

This section presents the main contributions of the thesis. The first contribution is the annotation of a multilingual dataset for epidemic event extraction. Due to the unavailability of dedicated epidemic event extraction datasets, we created a token-level dataset based on the **DAnIEL** dataset (Lejeune et al., 2015), a publicly available epidemiological dataset<sup>1</sup> that was initially annotated at document-level. We re-annotated the dataset and transformed it into a token-level dataset (Mutuvi et al., 2021)<sup>2</sup> assigning a label to each token in the text. The disease names and locations represent an epidemiological event in this dataset, a definition previously used in epidemic surveillance research (Arsevska et al. (2018)). The dataset consists of news articles in six different languages: French, Polish, English, Chinese, Greek, and Russian. While the languages have been extensively studied, in our context we considered them, except for French, as low-resource because they have less than 40 documents for the positive class. The positive class comprises relevant documents that contain epidemiological information.

Our second contribution proposes supervised learning approaches for epidemic event extraction. We formulate the problem of extracting epidemic events as a sequence labeling task, in which names of diseases and locations present in the text are identified and classified into sets of predefined classes (i.e., disease names and locations). The disease name and location are correct if their event type and offsets match those of a reference disease or location. Overall, we observe that models based on multilingual pre-trained language models had the best overall performance across all languages tested. The methods developed in this study could serve as a foundation

---

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

<sup>2</sup> The token-level annotated dataset is available at <https://doi.org/10.5281/zenodo.6024726>.

(baseline) for further research on multilingual epidemic event extraction. Further details about the deep learning approaches utilized in our study are provided in Appendix A.

Third, we present a detailed analysis of various factors that influence the performance of epidemic event extraction systems in multilingual and low-resource settings, such as training data size, noise, and document structure. In line with the inverted pyramid style (Piskorski et al., 2011) used in news writing, the content from the initial paragraphs of the news articles contributed the most to the performance, which corresponds to the journalistic writing style in which the most important information appears at the start of the article and the least newsworthy information is placed at the end. In addition, we demonstrate that domain adaptation via vocabulary expansion had a positive impact on low-resource languages. More specifically, the introduction of epidemiological domain-specific vocabulary was beneficial to the epidemic event extraction task in settings with limited labeled data resources.

Finally, we explore self-training and demonstrate that using unlabeled data, which is relatively easy to obtain, is a viable solution to the problem of labeled data scarcity in low-resource settings and domain-specific tasks such as epidemiological surveillance (Liang et al., 2022). Furthermore, we show that noisy text degrades model performance and that eliminating the errors could be critical in attaining performance improvements. While noise filtering is beneficial for self-training, alternative approaches to noise filtering need to be investigated to improve performance even further.

## 1.4 Structure of the Thesis

The thesis is organized as follows.

Chapter 2 presents the state of the art and related work. Existing approaches to event extraction, with a focus on epidemic event extraction, are described, as well as their strengths and limitations.

Chapter 3 describes the dataset that we used in our study and its preparation. More precisely, we adapted a specialized dataset for epidemic surveillance called the DANIEL dataset and re-annotated the dataset at the token level, making it suitable for sequence labeling tasks.

Chapter 4 investigates the suitability of various supervised learning algorithms for event-based epidemic surveillance. A dedicated surveillance baseline system was evaluated and compared to various data-driven learning techniques for epidemic event extraction.

Chapter 5 addresses the challenges facing epidemic event extraction in multilingual and low-resource settings. Through domain adaptation, unseen words are added to the tokenizer vocabulary, resulting in performance gains for the different languages considered.

Chapter 6 explores self-training, a semi-supervised approach that leverages labeled and unlabeled data for model training. In this chapter, we also investigate the impact of noise on model performance. The noise was filtered using topic modeling and the performance of self-training compared to performance before noise filtering.

Chapter 7 provides the conclusions of the thesis and makes suggestions for future work.

Finally, Appendix A highlights various concepts relevant to our study. We provide an overview of neural networks, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformer-based pre-trained language models that are utilized in this thesis.

## 1.5 Conclusions

In this chapter, we provided the context and objectives of the study and outlined various challenges facing the extraction of epidemic events from multilingual news text in low-resource settings. Limited annotated data, lack of effective multilingual epidemic event extraction techniques, and disparity and data imbalance in low-resource settings are the most important challenges for this task. We also highlight our contributions, which are focused on addressing the previously mentioned challenges. The next chapter presents the previous work related to epidemic event extraction. In addition, we review past work on event extraction in general, focusing on the concept of an event as articulated in the various studies.

## CHAPTER 2

---

### State of the Art and Related Work

---

This chapter provides a comprehensive review of the literature on event extraction, a subtask of information extraction that seeks to extract information describing an event from text. Based on the existing literature, we provide the definitions of an epidemiological event and an event in general. Moreover, we categorize the event extraction methodologies into two broad categories of systems: topic-based and template-based. These categories of systems were subdivided further by extraction type, architecture type, model type, paradigm, and degree of supervision. The review, in addition to synthesizing and consolidating the existing literature on epidemic and event extraction in general, identifies some research gaps that are addressed in this study.

#### **2.1 Event Extraction in Natural Language Processing**

Event extraction has long been a focus of information extraction research aimed at discovering event structures in natural language text (Ahn, 2006). It encompasses deducing specific knowledge concerning the incidents referred to in textual data. Since event extraction is domain-specific, there is currently no strict and precise definition of an event. As a result, the notion

of an event is tailored and adapted to the intended task or domain (Frisoni et al., 2021). For example, the concept of an event from the information extraction perspective differs from that of linguistics, where the same event can be expressed in various linguistic elements (Sprugnoli and Tonelli, 2017). The lack of consensus on the definition of an event limits the automatic comparison of event extraction approaches. In the field of IE, an event is generally defined as a specific occurrence at a specific time and place, involving one or more participants (Doddington et al., 2004). Therefore, the event extraction task broadly consists of two main subtasks: event detection and argument extraction (Grishman et al., 2005). The event detection task finds event triggers of specific types, whereas the argument extraction task identifies event arguments and their argument roles, which form an important component of an event. Event arguments describe a set of entities that play a particular role in a given event.

Over the years, research in event extraction has been shaped and advanced through a succession of evaluation campaigns and workshops that made available annotated corpora and accompanying evaluation tasks. The datasets provided through the campaigns are typically annotated manually by domain experts. The evaluation campaigns focused not only on contemporary news text but also defined domain-specific event definitions for social media text, clinical records, and biomedical-related documents (Bethard et al., 2017; Intxaurreondo et al., 2015). The first evaluation program was the Message Understanding Conference (MUC) (Grishman and Sundheim, 1996) campaign that was organized between 1987 and 1998 by the Defense Advanced Research Projects Agency (DARPA) and was primarily focused on the English language.

Initially, the program provided evaluations to foster research on the automated analysis of military messages about naval sightings and engagements. The participating teams were provided with sample messages, instructions for developing the systems, and test data to evaluate their proposed solutions. Reports on terrorist events were included in MUC-3, with more domains considered in later versions of the campaign. The majority of the suggested systems in the MUC evaluations were pipeline-based systems that relied on pattern-matching techniques. Besides NER and coreference resolution, which identify words or phrases that refer to the same real-world entity, the MUC campaign also proposed the scenario template (ST) task. The scenario template task involved filling event templates with extracted information about an event, such as



event type, participants, and attributes. The best performing system for the ST task recorded an F1 score of 0.51 (Aone et al., 1998). However, this performance was lower compared to that of NER and coreference resolution, which had F1 scores of 0.93 and 0.62, respectively, an indication that the ST task was a harder task than the others. The key determinants of performance for the ST task were the number of slots to be filled and the level of input from domain experts.

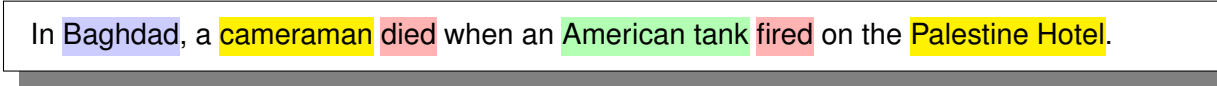
The Automatic Content Extraction (ACE) evaluation campaign (Doddington et al., 2004), convened by the National Institute of Standards and Technology (NIST) from 1999 to 2008, is another well-known program in which EE has been intensively researched. The campaign incorporated specifications for the event extraction task into the ACE 2005 annotation scheme (Walker et al., 2006). An event, in the context of ACE, is defined as an explicit occurrence involving one or more participants that occurs at a specific location and time. The ACE 2005 EE task defines eight (8) event types and 33 subtypes of events.

The following definitions describe an event structure in the ACE 2005:

- An **event mention** is a phrase or sentence that contains one or more event triggers and their corresponding arguments and describes an event.
- An **event trigger** is a word (verb or a noun) that most clearly expresses the occurrence of an event. In a schema-based extraction architecture, a trigger is classified into a predefined **event type**, which is a category to which a given event belongs.
- **Event arguments** are the attributes or participants with a specific role in an event. Event arguments can be either temporal expressions or mentions of entities and non-entities.
- **Argument roles** represent the relationship between the event arguments and the event triggers.

The event extraction task in ACE, therefore, entails discovering event mentions from text, detecting event triggers, and extracting the corresponding event arguments. The identification and classification of triggers into specific event types constitutes the event detection (ED) subtask. On the other hand, the argument extraction subtask includes argument identification and argument role classification, which extract words, phrases, or entities that play specific roles in a target event. In order to better distinguish between the event detection and argument extrac-

tion subtasks, we consider the sentence in Figure 2.1 that describes the **die** and **attack** events. An event extraction system should be able to detect, from the sentence, the trigger words **Died** and **Fired** for the **die** and **attack** event types, respectively. Further, the system should extract “Baghdad”, “cameraman” and “American tank” as arguments for the **Die** event, and classify their roles as **Place**, **Victim** and **Instrument** respectively. For the **Attack** event, “cameraman” and “Palestine Hotel” assume the **Target** event argument role while “Baghdad” and “American tank” should be classified into **Place** and **Instrument** event argument roles, respectively. Since the generation of the trigger-argument structures from text requires identification of trigger words and entities, NLP tasks such as named entity recognition (NER) (Lin et al., 2020), semantic parsing (Cao et al., 2020), and relation extraction (RE) (Ahmad et al., 2021) are important to the event extraction task.



In Baghdad, a cameraman died when an American tank fired on the Palestine Hotel.

Figure 2.1: An illustration of event extraction. The sentence has two events: **Die** and **Attack**. The **Die** event is triggered by “died” and its argument roles are **Place**, **Victim** and **Instrument**. The **Attack** event is triggered by “fire” with **Place**, **Target** and **Instrument** as the argument roles. Source: Sha et al. (2018)

Despite fostering the development of numerous supervised methods for event extraction, the ACE scheme presented various shortcomings. ACE annotations, like MUC, are confined to specific domains and have a limited number of predefined event types. As a result, ACE-based systems can only process events that fall into a predefined set of categories. Using predefined event types can constrain the development of models that can adapt to new domains and applications (Grishman, 2010). Another shortcoming of the ACE annotation scheme relates to its complexity, which hampered the creation of consistent labeled data in the absence of annotators with a solid background in linguistics. These drawbacks prompted the creation of alternative annotation schemes, among them the Entities, Relations, and Events (ERE) scheme that was proposed within the Deep Exploration and Filtering of Text (DEFT) program of DARPA (Song et al., 2015).

The ERE intends to rapidly and easily scale up data annotation efforts while facilitating the generation of consistently labeled multilingual data (Aguilar et al., 2014). There are two types

of ERE: Light ERE and Rich ERE. The Light ERE is a simplified version of the ACE annotation that includes fewer event types and attributes. Extracting events, like in ACE, involves detecting and classifying event triggers into specified event types, followed by identification of participating event arguments and their roles. In contrast to ACE, only actualized events are considered in ERE (Song et al., 2015). The Light ERE evolved into the Rich ERE by expanding the annotation of events and event arguments. It facilitated advanced tasks such as event-event relation extraction and cross-document and cross-lingual event coreference by providing a more complex representation of events. Rich ERE introduced the concept of an Event Hopper, which was required to deal with the granularity variation of event mentions and event arguments within and between documents. The event hopper allows a broad range of event mentions to be coreferential by grouping mentions referring to the same event occurrence (Song et al., 2015). In comparison to ACE and Light ERE, the use of event hoppers allows for a more inclusive and less strict definition of event coreference.

The Knowledge Base Population evaluation track of the Text Analysis Conference (TAC KBP) (Ellis et al., 2014), organized by the National Institute of Standards and Technology (NIST), introduced Event Nuggets (EN) to deal with multiword events. The TAC KBP 2016 event track included the Event Nugget Detection and Coreference task and the Event Argument Extraction and Linking (EAL) task (Song et al., 2016). The EAL task focused on extracting event arguments and linking arguments that belong to the same event, while the event nugget track aimed to evaluate system performance on the detection and coreference of sets of attributes referencing events in unstructured text. An event nugget comprises a semantically meaningful unit that is linguistically represented by multiple words that express the event in a sentence (Mitamura et al., 2015). For example, considering the sentence “*His death sentence was carried out.”, which describes an event of type **Justice** and subtype **Execute** (or of **Justice.Execute** type), the words “carried out” are hard to separate in terms of meaning. Dealing with such multi-words (continuous and discontinuous) required annotating the maximum extent of text that meets the definition of the event types and subtypes. Since the TAC KBP evaluation program was accessible only to the registered participants, researchers sought to have publicly available campaigns whose corpora would be freely available beyond the campaign period for further research in order to achieve a widespread impact.*

Among the efforts targeted at achieving this goal was the work by [Pustejovsky et al. \(2003\)](#), who introduced the Time Markup Language (TimeML), a specification language for events and temporal expressions that inspired improvements in the field of Temporal Processing. TimeML introduced temporal link annotations that provided expressive capability for capturing and representing the relations between events and temporal expressions and for determining the temporal order between events. TimeML specifications were used to annotate the TimeBank corpus that formed the basis of the training data utilized in the TempEval evaluation program ([Verhagen et al., 2007](#)). The TempEval training corpora contains 183 English news articles annotated with temporal information such as events, temporal expressions, and temporal links between them, denoted by EVENT, TIMEX3, and TLINK TimeML tags, respectively. The TempEval-2 ([Verhagen et al., 2010](#)) campaign featured a more complex and elaborate set of specifications, an additional set of subtasks (six rather than three), and was multilingual (English, French, Italian, Spanish, Chinese, and Korean).

The adaptation of event processing to different domains and languages (other than English) has also been explored, with domain-specific annotation guidelines and systems developed for various domains. The ISO-TimeML ([Pustejovsky et al., 2010](#)), a revised and interoperable version of the TimeML, was adapted for the clinical domain. The ISO-TimeML was the basis for the creation of the THYME (Temporal Histories of Your Medical Events) annotation guidelines used in the annotation of a corpus of clinical notes. In THYME specifications, an event includes diseases, medical treatments, and all states and actions relevant to the patient's clinical timeline ([Styler et al., 2014](#)). The Integrating Biology and the Bedside (i2b2) challenge ([Sun et al., 2013](#)) and the Clinical TempEval evaluation ([Bethard et al., 2017](#)) were based on the THYME guidelines and utilized clinical notes and pathology reports as sources of data. The best performing system (UTHealth) on the Clinical TempEval achieved an F1 score of 0.93 on span identification and 0.88 on event type classification. The system leveraged a support vector machines (SVM) model with rich features and embeddings obtained from domain-specific dictionaries. The BioNLP shared task is another domain-specific task that aims to extract biomolecular events from biomedical documents ([Nédellec et al., 2013](#)). An event in the BioNLP task is defined as a change in the state of biomolecular objects. This definition is anchored on the event specifications from the GENIA project ([Kim et al., 2008](#)). Among the language-specific evaluation tasks

is the EVENTI (EValuation of Events aNd Temporal Information) organized within the context of EVALITA (Caselli et al., 2014). The task assesses the performance of the temporal information processing systems on a corpus of Italian news articles. Besides event extraction, EVENTI aimed to promote the development of NLP tools for the extraction of temporal expressions and temporal relations from Italian texts.

Approaches to event extraction can broadly be categorized into topic-based and template-based systems. Topic-based systems extract unconstrained event types, while template-based methods extract events with predefined event types and schemas. Despite their operational differences, both categories aim to extract event types of interest from text and the constituent event argument.

### 2.1.1 Topic-based Event Extraction

For successful event extraction from text, techniques for event extraction with wider and more consistent coverage need to be developed. Topic-based methods, also known as open-domain event extraction (ODEE), aim to achieve this goal by extracting unconstrained types of events. ODEE considers events as a set of related descriptions of a topic, obtained through classification or clustering of similar events based on extracted event keywords, words, or phrases primarily describing the event. ODEE systems mainly employ unsupervised extraction strategies and induce universal event schemas from natural language text (Chau et al., 2019; Liu et al., 2019a; Mejri and Akaichi, 2017). As a result, the systems can be scaled to different domains and tasks, mitigating the adaptability pitfall of closed-domain event extraction. Their openness, however, can impede reproducible comparisons and benchmarks due to the lack of established annotated corpora that can serve as gold standard datasets (Frisoni et al., 2021).

Chau et al. (2019) proposed an ODEE method for extracting structured information about natural gas prices from public news headlines. Price and news data were fed into a convolutional neural network to infer the relationship between events and market movements. The work filtered irrelevant news headlines before performing the event extraction task. A similar study developed an unsupervised neural latent variable model for inducing event schemas and extracting events (Liu et al., 2019a). The experimental results show that latent variables generated by neural

networks provide rich representation compared to hand-crafted features. Furthermore, their study showed that news redundancy caused by different news agencies reporting the same events positively influenced event extraction.

Another line of research investigates open domain event extraction from the perspective of topic detection and tracking (TDT)<sup>1</sup>. In the TDT task definition, a topic is defined as an event or activity and its related events and activities (Strassel et al., 2000). Finding topically related documents from a continuous stream of texts involves clustering documents by topic (Allan et al., 1998; Sprugnoli, 2018). Using topic clustering for event extraction enables coarse-grained event detection on large data streams (Atefeh and Khreich, 2015). The application of TDT to text stream monitoring could facilitate the discovery of previously unreported events (topics) and establish the progress of the previously spotted events (Allan et al., 1998).

Topic detection and tracking include the following subtasks:

- **Story segmentation:** determining the boundary (i.e., beginning and ending) of a story in a news article.
- **First story detection:** detecting the story that discusses a new topic in the stream of text.
- **Topic detection:** categorizing the stories based on the topics they discuss.
- **Topic tracking:** identification of stories that discuss a previously known topic.
- **Story link detection:** determining whether two stories in different documents discuss the same topic.

Other works, in addition to the TDT task, have investigated the detection and clustering of open-domain events from news articles (Liu et al., 2008; Piskorski et al., 2011; Tanev et al., 2008; Yu and Wu, 2018). For global crisis surveillance, extraction of violent events from online news was performed based on keywords such as killed, injured, and kidnapped (Piskorski et al., 2011; Tanev et al., 2008). Yu and Wu (2018) proposed a dual-level clustering model that aggregates news articles reporting the same event into topic-centered news sets. The work by Liu et al. (2008) focused on performing clustering on news articles to generate significant events

---

<sup>1</sup> Linguistic Data Consortium (LDC) <https://www ldc.upenn.edu/> TDT task annotation guidelines

for topics such as politics, economics, society, sports, and entertainment.

### 2.1.2 Template-based Event Extraction

Template-based event extraction, also known as closed-domain event extraction (Ferguson et al., 2018; Sheng et al., 2021; Yang and Mitchell, 2016), searches for and extracts desired events from text using predefined event schemas. In this method, the extraction process assumes filling in predefined event templates with information about an event. Different event types necessitate different event templates, limiting the scalability of template-based event extraction systems. The event templates can be generated manually or automatically with the help of statistical models. Taking the sentence in Figure 2.1, an event extraction system should be able to extract and fill event templates for the two events present, namely, **Attack** and **Die** events. For example, considering the **Die** event, the template is filled with the trigger word “fired”. The arguments for this event are “Baghdad”, “cameraman” and “American tank”, that take the argument roles **Place**, **Victim** and **Instrument**, respectively. The filled event template for the **Die** event is shown in Figure 2.2.

<b>Type:</b>	Die
<b>Trigger:</b>	died
<b>Argument Place:</b>	Baghdad
<b>Argument Victim:</b>	cameraman
<b>Argument Instrument:</b>	American tank

Figure 2.2: A template of extracted event trigger and argument information for the **Die** event, from the example sentence presented in Figure 2.1 using closed-domain event extraction.

Several works have explored event extraction from a template-filling standpoint. Petroni et al. (2018) proposed a system for extracting breaking news events from news reports and social media. The system resolves seven event types, namely **floods**, **storms**, **fires**, **armed conflict**, **terrorism**, **infrastructure breakdown**, and **labor shortages**, and extracts their associated attributes. A financial event extraction system for stock market prediction and investment decision support was proposed by (Yang et al., 2018). The system extracts events at the document level and can handle various financial event types and their associated argument roles. The event

types include **Equity Pledge**, **Equity Freeze**, **Equity Trading**, **Equity Overweight**, and **Equity Repurchase**.

While appropriate for discovering well-defined event types, a drawback of template-based event extraction is the requirement to create event templates, which in most cases may require manual human input and domain expertise. As a result, the template design process can be time-consuming, costly, and error-prone. Furthermore, the predefined event types and event argument schemas limit the number of event types that can be extracted in template-based event extraction. Consequently, the ability to generalize the manually defined event types to new domains and languages is significantly constrained, limiting the real-world applications of template-based event extraction systems. For example, the event types **Transmit Virus** and **Treat Disease** are not present in the ACE 2005 event schema, which may necessitate adaptation of the ACE specification to include these epidemic events for successful extraction.

## 2.2 Event Extraction Approaches

The topic- and template-based event extraction approaches, discussed in Sections 2.1.1 and 2.1.2 respectively, are further classified based on extraction type, architectural characteristics, model type, paradigm, and degree of supervision.

### 2.2.1 Classification by Extraction Type

As regards extraction type, approaches to event extraction can be either knowledge-driven or data-driven. **Knowledge-driven**, also known as pattern-based event extraction, extracts desired events from unstructured text using patterns that express rules representing domain expert knowledge (Mejri and Akaichi, 2017). Specific event templates are first constructed, followed by template matching based on predefined or derived patterns to extract events and their corresponding arguments from the text. The patterns can either be lexico-syntactic (Hearst, 1992) or lexico-semantic patterns (Borsje et al., 2010). Lexico-syntactic patterns combine lexical representations and syntactical information, whereas lexico-semantic patterns use gazetteers or ontologies to incorporate semantic information. The patterns can be learned automatically from



training data or curated manually by domain experts. The involvement of domain experts in event pattern design results in well-defined and high-quality event patterns, which translates to high-performance event extraction systems (Valentin et al., 2020c). One of the first pattern-based was AutoSlog (Riloff and Shoen, 1995), which exploited a small set of predefined linguistic patterns and a manually annotated corpus to obtain event patterns for extracting terrorist events. An example of a linguistic pattern is “<subject> passive-verb”, which denotes a phrasal verb in passive form preceded by a subject. With the Autoslog training corpus annotated with a single argument for each event, the system could only extract events with a single event argument.

While various advantages stem from the usage of pattern-based systems for event extraction, such as the ability to define powerful expressions for extracting very specific information, the number of patterns generated corresponds to the level of human input. Manually defining and maintaining the patterns could require substantial domain knowledge, hampering the adaptability of these types of event extraction systems to different domains and languages. Automatic pattern construction, based on weakly supervised or bootstrapping methods, has been studied in order to reduce the amount of human effort required to obtain patterns. Among the systems that use automatically constructed patterns is the AutoSlog-TS (Riloff and Shoen, 1995) event extraction system. The system was proposed as an extension to Autoslog and employed the CIRCUS (Lehnert, 1990) sentence analyzer to extract more event patterns from untagged text corpora. Except for a small set of training examples pre-classified as relevant or irrelevant, the AutoSlog-TS system did not require manual annotations.

Machine learning algorithms have also been applied to automatic pattern construction, where new patterns are learned based on a few seed patterns. The NEXUS (News Cluster Event eXtraction Using Language Structures) system, for example, learns event patterns through an entropy maximization-based machine learning algorithm, followed by manual validation (Piskorski et al., 2011, 2007). Cao et al. (2015) introduced a pattern expansion technique based on active learning that involves importing frequent patterns from external corpora to improve ED performance. A system based on the expanded patterns achieved an F1 score of 70.4%, representing a 1.6% absolute improvement over the baseline.

Knowledge-driven event extraction has also been successfully utilized for biomedical event ex-

traction (Bui et al., 2013; Bui and Sloot, 2012; Cohen et al., 2009; Hunter et al., 2008). Cohen et al. (2009) uses biomedical ontology analysis to extract biomedical events by taking into account the semantics of domain concepts. The approach builds ontological templates for biomedical concepts and their properties using the OpenDMAP semantic parser (Hunter et al., 2008). Bui and Sloot (2012) builds a rule-based system that decomposes complex event structures into syntactic layers that form a structured representation that expresses the structures of biomedical events. Bui et al. (2013) further proposes a method comprising a learning phase that generates a dictionary and patterns and an extraction phase that exploits the dictionary and patterns to extract events from input text. On the GENIA event extraction task of the BioNLP 2013 shared task, the system's F1 score was 48.92 on strict matching and 50.68 on approximate span and recursive matching.

A semi-automatic method, which used lexico-semantic patterns and financial ontologies, was developed to extract financial events from news text (Borsje et al., 2010). The lexico-semantic patterns were found more useful and enabled the discovery of events more precisely than their lexico-syntactic counterparts. In a related study, the Business Events Extractor Component, based on the ONtology (BEECON) system, was developed for financial event extraction from newspaper text (Arendarenko and Kakkonen, 2012). The system used an ontology as the basis for a domain knowledge base that stores both verified facts and newly discovered information. The system achieved a performance of 95% precision and 67% recall on the extraction of business event types.

Despite knowledge-driven methods producing promising results, particularly in domain-specific problems, they strongly depend on the expression form of the text and manually constructed event patterns by domain experts, which poses a myriad of challenges. The manual creation of event patterns is labor-intensive, time-consuming, and can generate only a handful of event patterns. Therefore, not only is the scalability of pattern-based approaches constrained but also their portability to new domains.

**Data-driven** event extraction methods were proposed to address the bottlenecks of knowledge-based approaches. In contrast to knowledge-based methods, data-driven approaches rely on large amounts of data rather than linguistic resources, patterns, or rules to produce significant

performance gains. The methods use machine learning techniques to learn from large amounts of text. From a machine learning perspective, the extraction of events refers to the idea of feature selection and building text classifiers for event detection and argument role classification. Machine learning techniques can effectively capture the lexical and semantic information of triggers, arguments, and their relationships. Several works on data-driven event extraction exist in the literature. Using clustering and weighted undirected bipartite graphs, [Liu et al. \(2008\)](#) developed a method to extract entities and events from online news articles. [Okamoto and Kikuchi \(2009\)](#) used a hierarchical clustering method to detect occasional or local events from news text. In a similar study, [Tanev et al. \(2008\)](#) clustered together real-time news to identify events related to violence and disasters. The machine learning approaches rely on hand-crafted features. Hence, significant effort and time go towards feature engineering to achieve satisfactory performance.

In the recent past, neural-based approaches have been proposed to avert the need for manual feature engineering by making use of distributional embedding features ([Chen et al., 2015, 2018](#); [Hong et al., 2018](#); [Liu et al., 2018b](#); [Nguyen and Grishman, 2018](#); [Nguyen et al., 2016](#); [Nguyen and Grishman, 2015, 2016](#)). The distributional vector representation provides a more flexible way to represent semantics of natural language that takes into account the context in which a word *usually* ([Mikolov et al., 2013a,c](#); [Pennington et al., 2014](#)) and *currently* ([Deschacht et al., 2012](#); [Devlin et al., 2018](#); [Howard and Ruder, 2018](#); [Peters et al., 2018a](#); [Radford et al., 2019](#)) appears in the text. When large annotated datasets are available, supervised models perform well in the source domain or language. However, the models tend to degrade in performance when applied to new domains or languages with different feature space distributions. Transfer learning has been researched to deal with cross-domain and cross-lingual challenges by adapting event extractors from source to target setting ([Pan and Yang, 2009](#)). In this case, a model trained on a large amount of annotated data in a specific source setting is applied to target domains, modalities, or languages with little or no annotated data. [Ji and Voss \(2021\)](#) proposed a transfer learning approach for low-resource event extraction in which information extracted from source training data was encoded into a shared continuous semantic space using a combination of symbolic (compressed representation of the real world) and distributional embedding representations. An event extraction model was learned from the source representations and

applied to the target data. Therefore, the event structures that comprise event types and argument roles were effectively transferred from high-resource to low-resource settings, reducing data annotation requirements.

## 2.2.2 Classification by Architecture Type

From an architectural point of view, event extraction methods can be classified into pipeline-based and joint-based event extraction systems.

**Pipeline-based** event extraction methods formulate the event extraction task as a multi-stage classification problem, treating the trigger classification and argument classification independently (Hong et al., 2011; Liao and Grishman, 2010; Lu and Roth, 2012). As shown in Figure 2.3, the four event extraction subtasks, namely, trigger identification, trigger classification, argument identification, and argument role classification, are performed sequentially. Performing the subtasks in stages allows them to benefit from their shared dependencies, in which additional information from previous subtasks is transferred to subsequent ones.

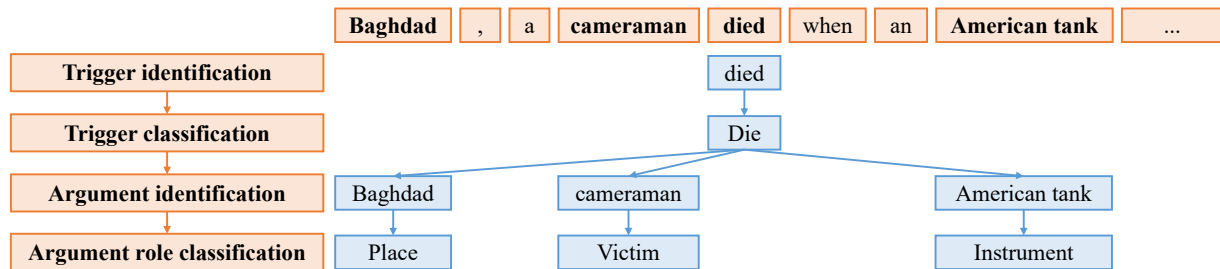


Figure 2.3: Pipeline-based Event Extraction. Source: Li et al. (2021)

Several studies investigated pipeline-based event extraction. For ACE event extraction, staged pipelines were used, with separate classifiers developed for trigger labeling and argument subtasks (Ahn, 2006; Chen and Ji, 2009; Grishman et al., 2005; Hong et al., 2011; Ji and Grishman, 2008; Li et al., 2012; Liao and Grishman, 2010). There also exist some pipelined approaches to event extraction that employ deep learning techniques (Björne and Salakoski, 2018; Chen et al., 2017, 2015; Huang et al., 2018; Li et al., 2020a; Nguyen and Grishman, 2018; Nguyen and Grishman, 2015, 2016). Chen et al. (2015) proposed a dynamic multi-pooling convolutional neural network (DMCNN) for trigger and argument classification. (Björne and Salakoski, 2018)

used dependency parsers to obtain features for a CNN model ensemble, while [Li et al. \(2020a\)](#) demonstrated the effectiveness of using tree-structured Long Short-Term Memory (LSTM) with knowledge bases for event extraction in the biomedical domain. Besides error propagation, another major drawback of the pipelined approach is the inability to adequately capture the relationships and interactions between the trigger and argument extraction tasks. As a result, the multi-staged execution of the subtasks makes it difficult to capitalize on their interdependencies.

**Joint-based** event extraction aims to address the shortcomings of pipeline-based methods by jointly predicting event triggers and arguments. Multistage sequential pipelines are prone to error propagation because independent classifiers are used for trigger and argument prediction ([Li et al., 2013](#)). The simultaneous classification of triggers and arguments, as shown in [Figure 2.4](#), minimizes the propagation of errors between the subtasks and facilitates the exploitation of the subtasks' interdependence.

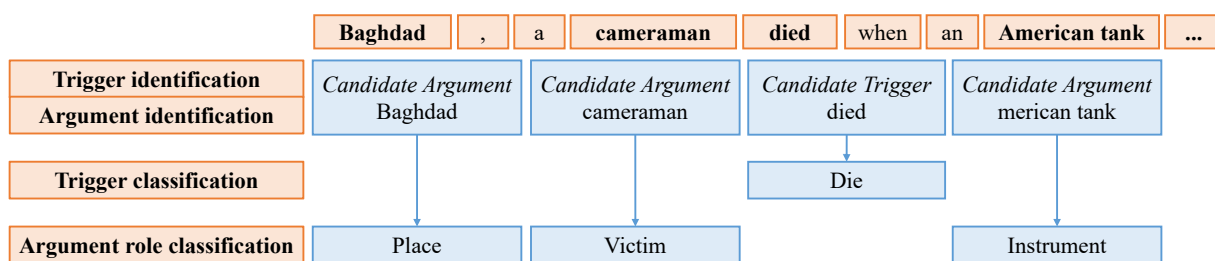


Figure 2.4: Joint-based Event Extraction. Source: [Li et al. \(2021\)](#)

Previously proposed joint event extraction approaches include methods based on Markov Logic Networks ([Poon and Vanderwende, 2010](#); [Riedel et al., 2009](#); [Venugopal et al., 2014](#)), dual decomposition [Riedel et al. \(2009\)](#); [Riedel and McCallum \(2011a,b\)](#) and structured perceptron ([Li et al., 2014, 2013](#)). Markov Logic, a general-purpose statistical relation learning language, was applied to learn a joint model of molecular event structure from given sentences ([Riedel et al., 2009](#)). Also based on Markov logic is work that jointly predicts events, arguments, and the dependency edges that compose the argument paths ([Poon and Vanderwende, 2010](#)). [Riedel and McCallum \(2011a\)](#) presented a joint model, which first jointly extracts triggers and arguments, then captures the correlations between events and provides a mechanism to ensure consistency between arguments for the same event.

Joint extraction of triggers and corresponding arguments using a structured perceptron model

was shown to improve performance significantly (Li et al., 2013). The beam search-based model learned both local and global features. Local features captured the characteristics of the trigger and argument role labeling tasks separately, whereas global features captured the interdependencies between event triggers and argument roles. A related study proposed a unified information network representation to extract events, entities, and relations using one single joint model based on structured prediction (Li et al., 2014). In addition to recording significant performance gains in event-argument extraction, the approach demonstrated the importance of FrameNet-based features in dealing with the sparsity problem in event trigger labeling. FrameNet is a lexical resource for semantic frames, which are conceptual structures that describe an event, relation, or entity and its participants (Baker and Sato, 2003). While English event extraction dominates existing research, (Chen and Ng, 2012) investigated a wide range of rich knowledge sources for Chinese event extraction that encode knowledge ranging from the character level to the discourse level.

Joint neural network models have been developed to reduce the complexity of manual feature engineering associated with traditional machine learning-based event extractors (Liu et al., 2018b; Nguyen et al., 2016; Zhao et al., 2021). Nguyen et al. (2016) presented a joint framework for learning rich text representations and jointly predicting event triggers and argument roles based on bidirectional recurrent neural networks (RNN). A memory matrix stored information regarding the interdependencies between the event triggers and arguments. The interdependence was beneficial to joint event modeling. (Liu et al., 2018b) proposed an attention-based Graph Convolutional Network (GCN) for joint extraction of multiple event triggers and arguments. Similarly, Zhao et al. (2021) developed an end-to-end framework for jointly extracting document-level biomedical events, which used dependency-based GCN and hypergraph to capture local and global contexts in biomedical documents, respectively.

### 2.2.3 Classification by Model Type

Statistical-based methods for event extraction can also be classified into machine learning and deep learning.

**Machine learning-based** methods were proposed to address the shortcomings of pattern-based

methods, which are highly dependent on the creation of event patterns. The main tasks in these methods include the selection of features and training of appropriate event classifiers. Lexical features, which include morphology features and part-of-speech (POS) tags, and contextual features, which include local information (sentence level), global information (document level), and external dictionaries, are the most common types of features.

The common statistical machine learning models used to build event classifiers include the maximum entropy model (Lee et al., 2015), hidden Markov model (Jiangde et al., 2007), conditional random field model (Llorens et al., 2010) and the support vector machine (SVM) model (Björne and Salakoski, 2011; Huang and Riloff, 2011; Saha et al., 2011). Björne and Salakoski (2011), for example, used SVMs to extract biomedical events, which are characterized by descriptions of biomolecular interactions, from research articles in a pipelined manner. Sakaki et al. (2010) uses an SVM classifier that was trained for real-time detection of earthquake events.

Patwardhan and Riloff (2009) proposed a model that applies Naive Bayes for plausible role-filler recognition and considers both Naive Bayes and SVMs for the sentential event recognition task. In this unified probabilistic model, the plausible role-filler recognizer identifies candidate event arguments from noun phrases, whereas the sentential event recognizer identifies the event mentions, that is, sentences in which the relevant event is present. Ahn (2006) proposed a modular system for ACE event extraction based on a combination of machine learning approaches. The extraction task was decomposed into a series of classification subtasks, including event anchor identification, argument identification, attribute assignment, and event co-reference. The task decomposition facilitated the evaluation of the contribution of each subtask to the overall task. The system achieved an F1 score of 60.1 on the event anchor detection task, which was the basis for the event mentions. Lexical, wordNet-based, and dependency features all contributed to the overall performance.

Although classical machine learning models have been largely successful, complicated feature engineering remains a significant bottleneck. Feature engineering is typically a manual process that necessitates linguistic intuition and domain knowledge. As a result, the performance and adaptability of the models based on manual feature extraction to a wide range of domains and languages are limited. Contrastingly, deep learning models automatically extract useful and

meaningful features from data, thus alleviating the problems of manual feature engineering.

**Deep learning-based** methods have gained prominence due to their ability to automatically and effectively extract crucial features from text. However, a key question has been how to design efficient deep learning-based event extraction systems, with researchers pursuing various approaches. In the context of event extraction, a neural network takes word embeddings as input, determines whether a given word is an event trigger or an event argument, and finally classifies them into event type or argument role. Word embedding techniques map a word or a phrase in the vocabulary into a low-dimensional and real-valued vector. The meaning of words is encoded such that words with similar meanings occur closer to each other in the vector space (Erk, 2009; Harris, 1954; Rubenstein and Goodenough, 1965; Sahlgren, 2008). Various neural network architectures have been considered for EE, including convolutional neural networks (Boros et al., 2021; Chen et al., 2015; Nguyen and Grishman, 2015), recurrent neural networks (Nguyen et al., 2016) and graph convolutional networks (Lai et al., 2020b; Nguyen and Grishman, 2018; Veyseh et al., 2019). Inspired by the success of BERT (Devlin et al., 2018), pre-trained language models (PLMs) have also been applied to event extraction (Tong et al., 2020; Wadden et al., 2019; Wang et al., 2019b,c ; Yang et al., 2019b). In particular, Nguyen and Grishman (2015) employs convolutional neural networks (CNNs) for domain adaptation and event detection. The CNN model demonstrated effectiveness and robustness across domains by learning rich features from pre-trained word embeddings, position embeddings, and entity type embeddings. Furthermore, because the feature extraction is automatic, the method avoids error propagation by reducing reliance on supervised preprocessing toolkits for features.

A Dynamic Multi-Pooling Convolutional Neural Network (DMCNN) that evaluates and extracts lexical-level and sentence-level features from the sentence was proposed by (Chen et al., 2015) to deal with multi-event sentences. Multi-event sentences are sentences that contain numerous events that share arguments. A skip-gram word model is used to capture lexical-level features from sentences. On the other hand, a dynamic multi-pooling layer encodes sentence-level features by returning the maximum value in each part of the sentence based on event triggers and arguments. Both the works by Nguyen and Grishman (2015) and Chen et al. (2015) only model the consecutive  $k$ -grams by concatenating the vectors of the  $k$  consecutive words in the sen-



tences.

Alternative approaches consider performing convolutional operations on all possible non-consecutive k-grams that could be necessary for trigger word prediction (Lei et al., 2015; Nguyen and Grishman, 2016). Nguyen and Grishman (2016) models the relative distances of words to trigger candidates in sentences using positional embeddings. On the other hand, Lei et al. (2015) aggregates the absolute distances between words in k-grams. A related study proposed a Dual-CNN technique to model event structures related to crisis situations (Burel et al., 2017). The model included a semantic layer to capture contextual information and extracted fine-grained crisis event information from social media data with 61% F-measure. Li et al. (2018) built a parallel multi-pooling convolutional neural network (PMCNN) for biomedical events, which used dependency-based embeddings to capture semantic and syntactic features from sentences. Experimental results show the method achieved 80.27% F-measure on the trigger identification task and 59.65% on the overall biomedical event extraction task. Despite the numerous efforts, CNN models fail to capture potential interdependencies between distant words and are thus incapable of efficiently dealing with long text sequences.

Recurrent neural networks (RNNs) were considered for event extraction due to their representational power and ability to capture long-term dependencies, which allow them to process arbitrary length sequences. (Ghaeini et al., 2018; Nguyen et al., 2016). Among the RNN-based models is a bidirectional recurrent neural network for trigger and augment role prediction developed by (Nguyen et al., 2016). The model uses memory matrices to store the prediction information during text labeling. Ghaeini et al. (2018) built a language-independent forward-backward recurrent neural network (FBRNN) to detect event nuggets from text. As a further enhancement, Sha et al. (2018) proposed a dependency bridge recurrent neural network (dbRNN). The dependency bridge played the role of connecting syntactically related words. Tree-structured long short-term memory network (Tree-LSTM) models (Tai et al., 2015) that exploited the syntactic dependency tree of a sentence were applied to extract financial events from news text (Duan et al., 2018) and for biomedical event extraction (Li et al., 2020a).

Other RNN variants, in addition to LSTM, have been evaluated for event extraction. (Zhang et al., 2019b) trained two bidirectional simple recurrent unit models (Bi-SRU): one for learning

word-level and another for character-level representations. [Nguyen and Nguyen \(2019\)](#) developed a bidirectional gated-recurrent unit (Bi-GRU) network model for jointly predicting entity mentions, event triggers, and arguments. Another category of models for event extraction used graphs. ([Ahmad et al., 2021](#)) created a Graph Attention Transformer Encoder (GATE) to learn the long-range dependencies and apply the dependencies in cross-lingual relation and event extraction. Similar work by [Liu et al. \(2018b\)](#) introduced an attention-based Graph Convolutional Network (GCN) that was able to extract multiple event triggers and arguments. More recently, [Zhao et al. \(2021\)](#) used a dependency-based GCN network to build a model that better captured the complex relationships between local and global contexts in biomedical documents.

Following the success of [PLMs](#) on a wide range of NLP tasks in the recent past, some studies have leveraged pre-trained models for event extraction ([Wang et al., 2020](#); [Yang et al., 2019b](#)). [Yang et al. \(2019b\)](#) presents the pre-trained Language Model-based Event Extractor (PLMEE), a framework consisting of a combination of extraction and generation models that rely on the feature representation of BERT ([Devlin et al., 2018](#)). The method can generate additional training data by defining adjunct tokens, which comprise all words in a sentence except the triggers and arguments. ([Wang et al., 2019b](#)) proposed a method that combines BERT representation and adversarial learning to explore weakly-supervised data for event trigger extraction. In addition to enhancing distantly supervised learning, the approach automatically constructs diverse and additional training data for semi-supervised event extraction.

([Wang et al., 2019c](#)) designed a hierarchical modular event argument extraction model that uses both CNNs and BERT to learn features. ([Zhang et al., 2019a](#)) introduced a generative adversarial imitation learning method based on ELMo [Peters et al. \(2018b\)](#) for joint entity and event extraction that was able to handle harder-to-extract events. The extended Dynamic Graph IE (DyIE++) ([Wadden et al., 2019](#)) framework extends a previous span-based IE method (DyIE) ([Luan et al., 2019](#)) for entity and relation extraction to include the event extraction task. The method uses BERT to encode candidate text spans, captures both local (within-sentence) and global context across sentences, and shares span representations across the tasks ([Wadden et al., 2019](#)).

### 2.2.4 Classification by Degree of Supervision

This section discusses related work on supervised, semi-supervised, and unsupervised machine learning methods for event extraction.

**Supervised learning-based** methods require human-annotated data to train event extraction models (Chen et al., 2015; Hong et al., 2011; Ji and Grishman, 2008; Lai et al., 2021a; Li et al., 2014; Nguyen et al., 2016). The methods assume a set of target event types and their corresponding event annotations are provided to train models that can efficiently generalize to unseen data. Several supervised learning-based methods for event detection have been proposed, including feature-based methods (Ahn, 2006; Hong et al., 2011; Huang and Riloff, 2011, 2012a,b; Ji and Grishman, 2008; Li et al., 2013; Liu et al., 2016; McClosky et al., 2011; Yang and Mitchell, 2016) that leverage manually designed features to detect event triggers and their types, and representation-based approaches (Chen et al., 2015; Ghaeini et al., 2018; Lin et al., 2018; Nguyen et al., 2016; Nguyen and Grishman, 2015) that rely on neural networks to automatically learn effective features from data. The methods achieve high performance due to their ability to model complex hidden interactions in data.

The performance of supervised event extraction methods is highly dependent on the amount of training data available. In real-world applications, in addition to new events emerging frequently, there are various event types, each with its own set of annotation rules. However, obtaining sufficient annotations is challenging and a key impediment to the portability and performance of event extraction systems. Furthermore, existing event extraction datasets, such as ACE (Grishman et al., 2005) and TAC-KBP (Mitamura et al., 2015), cover a limited number of events and are biased towards common event types, such as the attack and die events.

Furthermore, the identification of event instances primarily focused on monolingual clues of a specific language, while ignoring massive amounts of information provided by other languages. Some feature-based (Ji, 2009; Li et al., 2012; Wei et al., 2017) and representational-based (Li et al., 2012; Liu et al., 2018a) methods have been proposed for multilingual and cross-lingual event extraction. The Gated MultiLingual Attention (GMLATT) framework Liu et al. (2018a) used an attention mechanism to simultaneously address data scarcity and monolingual ambiguity problems in event detection tasks.

**Semi-supervised** methods, motivated by the need to address the shortcoming of supervised approaches, seek to minimize the amount of labeled data required to train models so that only a small set of labeled data is needed (Ferguson et al., 2018; Patwardhan and Riloff, 2007; Riloff, 1996; Yangarber et al., 2000). Semi-supervised methods reduce the amount of data annotation effort required by allowing the labeled data to be augmented with large amounts of unlabeled data to improve performance. Previous research includes a bootstrapping approach proposed by Yangarber et al. (2000) that extracts events by applying a set of seed patterns and incrementally and automatically discovering new patterns from the unlabeled text. Unlike the work by Riloff (1996), no pre-classified corpora with relevance judgments were provided in advance, so the system needed to discover documents that belong to the relevant set, which contains events of interest. A similar study by (Patwardhan and Riloff, 2007) performs event extraction by utilizing only a handful of seed patterns and a set of relevant and irrelevant documents. The extraction process involves first identifying relevant sentences in a document that describe an event, followed by applying extraction patterns to the identified sentences. Boros (2018) investigated a method for increasing training data by using dictionaries to generate morphological derivations and inflections.

Another line of research addresses the problem of data scarcity in event extraction by generating additional training examples through self-training, which involves training a classifier on a few labeled examples and then utilizing the classifier to predict an additional set of training instances from the unlabeled text. Ferguson et al. (2018) exploited self-training and took advantage of multiple mentions of the same event instances across multiple articles to extract events. In this method, generating additional data involved first identifying news article clusters referring to the same event, followed by detecting the events present in each cluster. Finally, articles in other clusters are scanned based on a given article in a cluster, with the most representative trigger in each article for the given event type selected. Wang et al. (2019b) used a set of words that serve as triggers in the gold dataset to identify unlabeled examples that could be expressing the same event. Before combining the extracted unlabeled examples, the authors applied an adversarial training mechanism to denoise and identify the most appropriate instances. A similar study by Munkhdalai et al. (2015) combined active learning and self-training for biomedical event extraction.

Some studies have investigated zero-shot and few-shot transfer learning strategies that require no or few training examples to adapt EE models to new domains and tasks without retraining the model. [Huang et al. \(2018\)](#) applied zero-shot transfer learning to predict unseen event types based on the structured definition of the unseen types and annotations of a few seen event types. Few-shot transfer learning, which allows generalization to new classes from only a small number of labeled examples, was also investigated for event extraction ([Liao and Grishman, 2010, 2011](#)). [Deng et al. \(2020\)](#) trained an event-type learner for few-shot event detection with dynamic memory networks using a few annotated training examples as an instantiation of the few shot learning. The training involved learning meta knowledge from event types and using the knowledge to predict new event types not seen during training. [Lai et al. \(2020a,b\)](#) proposed two metric-learning-based few-shot learning methods for event classification and event detection. The former extensively exploit the support set (consisting of examples from a small set of classes) during the training process to provide more training signals for the model, while the latter experiments with a variety of configurations and different encoder types (CNN, LSTM, and GCN). [Tuo et al. \(2022\)](#) proposed a prototypical network with a BERT encoder for event detection. More specifically, the authors optimized the use of the information contained in the different layers of a pre-trained BERT model. They demonstrate that simple strategies for combining BERT layers outperform the current state-of-the-art for event detection.

**Unsupervised learning** methods do not require event-annotated data and instead take an open domain approach to event extraction. The extraction of triggers and arguments is primarily based on the clustering of similar event instances and the detection and tracking of topics ([Chambers and Jurafsky, 2011](#)). The method obtained an F1-score of 40 when evaluated on the MUC-4 terrorism dataset ([Sundheim, 1991](#)). [Naughton et al. \(2006\)](#) proposed a method based on hierarchical agglomerative clustering to generate sentence clusters representing groups of sentences in a news article that refer to the same event. Vectorizing the sentences was accomplished using the bag-of-words encoding scheme. Besides word and sentence embeddings, the work by [Ribeiro et al. \(2017\)](#) considered additional information such as time, location, and content dimensions to enrich the text representation before applying the Markov clustering algorithm to cluster news articles describing the same event.

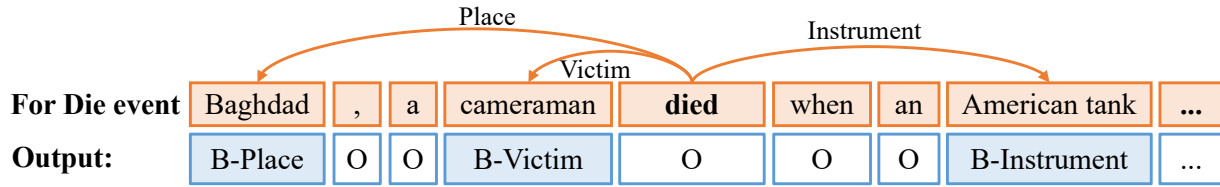


Figure 2.5: Sequence Labeling-based Event Extraction. Source: [Li et al. \(2021\)](#)

### 2.2.5 Classification by Paradigm Type

Event extraction can be modeled as a classification task, a sequence labeling task, or a machine reading comprehension task.

**Classification-based** methods were previously the most widely used approaches to event extraction. The methods transform the event extraction problem as a classification problem, with the goal of locating and categorizing event triggers and arguments into predefined event types  $n$  and argument roles  $[r_{i,1}, r_{i,2}, \dots, r_{i,l}]$  respectively, that characterize the event  $e_i$  ( $i \in [1, n]$ ) ([Ahn, 2006](#); [Chen et al., 2015](#); [Ghaeini et al., 2018](#); [Li et al., 2013](#); [Liu et al., 2016](#); [Nguyen et al., 2016](#); [Nguyen and Grishman, 2015](#)). Despite many advances, classification-based methods require large amounts of training data to achieve desirable performance ([Chen et al., 2017](#); [Li et al., 2013](#); [Liu et al., 2018a](#)) and are incapable of effectively dealing with unseen event types ([Huang et al., 2017](#)).

**Sequence labeling-based** methods tag the tokens in the text to produce the output. Thus, given the text, the argument role corresponding to each argument is labeled, as shown in [Figure 2.5](#). The BIO (Beginning, Inside, Outside) annotation scheme is the most commonly employed scheme in sequence labeling. Some studies formulate the event extraction task as a sequence labeling problem. [Lu and Roth \(2012\)](#) presented a sequence labeling event extraction model based on the latent-variable semi-Markov conditional random fields. The model predicts argument roles given information about the event mentions and types. The study used structured preference modeling to achieve effective learning, which involved assigning preferences to specific structures during the learning process. [Ramponi et al. \(2020\)](#) proposed the Biomedical Event Extraction as Sequence Labeling (BEESL) approach that employs a multi-label aware decoder with BERT to extract biomedical events.

**Machine reading comprehension** (MRC) has also been investigated for event extraction (Boros et al., 2022; Chen et al., 2019b; Du and Cardie, 2020; Li et al., 2020b; Liu et al., 2020). The goal is to extract answer spans (the starting and ending positions of the answer) from the text in response to questions about an event. In the work proposed by Du and Cardie (2020), two BERT-based QA models were developed, one for event trigger detection and the other for argument extraction tasks. Question templates were required to map the input sentence into the standard BERT input format. The models identified event triggers and event types as well as candidate event arguments, which are text spans in the input sentence. A dynamic threshold guides the extraction of arguments such that only arguments with a probability that meets the threshold are extracted. Li et al. (2020b) proposed a multi-turn question answering method that treats the event extraction task as a series of reading comprehension problems, by which triggers and arguments are successively extracted from a given sentence. The system attained an F1 score of 53.4% on the argument extraction task of ACE 2005. The machine reading comprehension-based (MRC) method introduced by Liu et al. (2020) consists of an unsupervised question generation process that transforms event schema into a set of natural questions and a BERT-based question-answering process to retrieve answers as EE results.

## 2.3 Epidemic Event Extraction Approaches

This section focuses on epidemic event extraction, a domain-specific task for extracting health-related events from text. We detail the approaches developed to date for this task, highlighting their strengths and limitations. The ever-increasing frequency and magnitude of disease outbreaks underscore the need for countries globally to develop efficient early warning systems for infectious disease outbreak detection. Moreover, the proliferation of digitized sources over the web and the increased availability of computational resources and techniques such as NLP are crucial factors that have influenced the adoption of data-driven disease surveillance (Paolotti et al., 2014). Public health stakeholders are increasingly adopting early detection systems that monitor multiple data sources to identify health threats. Early, accurate, and reliable identification of signals of public health threats from heterogeneous data sources has become an integral component of epidemic surveillance. Identifying critical information on emerging infectious

diseases from text, also known as epidemic intelligence (EI), allows for more rapid detection of potential health threats and prompt response. Epidemic intelligence describes the systematic and organized collection, analysis, and interpretation of information from diverse data sources to detect, verify, and investigate potential health risks.

The two broad categories of epidemic intelligence are indicator-based and event-based surveillance (Huff et al., 2016). The indicator-based surveillance methods depend on well-defined formal data from healthcare facilities (Dórea and Revie, 2021). As a result, these traditional surveillance methods are regarded as more reliable, as official reports from health facilities undergo rigorous verification before dissemination. While routinely collecting and analyzing data from well-defined sources translates to more reliable surveillance systems, indicator-based surveillance approaches are limited in terms of both promptness and geographical coverage (Brownstein et al., 2008; Jebara and Shimshony, 2006). A drawback of relying on official data sources to detect disease outbreaks is that it is prone to delays. Moreover, indicator-based surveillance is labor-intensive, requiring significant human input to implement. Typically, the approach predominantly relies on local health practitioners to identify infectious disease outbreaks, subject the suspected outbreak cases to laboratory tests for confirmation, and formally report the positive epidemic cases to relevant public health stakeholders.

Event-based surveillance (EBS), which leverages informal sources for disease surveillance, has been espoused as a feasible approach for complementing and addressing the limitations of indicator-based systems. EBS is defined as the organized, rapid capturing, monitoring, analysis, and interpretation of diverse data, mainly unstructured, to detect events of public health importance (Balajee et al., 2021; Hartley et al., 2010; Huff et al., 2016; Idubor et al., 2020; Velasco et al., 2014). In addition to having broad geographical coverage, EBS systems offer a decisive temporal advantage over the indicator-based systems by utilizing informal, unstructured data from online data sources such as news articles, social media, blogs, and discussion forums to detect infectious disease outbreaks (Freifeld et al., 2008; Valentin et al., 2020a; Woodall, 2001).

These digital data sources present numerous sources of real-time disease information that, when effectively harnessed, can help overcome the shortcomings of traditional surveillance methods.



For instance, online news reports that are continuously generated and often in real-time can be processed computationally, diseases detected early, and appropriate responses are taken to curb the further spread. Tracking online news can be a useful supplement to the information acquired from the community, facility, district, and regional components of the EBS system (Balajee et al., 2021). With the advancement of the Internet, news can easily be shared about outbreaks that begin in the most remote parts of the world and spread to countries far away, with global consequences. Such news reports originate from diverse locations globally, facilitating the development of event-based systems that cover a wider geographical area compared to their traditional surveillance counterparts. As a result, EBS holds the promise of complementing the traditional surveillance systems and creating a comprehensive surveillance strategy that significantly improves disease outbreak detection in terms of timeliness and geographical coverage.

Despite the potential benefits of increased digital data availability, the heterogeneous and constantly generated news text can be overwhelming. Devoid of efficient extraction systems, information overload could occur, obscuring critical and urgent details required to deploy appropriate control strategies. Furthermore, text data is associated with various complexities such as ambiguity, redundancy, and noise (Valentin et al., 2021). Another challenge relates to the multilingual nature of the data, particularly in low-resource scenarios where some languages are underrepresented. Natural language processing (NLP) approaches are increasingly being adopted to address these challenges while reducing the need for manual analysis of the continuous stream of free text (Hartley et al., 2010).

Epidemic event extraction systems can be classified into knowledge-based and data-driven systems (Arsevska et al., 2018). **Knowledge-based** systems extract epidemic events by matching text with pre-defined event templates. An example of a knowledge-based system is the ProMED-mail (Program for Monitoring Emerging Disease) (Woodall, 2001)<sup>2</sup> platform that tracks infectious disease outbreaks and acute exposures to toxins globally. Before being published, the PROMED news articles undergo moderation by infectious diseases experts, who review and verify the new reports based on some rules. Another pattern-based system is the Medical Information System (MediSys) (Linge et al., 2010), which is based on the Pattern-based Under-

---

<sup>2</sup> <https://promedmail.org/>

standing and Learning System (PULS) (Du et al., 2016). PULS extracts event attributes by applying patterns to the sentence structure of a news article. PULS can automatically learn new patterns through weakly-supervised learning, reducing the manual effort required to generate the patterns. The BioCaster, like MediSys, uses an ontology and the simple rule language (SRL) to generate matching patterns. The ontology includes epidemiological concepts (e.g., viruses, symptoms) and generic concepts (e.g., locations). However, relying on ontologies limits systems' ability to scale to new domains and languages because ontologies require regular manual updates. In general, pattern-based event surveillance methods are challenging to build and maintain and have poor portability and flexibility. The problem of coverage in pattern-based systems is primarily due to their rigid syntactic structure and limited vocabulary, which is dependent on the availability of expert knowledge (Valentin et al., 2020c). In order to address the limitations of pattern-based systems, data-driven surveillance approaches that use statistical machine learning have been considered.

The **data-driven** methods that use textual data, such as online news text, for surveillance purposes have gained traction among epidemic intelligence researchers. Arsevska et al. (2018) proposed the Platform for Automated extraction of Animal Disease from the Web (PADI-web)<sup>3</sup>, a surveillance system for monitoring online news sources to detect animal disease events. PADI-web retrieves Google news articles, classifies the articles, and extracts epidemiological information such as diseases, dates, symptoms, hosts, and locations (Valentin et al., 2020b). The classification component of PADI-web, which filters relevant news before performing information extraction, is based on supervised machine learning.

The Global Public Health Intelligence Network (GPHIN) is a multilingual surveillance system that collects and monitors news reports pertaining to emerging public health events. GPHIN gathers news documents in ten languages for analysis, which are then machine translated into English. The retrieved articles were assigned a relevance score to determine whether the articles were to be automatically published (high-relevance documents) on the GPHIN database or required further manual verification (low-relevance documents) by domain experts before publication.

---

<sup>3</sup> <https://padi-web.cirad.fr/en/>

**Unsupervised** surveillance systems have also been investigated. Among them is the DANIEL system, which considers text as a sequence of strings and does not depend on language-specific grammar analysis. As a result, the system is easily adaptable to a wide range of languages. The DANIEL extraction pipeline detects relevant documents before extracting event triggers. A similar system is the HealthMap system (Freifeld et al., 2008), which collects and extracts disease outbreak information such as geographic location, time, and infectious agent. Like the DANIEL system, the HealthMap relies on document structure, based on the hypothesis that the most relevant information appears at the beginning of a news report. Epidemiological entities (disease and location) are first searched in the title, then in the beginning paragraphs, and finally in the entire content. These types of systems yield low performance and are inefficient in extracting multiple events from an article with multiple spatial granularity levels.

## 2.4 Conclusions

In this chapter, we explored the existing approaches to event extraction in general and epidemic event extraction in particular. The most successful methods require substantial amounts of annotated data for training and adapting models to a new domain or language. However, most domain-specific event extraction tasks, such as the extraction of disease outbreak events, have few or no available annotations. The available gold-standard annotations cover a limited number of event types and exist for only a few languages, with a high bias towards high-resource languages like the English language. The huge disparity and imbalance between low-resource and high-resource languages poses a challenge to many information extraction tasks and is not an exception for the epidemic event extraction task. A significant performance difference exists between extraction models in high-resource and low-resource language settings, necessitating further research to improve low-resource event extraction.

In addition to data scarcity, existing epidemic event extraction approaches focus on extracting desired events, primarily from the English text. However, the news reports predominantly used in event-based surveillance originate from different parts of the world and a variety of languages and dialects. Creating methods and models that work well across languages remains a challenging task. The mainstreaming of transfer learning and cross-lingual embeddings into NLP tasks

has been significantly successful. In particular, multilingual pre-trained language models such as Multilingual BERT (mBERT) allow for a general-purpose multilingual representation of text, which presents enormous potential for learning across different languages.

While advanced NLP approaches such as deep learning techniques based on pre-trained language models have been successfully applied to a wide range of tasks, minimal work exists on the use of these techniques for the extraction of events in multilingual and low-resource settings. Generally, machine and deep learning-based methods remain largely underutilized in the domain-specific task of epidemic surveillance, which is characterized by the lack of adequate annotated data to train and evaluate the systems (Valentin et al., 2021). To the best of our knowledge, learning approaches based on pre-trained language models which have the ability to learn rich text representation have yet to be adequately explored for the epidemic surveillance domain. We provide a discussion of the pre-trained language models in Section A.2. In addition, existing approaches to epidemic surveillance are predominantly based on pattern matching or classical machine learning approaches. As a result, the systems require significant domain knowledge, and manual feature engineering, have poor cross-lingual applicability, and do not scale well in low-resource settings with constrained labeled training data. Accordingly, there is a need for further development of approaches for efficient extraction of health-related events from unstructured text. This study seeks to advance research in data-driven event surveillance towards attaining optimal implementation of event-based epidemic surveillance systems.

In the next chapter, we describe the dataset used in the various experiments performed in this study, both for supervised and semi-supervised learning experiments. We also provide the definition of an epidemic event and what it entails to extract an epidemic event from the dataset.

---

### Epidemiological Event-based Dataset

---

In this chapter, we describe the dataset used in the experiments performed in this thesis, including the dataset’s annotation process. Despite playing a crucial role in the development and evaluation of event extraction systems, benchmark datasets provided by the ACE ([Doddington et al., 2004](#)) and TAC KBP ([Ellis et al., 2014](#)) evaluation campaigns have various drawbacks. First, due to the inherent complexity of data annotation processing, the existing datasets are small in size and insufficient for training models with satisfactory performance. Second, the benchmark datasets have limited event types and low coverage, limiting their applicability to diverse domains such as epidemic surveillance. Because the datasets are derived from non-epidemiological sources, their suitability for the epidemic event extraction task may necessitate expanding the dataset by incorporating health-related documents with epidemiological entities.

We, thus, present a more relevant and suitable dataset, specific to epidemic surveillance, the multilingual dataset used to evaluate the DANIEL system proposed by [Lejeune et al. \(2015\)](#). However, the dataset has few entries per language, with only the French language having more than 500 documents. In addition, the dataset was annotated at the document level with one

primary event per document, which differs from the typical datasets (sentence or token-level annotations) used in event extraction research, such as the [MUC](#) and [ACE](#) datasets. The document-level dataset was re-annotated and transformed into a token-level dataset, with each token in the text assigned a pre-defined category. The categories are **DIS**, **LOC**, and **O**, denoting disease, location, and other, respectively.

### 3.1 DANIEL Dataset

The DANIEL<sup>1</sup> dataset ([Lejeune et al., 2015](#)) is a multilingual epidemic-based dataset composed of health-related news articles from online news outlets such as Google News and other major newspapers<sup>2</sup>. The news documents are of varying lengths in terms of paragraphs and characters. The dataset assumes a single event type (disease outbreak), and the news headlines provide cues for identifying the main event trigger and its arguments. The document-level annotation of the dataset involved enlisting native speakers of each language represented in the dataset to identify the relevant documents describing an infectious disease outbreak.

```
"15962": {
  "annotations": [
    [
      "Polio",
      "India",
      "unknown"
    ]
  ],
  "comment": "",
  "date_collecte": "2012-01-12",
  "langue": "en",
  "path": "doc_en/20120112_www.smh.com.au_2a21025f6f4dc13c9eb8ebf3d249f3",
  "url": "https://www.smh.com.au/healthcare/polio-is-one-nation-closer-to-being-wiped-out-20120112-1pxho.html"
}
```

Listing 3.1: Example of an event annotated document in the DANIEL dataset.

The annotators were also required to identify the disease name and location that evoked the epidemic event. Therefore, the task of extracting an epidemic event from the document-level

<sup>1</sup> <https://github.com/NewsEye/event-detection/tree/master/event-detection-daniel>

<sup>2</sup> “Gazeta”, “Gazeta polska”, etc. for Polish

annotated dataset entails identifying articles that contain an event described by the disease name and location pairs. The number of outbreak cases, which quantifies the number of victims affected by the disease, was included in some annotations. Listing 3.1 shows an example event annotated document with “Polio” as the disease name and “India” as the location. The number of outbreak cases (victims) in this document is unknown. The DANIEL corpus and associated annotation guidelines are publicly available on the DANIEL website<sup>3</sup>.

### 3.1.1 DANIEL Epidemiological Event Definition

An epidemiological event is defined as an incident of public health importance, such as infectious disease outbreaks, chemical spills, or radiation leaks (Balajee et al., 2021). The events are described by a set of epidemiological entities (attributes) such as disease names, locations, dates, hosts, and the number of cases cited in the text (Valentin et al., 2017). Although the levels of analysis performed differ, existing event-based surveillance systems extract at least the disease name. The BioCaster (Collier et al., 2008), GPHIN (Mawudeku and Blench, 2005) MedISys (Linge et al., 2010), and PADI-web (Arsevaska et al., 2017; Valentin et al., 2017) systems extract symptoms in addition to the disease name. PADI-web can also detect the host species and the number of cases using regular expressions.

In this dataset, an epidemic event is described by the disease name and location of the reported epidemic. Therefore, extracting an event from the news text involves finding the relevant documents and extracting the disease name and location. Relevant documents are the news articles containing mentions of disease outbreaks. The temporal aspects of epidemic events were not considered in this study but can be explored in future work. As illustrated in the extract<sup>4</sup> from a relevant document in the DANIEL dataset in Figure 3.1, the mentioned disease is “Polio”, which was reported in “Howrah” region near “Kolkata” in “India”.

In contrast to other event extraction datasets such as the ACE 2005 dataset, which defines multiple event types, the DANIEL dataset defines a single type of event (disease outbreak), with disease name and location as the event arguments. While extracting a single event type is a

---

<sup>3</sup> <https://daniel.greyc.fr/public/index.php?a=corpus>

<sup>4</sup> The article 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>.

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.

Figure 3.1: 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>.

relatively simple task, the extraction of epidemic events in our study is more difficult due to the multilingual and low-resource nature of the dataset.

### 3.1.2 DANIEL Languages

The epidemiological corpus comprises news articles from diverse language families: Germanic: English (en), Hellenic: Greek (el), Romance: French (fr), Slavic: Russian and Polish, and Chinese from the Sino-Tibetan family. The language families belong to the broader Indo-European family, which is (arguably) the most well-studied language family and consists of a number of the highest-resource languages in the world. English and French, being Indo-European languages, are similar and share some basic typological traits like their subject-verb-object word order. They share a fair amount of vocabulary because of historical language contacts ([Finkenstaedt and Wolff, 1973](#)). While the languages in the corpus are generally considered high-resource languages with sufficient data resources available, all but the French language had fewer than 500 training examples (of which less than 40 were relevant) in the DANIEL dataset. As a result, we considered these languages as low-resource in the various experiments performed in our study.

The **Greek** language, whose writing system is the Greek alphabet, is an independent branch of the Indo-European family of languages, with a morphology that depicts an extensive set of productive derivational affixes. On the other hand, **Russian** (East-Slavic) and **Polish** (West-Slavic) are Balto-Slavic languages with highly inflectional morphology, which can cause serious problems when detecting nouns or entities (disease names and locations, in our case). In these



languages, there are cases where it is impossible to determine the base form of an entity without appealing to data such as verb-frame subcategorization, which is usually beyond the scope of tasks such as ours. These languages are morphologically rich, with substantial information about syntactic units and relations expressed at the word level (Tsarfaty et al., 2010). Notably, in the DANIEL dataset, the Greek, Russian, and Polish languages have only about 38% lexical overlap.

### 3.1.3 Extending the DANIEL Dataset

Among the models evaluated in this research are text classification models, focused on finding the relevant documents (those reporting disease outbreaks) from the news text. These models, especially neural network-based models, require substantial amounts of data to train and evaluate. Motivated by this, we extended the DANIEL dataset by including additional news documents for the English language, increasing their number from 474 to 3,562, as shown in Table 3.1. The news articles were obtained from a variety of online news sources, with articles relevant to disease outbreaks primarily obtained from the Program for Monitoring Emerging Diseases (ProMED)<sup>5</sup> platform (Carrion and Madoff, 2017). The ProMED program is an initiative of the International Society for Infectious Diseases<sup>6</sup> that tracks infectious disease outbreaks and acute exposures to toxins globally. The platform is supported by a global network of experts who identify and share disease outbreak reports. Before being published on the platform, moderators review and validate the reports.

Table 3.1: Dataset statistics for the extended dataset used for the document classification task.

Language	#Documents	#Sentences	#Tokens
English (en)	3,562	117,190	2,692,942
French (fr)	2,415	70,893	1,959,848
Polish (pl)	341	9,527	151,901
Russian (ru)	426	6,865	133,905
Chinese (zh)	446	4,555	236,707
Greek (el)	384	6,840	183,373

<sup>5</sup> <https://promedmail.org/>

<sup>6</sup> <https://isid.org/>

The first step in the data collection process was the retrieval of ProMED news articles published between August 1, 2013 and August 31, 2019. While some of the original files were no longer available online, approximately 80% of the articles were retrieved. The retrieved documents were processed and saved in JSON format, allowing the corpus to be easily reused and reproduced. The title, description of the reported disease, location, date, and source Uniform Resource Locator (URL) are all annotated in the articles. Using the source URLs, where the article was originally published, the corresponding source documents were downloaded, forming the relevant documents of the dataset. K-means clustering was used for language filtering to ensure that only documents belonging to the languages of interest were retained. Further text preprocessing tasks include removing boilerplate content from downloaded HTML documents such as navigation links, headers, and footers. De-duplication<sup>7</sup> was also required to eliminate perfect duplicate and near-duplicate content, ensuring only a single instance of each document was preserved.

On the other hand, the irrelevant<sup>8</sup> news articles consist of general health-related news but without direct or indirect mentions of disease outbreaks (e.g., “plague”, “cholera”, “cough”), as well as general news like politics and sports. The news articles cover a wide range of topics, including culture, politics, wellness, healthy living, sports, and entertainment. Finally, having collected a total of 7,574 articles, we split the data into training, validation, and testing sets. The training set had a total of 5,074 documents, while the remaining documents were divided evenly between the validation and the testing sets, that is, 1,250 documents for validation and 1,250 documents for testing, stratified by language, as shown in Table 3.2. These data splits were used for training and evaluation of classification models, as described in Chapter 4.

### 3.1.4 Token-level DANIEL Data Annotation

This section describes a token-level annotated dataset (Mutuvi et al., 2021) for epidemic event extraction. Due to the lack of dedicated datasets for multilingual epidemic event extraction, we adapted the DANIEL dataset (Lejeune et al., 2015). However, the DANIEL dataset, in

---

<sup>7</sup> De-duplication was achieved using the Onion (ONE Instance ONLY) tool: <https://corpus.tools/wiki/Onion>

<sup>8</sup> Most of the irrelevant documents were obtained from the News Category Dataset (Misra, 2018) comprising HuffPost<sup>9</sup> news articles for the period 2012 to 2018.

Table 3.2: The number of documents (percentage of relevant documents) per dataset split. The dataset consists of both relevant and irrelevant documents and was used for the document classification task.

	All	Polish	Chinese	Russian	Greek	French	English
Train	5,074 (10.8)	241 (7.4)	300 (2.6)	296 (9.45)	253 (6.7)	1,593 (10.9)	2,365 (11.7)
Validation	1,250 (10.9)	54 (7.4)	71 (2.8)	60 (10.0)	68 (10.2)	388 (13.4)	583 (12.6)
Test	1,250 (10.5)	46 (13.0)	75 (6)	70 (10.0)	63 (4.7)	434 (12.4)	614 (12.8)

its original form, was annotated at document level, which distinguishes it from typical datasets (token or word level annotations) used in research for the event extraction task (e.g., ACE 2005<sup>10</sup>, TAC KBP 2014-2015<sup>11</sup>). An epidemiological event in this dataset is represented by a disease name and the location of the reported event. Thus, a document is either reporting an event of interest (disease and location name appear in a relevant document) or it is not (an irrelevant document). An example relevant document contains 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 [ . . . ]*. In this case, the document is annotated with “Tuberculosis” as the disease name and “India” as the location of an epidemic event.

The annotation process began with sentence segmentation to obtain individual sentences from the text corpus. For annotation, we chose to use Doccano<sup>12</sup>, a collaborative annotation tool that provides annotation features for various tasks, including text classification, sequence-to-sequence tasks, and sequence labeling. Besides being an open source platform, Doccano provides an intuitive and user-friendly interface that allows faster annotation of datasets. We defined entity types (DIS, LOC, and O) to represent disease, location, and other tokens. Furthermore, we defined the annotation guidelines, which required annotators to identify and mark the entity span from the text. An epidemic event is distinguished by mentions of the disease name and the location of the disease outbreak. Therefore, the annotators were required to identify and label disease and location mentions present in the text using DIS and LOC tags, respectively. Three annotators, who are native speakers of their respective languages, were selected for each

<sup>10</sup><https://catalog ldc.upenn.edu/LDC2006T06>

<sup>11</sup><https://catalog ldc.upenn.edu/LDC2020T13>

<sup>12</sup><https://github.com/doccano/doccano>.

language.

After completing the annotation, we converted the annotations into IOB (Inside, Outside, Beginning) tagging scheme. For example, based on the spans, each token of a disease name is assigned the tags B-DIS, I-DIS, and O, indicating the beginning (B-), intermediate (I-), and out-of-span markers (O) for a disease. We also computed the Inter-Annotator Agreement (IAA) using Cohen's kappa coefficient (Cohen, 1960), for annotation quality check. By measuring how well multiple annotators make the same annotation, the IAA demonstrates the reliability of the annotation process. Table 3.3 presents the per language inter-annotator agreement score, with the average IAA being 0.66. A kappa value in the range 0.61 - 0.80 is categorized as substantial agreement (Landis and Koch, 1977). Therefore, we posit that our token-level annotated multi-lingual epidemiological dataset<sup>13</sup> (Mutuvi et al., 2021) is highly reliable as the ground truth.

Table 3.3: Inter-annotator agreement score (Kappa coefficient) per language for the token-level corpus

Language	Score
French	0.857
English	0.807
Chinese	0.301
Polish	0.523
Greek	0.815
Russian	0.683
<b>Average</b>	<b>0.660</b>

Finally, we randomly split the token-level annotated dataset into training, validation, and testing sets based on the 80:10:10 ratio. The statistics for the entire dataset (relevant and irrelevant) are presented in Table 3.4 while Table 3.5 shows the statistics for only the relevant documents in the dataset.

<sup>13</sup>The token-level annotated dataset is available at <https://doi.org/10.5281/zenodo.6024726>.

Table 3.4: Statistics of the token-level dataset. The terms DIS and LOC represent the number of disease and location mentions, respectively.

	<b>Partition</b>	<b>Documents</b>	<b>Sentences</b>	<b>Tokens</b>	<b>Entities</b>	<b>DIS</b>	<b>LOC</b>
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

## 3.2 Conclusions

This chapter described the DANIEL dataset, a multilingual epidemiological dataset used in this research. The dataset, which consists of online news articles, was originally annotated at the document level. We re-annotated the dataset at the token level, as is typical for event extraction datasets. According to the statistics in Tables 3.2 and 3.4, the DANIEL dataset presents various

Table 3.5: Number of tokens and sentences for the relevant documents per language.

<b>Split</b>	<b>Sentences</b>	<b>Tokens</b>	<b>French</b>	<b>English</b>	<b>Polish</b>	<b>Chinese</b>	<b>Greek</b>	<b>Russian</b>
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

particularities and challenges. Besides being multilingual, the dataset is also imbalanced, with only around 10% of the documents containing events. This simulates the real-world scenario of news reporting, where only a small portion of the reported news is relevant to epidemiology. The number of documents in each language is relatively balanced, except for French, which has approximately five times the number of documents as the other languages. Besides the availability of many annotators for the French language, the other reason for the large number of documents for this language relates to the ease of obtaining French news articles. While the other languages required crawling and collecting articles from Google news, sufficient epidemiological news reports were readily available for the French language, in a database containing facts extracted from ProMED-Mail<sup>14</sup> (Yangarber et al., 2005). In addition to being a relatively small dataset, the majority of the languages present in the dataset are regarded as low-resource languages. Polish, Greek, and Russian are morphologically rich, with significant information about syntactic units and relations expressed at the word level (Tsarfaty et al., 2010).

In the next chapter, we use the token-level dataset created in this chapter to develop supervised baseline models for epidemic event extraction. Such baselines are currently unavailable for the sequence-labeling tasks related to the extraction of epidemiological events. Various machine and deep learning models described in Appendix A were evaluated. The proposed machine and deep learning-based sequence labeling baselines could be applied in subsequent epidemic event extraction research.

---

<sup>14</sup><https://promedmail.org/>

## CHAPTER 4

---

# Supervised Learning for Multilingual Epidemic Event Extraction

---

This chapter proposes sequential labeling baseline models for epidemic event extraction. We explore the applicability of various supervised learning algorithms for event-based surveillance using the token-level annotated training data presented in Chapter 3. We compare the supervised learning approaches to the DANIEL system, a dedicated, unsupervised multilingual system for epidemic event extraction. Furthermore, we evaluate models based on pre-trained language models that have been successful on a wide range of NLP tasks (Qiu et al., 2020). However, these methods have received little attention in the specialized domain of epidemic surveillance. More specifically, we aim to answer the following questions: 1) To what extent do conventional machine and deep learning models improve the performance of event extraction in specialized domains? 2) How applicable are pre-trained language models to the epidemiological event extraction task?

Supervised learning methods depend on labeled data, with predictor features and known targets (labels), to train models (Rao and Gudivada, 2018). The trained models are then evaluated on

the unlabeled test dataset to determine their performance in categorizing the data into predefined classes (Uddin et al., 2019). Models judged as achieving desired performance make predictions within an acceptable range on the test set. Supervised machine learning methods have dominated a wide range of tasks, including disease risk prediction (Uddin et al., 2019). Uddin et al. (2019) compared the performance of different supervised machine learning algorithms, including Support Vector Machine (SVM), Naive Bayes, Random Forest, Artificial Neural Network (ANN), and Logistic Regression, on a task that categorizes patients as either low-risk or high-risk. The health data used to train the models was obtained from the Scopus and PubMed databases, with only articles written in English being considered. The Random Forest algorithm demonstrated superior performance comparatively, attaining overall best performance in 53% of the evaluations it was considered in, followed by SVM with 41%.

Data-driven disease surveillance is a multistage task that includes document classification and epidemic event extraction (Joshi et al., 2019), as shown in Figure 4.1. The document classification task distinguishes between irrelevant and relevant disease outbreak-related documents. The event extraction task extracts disease names and their reported locations from the news text. Three different scenarios were considered when evaluating the models: For the document classification and epidemic event extraction tasks in the first scenario, we use all data instances comprising both relevant and irrelevant documents. The second scenario entailed extracting events from the predicted relevant documents returned by the document classification step. Finally, we performed event extraction using only the ground-truth relevant documents.

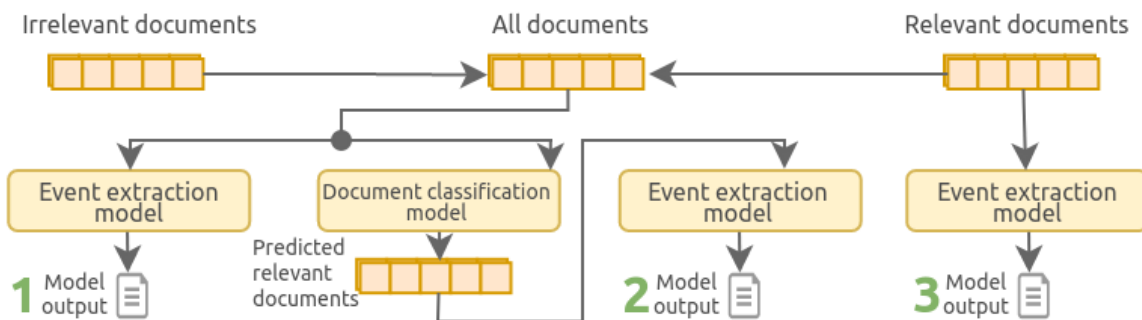


Figure 4.1: Illustration of the types of experiments carried out: (1) using all data instances (relevant and irrelevant documents), (2) testing on the predicted relevant documents provided by the document classification step, (3) using only the ground-truth relevant documents.



## 4.1 Document Classification

With a significant number of news reports being generated continuously, selecting the relevant documents is critical for efficient epidemic event extraction. Relevant documents describe a specific disease event and related information such as prevention and control measures (Valentin et al., 2020b). The purpose of the document classification phase is to filter and select the relevant news reports from the continuous text stream. We hypothesize that prioritizing the selection of relevant documents enhances the performance of epidemic event extraction models. The assumption made was that extraction of epidemiological entities (e.g., disease name and location) on pre-classified documents encounters minimal noise (errors) compared to when using all text instances (a combination of both relevant and irrelevant documents).

### 4.1.1 Experimental Setup

We compared the performance of a specialized baseline system, machine learning, and deep neural network models in classifying news documents into relevant or irrelevant classes. Relevant news articles mention and provide details regarding a disease outbreak, whereas the rest of the documents are irrelevant. A description of the neural networks models investigated in this study, which include Convolutional, Recurrent, and Transformer-based neural networks, is provided in Appendix A.

**Baseline System.** We chose the DANIEL system (Lejeune, 2013; Lejeune et al., 2015), a multilingual epidemic event extraction system with a complete unsupervised learning pipeline that detects relevant documents before extracting event triggers. By adopting the inverted-pyramid writing style used in journalistic writing, the system takes advantage of the unique role of document structure (Piskorski et al., 2011). The most important parts of the story are placed at the beginning of the document, while the least important facts are placed at the end. The DANIEL system does not use language-specific grammar analysis and treats text as a sequence of strings rather than words. Consequently, the system can be easily adapted to operate in any language and extract crucial information early on, significantly improving decision-making processes. This is critical in epidemic surveillance since timeliness and geographical coverage are crucial, and

more often than not, initial medical reports are in the vernacular language where patient zero appears (Lejeune et al., 2015). We did not evaluate similar systems such as the BIOCASTER because only the ontology is publicly available and covers a limited number of languages. Similarly, for the PULS (Von Etter et al., 2010) extraction system, its IE component utilizes a large set of linguistic patterns that depend on a large scale public health ontology.

**Machine Learning models.** We sought to investigate supervised text classification models on multilingual datasets with varying data characteristics, such as ours. Classification models take an input vector comprising  $n$ -features and map features to associated target values (class labels) (Kirasich et al., 2018). We specifically evaluated the Logistic Regression (LR) (Wright, 1995), Random Forest (RF) (Breiman, 2001), Support Vector Machine (SVM) (Cortes and Vapnik, 1995; Wang, 2005), and Naive Bayes (NB) (Lewis, 1998; Rish et al., 2001) models using their default hyperparameters. The first three models are classified as discriminative, while Naive Bayes is a generative model. Discriminative models make no assumptions about the probability distributions underlying a dataset (Jebara, 2012). They determine, from the input, the most useful features to find decision boundaries and discriminate between the various possible classes. Generative models, in contrast, make assumptions about the underlying probability distributions and are capable of generating new data points (Jebara, 2001). The Naive Bayes method makes the simplified (naive) assumption that input variables are conditionally independent (Lewis, 1998). While assumptions can greatly simplify the learning process, they could limit the learning abilities of models.

Despite the three selected discriminative classifiers sharing a similar learning approach, the models exhibit a varying degree of architectural complexity. As a parametric model, Logistic Regression has a finite number of (fixed-size) parameters (weight coefficients), which are independent of the number of training instances. In contrast, random forest and non-linear SVMs, which can learn non-linear decision surfaces, are considered non-parametric learning algorithms because the number of parameters changes as the training data size changes. The SVM utilizes a maximum margin hyperplane to separate the classes found in the training examples (Cortes and Vapnik, 1995). The random forest is an ensemble of decision trees that use the bagging concept (bootstrapping followed by aggregation) (Breiman, 2001). Bootstrapping involves training sev-

eral individual decision trees in parallel on various subsets of the training dataset using different subsets of available features. Therefore, each individual decision tree in the random forest is distinct, lowering the overall variance of the **RF** classifier. The classifier’s final decision is formed by aggregating the decisions of individual decision trees. This significantly reduces the risk of overfitting, effectively improving the generalization ability of **RF** models.

**Deep Learning models** have the advantage of being able to learn features from text automatically, eliminating the need for manual feature engineering (Liang et al., 2017). We evaluated various deep learning models for document classification, which are described in Appendix A, beginning with a CNN and a BiLSTM with FastText (Joulin et al., 2016) word representations. We considered FastText embeddings for English, French, Polish, Russian, Chinese, and Greek languages, with an embedding dimension of 300. For the CNN, which is described in Section A.1.1, a sequence of word embeddings was passed through a convolution of kernel size 3 and filter size 250, while the Bidirectional LSTM (**BiLSTM**) passed the word embeddings through a bi-directional LSTM with a cell size of 128. The LSTM is a type of a RNN model whose architecture is explained in Section A.1.2. Additional hyperparameters for the models included a batch size of 32, a learning rate of  $1 \times 10^{-2}$ , and a total of 15 epochs with an early stopping of 3 to avoid overfitting. We also test a graph convolutional networks (GCN) based-approach that augments BERT with graph embeddings (Lu and Nie, 2019). Combining the capabilities of BERT with GCNs has been shown to be effective in capturing both local and global information.

Furthermore, we chose to conduct experiments with various BERT-based architectures for the sequence classification task presented by Devlin et al. (2018). We used the default hyperparameters; a learning rate of  $2 \times 10^{-5}$ , and a maximum length of 512 tokens, with the longer sentences truncated to the defined maximum length. In particular, the pre-trained BERT models used were the `bert-base-multilingual-cased` and `uncased`. Finally, we also evaluated the CNN/BiLSTM described earlier in this section, but this time utilizing BERT features.

### 4.1.2 Results and Analysis

Choosing the most appropriate performance metrics is critical, especially when dealing with class-imbalanced datasets such as ours. Standard evaluation metrics (e.g., accuracy) fall short

of being sufficiently reliable or could even be misleading in the presence of a skewed class distribution (Chawla, 2009; Hossin and Sulaiman, 2015). Among the most appropriate evaluation metrics when data is imbalanced include precision, recall, and F1-score, which we chose for our study. Precision measures the proportion of relevant (positive) predictions that were correct, while recall is the fraction of relevant documents that were correctly predicted (Powers, 2020). In the context of epidemic surveillance, measuring recall is crucial because of the risk posed by failure to identify all the relevant cases of disease outbreaks. The F1-score represents the harmonic mean of precision and recall values.

Besides appropriate metrics, the choice of the most suitable document classification approach is less clear, especially in multilingual settings where imbalances manifest due to differences in data sizes among the represented languages. In such circumstances, classical machine learning and deep learning approaches present valid options. Typically, deep learning approaches require large amounts of data to train models that achieve desired performance. However, the models tend to perform dismally when only a few labeled examples per class exist, typically 100 to 1000, as is the case of the low-resourced languages present in the dataset used in this study (LeCun et al., 2015).

In order to determine the most appropriate model for the classification of epidemic-related documents in multilingual and low-resource settings, we experimented with different machine learning and deep learning models, using the dataset configurations presented in Table 3.2. The evaluation results are shown in Table 4.1. From the results, we observed that the performance of SVM, in terms of the F1-score, was slightly higher compared to LR, RF, and NB. The RF attained the highest precision (95.70%) not only among the classical machine learning models but also the highest among all models tested. However, there was a substantial difference between the precision and recall values generated by the machine learning models (LR, RF, SVM, NB). The models registered high precision and low recall values, which could be detrimental to the interests of an epidemiological detection system because such systems could fail to capture crucial information regarding emerging and evolving epidemic events. Compared to the baseline results provided by the DANIEL system, we noted the recall was higher than precision, demonstrating the specialized nature of the system, despite having the lowest recall of all the

Table 4.1: Evaluation scores of the analyzed models for the relevant documents for all languages. The models evaluated, using the dataset configurations presented in Table 3.2, were Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB). Additionally, pre-trained BERT<sub>base-multilingual</sub> models were evaluated, both the †fine-tuned and models trained on BERT features.

Models	Precision %	Recall %	F1 %
DANIEL	33.9	60.61	43.48
LR	93.81	68.94	79.48
RF	<b>95.70</b>	67.42	79.11
SVM	91.26	71.21	80
NB	84.55	70.45	76.86
CNN+FastTtext	86.11	70.45	77.5
BiLSTM+FastTtext	77.44	78.03	77.74
BERT (cased) <sup>†</sup>	88.62	82.58	85.49
CNN+BERT (cased) <sup>†</sup>	88.79	71.97	79.5
BiLSTM+BERT (cased) <sup>†</sup>	90.20	69.70	78.63
BERT (uncased) <sup>†</sup>	84.67	87.88	<b>86.25</b>
CNN+BERT (uncased) <sup>†</sup>	82.14	87.12	84.56
BiLSTM+BERT (uncased) <sup>†</sup>	83.72	81.82	82.76
BERT (cased)	80.71	85.61	83.09
CNN+BERT (cased)	86.67	78.79	82.54
BiLSTM+BERT (cased)	75.95	<b>90.91</b>	82.76
BERT (uncased)	88.52	81.82	85.04
CNN+BERT (uncased)	86.07	79.55	82.68
BiLSTM+BERT (uncased)	81.51	73.48	77.29
VGCN+BERT	87.18	77.27	81.93

evaluated methods.

On the other hand, the models based on either a CNN or a BiLSTM with FastText embeddings had lower F1-scores than the classical machine learning methods (LR, RF, SVM, NB). This could be attributed to the small volume of training data available, which was insufficient to train models with the ability to better distinguish between relevant and irrelevant documents. Typically, training neural network models requires a large number of labeled examples, which is

challenging to acquire in specialized domains such as epidemiological surveillance (Yang et al., 2019d). In such scenarios with limited labeled instances, transfer learning could provide a feasible solution that minimizes the need for extensive data labeling. Therefore, we sought to explore deep transfer learning where the BERT pre-trained language model was either fine-tuned or used directly as a feature extractor for the text classification task. The results show that the BERT-based models achieved a higher F1-score compared to all other models. Furthermore, we see that the BERT-based models were able to balance recall and precision (precision remains consistent despite the increase in recall). Besides having been trained on massively large corpora, the robustness of BERT can also be attributed to the wordpiece-based sub-word tokenization method (Schuster and Nakajima, 2012), which enables BERT to handle out-of-vocabulary (OOV) words with significant success. Further examination of BERT revealed that fine-tuning BERT on the epidemic text classification task yielded slightly better performance than when the BERT encoder was used only for feature extraction. However, adding a CNN or LSTM layer on top of the BERT encoder resulted in a notable performance decline. Overall, the findings indicate that deep learning approaches can learn rich representations, thus allowing the models to more effectively utilize previously learned features for newer documents, even when the languages of the documents differ.

A comprehensive analysis of the per language performance revealed that the discriminative machine learning models (LR, RF, and SVM) exhibited similar trends in their unequal performance across the languages, failing to detect (having F1-score of zero) the relevant documents in Polish and Chinese, as presented in Table 4.2. In contrast, the Naive Bayes (NB) model, which is a generative model, was able to detect some of the relevant documents in the languages (Polish, Chinese, and Greek languages) that the discriminative models were unable to detect, albeit with a lower F1-score when evaluated on all languages combined. The inability of the discriminative models to detect relevant documents could be due to the small size of the training data for these specific languages. Similarly, the CNN-based and BiLSTM-based models with pre-trained FastText embeddings were unable to correctly classify the relevant documents for the low-resource languages (Polish, Chinese, Russian, and Greek). The F1-scores for Chinese tend to be consistent for all BERT-based models, while the performance for Polish remarkably varies from one model to the other. VGCM+BERT had the highest F1-scores for the majority of

the low-resourced languages, namely, Polish, Chinese, and Russian, and the second highest for Greek.

Table 4.2: F1-scores of the analyzed models for the relevant documents per language. The models evaluated, using the dataset configurations presented in Table 3.2, were Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB). Additionally, pre-trained BERT<sub>base-multilingual</sub> models were evaluated, both the †fine-tuned and models trained on BERT features.

Models	Polish	Chinese	Russian	Greek	French	English
DANIEL	40	80	33.33	33.33	71.43	32.23
LR	0	0	66.67	66.67	84.21	80
RF	0	0	40	66.67	86.84	78.83
SVM	0	0	33.33	0	87.18	81.38
NB	33.33	66.67	75.00	66.67	82.76	75.76
CNN+FastText	0	0	0	0	84.21	81.88
BiLSTM+FastText	0	0	0	0	73.12	85.71
BERT (cased) <sup>†</sup>	50	80	66.67	66.67	<b>94.12</b>	82.89
CNN+BERT (cased) <sup>†</sup>	50	80	66.67	40	86.05	86.75
BiLSTM+BERT (cased) <sup>†</sup>	0	80	40.00	66.67	87.36	86.27
BERT (uncased) <sup>†</sup>	57.14	80	50	<b>100</b>	91.95	86.08
CNN+BERT (uncased) <sup>†</sup>	50	80	66.67	40	86.05	86.75
BiLSTM+BERT (uncased) <sup>†</sup>	0	80	40	66.67	87.36	86.27
BERT (cased)	33.33	80	50	66.67	87.50	85.54
CNN+BERT (cased)	0	0	40	66.67	83.33	86.45
BiLSTM+BERT (cased)	0	80	22.22	28.57	85.11	<b>88.37</b>
BERT (uncased)	0	66.67	85.71	66.67	87.18	86.25
CNN+BERT (uncased)	0	50	40	66.67	82.35	86.45
BiLSTM+BERT (uncased)	0	0	33.33	0	72.94	84.42
VGCN+BERT	<b>71.43</b>	<b>88.89</b>	<b>88.89</b>	80	87.80	78.26

#### 4.1.2.1 Effect of Article Structure

Motivated by the journalistic writing style (inverted pyramid format) of news articles, in which the most important, newsworthy information appears at the beginning of an article, we investi-

gate the influence of article structure on the performance of document classification models. The work by [Lejeune \(2013\)](#) regarded a document as the primary unit with language-independent organizational properties. The assumption is that the document-detectable features at the document granularity offer robustness at the multilingual scale. The author suggests using the text as a minimal unit of analysis beyond its relation to the genre from which it came. The press article follows precise rules: the structure of the press article and the vocabulary used are established, and there are well-defined communication aims known to the source as well as the target of the documents. These rules, at a higher level than the grammatical rules, are very similar in different languages, and from the knowledge of these rules, remarkable positions are defined that are independent of languages. To exploit particular zones of news article content, we perform experiments similarly to [Lejeune et al. \(2010\)](#) and [Lejeune \(2013\)](#), inspired by the work on genre invariants done by [Giguet and Lucas \(2004\)](#) and [Lucas \(2009\)](#). The following are the various sections of the text that were examined:

- Beginning of the article: ideally composed of the title of the article
- Beginning of the text body: containing the first two paragraphs
- End of the text body (foot): comprising the last two paragraphs
- Rest of the text body: made up of the rest of the textual elements (e.g., paragraphs)

The results, as presented in [Table 4.3](#), indicate that the combination of the beginning and ending text of news articles provided the best features for classifying news documents as relevant or irrelevant to a disease outbreak. When evaluated independently, the body and conclusion scored the lowest, while utilizing the beginning text only provided the highest performance. This is consistent with the inverted pyramid news writing style, in which the most important information about the reported disease outbreak regularly appears at the beginning of relevant documents.

#### 4.1.2.2 Effect of Training Data Size

Previous studies show that the performance of models scales with increasing training data size ([Lei et al., 2018](#)). To investigate the impact of data size on the news document classification task, we split the training data at an interval of ten percent. Subsequently, we evaluated the model



Table 4.3: Performance based on the sections of the documents using fine-tuned multilingual BERT (`uncased`) and VGCN with BERT. All positions of text have a limit of 512 tokens.

Text Position	Models	Precision %	Recall %	F1 %
Beginning	VGCN+BERT	<b>87.18</b>	77.27	81.93
	BERT ( <code>uncased</code> ) <sup>†</sup>	84.67	87.88	86.25
Body	VGCN+BERT	79.83	71.97	75.70
	BERT ( <code>uncased</code> ) <sup>†</sup>	75.71	80.30	77.94
End	VGCN+BERT	72.93	73.48	73.21
	BERT ( <code>uncased</code> ) <sup>†</sup>	76.12	77.27	76.69
Beginning+End	VGCN+BERT	86.61	83.33	84.94
	BERT ( <code>uncased</code> ) <sup>†</sup>	85.61	<b>90.15</b>	<b>87.82</b>

with the best overall performance, that is, the fine-tuned BERT (`multilingual-uncased`) model, to determine how different data sizes affect performance. As illustrated in Figure 4.2, there is a general upward trend, with the F1-score performance improving when trained on increasingly large datasets. The model achieves an F1-score performance that is comparable to that of the classical machine learning models when only 10% of the data is used and plateaus at 30% of the data. It is worth noting that the model achieves F1-score of 64.03 while using only 5% of the training data, which is an impressive performance given the small sample size.

#### 4.1.2.3 Zero-shot Transfer Learning

We also investigated zero-shot transfer learning, in which a model is first trained on the source language, and then evaluated directly on multiple target languages that were unseen during the training stage (Lauscher et al., 2020; Mutuvi et al., 2020; Norouzi et al., 2013; Wang et al., 2018; Wang et al., 2019d; Wu and Dredze, 2019). The aim was to determine how well a model trained with text from one language could predict the target text in another language without having seen any of the candidate labels. In addition, the evaluation sought to demonstrate the extent to which languages with a greater number of documents (French and English, approximately 2000 news articles) influence the classification of low-resource languages. We considered each language as the source language and the other five languages as target languages. At every iteration, the fine-tuned BERT (`multilingual-uncased`) model, which was the best performing model from

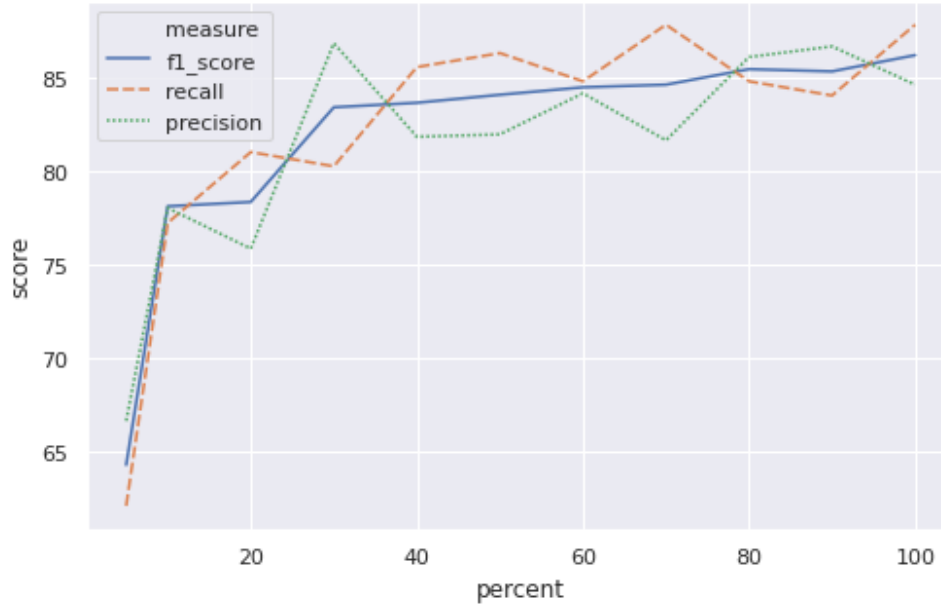


Figure 4.2: Impact of data size on performance of the best performing model: fine-tuned BERT (multilingual-uncased).

our previous experiments, was trained on the data in the source language and applied directly to every target language. For instance, taking the English language as the source language, the fine-tuned BERT (multilingual-uncased) was trained on English text and tested on text in other languages (Polish, Chinese, Russian, Greek, and French). As shown in Table 4.4, the performance of models trained on the English and French documents is consistently higher than models trained in the other languages. The Greek language was only able to predict target text in English and Greek but was unable to classify unseen classes from the other languages.

#### 4.1.2.4 Discussion and Conclusion

With the rapid and high volume of online news generated daily, identifying relevant articles that report events of public health importance can be challenging. Document classification using learning-based approaches can facilitate the timely identification of disease outbreaks at scale. Given that disease outbreaks are reported from different parts of the world and in different languages, successful epidemic surveillance necessitates systems that are not only accurate but also have broad geographical coverage. Therefore, we extensively evaluated various models to determine their effectiveness in multilingual online news text classification. Accuracy and

Table 4.4: Evaluation scores of the BERT (multilingual-uncased)<sup>†</sup> model fine-tuned on the relevant documents in a zero-shot transfer learning setting.

<b>Train \ Test</b>	<b>Polish</b>	<b>Chinese</b>	<b>Russian</b>	<b>Greek</b>	<b>French</b>	<b>English</b>
<b>Polish</b>	40	0	66.67	66.67	76.92	85.71
<b>Chinese</b>	0	80	60	0	70.97	81.08
<b>Russian</b>	33.33	0	33.33	66.67	62.86	88.61
<b>Greek</b>	0	0	0	66.67	0	63.05
<b>French</b>	0	66.67	57.14	0	91.95	85.90
<b>English</b>	50	0	33.33	66.67	39.29	84.35

scalability are essential attributes of data-driven epidemic surveillance systems, hence the need to develop NLP methods that achieve high performance and scale across multiple languages.

Pre-trained models based on BERT outperformed the other models in both recall and F1-score. The high performance can be attributed to the use of large corpora to pretrain the Transformer-based models, which allows the models to learn universal language representations that are beneficial to downstream tasks. Notably, the fine-tuned BERT models performed better than the feature-based CNN and BiLSTM classifiers that utilized FastText and BERT embeddings as input features. The high precision and low recall observed for machine learning models suggest that the models were unable to detect the relevant class well but are highly reliable when they do. Therefore, while the classifiers provided reliable results, the machine learning models had a high false-negative rate, resulting in only a subset of all relevant results being returned. The approaches based on fine-tuned BERT uncased achieved a good balance of precision and recall.

Regarding the performance of individual languages, the performance of models on English and French documents was consistently higher than that of any of the models trained on text in the other languages. This can mainly be attributed to the larger quantity of annotated data (> 2000 documents) for the two languages compared to the other languages. Also, English typology more closely resembles French typology as it has had more recent influence from French and other Romance languages. The two languages share lexical similarities and cognate words. While Russian and Polish are both Slavic languages, we noticed that the performance varied

greatly in the case of Polish and less in the case of Russian. Considering the quantity of training data, the difference of only about 50 more documents in the training set for Russian compared to Polish seems to influence the performance. VGCN+BERT performed particularly well for Polish, Chinese, and Russian. The model utilizes graph embeddings produced by integrating local information captured by BERT and global information from the vocabulary graph, which is based on word co-occurrence information. During the learning process, both local and global information interact with each other via a self-attention mechanism. The interaction introduces useful global information to BERT that helps improve performance across all languages, including those with limited resources.

In terms of the impact of various document segments on performance, the results show that the combination of the beginning and ending text yielded the highest recall and F1-score. This was particularly true for BERT-based models such as VGCN+BERT and the fine-tuned BERT model. The fact that the first paragraphs of an article frequently capture the most important information about the story and the last section usually serves as a summary of the article provides useful information that improves the model's performance.

The performance of the model improved proportionately with the size of the training data. Neural network models, in particular, require large amounts of data to train and evaluate. The competitive performance recorded by pre-trained language models, despite fine-tuning the models on limited labeled training data, could be due to the ability of the models to transfer knowledge obtained from pre-training on large-scale training data to the specific task of classifying epidemic text. This was also evident in the results of zero-shot transfer learning, which involves transferring knowledge from the source language to the target language. Therefore, transfer learning can benefit the process of extracting useful information from multilingual epidemiological texts.

In conclusion, the evidence presented in this work suggests that the models based on fine-tuned language models and/or graph convolutional networks achieved very good performance (> 90%) on the classification of multilingual epidemiological text, not only for high-resource languages but also for low-resource languages. The next section delves into the task of epidemic event extraction in similar settings (multilingual and low-resource settings).

## 4.2 Epidemic Event Extraction

This section discusses the extraction of epidemic events, another subtask of the multi-step event-based surveillance process. While document classification is important in filtering the most relevant news articles, epidemic event extraction models the disease outbreak at a finer-grained (more granular) level, identifying more specific disease outbreak information. Concretely, this entails detecting all the occurrences of the disease name and the locations of the reported events in online news text. For example, given an excerpt from an English article, shown in Figure 3.1, an epidemiological event extraction system should then detect the “polio” as the disease name (DIS) along with the mentioned locations (LOC), “India, Howrah, and Kolkata”. In this case, the disease and location pair represent the epidemic event reported in the news article. We formulate the epidemic event extraction task as a sequence labeling task that assigns a label to every token in the input sequence. We train models to predict the text tokens using the token-level annotated dataset described in Table 3.4

### 4.2.1 Model Selection and Evaluation

We selected the DANIEL system (Lejeune et al., 2015) described in Section 4.1.1 as the baseline model for epidemic event extraction. The baseline model was compared to CNN and LSTM neural network models and also to models based on pre-trained language models, namely BERT (Bidirectional Encoder Representations from Transformers) and XLM-Roberta<sup>1</sup> (Conneau et al., 2020). The pre-trained models, which are generally easy to fine-tune, performed impressively on the document classification task and were deemed suitable for token classification. More specifically, we considered the two versions of multilingual BERT pre-trained language models, BERT-multilingual-cased<sup>2</sup> and BERT-multilingual-uncased<sup>3</sup>. Due to the multilingual nature of our dataset described in Section 3.1.4, these models, along with XLM-Roberta models, were considered appropriate for the epidemic event extraction task. The predicted relevant

---

<sup>1</sup> XLM-RoBERTa-base was trained on 2.5TB of newly created clean CommonCrawl data in 100 languages.

<sup>2</sup> <https://huggingface.co/bert-base-multilingual-cased>. This model was pre-trained on the top 104 languages having the largest Wikipedia articles using the masked language modeling (MLM) and Next Sentence Prediction objectives.

<sup>3</sup> <https://huggingface.co/bert-base-multilingual-uncased>. This model was pre-trained on the top 102 languages having the largest Wikipedia articles using the masked language modeling (MLM) and Next Sentence Prediction objectives.

documents used in the second scenario of our experiments were obtained using the fine-tuned BERT-multilingual-uncased model, which was the best performing classifier overall in the classification experiments performed in Chapter 4. The model achieved F1-score of 86.25% on the document classification task when using training data from all the languages. The performance of the model in the individual languages for the relevant documents was 57.14% (Polish), 80% (Chinese), 50% (Russian), 100% (Greek), 91.95% (French), and 86.08% (English).

We evaluated the epidemic event extraction models at two levels: coarse-grained and fine-grained. For the coarse-grained evaluation, the entity was the reference unit (Makhoul et al., 1999), with the models evaluated quantitatively using holistic metrics, namely precision (P), recall (R), and F1-measure (F1). We focused on the micro-level F1-score, which takes into account label imbalance, as well as all error types across all documents. Therefore, the micro-level evaluation is considered more appropriate in settings such as ours where there is a class imbalance. In addition to holistic evaluation, described in Section 4.2.2, a fine-grained evaluation was performed that comprised model-wise and attribute-wise analysis, as discussed in Sections 4.2.3 and 4.2.4.

## 4.2.2 Holistic Analysis

When evaluating the models using the ground-truth relevant documents (scenario 3), the task is relatively easier and has significantly higher precision than the other scenarios, as shown in Table 4.5. When we tested on the predicted relevant documents (scenario 2), we noticed a significant drop in precision and F1-score. The amount of errors propagated to the event extraction step is substantial, which reduces the F1-scores by more than 20 percentage points for all the models. Following the classification step, the ratio of relevant instances to retrieved instances is altered, significantly reducing the number of relevant examples. As a result, we observed that not only does the F1-score drop substantially across all models but also the precision in comparison with the ground-truth results. The decrease in precision is due to several relevant documents being discarded following the classification phase. When all data instances were considered (scenario 1), we observed that the BERT-multilingual-uncased model attained the highest F1-score of 80.99%. The neural network models improved significantly over the se-

lected baseline (DANIEL), with a performance gain of approximately 30 percentage points for the lowest performing model (BiLSTM+CNN), using all the data instances (relevant and irrelevant documents), as shown in Table 4.5.

It is worth noting that the BERT-multilingual-uncased model had the highest precision, recall, and F1-score when using the predicted relevant documents (scenario 2). However, we observed a reduction in the overall performance when using the predicted relevant documents after the classification task, compared to when using all data instances without document filtering (scenario 1). As a result, we decided to continue the per-language and fine-grained evaluations (model-wise and attribute-wise analysis) using the initial dataset comprising both relevant and irrelevant documents (used in scenario 1). The model-wise and attribute-wise analyses are presented in subsections 4.2.3 and 4.2.4, respectively.

The results of the per-language evaluation, shown in Table 4.6, revealed that BERT-multilingual-uncased obtained the highest scores for three out of the four low-resource languages, while BERT-multilingual-cased was better suited for Polish. The higher results for the low-resource languages, namely, Greek, Chinese, and Russian, could be explained by considering the experiments described in the paper by [Conneau et al. \(2020\)](#) that introduced the XLM-RoBERTa model. The authors observed that by initially pre-training XLM-RoBERTa on a relatively small number of languages (between 7 and 15), the model was able to take advantage of positive transfer, which improved performance, particularly for low-resource languages.

When the number of languages increases, the *curse of multilinguality* ([Conneau et al., 2020](#)) occurs and triggers per-language capacity dilution that manifests in the degradation of overall downstream performance for low-resource languages as more similar high-resource languages are added during pre-training. The decrease in per-language capacity ultimately degrades the performance when all languages are jointly evaluated. Therefore, a trade-off between positive transfer and capacity dilution is required for optimal performance. Increasing model capacity can help alleviate the *curse of multilinguality* problem, thus positively influencing the performance of models, especially for the low-resource languages ([Conneau et al., 2020](#)).

Table 4.5: Evaluation results for the detection of disease names and locations on all languages and all data instances (relevant and irrelevant documents).

<b>Models</b>	<b>P</b>	<b>R</b>	<b>F1</b>
DANIEL Baseline	38.97	47.32	42.74
<b>Relevant and irrelevant documents (scenario 1)</b>			
BiLSTM+LSTM	79.68	70.07	74.57
BiLSTM+CNN	73.38	71.00	72.17
BERT-multilingual-cased	80.66	79.72	80.19
BERT-multilingual-uncased	82.25	<b>79.77</b>	<b>80.99</b>
XLM-RoBERTa-base	<b>82.41</b>	76.81	79.52
<b>Predicted relevant documents (scenario 2)</b>			
BiLSTM+LSTM	53.35	87.40	66.26
BiLSTM+CNN	50.84	86.18	63.95
BERT-multilingual-cased	52.13	89.43	65.87
BERT-multilingual-uncased	<b>53.66</b>	<b>92.28</b>	<b>67.86</b>
XLM-RoBERTa-base	53.10	90.65	66.97
<b>Ground-truth relevant documents (scenario 3)</b>			
BiLSTM+LSTM	<b>91.32</b>	85.38	88.25
BiLSTM+CNN	87.29	84.45	85.85
BERT-multilingual-cased	85.40	<b>90.95</b>	88.08
BERT-multilingual-uncased	87.16	89.79	88.46
XLM-RoBERTa-base	88.53	89.56	<b>89.04</b>

Table 4.6: Evaluation scores (F1) of the analyzed models for the predicted relevant documents per language, found by the classification model.

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

### 4.2.3 Model-wise Analysis

With different models performing differently on different datasets, we sought to go beyond the holistic score assessment (entity F1-score) and compare the strengths and weaknesses of



the models at a fine-grained level. This was achieved through the use of Upset plots (Lex et al., 2014) to visualize the individual performance of the models and the intersections of their predicted outputs. The Upset plot generalizes a Venn diagram by indicating the overlapping sets with filled dark circles at the bottom and the size of the intersections with the bar charts. Elements not included in any of the sets are represented by non-filled (light-gray) circle. For example, as seen in Figure 4.3 (a), there are approximately 70 positive instances that none of the systems was able to find. In the same diagram, the highest intersection represents the true positives jointly found by the three systems and consists of approximately 340 instances.

BERT-multilingual-cased was able to find a higher number of unique true positive instances, instances not detected by the other models. The BERT-multilingual-uncased had the second-highest number of distinct true positives and the fewest false positives. This demonstrated the ability of the BERT-multilingual-uncased model to find the relevant examples in the dataset and to correctly predict a large proportion of the relevant data points, hence the high recall and precision, and overall F1 performance. The overall performance is generally impacted by the equally higher number of false positive and false negative results, as presented in Figures 4.3 (b) and 4.3 (c). The XLM-RoBERTa-base model had the highest false negative rate and the lowest number of true positive instances, which explains the low recall and F1 scores for this particular model.

#### 4.2.4 Attribute-wise Analysis

For the attribute-wise analysis, we adopted an evaluation framework for interpretable evaluation<sup>4</sup> of the named entity recognition (NER) task (Fu et al., 2020a) that proposes a fine-grained analysis of entity attributes and their impact on the overall performance of information extraction systems. The attributes characterize the properties of an entity, such as entity length, which may be correlated with performance. The framework defines a wide range of entity attributes, namely, entity length (eLen), sentence length (sLen), entity frequency (eFreq), token frequency (tFreq), out-of-vocabulary density (oDen), entity density (eDen) and label consistency. Label consistency describes the degree of label agreement for an entity in the training set. We con-

---

<sup>4</sup>The code (Fu et al., 2020a) is available here: <https://github.com/neulab/InterpretEval>.

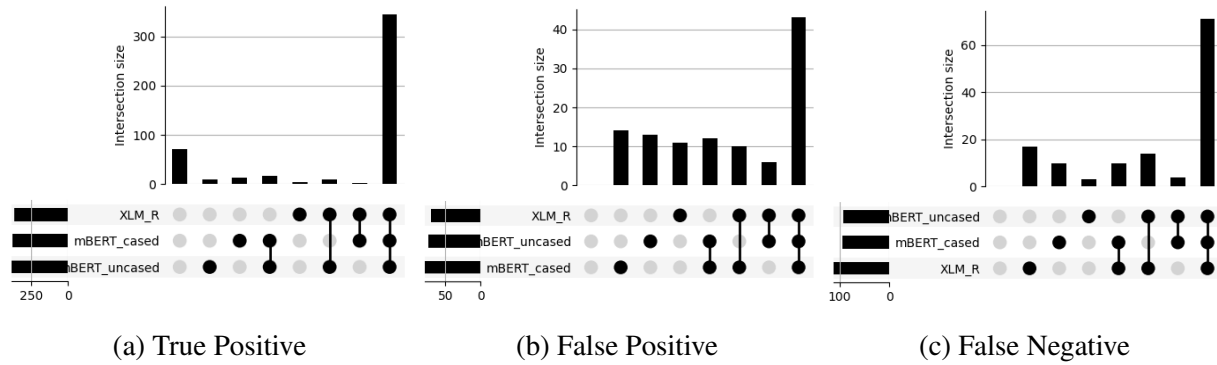


Figure 4.3: Intersection of models predictions. The figures represent (from left) the true positive, false positive, and false negative intersection sizes. The x-axis is interpreted as follows; from left to right, the first bar represents the number of instances that no system was able to find, the next 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.

sider entity and token label consistency, denoted as eCon and tCon, respectively. eCon and tCon quantify how consistently a given span (entity or token) is labeled with a particular label across the dataset. The attribute-aided analysis involved a bucketing procedure that decomposed the holistic performance into different portions (Fu et al., 2020a,b ; Neubig et al., 2019). This was realized by partitioning the test entities into different buckets and computing the corresponding performance for each bucket. Entities in different buckets could yield different performance scores on average. For instance, lower performance could be observed for entities with more words compared to those represented by fewer or atomic words. Using the default parameters as specified by the evaluation framework, we partition the entities into  $m=4$  discrete parts. For example, for entity length (eLen), entities in the test set with lengths of (1, 2, 3) and greater than 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 in each bucket.

Table 4.7 presents the results, which show that for our dataset the performance of all models varies considerably and is highly correlated with oDen, eCon, tCon, and eLen. The results indicated that the prediction of epidemic event is impacted most 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

Table 4.7: Attribute-wise F1-measures (%) 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			<b>2.48</b>	<b>12.97</b>	16.69	<b>31.14</b>	4.48	5.76	<b>4.99</b>	<b>36.80</b>
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	<b>34.77</b>
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	<b>17.32</b>	6.86	<b>4.83</b>	<b>6.61</b>	3.70	<b>36.30</b>

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, resulting in an F1 score of 0. 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), shown in Figure 4.4. This is an indication that this model is not only the best performing, but it is also the most stable and recommendable.

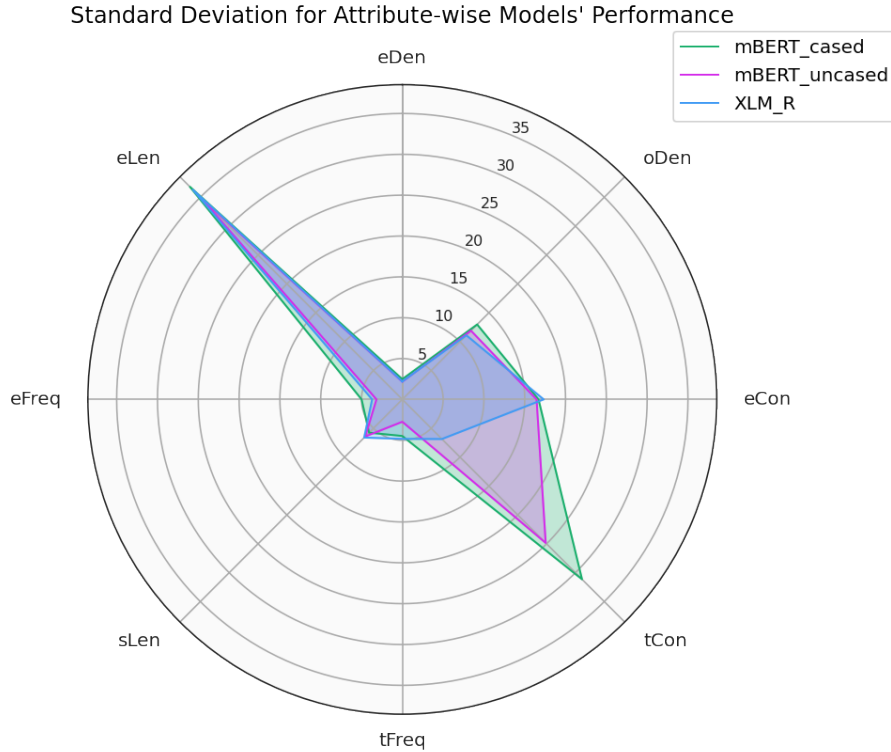


Figure 4.4: Attribute-wise analysis of performance

### 4.3 Conclusions

In this chapter, we developed and evaluated supervised learning techniques for epidemic event extraction. Concretely, we formulated the extraction task as a sequence labeling task and proposed machine and deep learning models for the task. Besides producing superior results than the specialized DANIEL system, the results from our experiments show that building surveillance systems through supervised learning can produce competitive results in both the tasks of relevant document classification and epidemic event extraction, particularly when sufficient annotated data is available. More importantly, approaches based on pre-trained language models were found to produce the best overall performance in both the document classification and event extraction tasks. However, poor performance was observed for the under-resourced languages (Polish, Greek, Chinese, and Russian) with limited annotated training examples. As observed in Figure 4.2, the performance increases with the increase in data size. Therefore, an easier approach to scaling and improving the performance of the systems would be to obtain additional human-annotated data. However, this is time-consuming and expensive, particularly

in specialized domains where annotators need to be domain experts.

Alternative approaches for dealing with labeled data scarcity in the epidemiological domain are investigated in subsequent chapters of this thesis, namely, domain adaptation, which is discussed in the next chapter, and semi-supervised learning, which is described in Chapter 6. Domain adaptation entails injecting domain-specific knowledge into the vocabulary of generic (general domain) pre-trained language models, while semi-supervised learning leverages both labeled and unlabeled data to train and evaluate models. A model trained on the labeled data is used to generate labels for the unlabeled examples, which are then added to the set of labeled examples, effectively increasing the labeled data size.

## CHAPTER 5

---

# Domain Adaptation for Low-resource Epidemic Event Extraction

---

Supervised learning approaches, particularly those based on pre-trained language models, have been widely successful in settings with sufficient training data. However, this is not the case for low-resource multilingual settings as well as in specialized domains such as epidemic surveillance. In practice, these settings and domains face constraints related to labeled data scarcity. Furthermore, in multilingual settings, some languages (low-resource languages) are likely to have a relatively small amount of annotated data for model training and evaluation, compared to their high-resource counterparts.

Despite the remarkable success achieved by the pre-trained language models through transfer learning (cross-domain and cross-lingual transfer), the models capture only the general language representation learned from the large-scale training corpora (Sun et al., 2021), limiting the adaptability of the models to new domains. Specialized domains have a distinct vocabulary that is underrepresented in the vocabulary of general domain pre-trained models. While effort has been made to address the problem of data scarcity in event extraction via domain adaptation

([Yang et al., 2020](#)), empirical work is lacking for the domain-specific task of epidemic event extraction in multilingual, low-resource settings. In this chapter, we investigate domain adaptation of pre-trained language models to the task of epidemic event extraction. The empirical results show improvement in model performance and adaptability.

## 5.1 Epidemiological Domain Adaptation

Recent multilingual pre-trained language models (e.g., mBERT ([Devlin et al., 2018](#)), XLM-RoBERTa ([Conneau et al., 2020](#))) demonstrated great performance on various tasks and languages with sufficiently large pretraining corpora, and between close languages. Besides, the models have also been applied to zero-shot cross-lingual transfer tasks ([Lauscher et al., 2020](#); [Mutuvi et al., 2020](#); [Norouzi et al., 2013](#); [Tian et al., 2021](#); [Wang et al., 2018](#); [Wang et al., 2019d](#); [Wu and Dredze, 2019](#)), where a model is able to predict data not seen during training. However, the pre-trained models tend to underperform when applied to specialized domains due to the domain shift problem resulting from the numerous unique in-domain terms (different from generalized vocabulary in standard pre-trained models) in these domains. One approach to dealing with the domain shift problem is domain adaptation.

Typically, domain adaptation of pre-trained language models is achieved by unsupervised pre-training from scratch on target-domain text. For example, BioBERT ([Lee et al., 2020](#)) was initialized from general-domain BERT and then pre-trained on biomedical corpora comprising scientific publications. ClinicalBERT ([Huang et al., 2019](#)) was trained on clinical text from the MIMIC-III database ([Johnson et al., 2016](#)) while PubMedBERT was pre-trained from scratch using abstracts from PubMed ([Gu et al., 2021](#)). [Beltagy et al. \(2019\)](#) proposed SciBERT, a pre-trained language model based on BERT, which relied on a large corpus of scientific text to pre-train. In contrast to the aforementioned models, SciBERT uses an in-domain vocabulary (SciVOCAB) while the other models use the original BERT vocabulary. Generally, the pre-training approach to domain adaptation requires training the model from scratch, which is prohibitively expensive in terms of time and computational cost.

An alternative approach involves extending the vocabulary of the pre-trained models by includ-

ing domain-specific words in their vocabulary. Utilizing a combination of domain-specific and general vocabulary is beneficial to downstream tasks in new domains (Garneau et al., 2018; Tai et al., 2020). It alleviates the problem of directly using the general pre-trained vocabulary on new domains, where previously unseen domain-specific words get split into several subwords by the tokenizer, making the training more challenging (Hong et al., 2021; Tai et al., 2020). Therefore, extending the vocabulary of pre-trained language models could improve tokenization quality, ultimately improving performance, particularly in low-resource settings.

Some studies explored the extension and adaptation of pre-trained models from general domains to specific domains with new additive vocabulary. Tai et al. (2020) proposed the extended BERT (exBERT), which adapts BERT to the biomedical domain by using an extension module to add new vocabulary. While an extension module is required, the weights of the original BERT model remain fixed, thus significantly reducing the amount of training resources required. A similar study by Poerner et al. (2020) focused on the named entity recognition (NER) task. The authors train Word2Vec (Mikolov et al., 2013b) on the target-domain text and align the resulting word vectors with the wordpiece vectors of a general-domain PTLM. Thus, the PTLM gains domain-specific knowledge in the form of additional word vectors. Similarly, Hong et al. (2021) presented the Adapt the Vocabulary to downstream Domain (AVocaDo), an approach that requires only the downstream dataset to generate domain-specific vocabulary, which is then merged with the original pre-trained vocabulary for in-domain adaptation. By selecting a subset of domain-specific vocabulary while considering the relative importance of words, the method could be applied to a wide range of NLP tasks and diverse domains (i.e., biomedical, computer science, news, and reviews).

Despite these recent efforts, research on domain adaptation of pre-trained language models to low-resource multilingual datasets is largely lacking. Moreover, an in-depth analysis of model performance that would take into account salient attributes of the extracted information and the quality of pre-trained model tokenization is needed. While previous studies have investigated the importance of high-quality subword-based tokenizers on performance, the focus was on models based on WordPiece (Schuster and Nakajima, 2012) tokenization, such as the multilingual BERT (mBERT). In this study, we, in addition, explore the XLM-RoBERTa model that uses byte-pair



Table 5.1: Statistics of the DANIEL dataset. DIS = disease, LOC = location.

Language	Doc	Sent	Token	Entity	DIS	LOC
English	474	9,090	255,199	563	349	214
Greek	390	6,402	191,890	300	166	134
Polish	352	9,240	160,887	638	444	194
Russian	426	6,415	138,396	327	224	103

encoding (BPE) (Gage, 1994; Sennrich et al., 2016) subword tokenization algorithm. Unlike BPE, which relies on the frequency of the symbol pair to determine if it is to be added to the vocabulary, WordPiece selects the pair that maximizes the language-model likelihood of the training data. In both algorithms, commonly utilized words remain unsegmented, but rare words are decomposed into known subwords. The BPE and Wordpiece sub-word tokenization techniques are discussed in detail in Section A.3.1.

## 5.2 Tokenizer Analysis

We hypothesize that an epidemic event extraction system in a multilingual and low-resource setting could be impacted by the type of pre-trained language model used and the quality of the applied pre-trained tokenizer. To analyze tokenizer quality, we utilize the token-level multilingual dataset described in Table 5.1. While the dataset comprises text from six languages, we excluded French and Chinese because, as shown in Table 3.4, Chinese has only 7 test entities (only 2 for location), while French has 300, which is far too many in comparison to the other languages. The number of test instances for the other languages range between 20 and 70. The size of the test set reflects the amount of data resources available in each language to train and evaluate models. Therefore, we took into account English, Greek, Russian, and Polish and treated them as low-resource languages. All of these languages belong to the Indo-European family, and the genera for the languages are Germanic, Balto-Slavic, Romance, and Greek, respectively. The Indo-European family is (arguably) the most well-studied language family, containing a few of the highest-resource languages in the world, and thus, large pre-trained language models are generally likely to be biased towards such high-resource languages. Greek language, whose

writing system is the Greek alphabet, is an independent branch of the Indo-European family of languages, with a morphology that depicts an extensive set of productive derivational affixes. On the other hand, Russian (East-Slavic) and Polish (West-Slavic) are Balto-Slavic languages with highly inflectional morphology. As a result, detecting nouns or entities (disease names and locations, in our case) in these languages is difficult, since there are cases where it is impossible to determine the base form of an entity without appealing to data such as verb-frame subcategorization, which is usually out of scope for the kinds of tasks as ours. While English is typically considered a high-resource language, we include it in our experiments because our dataset contains fewer English documents.

Table 5.2: Language family, vocabulary size of individual language-specific models, and the size of vocabulary shared with mBERT and XLMR.

Language	Family	Model	Vocab. Size	Shared mBERT(%)	Shared XLMR(%)
English	Germanic	bert-base-uncased <sup>1</sup>	30,522	17.814	3.683
Greek	Hellenic	bert-base-greek-uncased-v1 <sup>2</sup>	35,000	4.734	2.076
Polish	Balto-Slavic	bert-base-polish-uncased-v1 <sup>3</sup>	60,000	11.490	5.077
Russian	Balto-Slavic	rubert-base-cased <sup>4</sup>	119,547	15.981	11.487

**Pre-trained multilingual models.** We comprehensively compare two multilingual models, multilingual BERT (mBERT) (Devlin et al., 2018) and XLM-RoBERTa (XLMR) (Conneau et al., 2020), against their language-specific (monolingual) counterparts for all the low-resource languages of our dataset. Table 5.2 describes the language-specific models used in our experiments, including information about their vocabulary sizes. The language-specific models are bert-base-uncased, bert-base-greek-uncased-v1, rubert-base-cased and bert-base-polish-uncased-v1<sup>5</sup>. The multilingual BERT, which is pre-trained on Wikipedia, has a vocabulary size of 119,547 tokens (same as the monolingual model for Russian) shared across 104 languages. On the other hand, XLMR is pre-trained on 2.5TB of CommonCrawl data covering 100 languages. Compared to Wikipedia data, using CommonCrawl substantially increases the amount of monolingual resources, principally for low-resource lan-

<sup>5</sup> All models can be found at HuggingFace website: <https://huggingface.co>.

guages (Conneau et al., 2020; Wu and Dredze, 2020b). Despite most of the languages having already been covered by the multilingual models, the justification for the language-specific models has been the lack of capacity by the multilingual models to represent all languages in an equitable way (Rust et al., 2021). The models have been pre-trained on either smaller or lower quality corpora, particularly for most of the low-resource languages. As a result, multilingual models often underperform their monolingual counterparts.

**Quality of pre-trained tokenizer.** While subword tokenizers (Schuster and Nakajima, 2012; Sennrich et al., 2016) provide an effective solution to the out-of-vocabulary (OOV) problem, it has been demonstrated that pre-trained language models still struggle to understand rare words (Schick and Schütze, 2020). The pre-trained tokenizers, in most instances, fail to represent such words with single tokens but instead do it via a sequence of subword tokens, thus limiting the learning of high-quality representations for rare words. Ensuring effective tokenization for low-resource and morphologically complex languages, which leads to improved quality of the corresponding models, could be difficult due to the limited availability of resources to build tokenizers with sufficient vocabulary in such languages. We sought to better understand the impact of tokenizer quality on the performance of pre-trained models in low-resource settings. First, we tokenized the text using the respective tokenizers for the language-specific, mBERT, and XLMR models. We then computed the two metrics proposed by Rust et al. (2021) and Ács (2019): **fertility** and **continued words**. Both metrics assess the suitability of a tokenizer for texts from different languages.

**Fertility of the pre-trained tokenizer.** Fertility measures the average number of subwords generated per tokenized word, with a lower fertility value indicating that a tokenizer splits the tokens aggressively. For example, the word “norovirus” is tokenized into [“nor”, “##ovi”, “##rus”], which denote a fertility of 3. A minimum fertility score of 1 indicates that the tokenizer’s vocabulary contains nearly every word in the dataset. Figure 5.1 shows the fertility values for the pre-trained tokenizers per language. As expected, the language-specific models have the lowest fertility scores, followed by XLMR for all the languages. Overall, English recorded the lowest fertility score, which indicates that a large number of text tokens in this

language were represented, to a greater extent, in the vocabulary of the tokenizers evaluated.

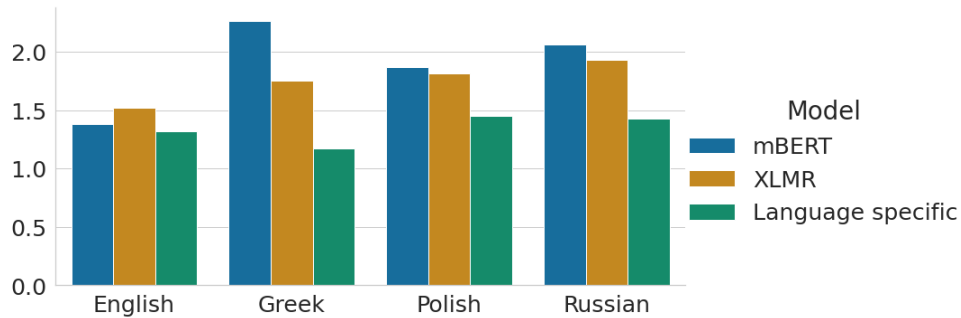


Figure 5.1: Tokenizer fertility score per language and model type.

**Continued words.** This measure describes words not found in the vocabulary of the pre-trained models and which are tokenized into multiple subword tokens. For example, because the word “norovirus” does not exist in the vocabulary of the BERT tokenizer, it is tokenized into “nor”, “##ovi” and “##rus” word pieces, which are present in the tokenizer’s vocabulary. The continuation symbol ## means the prefixed token should be attached to the previous one. In the case of XLM-RoBERTa, the word “norovirus” is tokenized into “\_no”, “ro”, “virus”, where the first token is prefixed with an underscore ( ) and the rest of the tokens (without the underscore symbol) could be appended to form the complete word.

Therefore, the proportion of continued words indicates how often a tokenizer splits words on average. As with fertility, a low rate of continued words is desired, which indicates that the majority of the tokens are present in the vocabulary of the tokenizer. Figure 5.2 shows the proportion of continued words per language and model type. Similar to fertility, we observed the desired low rate of continued words for the language-specific models. The superior results produced by language-specific tokenizers as compared to their multilingual counterparts could be attributed to the fact that language-specific models have a higher parameter budget compared to the parameters allocated to the various languages in the vocabulary of multilingual models. Additionally, the models are typically trained by native-speaker experts who are aware of relevant linguistic phenomena exhibited by the respective languages.

Further examination of the per-language results shows that the English language had the best performance in terms of tokenizer fertility and the proportion of continued words. This implies that

the tokenizer mostly keeps English tokens intact while generating different token distributions in morphologically rich languages. Besides English being morphologically poor, this performance could also be attributed to the models having seen the most data in this language during pre-training, being a high-resource language. Notably, a high rate of continued words was observed for the Russian language on both multilingual models (mBERT and XLMR). Moreover, we add the ratios of **continued entities** that is, the locations and disease names in the training set tokenized into multiple subwords. We observe, for all the languages, a high percentage of continued entities.

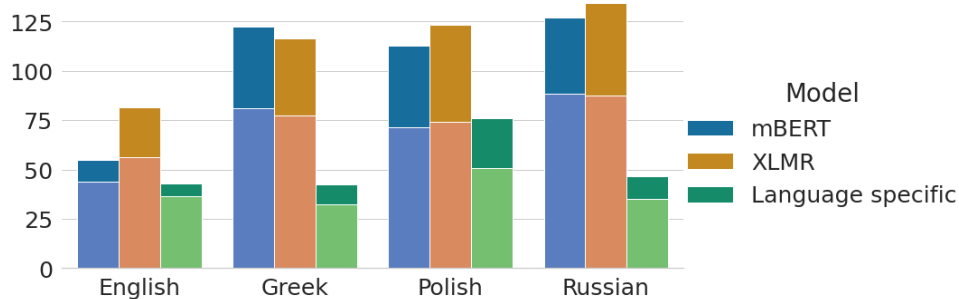


Figure 5.2: Proportion of continued words per language and model type.

We also provide the ratios of out-of-vocabulary (OOV) words in Figure 5.3. OOV are words that cannot be tokenized, thus the entire word is mapped to a special token <UNK>. Generally, the proportion of OOVs should be extremely low, implying that the tokenizers should typically split the words into known subwords. We noticed that mBERT generated a large number of OOVs, especially for the English and Polish languages. A closer examination of the OOV words reveals that the majority of the words, particularly in English language text, contain backticks and curly double quotation marks, which could easily be handled through text preprocessing. The XLMR tokenizer, on the other hand, was able to recognize and split all tokens in our dataset without the need for the <UNK> symbol. This is because byte-level BPE tokenization used in XLMR, covers any UTF-8 sequence with just 256 characters, ensuring that every base character is included in the vocabulary.

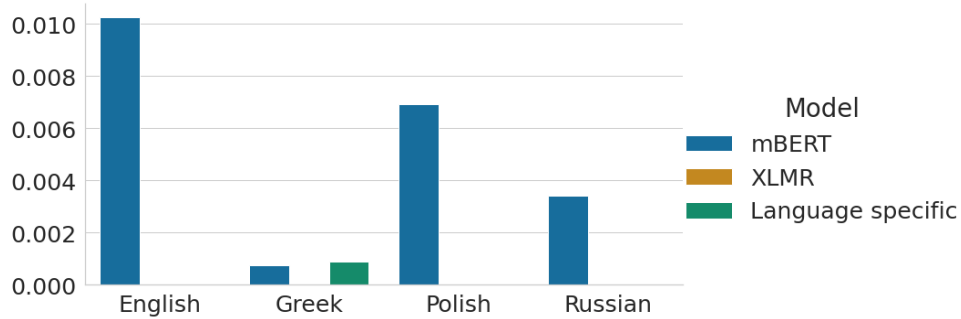


Figure 5.3: The ratios of unknown words per language and model type.

### 5.3 Results and Analysis

To analyze the impact of tokenizer quality, we fine-tuned both the multilingual and language-specific models on the epidemic event extraction dataset. We trained<sup>6</sup> and evaluated the models by averaging the precision (P), recall (R), and F1 score over five runs using different seed values and reporting the standard deviation ( $\pm$ ). As presented in Table 5.3, the language-specific models attained the best overall performance, recording the highest scores in all languages considered except for Russian. The high performance for the language-specific model is consistent with the lower fertility and continued word values when compared to the multilingual models, as shown in Figures 5.1 and 5.2, respectively. In terms of per-language performance, the Russian language had the lowest performance, which could be attributed to high values of continued words and subword fertility.

In order to determine the correlation between the number of continued words, entities, and the fertility of the tokenizer and the F1 scores for all the languages considered, we drew a scatterplot, fitted a regression model, and plotted the resulting regression line with a 95% confidence interval. As shown in Figure 5.4, we observed that for English, the results were generally stable and the performance was negatively impacted by fertility and continued words. Furthermore, the performance score (F1) decreases with an increase in the proportion of unseen entities (continued entities). English has a higher amount of shared vocabulary (e.g., 17% for mBERT), as shown in Table 5.2, so we would expect that further addition of new entities would have a minimal impact on performance. Similarly, while rather unstable, the results for the Greek and Polish

<sup>6</sup> In all experiments, we used AdamW (Kingma and Ba, 2014) with a learning rate of  $1e - 5$  and for 20 epochs. We also considered a maximum sentence length of 164 (Adelani et al., 2021).

Table 5.3: Model performance per language and model type.

Language	Model	P	R	F1
English	mBERT	76.18±1.21	64.52±3.95	69.88±2.36
	XLMR	69.50±9.68	55.48±3.53	61.59±5.16
	Language-specific	<b>76.83±3.69</b>	<b>68.39±5.30</b>	<b>72.30±3.67</b>
Greek	mBERT	<b>83.94±7.00</b>	73.85±5.01	78.39±4.07
	XLMR	81.45±7.73	70.00±6.32	74.85±2.52
	Language-specific	72.78±3.57	<b>91.54±5.02</b>	<b>80.96±1.95</b>
Polish	mBERT	88.02±5.70	89.41±2.63	88.60±3.05
	XLMR	86.94±4.33	87.45±2.24	87.14±2.49
	Language-specific	<b>89.62±6.45</b>	89.41±3.56	<b>89.32±2.54</b>
Russian	mBERT	<b>67.18±3.84</b>	<b>67.28±3.32</b>	<b>67.08±0.76</b>
	XLMR	57.35±11.23	53.34±11.86	54.47±3.55
	Language-specific	51.83±4.30	65.94±4.16	58.14±1.75

were also impacted by fertility, continued words, and entities. Russian does not seem to be impacted, which could be attributed to the vocabulary size of its corresponding language-specific tokenizer. Besides, the proportion of shared words or entities is relatively higher for Russian, as shown in Table 5.2, implying that the language is fairly well represented in the vocabulary of the pre-trained multilingual models.

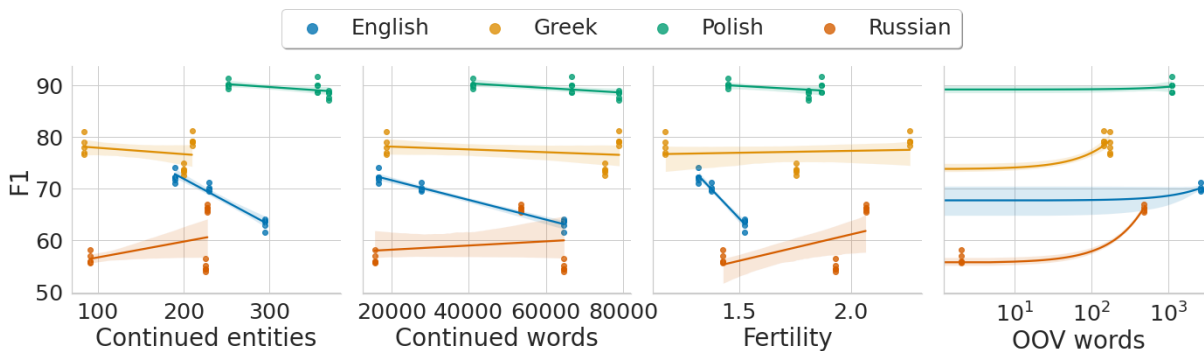


Figure 5.4: Relationship between tokenization quality and F1 performance. The languages considered were English, Greek, Polish, and Russian. The assessment of tokenizer quality was based on the real values of continued entities, continued words, OOVs, and fertility.

Based on these findings, we could expect all languages' performance for mBERT and XLMR to

improve if the continued entities are correctly tokenized. We would also anticipate a change in Russian performance for all models, though it is unclear whether this would be a lower or higher performance.

Table 5.4: Comparison between the default pre-trained model and the extended model (E-Model). The E-model is obtained by enriching the tokenizer vocabulary with domain-specific words (entities) that the tokenizer splits into subwords.

Language	Model	Default	E-Model
English	mBERT	<b>69.88</b> $\pm$ 2.36	50.11 $\pm$ 2.30
	XLMR	<b>61.59</b> $\pm$ 5.16	55.30 $\pm$ 1.73
	Language-specific	<b>72.30</b> $\pm$ 3.67	56.79 $\pm$ 2.20
Greek	mBERT	<b>78.39</b> $\pm$ 4.07	76.47 $\pm$ 2.61
	XLMR	74.85 $\pm$ 2.52	<b>79.88</b> $\pm$ 2.73
	Language-specific	80.96 $\pm$ 1.95	<b>83.95</b> $\pm$ 4.08
Polish	mBERT	88.60 $\pm$ 3.05	<b>91.43</b> $\pm$ 2.90
	XLMR	87.14 $\pm$ 2.49	<b>87.67</b> $\pm$ 2.92
	Language-specific	89.42 $\pm$ 2.54	<b>91.76</b> $\pm$ 2.39
Russian	mBERT	<b>67.08</b> $\pm$ 0.76	60.40 $\pm$ 0.09
	XLMR	54.47 $\pm$ 3.55	<b>58.77</b> $\pm$ 9.90
	Language-specific	58.14 $\pm$ 1.75	<b>61.81</b> $\pm$ 3.22

**Repairing Continued Entities.** On the basis of the preceding analyses, we next evaluate the impact of increasing the model capacity by adding words from the epidemiological domain into an existing tokenizer vocabulary<sup>7</sup>. Concretely, continued (unseen) entities, which were entities from the training corpus of the specialized domain that were not already present in the existing vocabulary and which were split into multiple subwords by the pre-trained tokenizer, were used as the extension vocabulary. We note that it is critical to add to an existing subword tokenizer only whole words rather than their respective subwords. This is because adding subwords instead of whole entity words introduces errors to the vocabulary, considering that tokens in the

<sup>7</sup>The HuggingFace<sup>8</sup> library provides the function to add the continued entities to the existing vocabulary of the tokenizer. The function implements a mechanism for discarding tokens in the extension vocabulary that are also present in the original pre-trained vocabulary, ensuring that the extension vocabulary is an absolute complement to the original vocabulary. The size of the extension vocabulary varies depending on the language and pre-trained model.



Table 5.5: Number of entities added to the vocabulary per language and model. DIS (%) and LOC (%) denote the percentage of unseen disease and location entities, respectively, which were not found in the tokenizer vocabulary.

Language	Model	Unseen	DIS (%)	LOC (%)
English	mBERT	58	60.34	41.38
	XLMR	75	52.00	49.33
	Language-specific	46	63.04	39.13
Greek	mBERT	36	58.33	41.67
	XLMR	33	63.64	36.36
	Language-specific	14	100.0	0.00
Polish	mBERT	146	66.44	34.93
	XLMR	157	61.78	39.49
	Language-specific	124	68.55	33.06
Russian	mBERT	76	56.58	43.42
	XLMR	74	54.05	45.95
	Language-specific	39	74.36	25.64

tokenizer vocabulary are ordered by frequency. Table 5.5 shows the number of previously unseen entities per language that were included in the vocabulary of the considered pre-trained models. Polish has the highest number of unseen tokens among all the languages, while the Greek language has the lowest. The “DIS” entity accounted for the greatest number of unseen entities in comparison to the “LOC” entity. For instance, all the 14 unseen entities identified when a language-specific tokenizer was applied to the Greek text were “DIS” entities.

From our analysis, we expected the models to obtain increased performance values for Greek, Polish, and Russian as a result of augmenting the pre-trained tokenizer vocabulary with entity words specific to our corpus. The performance improves across the low-resource languages and models, as shown in Table 5.4 (the model with improved performance is called Extended Model or, in short, E-model). Notable performance improvement was observed for the Polish language on all the models, while for the Greek and Russian languages, only mBERT was unable to record a performance gain. A significant performance drop is observed for the English language, which can be due to either the negative influence of rare words or due to the vocabulary of pre-trained

and fine-tuned models becoming too distant and the model losing previously learned knowledge, thus hindering the generalization performance of the models. This phenomenon is referred to as catastrophic forgetting (Chen et al., 2019a; Kirkpatrick et al., 2017). While the addition of entities to the vocabulary marginally improves the performance of the different model types evaluated, we presume that the performance could substantially improve when the vocabulary is extended with a sizable number of entities. In our case, as presented in Table 5.5, only a small number of continued entities was available for each language in the dataset.

## 5.4 Conclusions

In this chapter, we investigated domain adaptation through the enhancement of tokenizer vocabulary. First, we performed an analysis of tokenizer output quality by measuring the proportion of unknown words, continued words, and entities, and tokenizer fertility for multilingual epidemic surveillance. In the second step, based on the analysis, unseen entities related to the epidemiology domain are identified and utilized to extend the vocabulary of the mBERT and XLM-R, two popular pre-trained multilingual models. Results suggest that in-domain vocabulary plays an important role in adapting pre-trained models to our specific domain and is particularly effective in low-resource settings. More specifically, we make the following key observations: The quality of the tokenizer, measured in terms of subword fertility, the proportion of continued words, and the OOV words, influences performance, particularly for low-resource languages. The extension of pre-trained tokenizer vocabulary positively impacted the performance of models. Furthermore, a comparison of the pre-trained models reveals that, for the task of epidemic event extraction, language-specific models outperform their multilingual counterparts. The next chapter considers self-training, which utilizes both labeled and unlabeled instances to improve the performance of the epidemic event extraction task.

---

# Self-training with Topic Modeling for Event Extraction in Noisy Settings

---

This chapter evaluates the applicability of the self-training technique to the epidemic event extraction task. While supervised learning has had remarkable success, including in epidemic event extraction, as demonstrated in this study, the methods are still prone to overfitting, particularly in specialized domains where labeled data resources are scarce ([Karisani and Karisani, 2021](#)). A typical approach to improving the performance of supervised models involves labeling additional data. However, the data annotation process in specialized domains (e.g., epidemiological surveillance) requires experts with sufficient domain knowledge, making the annotation a labor-intensive and time-consuming task ([Lai et al., 2021b](#)). Consequently, the substantial labeling cost may limit the wide-spread adoption and scalability of domain-specific event extraction systems.

Various techniques, such as semi-supervised learning ([Berthelot et al., 2019](#); [Olivier et al., 2006](#); [Rosenberg et al., 2005](#); [Van Engelen and Hoos, 2020](#); [Zhu and Goldberg, 2009](#); [Zhu, 2005](#)), have been employed to relieve the labeling requirements and address the problem of training models

with limited labeled training data. The goal of semi-supervised learning is to minimize labeling efforts while at the same time maximizing model performance by combining unlabeled data with a typically smaller set of labeled data during training. By leveraging the readily available unlabeled data, which is relatively easy to acquire (Zhu and Goldberg, 2009), semi-supervised learning provides a cost-effective way to train models that generalize well to unseen data in settings with limited labeled training data.

## 6.1 Mean Teacher Self-training

Among the semi-supervised learning techniques is self-training (McClosky et al., 2006; Yarowsky, 1995), which provides a simple mechanism for incorporating unlabeled data into the training by generating pseudo-labels for the unlabeled data. In self-training, a model is trained on the labeled data and subsequently used to predict and obtain pseudo-labels on the unlabeled data. The pseudo-labels from the most confident predictions are selected and incorporated into the training set, effectively increasing the data size. The standard self-training method, however, is susceptible to confirmation bias (Arazo et al., 2020), in which the generated pseudo labels remain fixed during training. As a result, incorrect pseudo labels are not corrected during the semi-supervised learning process, which could potentially influence the performance of models negatively (Rosenberg et al., 2005).

Overcoming confirmation bias necessitates more robust techniques, such as the mean teacher self-training (Liang et al., 2020; Tarvainen and Valpola, 2017), a teacher-student approach that maximizes the consistency between the student and the teacher classifier. The teacher model is an average of student model weights rather than label predictions (Tarvainen and Valpola, 2017). The training procedure begins with fine-tuning a pre-trained language model (both XLM-R and mBERT were considered in our study) on labeled data  $L$ , followed by using the model to estimate pseudo labels for unlabeled instances  $U$ . Both the labeled and pseudo-labeled data are utilized in training the final model. The mean teacher self-training procedure is illustrated in Figure 6.1 and is further described by Algorithm 1. The mean teacher (MT) technique consists of two models, teacher ( $\Theta_{teacher}$ ) and student ( $\Theta_{student}$ ), with identical architectures but different parameters (weights). While the weights for  $\Theta_{student}$  are learned through standard backpropaga-

tion, the weights for the  $\Theta_{teacher}$  comprises the exponential moving average (EMA) of  $\Theta_{student}$  weights<sup>1</sup>. The labeled instances, augmented with pseudo-labels of the unlabeled data, form the input to both  $\Theta_{teacher}$  and  $\Theta_{student}$ . The cost function, which measures the performance of the model, is a linear combination of two different types of costs: classification (categorical cross-entropy) and Kullback–Leibler (KL) divergence consistency cost (Pérez-Cruz, 2008)<sup>2</sup>. The consistency cost is used for unlabeled data points and aims to minimize the differences in predictions between the teacher and the student models.

---

**Algorithm 1** Mean Teacher Self-training.
 

---

**Input:** the labeled set  $L = \{x_i | i = 1, \dots, n\}$ , the unlabeled set  $U = \{x_i | i = 1, \dots, m\}$

```

P ← {0.2, 0.3, 0.4, 0.5, 1.0}                                ▷ For each of our training settings
for p ∈ P do
  Ls ← Ls * p                                             ▷ Sample a subset
  Θteacher ← Θteacher(Ls)                                  ▷ Fine-tune Θteacher until convergence using Ls
  Us ← Θteacher(U)                                         ▷ Generate the pseudo labels for self-training
  Us = Ls ∪ Us                                           ▷ Combine labeled and pseudo-labeled data
  Θstudent ← Θteacher                                       ▷ Identical architectures for teacher and student
  while not converged do
    for t ∈ T do
      Bk ⊂ Us                                             ▷ Sample a batch
      Yk = Θteacher(Bk)                                   ▷ Generate pseudo labels Yk = {xi | i = 1, ..., k}
      Θstudent(t, k) = (Θstudent(t, k - 1), Yk)          ▷ Update the student model
    end for
    Θteacher ← Θstudent(t) ← Θstudent(t, T)                ▷ Update the teacher and student
  end while
end for

```

---

## 6.2 Selection of Unlabeled Data using Topic Modeling

Noisy training data can significantly impact the performance of models (Atla et al., 2011). In the context of epidemiological event extraction from online news text, articles with or without mention of epidemiological entities (disease name and location) not linked to an event are considered noisy (Valentin, 2020). Prior to performing event extraction, document selection could aid in filtering out the noisy ones, thus retaining the documents that contribute the most to the

<sup>1</sup> In all our experiments, we use an EMA decay rate of 0.999.

<sup>2</sup> Our consistency cost weight is 1.0.

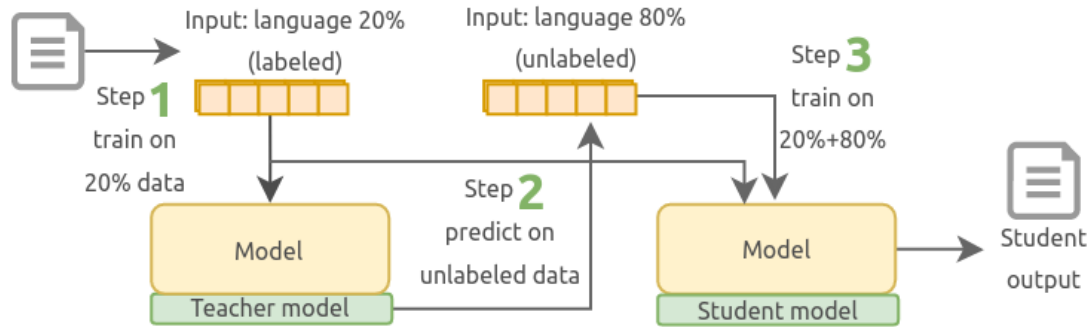


Figure 6.1: An illustration of the mean-teacher self-training approach using 20% labeled and 80% unlabeled data few-shot setting.

performance. In this section, we evaluated the use of topic modeling to select the set of the most relevant unlabeled documents for self-training. The problem of selecting the most optimal data for self-training has yet to be adequately examined (Lai et al., 2021b). Most conventional self-training methods do not filter out insignificant unlabeled samples but instead use all unlabeled data for learning (Jeong et al., 2020). We hypothesize that the performance of self-training improves after the filtering procedure has eliminated the documents deemed to be irrelevant and noisy, leaving only the most relevant documents. Selecting the most relevant documents from the unlabeled data could result in the generation of pseudo-labels with higher prediction confidence. However, it is worth noting that document filtering risks significantly reducing the amount of training data, negatively impacting model performance (Jeong et al., 2020).

We propose a simple topic modeling-based selection approach for selecting the most optimal unlabeled documents from the DANIEL dataset. Topic modeling provides an approach to extract interpretable themes and topics from a large text corpus (Kayi et al., 2013; Tobius et al., 2022). Topic models have previously been applied to document classification without the requirement of acquiring gold standard labels (Miller et al., 2016). The topic models considered were the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and the BERTopic (Grootendorst, 2022), which leverages pre-trained Transformer-based language models, described in Section A.2, to generate robust text representation. Concretely, BERT topic employs the BERT model to generate document embeddings, clusters these embeddings, and then, using a class-based TF-IDF procedure, generates topic representations.

While the DANIEL dataset comprised news articles from six diverse languages, only the languages considered as low-resource in our study were evaluated, namely, Greek, Polish, and Russian. We also investigated the English language since the number of documents present in the dataset for the language was less than 500. To simulate the labeled and unlabeled data settings, we split the training data into labeled  $L = \{x_i | i = 1, \dots, n\}$  and unlabeled set  $U = \{x_i | i = 1, \dots, m\}$ , starting from 10%, and incrementing by 10 percentage points until 50%<sup>3</sup>, as shown in Table 6.1. In the 20% scenario, for example, 20% of the data is considered annotated, while the remaining 80% had their labels removed to form the unlabeled set.

For both the LDA and BERTopic models, we generated a total of ten terms that best describe a particular topic. The topic with the most relevant documents was used as the basis for selecting the unlabeled examples to be used for self-training. We made the assumption that, since topic models detect and cluster word groups and similar expressions that best characterize a set of documents, the topic with the most relevant documents would most likely contain the most optimal unlabeled instances that could contribute the most towards improving self-training performance. The selected unlabeled examples and the labeled instances of a given split were combined and used to train the model. The procedure was repeated for the remaining data splits (20%, 30%, 40%, and 50%).

Table 6.1: The studied scenarios, with an increasing number of training sentences and the number of disease names (DIS) and locations (LOC) entities per scenario and language.

Setting	Greek			Russian			Polish			English		
	Sentences	DIS	LOC	Sentences	DIS	LOC	Sentences	DIS	LOC	Sentences	DIS	LOC
10/90%	409	6	1	475	16	8	627	8	8	727	48	29
20/80%	855	35	19	1,116	44	22	1,487	56	28	1,545	71	51
30/70%	1,316	53	32	1,618	62	35	2,502	64	29	2,276	87	56
40/60%	1,802	69	46	2,032	72	40	3,079	83	47	2,947	113	71
50/50%	2,229	87	58	2,489	95	50	3,698	112	56	3,774	122	79
100%	4,947	144	115	5,249	151	84	7,287	231	134	7,311	271	173

<sup>3</sup> We stop at 50% so that we always maintain fewer labeled than unlabeled instances.

### 6.3 Results and Analysis

The results of self-training multilingual pre-trained models (mBERT and XLM-R) on the task of epidemic event extraction, shown in Figure 6.2, revealed a marginal improvement in performance over the baseline model. The baseline model was obtained by fine-tuning the pre-trained multilingual language models on the labeled data splits in Table 6.1.

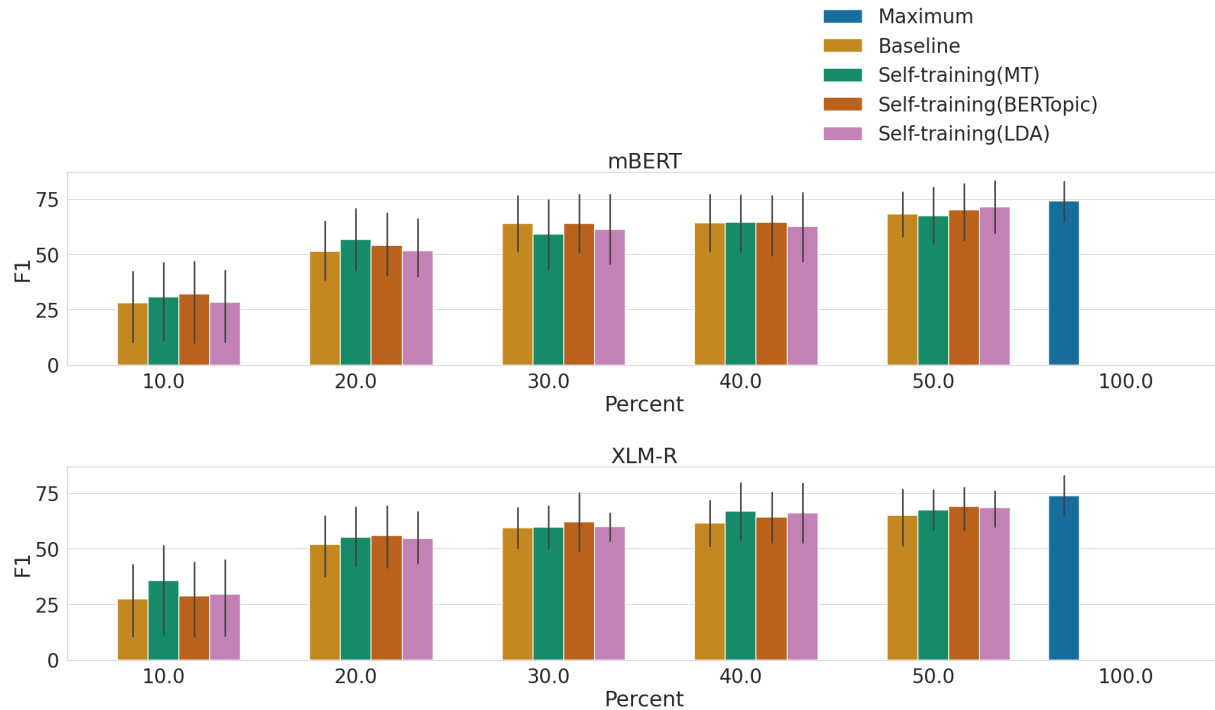


Figure 6.2: Performance (F1-scores) comparison of mBERT (Multilingual BERT) and XLM-R (XLM-RoBERTa) models per data split. Baseline denotes a model trained using supervised learning, self-training (MT) is mean teacher self-training, self-training (BERTopic) and self-training (LDA) represent a scenario where topic modeling techniques (BERTopic and LDA) were employed for noise filtering before self-training.

Self-training with the filtered documents obtained through topic modeling (BERTopic and LDA) outperformed the standard mean teacher self-training approach on three of the five data splits for both mBERT and XLM-R. This could be attributed to reduction of amount of noise in the data following the selection of unlabeled examples using topic modeling.

A comparison of the results of self-training on the unlabeled instances selected using either BERTopic or LDA showed that BERTopic yielded better performance than LDA in four out of



the five data splits for mBERT and three out of the five splits for the XLM-RoBERTa model. The superior performance for BERTopic could be due to the learning of better (contextualized) representation by the underlying BERT model. While performance improvement was observed in the majority of the data splits, the performance of self-training declined when using the 30% data split and was lower than the performance of the mBERT-based baseline model. Furthermore, we observed that the F1-score performance improved as the data size increased for both models. The 50 percent split produced the best results, demonstrating the importance of having enough training data to train supervised models (the baseline model in our case) and generate high-quality pseudo labels for self-training.

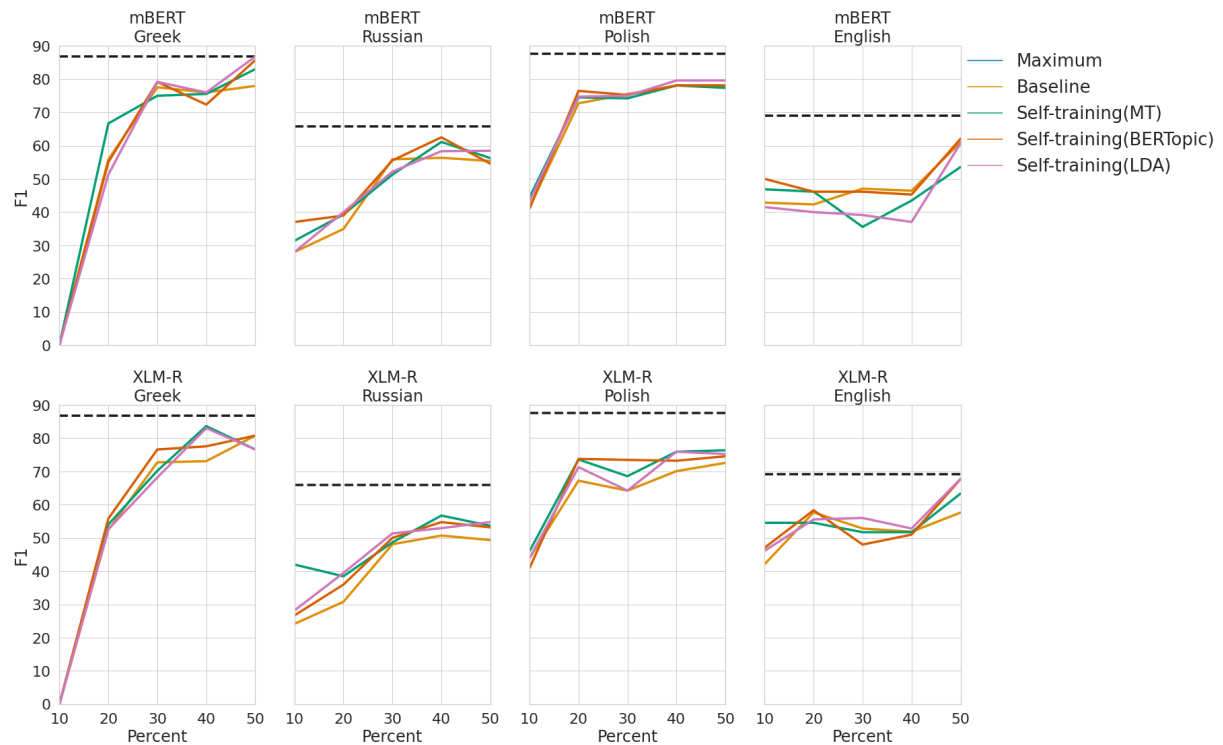


Figure 6.3: Performance (F1-scores) of mBERT (Multilingual BERT) and XLM-R (XLM-RoBERTa) models per language. The results show how performance scores change across the data splits for the baseline, mean teacher self-training, self-training with BERTopic, and self-training with LDA. The dotted line denotes the performance attained by the model when trained on the entire training dataset.

In Figure 6.3, we compare the performance of the models across different languages. Notably, for all the languages, the performance increases with the increase in data size for the supervised

baseline model and models trained using the mean teacher self-training approach. The baseline model was trained using only the labeled examples. High F1-score performance was observed for the Greek and Polish languages, with low scores recorded for English and Russian. We note that mBERT-based self-training yielded lower performance than the baseline for the English language. While self-training produced better overall performance than the baseline, we observed some occasions where mean teacher self-training (without noise filtering) performed better than self-training with noise filtering. We posit that the process of document selection reduces the training data sizes, therefore, in some cases, negatively impacting the performance of the models.

In addition to evaluating the performance of the individual languages independently, we also considered `joint self-training`, which involved training a model on a concatenated dataset from all languages ( $\text{Greek} \cup \text{Russian} \cup \text{Polish} \rightarrow \forall \text{language} \in \text{languages}$ ) and testing the model separately on each of the target languages. We observe from the results presented in Table 6.2 that joint training resulted in an increase in performance (approximately 5%) for Polish compared to when the languages were evaluated separately. The performance change in Russian was marginal but dropped significantly (approximately 20%) for the Greek language. We attribute the competitive performance for Polish and Russian to the sharing of subwords across the languages, which is deemed necessary for multilingual pre-trained language models to work (Pires et al., 2019). The significant drop in performance when Greek was jointly trained with Polish and Russian could be due to the major dissimilarities between Greek and the languages.

Another plausible explanation for the performance decline is that the Greek language could be under-represented in the multilingual pre-trained language model. While joint training achieved strong performance overall on the `Baselines`, the joint self-trained model did not, with only the Russian language showing improvement over the baseline for all data splits except the 20% split. This could be due to the noise introduced by the generated pseudo labels for the various languages.

Regarding zero-shot training, we considered English as the source language and Greek, Russian, and Polish as the target languages. Table 6.2 shows the results of transferring from English to an unseen language. The drop in transfer performance primarily depends on the language

(dis)similarity and is much more pronounced for languages that are less similar to English. One of the most relevant differences between Greek and English is that Greek articles and nouns take different forms depending on the context. In English, the subject-verb-object word order is used to determine who is doing what to whom, and the form of an article or a noun does not change when it is used as an object. Cross-lingual transfer for such languages has been previously investigated (Ahmad et al., 2018; Dehouck and Denis, 2019) with the conclusion that an order-agnostic model will perform better when transferring to distant foreign languages. Moreover, the zero scores for Greek in the 100% and 50% are rather unexpected, which might indicate a data artifact.

Zero-shot self-training (MT) underperforms the baseline, which clearly indicates that the amount of annotated English training (less than 500 articles) is insufficient to cover the trade-off between high-resource and low-resource languages in the pre-trained model. Moreover, self-training (MT) likely increases the amount of noisy annotated data to the point where the mean teacher technique barely handles the heterogeneous nature of generated pseudo labels, resulting in scores of zero.

Table 6.2: The results of joint self-training on low-resource languages (F1%) and zero-shot learning with English as the source language. Baseline represents a model finetuned on labeled data while self-training (MT) represents a model trained using the mean teacher self-training method on both labeled and unlabeled data instances.

Language		Baseline				Self-training (MT)			
	100%	20%	30%	40%	50%	20%	30%	40%	50%
<i>Joint Training</i>									
Greek	69.23	68.00	70.37	<b>71.11</b>	<b>81.63</b>	<b>77.55</b>	<b>74.07</b>	70.83	77.78
Russian	65.67	<b>47.89</b>	50.00	59.15	54.79	43.33	<b>51.52</b>	<b>61.33</b>	<b>62.86</b>
Polish	91.43	<b>70.91</b>	74.75	<b>78.85</b>	<b>78.43</b>	70.48	<b>77.55</b>	78.10	78.10
<i>Zero-shot Learning</i>									
Greek	0.00	6.06	<b>35.56</b>	<b>12.90</b>	0.00	6.06	25.00	7.14	0.00
Russian	4.76	0.00	<b>7.84</b>	0.00	0.00	0.00	0.00	0.00	<b>7.55</b>
Polish	3.33	<b>7.55</b>	<b>6.78</b>	<b>3.70</b>	<b>3.70</b>	7.41	3.64	0.00	3.51

## 6.4 Evaluating the Impact of Errors on Topic Modeling

Besides the noise related to documents with epidemiological entities not linked to an event (epidemiological noise), we sought to explore other forms of errors and their impacts on topic detection and tracking. In particular, we evaluated the impact of OCR errors on topic modeling. The OCR errors, which are inherently introduced during the digitization process, are mainly due to factors such as font variation across different materials, the same words spelled differently, and document deformations that alter the material quality (Silfverberg and Rueter, 2015). We sought to compare the performances of LDA (Blei et al., 2003) and Non-negative Matrix Factorization (NMF) (Lee and Seung, 2000; Lee and Seung, 1999) topic modeling algorithms in the presence of OCR noise. The LDA is a probabilistic topic modeling method that extracts topics from a collection of documents by considering a topic as a probability distribution over a fixed vocabulary and a document as a mixture of topics. By leveraging the co-occurrence of the words in the documents, LDA can infer the latent topics and topic mixtures of text data. The NMF is a linear algebraic optimization algorithm that learns topics by decomposing the term-document matrix into a weighted combination of a set of topic distributions. The term-document matrix is mostly a distributional (Bag-of-Words or TF-IDF) representation of the text corpus. The performance of the models is evaluated by measuring the stability and coherence of topics generated over different runs.

We apply topic modeling to a large corpus of historical documents (Chiron et al., 2017) consisting of twelve million OCRed characters, denoted as OCRedtoInput, along with the corresponding Gold Standard (GS). Specifically, the OCRedtoInput represents the raw OCRed text, whereas GSaligned is the corresponding aligned ground truth. The OCRaligned represents the corrected version of the text corpus with the “@” symbols added as padding symbols to ensure uniform data length. The “#” symbol represents the absence of GS due to alignment uncertainties or unreadable characters in the source document. The dataset includes documents in a variety of formats, such as monographs and periodical documents, and has an equal share of English- and French-written documents spanning four centuries. The documents originate from various digital collections, including the National Library of France (BnF) and the British Library (BL). The corresponding GS is derived from BnF internal projects and external initiatives

such as Europeana Newspapers, IMPACT, Project Gutenberg, Perseus, Wikisource, and Bank of Wisdom (Chiron et al., 2017). Only the English documents from the dataset were considered in our experiments.

The experiment process involved applying LDA and NMF (Lee and Seung, 2000; Lee and Seung, 1999) to the corpus to obtain topics. We used Gensim<sup>4</sup> (Rehurek and Sojka, 2011) with default parameters to implement both the LDA and NMF topic models. For each of the three sets of text corpora (OCRtoInput, GS, and OCRaligned), fifty different iterations of both the NMF and LDA models were executed to generate topics, and the performance was measured based on model stability and coherence scores. A high level of agreement between term rankings generated by multiple runs of the same model indicates that the topic is stable (Greene et al., 2014). The average topic stability results are shown in Table 6.3. When compared to raw OCR text, both models recorded higher average topic stability on aligned text. The mean stability on the Gold Standard text was 0.265 and 0.414, while for the noisy OCR text was 0.252 and 0.383 for LDA and NMF topic models, respectively.

Table 6.3: Topic Stability and Coherence performance of LDA and NMF of OCR generated text.

Model	Dataset	Mean Stability	Mean Coherence
LDA	GSaligned	0.265	0.3622
LDA	OCRaligned	0.256	0.3585
LDA	OCRtoInput	0.252	0.3529
NMF	GSaligned	0.414	0.4748
NMF	OCRaligned	0.384	0.4737
NMF	OCRtoInput	0.383	0.4720

In addition, we computed the coherence of the topic descriptors to assess the quality of the topics generated by the models. Distinctively, the most coherent topics tend to have their top terms co-occurring within the corpus. The results of the average coherence score for LDA and NMF algorithms on the noisy and corrected data are presented in Table 6.3. The mean coherence score on the aligned OCR text was 0.4737 and 0.3585 for the NMF and LDA algorithms, respectively. The mean coherence on raw OCR text, on the other hand, was marginally lower, with 0.4720

<sup>4</sup><https://radimrehurek.com/gensim/>

for NMF and 0.3529 for the LDA topic model. Overall, we find that both LDA and NMF topic models perform poorly on the raw, noisy OCR text when compared to the aligned text (corrected) and the ground-truth data. We conclude that the presence of noise in the text has a negative impact on the extraction of latent topics and topic structures from the text.

## 6.5 Conclusions

In this chapter, we investigated the suitability of self-training for epidemic event extraction in multilingual and low-resource settings. Such settings face insufficient labeled data to train and adapt models to the target domain. Typically, it is challenging to build a successful learning system, if only a few labeled samples are available. Therefore, it is desirable to leverage unlabeled data, which is usually readily available or could inexpensively be obtained, to improve the learning performance given limited labeled training examples. The results obtained in this study indicate that self-training was beneficial to the epidemic event extraction tasks, with improved performance observed in the different scenarios when compared to the baseline results.

Furthermore, we examined the impact of noise on performance by evaluating topic modeling algorithms on OCR text. The results showed that the OCR errors present in the text influenced the performance of the models as measured by topic stability and coherence of the generated topic. Against the backdrop of the results indicating errors to be impacting performance, we evaluated and compared the performance of self-training before and after filtering the unlabeled instances using topic modeling. The evaluation aimed to test whether the selection of the most suitable unlabeled examples to use in the self-training. However, while it was expected that noise filtering could enhance the performance of self-training after topic modeling, the performance in some cases was lower than for the default mean teacher self-training. The performance drop could be attributed to the reduction of the training data after the data selection process. Typically, neural networks require large amounts of data to train and achieve satisfactory performance.

The next chapter provides the conclusions of the thesis and summarizes the work performed in the experiments and the thesis contributions. Additionally, the chapter highlights recommendations for future work.

---

## Conclusions and Future Work

---

In this chapter, we provide the conclusions of the thesis and highlight some of the possible future research directions.

### 7.1 Conclusions

This dissertation investigated data-driven approaches to epidemic event extraction in multilingual and low-resource settings. Data-driven surveillance promises to complement conventional disease surveillance, improving timeliness and geographical coverage. We explored three major challenges facing the extraction task, as presented in Section 1.1 and proposed solutions to the problems. The conclusions of this thesis with respect to the contributions described in Section 1.3 are as follows:

**Token-level multilingual epidemic dataset.** In Chapter 3, we presented a token-level multilingual dataset for epidemic event extraction. We adapted and re-annotated the DANIEL dataset, a freely available epidemiological dataset originally annotated at the document level. The mul-

tilingual epidemic dataset consists of online news text from six different languages, namely French, English, Polish, Chinese, Russian, and Greek. The French, English, and Chinese languages are regarded as high-resource languages, while the others are low-resource languages. The low-resource languages have limited or no available annotated training data, particularly in specialized domains such as epidemiological ones. This contribution, therefore, fills the data scarcity gap, and we hope the corpus, which we made publicly available, will ease and possibly foster further research.

**Supervised neural-based approaches.** Chapter 4 evaluated supervised machine learning-based models for epidemic event extraction. Our experiments showed that the proposed neural-based methods perform better than the baseline DANIEL system, a dedicated, unsupervised multilingual system for epidemic event extraction. In particular, the models based on pre-trained language models achieved the best overall performance. While the models based on supervised learning produced satisfactory results, the small training data size, particularly for the low-resource language, could create a bottleneck during model training. To the best of our knowledge, this is the most extensive analysis of epidemiological event extraction so far. Furthermore, this is the first effort to formulate epidemic event extraction as a sequence labeling task, leading to strong supervised learning baselines that can be the basis for further epidemic event extraction research.

**Epidemiological Domain Adaptation.** In order to deal with the data scarcity problem, Chapter 5 investigated domain adaptation to improve the performance of epidemic event extraction models. Specifically, the proposed domain adaptation approach entailed incorporating epidemiological domain-specific tokens to the vocabulary of the pre-trained language models. The results reveal the viability of vocabulary expansion in adapting pre-trained models for epidemic event extraction tasks in low-resource languages. While different domain-adaptation approaches have previously been investigated, we demonstrate in this contribution that adding domain-specific entities to the vocabulary of Transformer-based pre-trained models can be useful in adapting the models to domain-specific tasks such as epidemic event extraction.

**Self-training with Topic Modeling for Event Extraction in Noisy Settings.** In Chapter 6, we performed an analysis to determine the suitability of self-training for epidemic event ex-



traction. Self-training is a semi-supervised method that uses unlabeled examples alongside the few available labeled instances to train models. The results show that self-training achieved a marginal improvement over a baseline model trained using supervised learning on the epidemic event extraction task. Furthermore, we examined the influence of noisy text by first considering the impact of OCR errors on topic modeling. The findings indicate that OCR errors negatively impact the generation of topics. We further evaluated the extent to which noise filtering via unlabeled data selection impacted the performance of self-training. To our knowledge, this is the first study that investigates self-training for epidemic event extraction while also taking into account the influence of noise on the extraction task. Contrary to our expectations, performing self-training after noise filtering yielded lower performance compared to the mean teacher self-training in some instances. This could be due to data loss following the document selection process.

## 7.2 Future Work

One of the possible future research directions is the **enrichment of the event annotated dataset** to capture the number of reported cases and the temporal information to facilitate temporal reasoning and inference over epidemiological data. Furthermore, additional languages could be included, and the data size for each language enhanced.

Second, **incorporating external knowledge into epidemic event extraction** should be examined. While using domain-specific information from the target dataset yielded promising results in our study, incorporating external knowledge (from the Web or Wordnet, for example) could further improve the generalizability of epidemic event extraction models. Moreover, besides the simple approach employed for vocabulary enrichment, alternative approaches to domain adaptation that could potentially result in significant performance gains need to be explored. Finally, **distantly supervised epidemic event extraction** that involves weakly labeled training sets could be explored. Weakly labeled data describes training data that is automatically labeled based on heuristics and existing knowledge bases. While this thesis focused on semi-supervised learning models to mitigate the issue of data sparsity, distant supervision could also be studied to determine the suitability of distantly-labeled data for epidemic event extraction. The auto-

matic acquisition of labeled data by matching event information in raw text together with that held in knowledge bases can minimize the amount of human annotation effort required. However, like self-training, distant labeling can invariably induce incomplete and noisy labels, thus necessitating additional scrutiny.

---

## Publications

---

This thesis has led to the following publications:

### Conferences

- **Mutuvi S.**, Boroş E., Doucet A., Jatowt A., Lejeune G., and Odeo M. "Fine-tuning de modèles de langues pour la veille épidémiologique multilingue avec peu de ressources (Fine-tuning Language Models for Low-resource Multilingual Epidemic Surveillance)." *In Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Volume 1: conférence principale*, pp. 345-354. 2022.
- **Mutuvi S.**, Boroş E., Doucet A., Lejeune G., Jatowt A., and Odeo M. "Multilingual Epidemic Event Extraction." *In International Conference on Asian Digital Libraries (ICADL)*, pp. 139-156. Springer, Cham, 2021.
- Zosa E., **Mutuvi S.**, Granroth-Wilding M., and Doucet A. "Evaluating the Robustness of Embedding-Based Topic Models to OCR Noise." *In International Conference on Asian Digital Libraries (ICADL)*, pp. 392-400. Springer, 2021.
- **Mutuvi S.**, Boroş E., Doucet A., Lejeune G., Jatowt A., and Odeo M. "Token-level multilingual epidemic dataset for event extraction." *In International Conference on Theory and*

*Practice of Digital Libraries* (TPDL), pp. 55-59. Springer, 2021.

- **Mutuvi S.**, Boroş E., Doucet A., Lejeune G., Jatowt A., and Odeo M. "Étude comparative de méthodes de classification multilingue appliquées à l'épidémiologie." *In Conférence en Recherche d'Informations et Applications (CORIA), French Information Retrieval Conference*. 2021.
- **Mutuvi S.**, Boroş E., Doucet A., Lejeune G., Jatowt A., and Odeo M. "Multilingual epidemiological text classification: a comparative study." *In International Conference on Computational Linguistics (COLING)*, pp. 6172-6183. 2020.
- **Mutuvi S.**, Boroş E., Doucet A., Lejeune G., Jatowt A., and Odeo M. "A dataset for multi-lingual epidemiological event extraction." *In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*, pp. 4139-4144. 2020.
- **Mutuvi S.**, Boroş E., Doucet A., Lejeune G., Jatowt A., and Odeo M. "Evaluating the impact of OCR errors on topic modeling." *In International Conference on Asian Digital Libraries (ICADL)*, pp. 3-14. Springer, 2018.

# Appendices



---

## Overview of Neural Network Approaches for Information Extraction

---

This chapter presents the background information necessary to understand the rest of the thesis. We discuss and explain the underlying concepts such as neural networks and pre-trained language models that model textual features for information extraction.

### **A.1 Neural Networks for Sequential Data Modeling**

Neural network (NN) models have been proven to be remarkably successful both in core areas of NLP and related applications aimed at achieving practical and useful objectives. More specifically, the networks have been applied in NLP tasks such as text classification (Adhikari et al., 2019; Kim, 2014), question answering (Wang et al., 2017; Yang et al., 2019c), relation extraction (RE) (Lin et al., 2019; Zeng et al., 2014; Zheng et al., 2017), named entity recognition (NER) (Akbik et al., 2018; Hammerton, 2003; Lample et al., 2016; Santos and Guimaraes, 2015) and event extraction (Boros et al., 2021; Chen et al., 2015; Nguyen and Grishman, 2015;

Wang et al., 2019b; Yang et al., 2019b). An excellent characteristic of neural networks is their ability to automatically learn feature representation from both structured and unstructured text without the need for manual feature engineering, which is time-consuming and labor-intensive (Liang et al., 2017). Due to a large number of parameters, deep neural networks require massive amounts of data to optimize the parameters to extract high-quality features.

Conventionally, text sequences were represented as a bag of tokens, such as BoW and TF-IDF in NLP tasks. These schemes result in high-dimensional, sparse (mostly zero) text representations that are unable to capture similarities between words. In the recent past, pre-trained word embedding techniques, such as Word2vec (Mikolov et al., 2013c) and GloVE (Pennington et al., 2014) have been utilized. The two types of Word2vec models are the Continuous Bag-Of-Words (CBOW) that predict a target word given its context and the Skip-gram that predicts the context words given a target word using a simple feed-forward neural network. In contrast, the GloVe not only uses the local context windows but also incorporates global word co-occurrence counts.

Pre-trained word embeddings use dense vector representations for words, and words in the same contexts tend to have similar meanings. By being able to capture the semantic similarity of words, word embeddings have enabled deep learning techniques, such as convolutional neural networks (Krizhevsky et al., 2012) and recurrent neural networks (Schuster and Paliwal, 1997), to better capture the syntactic and semantic structures of sentences, thus modeling text sequences with great success. Transformer-based neural networks (Vaswani et al., 2017) use contextualized embeddings that assign meanings to words based on their context. This further bolsters the generalizability of the models on sequential data, including text streams.

### A.1.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) are among the most commonly used neural network models, which consist of multilayer, fully connected neurons (i.e., each neuron in a lower layer is connected to all neurons in its upper layer). While convolutional neural networks (CNNs) have predominantly been used for computer vision (CV) tasks such as image classification, these models have previously been successfully applied to a wide range of NLP tasks (Bai et al., 2018; Chen, 2015; Chen et al., 2015; Nguyen and Grishman, 2015; Zhang et al., 2015; Zhang



and Wallace, 2015). Sentences are tokenized and represented as a matrix, with each row of the matrix corresponding to a token (word or character). The mapping from tokens to vectors can be accomplished using pre-trained word embeddings (e.g., word2vec and GloVe).

In contrast to CV, which involves sliding the kernel over local patches of an image, feature extraction in text involves convolution (filter sliding) over the word vector representations (full rows of the matrix). The width of the filter corresponds to the width of the input matrix (dimensionality of the word vectors), whereas the height of the filter (number of rows that can be handled concurrently), also known as the region size, varies. The CNN architecture for sentence classification shown in Figure A.1 uses three region sizes (2,3,4). Consequently, the dimensionality of the resulting feature maps varies since it is a function of the input sentence length and region sizes of the respective filters. A pooling operation is performed on the variable-length feature maps to generate a fixed-length feature vector, which is then provided as input to a softmax layer to obtain the final classification predictions.

### A.1.2 Recurrent Neural Networks

Feed-forward networks, such as convolutional neural networks, assume a fixed length of input and output vectors. However, it is difficult to define optimal fixed dimensions a-priori for many natural language problems, such as machine translation and information extraction. In addition, feed-forward networks do not preserve the sequential order of text sequences. Recurrent neural networks (RNNs) are able to model dependencies of inputs over time by leveraging a recurrent connection where the last hidden state is passed as an input to the next state. The capacity to memorize previous computed results and use that information to condition the current computation enables RNNs to better handle sequential data by modeling context dependencies in inputs of arbitrary length. As shown in Figure A.2, the computation of the hidden variable (state) is recurrent since the hidden state uses the same definition as the previous time step in the current time step. Therefore, the transition function in a vanilla RNN is a linear transformation of the hidden state and the input, followed by a pointwise non-linearity.

$$\mathbf{h}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b}) \quad (\text{A.1})$$

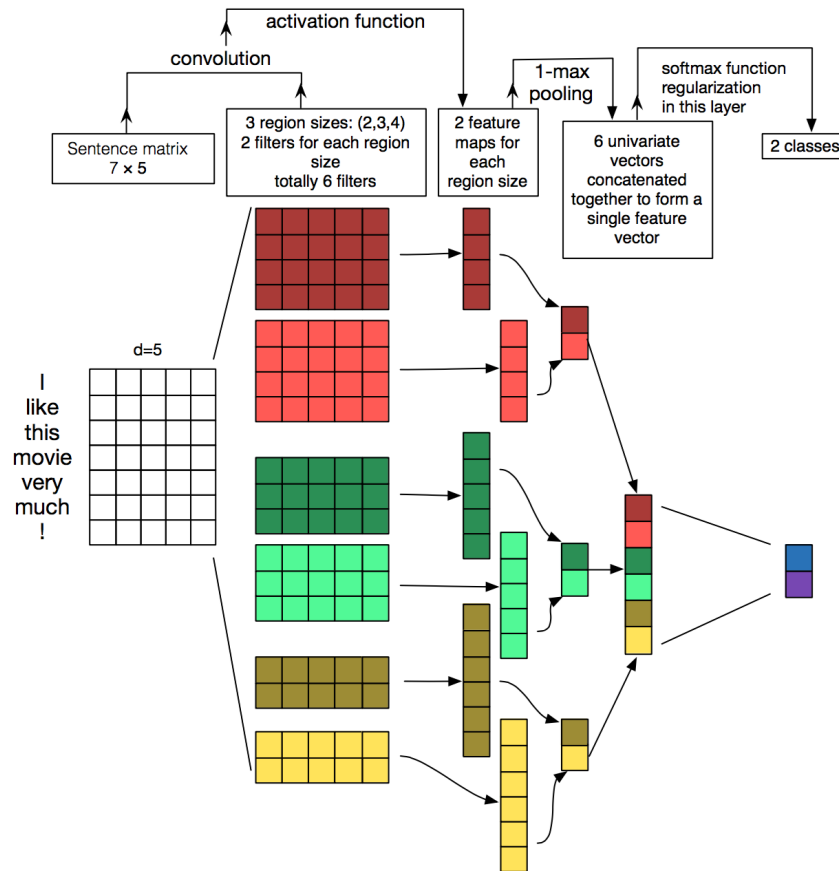


Figure A.1: Convolutional Neural Networks for sentence classification. Source: [Zhang and Wallace \(2015\)](#)

where  $\mathbf{x}_t \in \mathbb{R}^d$  and  $\mathbf{h}_t \in \mathbb{R}^n$  denote the input and hidden state at time step  $t$  respectively.  $\mathbf{W} \in \mathbb{R}^{n \times d}$ ,  $\mathbf{U} \in \mathbb{R}^{n \times n}$  are the input and hidden weights, and  $\mathbf{b} \in \mathbb{R}^n$  represent the bias term.  $\sigma$  is an element-wise activation function of the neurons, and  $N$  is the number of neurons in this RNN layer.

The Backpropagation Through Time (BPTT) training algorithm is used to update RNN weights. Unlike in regular backpropagation, the chain rule must be applied recursively in BPTT, and the gradients are summed through the network. The training of the RNNs, however, suffers from the vanishing and exploding gradient problems. The vanishing gradient problem occurs when the gradients used to update the weights of RNN shrink and become insignificant, which stalls the

<sup>1</sup> Source: <https://www.oreilly.com/library/view/deep-learning-for/9781788295628/20f639d4-8079-4137-b613-c8a479e6f2cf.xhtml>

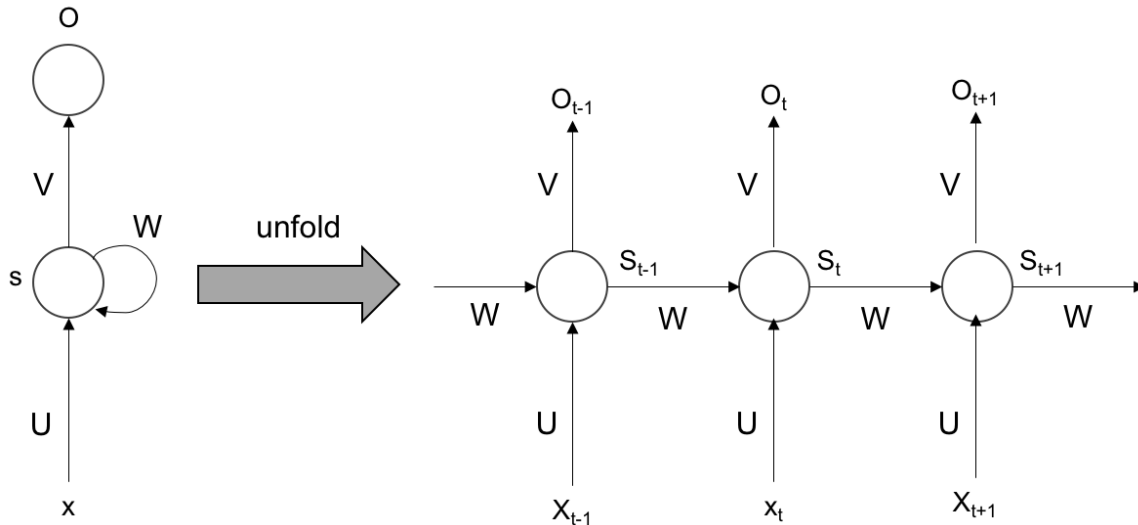


Figure A.2: Recurrent Neural Network Architecture<sup>1</sup>.  $U$  and  $V$  are the weights of the hidden layer and output layer respectively, while  $W$  represents the transition weights of the hidden state.  $x_t$  and  $O_t$  are the input vector and output result at time  $t$ , respectively.

learning process (no real learning happens). With the exploding gradient problem, the gradients grow exponentially, resulting in the learning becoming unstable. Among the techniques for dealing with the exploding and vanishing gradients problem include proper initialization of weight matrices, gradient clipping, and using L1 or L2 penalties on the recurrent weights. Furthermore, while RNNs can access the preceding sequence, the information encoded in hidden states tends to be fairly local and more relevant to the most recent parts of the input sequence. However, not only is local but also distant information important in most real-world language applications, where modeling of long-range contextual information is necessary. To address the problems of vanilla RNNs, gated RNNs, such as the Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), have been proposed.

### A.1.2.1 Long short-term memory (LSTM)

The Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), illustrated in Figure A.3, solves the problem of learning long-term dependencies by introducing a gating mechanism within each LSTM cell that regulates and allows gradients to flow uninterrupted during training. The memory cell consists of a forget gate, an input gate, and an output gate that control

information flow. The forget gate maintains the relevant context over time by utilizing a sigmoid activation function to determine the information to be retained in the cell state. The gate takes the current input  $X_t$  and the previous hidden state  $H_{t-1}$  and generates values,  $F_t$ , between 0 and 1. The values close to 1 are retained, while values close to 0 are eliminated. The input gate determines the new information to be stored in the cell state. This is accomplished through the use of a Sigmoid layer, which determines the values to let through, and a hyperbolic tangent (tanh) activation function, which generates a vector of new candidate values  $\tilde{C}_t$ . The outputs from both the Sigmoid and Tanh functions are then passed through a pointwise multiplication operation. At this point, the outputs of the forget gate and input gate provide the required information to update the cell state. The process of updating the state involves first a pointwise multiplication of the previous cell state  $C_{(t-1)}$  by the forget vector ( $F_t$ ), followed by a pointwise addition of the output from the input gate. This operation adds useful information to the cell state, resulting in a new cell state  $C_t$ . Finally, the output gate decides how much information from the current state should be transferred to the next hidden state. First, a Sigmoid layer determines which parts of the cell state make it to the output. The cell state is then passed through Tanh to push the values to be between -1 and 1, which are then multiplied by the Sigmoid function output.

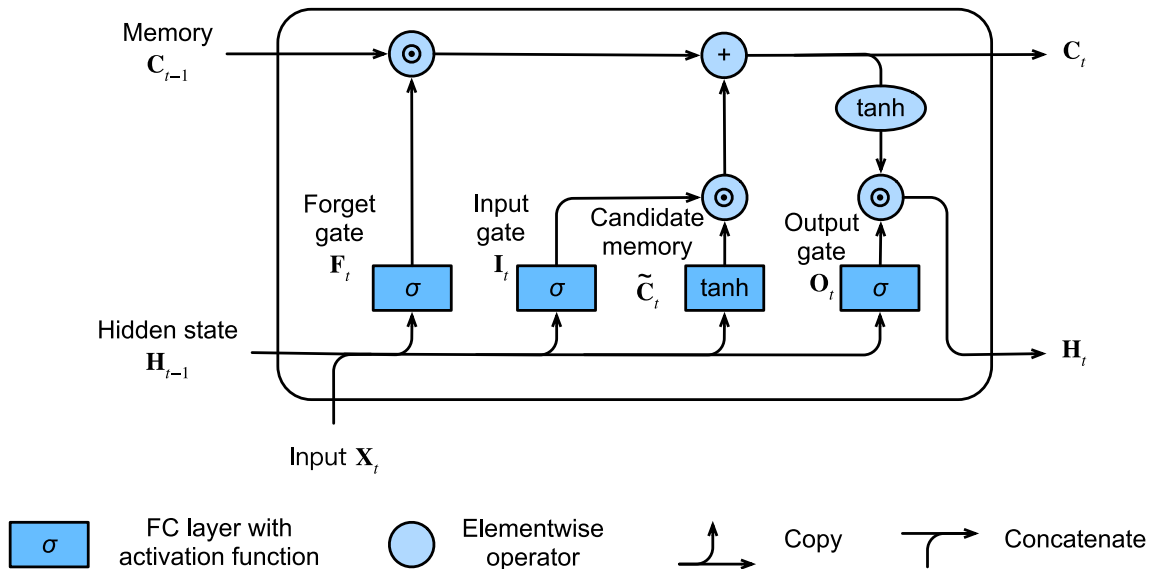


Figure A.3: Long Short-term Memory. Source: <https://www.deeplearningbook.org/contents/rnn.html>.

### A.1.3 Transformer-based Neural Networks

The sequential nature of RNN-based models limits parallelization due to their inherent dependence on the previous state (Vaswani et al., 2017; Yang et al., 2019a). Attention-based methods solve this issue by applying the attention mechanism to draw global dependencies between input and output and therefore being able to better model long sequences. The intuition behind the attention mechanism is that to predict an output word, the model only uses parts of the input where the most relevant information is concentrated instead of the entire sequence. Previously, the most successful approaches to sequential modeling leveraged recurrent or convolutional neural networks that were based on the encoder-decoder architecture with attention-mechanism. The encoder network transforms an entire source sentence into a fixed-length vector representation while the decoder network uses the representation from the encoder to produce the output. The attention mechanism improves the performance of the encoder-decoder models by permitting the decoder to utilize the most relevant parts of the input sequence, which are determined through differential weightings. The vectors representing words in the sentence that are most critical to the interpretation of the sentence are assigned the highest weights.

Recently, the Transformer architecture (Vaswani et al., 2017), which is entirely based on the self-attention mechanism and dispenses with recurrence and convolutions, was proposed. The Transformer's multi-head self-attention mechanism allows each token to attend to all of the tokens in the input sequence, which facilitates understanding of words that are relevant to the one currently being processed. As a result, the Transformers can effectively model long-range sequence dependencies without requiring sequence-aligned recurrence or convolution. Furthermore, because Transformer-based encoder-decoders perform highly parallelizable matrix multiplication operations, they are orders of magnitude more computationally efficient than RNN-based encoder-decoders. As illustrated in Figure A.4, the Transformer model adopts an encoder-decoder architecture with self-attention and does not rely on recurrence mechanism to generate an output.

**The Encoder-Decoder Stacks:** The encoder is responsible for mapping the input sequence into high-dimensional space (sequence of continuous representations). Each encoder layer

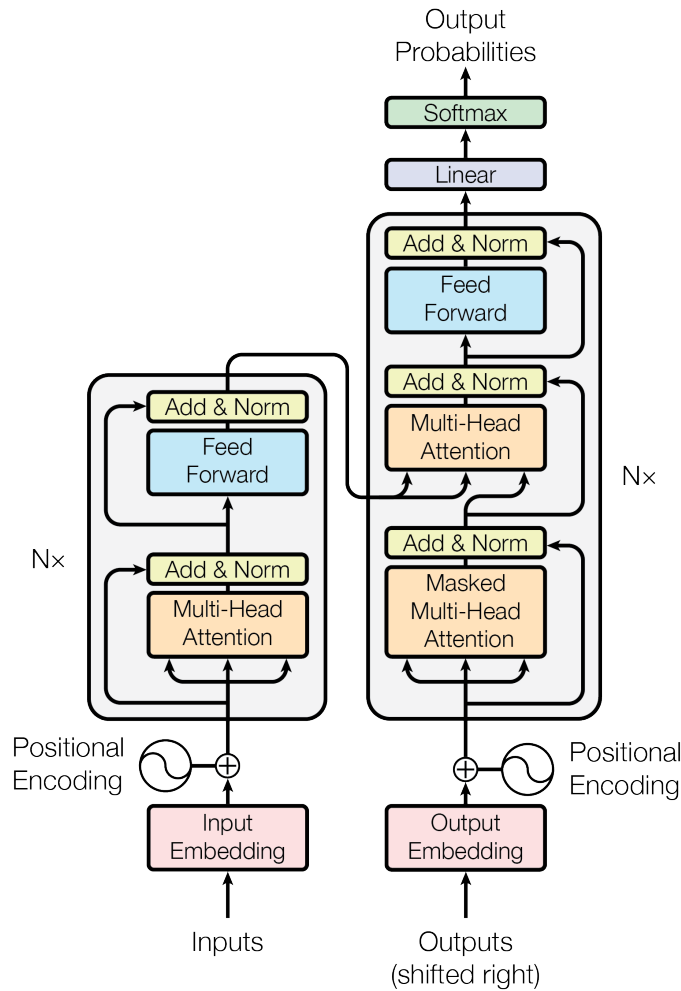


Figure A.4: The Transformer architecture. Source: [Vaswani et al. \(2017\)](#).

is composed of a multi-head self-attention module and a position-wise feed-forward network (FFN). A layer normalization (Ba et al., 2016) and residual connection (He et al., 2016) are applied around each module in order to train deeper model. The decoder includes an encoder-decoder attention layer, which performs multi-head attention over the encoder stack output. Furthermore, the decoder self-attention is modified to include a mask so that each position only attends to all previous positions in the decoder and not subsequent positions. By depending on the previously generated tokens to predict the next, masked attention preserves the autoregressive property of the decoder. Figure A.4 illustrates the architecture of the Transformer model.

**Attention Modules:** The Transformer uses multi-head attention, which comprises multiple scaled dot-product attention performed in parallel. Using a Query-Key-Value (QKV) model, which is a representation of queries and keys of dimension  $d_k$  and values of dimension  $d_v$ , the scaled dot-product attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (\text{A.2})$$

Dividing the dot products of the queries and keys by  $\sqrt{d_k}$  keeps the magnitude of the dot product small, which helps deal with the gradient vanishing problem (extremely small gradients). The softmax function ensures that the dot product weights are between 0 and 1 and add up to 1. The function normalizes the weights to a probability distribution that assigns higher values to more relevant keys and smaller weights to less important keys. Instead of the single attention function, the Transformer employs multi-head attention, which allows attention function to be applied in parallel. Thus, the queries, keys, and values are linearly projected to  $d_k$ ,  $d_k$  and  $d_v$  dimensions multiple times. The multi-head attention is represented as follows:

$$\text{MultiHeadAttn}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (\text{A.3})$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

**Position-wise Feed-Forward Networks:** In addition to the attention sub-layer, the encoder and decoder layers each have a fully connected feed-forward network (FFN) that is applied to every attention vector, with the main purpose of transforming the attention vectors into a form that is acceptable to the next encoder or decoder layer. The FFN consists of two linear transformations with a ReLU activation in between and is defined as follows:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (\text{A.4})$$

where  $W_1$ ,  $W_2$ ,  $b_1$  and  $b_2$  are learnable parameters. The linear transformations are the same across different positions but their parameters vary from layer to layer.

**Positional Encoding:** Positional information about the relative or absolute position of tokens is essential for maintaining the order of tokens in a given sequence. A change in the position and order of words could alter the meaning of the entire text sequence. While RNNs take word order into account by default since they process sentences sequentially, Transformer attention-based models treat the data points independently of each other. The words of a given sentence are fed simultaneously into the encoder-decoder stack of the Transformer, with positional encoding utilized to add a vector to each embedding input to account for the order of the words in the input sequence. Specifically, the Transformer uses an absolute positional encoding based on the sine and cosine functions defined in Equation A.5:

$$\begin{aligned} PE_{(pos,2i)} &= \sin(pos/10000^{2i/d_{model}}) \\ PE_{(pos,2i+1)} &= \cos(pos/10000^{2i/d_{model}}) \end{aligned} \tag{A.5}$$

where  $pos$  is the position vector,  $i$  is the index of the position vector, and  $d_{model}$  is the input dimension. In the case of the original Transformer, the values generated by the sine and cosine functions are interwoven in order to form the positional encoding vectors.

## A.2 Pre-Trained Language Models

Motivated by the Transformers, text representation schemes that assign different representations to words in a given input sequence based on context have recently been explored. Previous representation learning schemes learned either sparse embeddings that could encode little semantic meaning (one-hot encoding and TF-IDF) or fixed embeddings (word2vec, glove), where a given word has the same meaning regardless of context. These context-independent word embeddings disregard contextual information despite its importance in capturing word meanings. On the other hand, context-dependent embeddings go beyond obtaining a single global representation for a word and consider the dynamic nature of word meanings. Concretely, instead of depending on a lookup table of word embedding matrices, the contextual embeddings assign



each word a representation based on its context. A given word can thus be assigned different low-dimensional vectors depending on its context. As a result, these embeddings adequately capture the uses of words across varied contexts, which could translate to performance improvement on downstream tasks. A previous study demonstrated that context influences the detection of relevant diseases and location in an epidemic event extraction task (Valentin, 2020).

While supervised contextual embedding methods such as CoVe (McCann et al., 2017) and InferSent (Conneau et al., 2017) have also been previously studied, pre-trained contextual embeddings based on self-supervised learning have recently received increased attention from researchers and practitioners alike. Self-supervised learning leverages the underlying structure of typically large unlabeled data to obtain supervisory signals. This is important in real-world applications where the acquisition of adequate labeled data required to train deep neural networks is challenging. Self-training has enabled the development of large pre-trained language models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2021), which have significantly outperformed their supervised counterparts, as well as the context-independent embeddings. The models are pre-trained on large-scale unlabeled text and then fine-tuned on usually smaller task-specific labeled data Kalyan et al. (2021); Rogers et al. (2020). Effectively, the models facilitate transfer learning by learning generic latent feature representations from the large unannotated text and transferring the acquired knowledge across a broad range of downstream NLP tasks.

### **A.3 Bidirectional Encoder Representations from Transformers (BERT)**

The BERT model, one of the most popular pre-trained language models, learns to predict a word based on the previous words using the masked language modeling (MLM) objective. MLM trains a model to reconstruct text by predicting randomly masked tokens, given their surrounding context. Both the left and right contexts are considered when predicting masked tokens, resulting in the pre-training of deep bidirectional representations. Besides the MLM objective, BERT employs the Next Sentence Prediction (NSP) objective to predict whether a pair of input sentences are adjacent to each other. This is useful for tasks that require understand-

ing of sentence relations, such as question answering and natural language inference. Being a Transformer-based model, the architecture of BERT comprises a stack of encoder layers with multiple self-attention heads. Each head generates key, value, and query vectors for every input token, resulting in a weighted representation of the input sequence. The outputs of all attention heads in the same layer are combined and run through a fully connected layer. The encoder layers are wrapped with a skip connection and layer normalization.

BERT input embeddings are fixed-length vector representations that encapsulate word meanings and consist of the sum of token embeddings, segmentation embeddings, and position embeddings. The segment and position embeddings contribute to preserving the order of words. The segmentation embedding encodes the sentence number into a vector, which is essential in determining the sentence in which a particular token belongs. The position embedding provides information about the position of the token in a sentence. A special token [CLS] is added at the beginning of the sentence to denote a classification prediction, and a [SEP] token separates the input segments (separates two sentences). BERT has achieved remarkable success in token-level NLP tasks such as named entity recognition, where models are trained and evaluated on fine-grained token-level annotations (Rajpurkar et al., 2016; Sang and De Meulder, 2003). In addition to monolingual models, BERT provides a multilingual model (mBERT) (Devlin et al., 2018) that is pre-trained on Wikipedia data for 104 languages. The model outperforms static non-contextualized word embeddings in most cross-lingual transfer tasks.

RoBERTa (Liu et al., 2019b), another Transformer-based model, optimizes the training procedure of BERT for improved performance. Similar to BERT, RoBERTa is trained using a dynamic masking strategy but drops the next-sentence pretraining objective. In addition, RoBERTa modifies key hyperparameters in BERT, such as the use of larger mini-batches, training for a longer time, and scaling the training data to an order of magnitude more data than BERT. Besides being larger, the CommonCrawl News dataset<sup>2</sup> used for RoBERTa pre-training was more diverse than the Wikipedia text used in BERT. The Wikipedia text for the lower-resource languages is relatively small in size. XLM-RoBERTa, the multilingual version of RoBERTa, was pre-trained on 2.5 TB of filtered CommonCrawl data containing 100 languages.

---

<sup>2</sup> <https://commoncrawl.org/>

### A.3.1 Tokenization in Transformer-based Models

Prior to generating token embeddings, Transformer-based language models use subword-based tokenization algorithms to split the text into subword units. Other than enabling the model to process rare words by decomposing them into known subwords, subword tokenization enables the model to create a reasonable vocabulary size. The most commonly used subword tokenization algorithms are the WordPiece (Schuster and Nakajima, 2012) and Byte Pair Encoding (BPE) (Sennrich et al., 2016) algorithms, used in the BERT and RoBERTa models, respectively. While tightly coupled, the tokenization process consists of vocabulary construction and tokenization procedures. The vocabulary construction procedure generates a vocabulary of the desired size from a given text corpus, while tokenization takes the constructed vocabulary and applies it to new text to generate tokens.

**BERT WordPiece** tokenization begins with the creation of a base vocabulary that consists of the individual characters present in the training data. Rather than considering the frequency of symbol pairs as with byte-pair encoding (BPE), merge rules are iteratively applied to join the characters that maximize the likelihood score of language modeling, that is, when the probability of the merged symbol divided by the individual probabilities of the symbols is greater than any other symbol pair. Being a subword tokenization algorithm, WordPiece does not split frequently used words into smaller subwords, but rare words are decomposed into meaningful subwords. The decomposition of tokens into multiple subwords adds a prefix (double # character) to all the characters inside the word, which indicates that a given token is part of the previous word. For example, the word “Swine flu” will produce four tokens, namely “Sw”, “##ine”, “fl” and “##u”. Such subwords form the initial alphabet and, together with special tokens used by the model, create a small vocabulary used to train the WordPiece tokenizer. Whenever a particular subword is not found in the vocabulary, the entire word is tokenized as “UNK” (unknown).

**RoBERTa Byte-pair Tokenization:** RoBERTa tokenization is based on byte-pair encoding (BPE) algorithm, a subword tokenization technique that merges token pairs with the highest frequency count. Instead of using unicode characters, the Byte-pair tokenization uses bytes as the base subword units. Unlike unicode characters, bytes result in a small base vocabulary of size 256. Despite learning a modest size base vocabulary, byte-level BPE ensures encoding

of any input text without getting “unknown” tokens. Similar to WordPiece, BPE tokenization starts by building a base vocabulary comprising all the symbols that constitute the unique set of words of a given corpus. This is followed by learning merge rules, involving two symbols from the base vocabulary, to form a new symbol. This is done until the desired vocabulary size, a hyperparameter defined before training the tokenizer, is achieved. Based on the base vocabulary, the frequency of each possible symbol pair is determined, and the most frequently occurring pairs are selected, merged, and added to the vocabulary. Like the vocabulary size, the number of merges is a hyperparameter that has to be set.

## **A.4 Conclusions**

Various concepts related to sequential data modeling are discussed, including a theoretical description of different neural networks and pre-trained language models that were utilized in this study. More specifically, we focused on two variants of pre-trained Transformer-based multilingual models, mBERT and XLM-RoBERTa, which facilitate, with great success, the transfer of knowledge from across languages and tasks. Furthermore, we described the subword tokenization methods, namely WordPiece and BPE, applied in BERT and RoBERTa models respectively.

---

## Bibliography

---

Ács, J. (2019). *Exploring bert’s vocabulary* (cited on p. 79).

Adelani, D. I., Abbott, J., Neubig, G., D’souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S. M., Gwadabe, T. R., Ezeani, I., Niyongabo, R. A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E. P., Oyerinde, S., Siro, C., Bateesa, T. S., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., MBOUP, M., Gebreyohannes, D., Tilaye, H., Nwaike, K., Wolde, D., Faye, A., Sibanda, B., Ahia, O., Dossou, B. F. P., Ogueji, K., DIOP, T. I., Diallo, A., Akinfaderin, A., Marengereke, T., and Osei, S. (2021). “MasakhaNER: Named Entity Recognition for African Languages”. *Transactions of the Association for Computational Linguistics*, 9, pp. 1116–1131 (cited on p. 82).

Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019). “Docbert: Bert for document classification”. *arXiv preprint arXiv:1904.08398* (cited on p. 107).

Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., and Ellis, J. (2014). “A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet anno-

- tation standards”. In: *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pp. 45–53 (cited on p. 14).
- Ahmad, W. U., Peng, N., and Chang, K.-W. (2021). “GATE: graph attention transformer encoder for cross-lingual relation and event extraction”. In: *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*. Vol. 4, pp. 74–75 (cited on pp. 14, 30).
- Ahmad, W. U., Zhang, Z., Ma, X., Hovy, E., Chang, K.-W., and Peng, N. (2018). “On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing”. *arXiv preprint arXiv:1811.00570* (cited on p. 95).
- Ahn, D. (2006). “The stages of event extraction”. In: *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pp. 1–8 (cited on pp. 11, 24, 27, 31, 34).
- Akbik, A., Blythe, D., and Vollgraf, R. (2018). “Contextual string embeddings for sequence labeling”. In: *Proceedings of the 27th international conference on computational linguistics*, pp. 1638–1649 (cited on p. 107).
- Allan, J., Papka, R., and Lavrenko, V. (1998). “On-line new event detection and tracking”. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 37–45 (cited on p. 18).
- Aone, C., Halverson, L., Hampton, T., and Ramos-Santacruz, M. (1998). “SRA: Description of the IE2 system used for MUC-7”. In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998* (cited on p. 13).
- Arazo, E., Ortego, D., Albert, P., O’Connor, N. E., and McGuinness, K. (2020). “Pseudo-labeling and confirmation bias in deep semi-supervised learning”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8 (cited on p. 88).
- Arendarenko, E. and Kakkonen, T. (2012). “Ontology-based information and event extraction for business intelligence”. In: *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer, pp. 89–102 (cited on p. 22).

- Arsevska, E., Falala, S., Hervé, J. De Goër de, Lancelot, R., Rabatel, J., and Roche, M. (2017). “PADI-web: Platform for Automated extraction of animal Disease Information from the Web”. In: Adam Mickiewicz University (cited on p. [43](#)).
- Arsevska, E., Valentin, S., Rabatel, J., Hervé, J. De Goër de, Falala, S., Lancelot, R., and Roche, M. (2018). “Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System”. *PLoS One*, 13(8), e0199960 (cited on pp. [3](#), [8](#), [37](#), [38](#)).
- Atefeh, F. and Khreich, W. (2015). “A survey of techniques for event detection in twitter”. *Computational Intelligence*, 31(1), pp. 132–164 (cited on p. [18](#)).
- Atla, A., Tada, R., Sheng, V., and Singireddy, N. (2011). “Sensitivity of different machine learning algorithms to noise”. *Journal of Computing Sciences in Colleges*, 26(5), pp. 96–103 (cited on p. [89](#)).
- Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). “Layer normalization”. *arXiv preprint arXiv:1607.06450* (cited on p. [114](#)).
- Bahk, C. Y., Scales, D. A., Mekaru, S. R., Brownstein, J. S., and Freifeld, C. C. (2015). “Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting”. *BMC infectious diseases*, 15(1), pp. 1–6 (cited on p. [3](#)).
- Bai, S., Kolter, J. Z., and Koltun, V. (2018). “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling”. *arXiv preprint arXiv:1803.01271* (cited on p. [108](#)).
- Baker, C. F. and Sato, H. (2003). “The framenet data and software”. In: *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pp. 161–164 (cited on p. [26](#)).
- Balajee, S. A., Salyer, S. J., Greene-Cramer, B., Sadek, M., and Mounts, A. W. (2021). “The practice of event-based surveillance: concept and methods”. *Global Security: Health, Science and Policy*, 6(1), pp. 1–9 (cited on pp. [2](#), [36](#), [37](#), [43](#)).

- Beltagy, I., Lo, K., and Cohan, A. (2019). “SciBERT: A Pretrained Language Model for Scientific Text”. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3615–3620 (cited on p. 75).
- Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019). “Mixmatch: A holistic approach to semi-supervised learning”. *Advances in neural information processing systems*, 32 (cited on p. 87).
- Bethard, S., Savova, G., Palmer, M., and Pustejovsky, J. (2017). “SemEval-2017 Task 12: Clinical TempEval”. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 565–572 (cited on pp. 12, 16).
- Björne, J. and Salakoski, T. (2011). “Generalizing biomedical event extraction”. In: *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 183–191 (cited on p. 27).
- (2018). “Biomedical event extraction using convolutional neural networks and dependency parsing”. In: *Proceedings of the BioNLP 2018 workshop*, pp. 98–108 (cited on p. 24).
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). “Latent dirichlet allocation”. *Journal of machine Learning research*, 3(Jan), pp. 993–1022 (cited on pp. 90, 96).
- Bloom, D. E. and Cadarette, D. (2019). “Infectious disease threats in the twenty-first century: strengthening the global response”. *Frontiers in immunology*, 10, p. 549 (cited on p. 1).
- Boros, E. (2018). “Neural Methods for Event Extraction”. PhD thesis. Université Paris Saclay (COMUE) (cited on p. 32).
- Boros, E., Besançon, R., Ferret, O., and Grau, B. (2021). “The importance of character-level information in an event detection model”. In: *International Conference on Applications of Natural Language to Information Systems*. Springer, pp. 119–131 (cited on pp. 28, 107).
- Boros, E., Moreno, J. G., and Doucet, A. (2022). “Exploring Entities in Event Detection as Question Answering”. In: *European Conference on Information Retrieval*. Springer, pp. 65–79 (cited on p. 35).



- Borsje, J., Hogenboom, F., and Frasinca, F. (2010). “Semi-automatic financial events discovery based on lexico-semantic patterns”. *International Journal of Web Engineering and Technology*, 6(2), pp. 115–140 (cited on pp. [20](#), [22](#)).
- Breiman, L. (2001). “Random forests”. *Machine learning*, 45(1), pp. 5–32 (cited on p. [54](#)).
- Brownstein, J. S., Freifeld, C. C., Reis, B. Y., and Mandl, K. D. (2008). “Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project”. *PLoS medicine*, 5(7), e151 (cited on p. [36](#)).
- Bui, Q.-C., Campos, D., Mulligen, E. van, and Kors, J. (2013). “A fast rule-based approach for biomedical event extraction”. In: *proceedings of the BioNLP shared task 2013 workshop*, pp. 104–108 (cited on p. [22](#)).
- Bui, Q.-C. and Sloot, P. M. (2012). “A robust approach to extract biomedical events from literature”. *Bioinformatics*, 28(20), pp. 2654–2661 (cited on p. [22](#)).
- Burel, G., Saif, H., Fernandez, M., and Alani, H. (2017). “On semantics and deep learning for event detection in crisis situations” (cited on p. [29](#)).
- Busser, R. de and Moens, M.-F. (2006). “Information extraction and information technology”. In: Berlin, Heidelberg: Springer, pp. 1–22 (cited on p. [3](#)).
- Cao, K., Li, X., Fan, M., and Grishman, R. (2015). “Improving event detection with active learning”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 72–77 (cited on p. [21](#)).
- Cao, R., Zhu, S., Yang, C., Liu, C., Ma, R., Zhao, Y., Chen, L., and Yu, K. (2020). “Unsupervised dual paraphrasing for two-stage semantic parsing”. *arXiv preprint arXiv:2005.13485* (cited on p. [14](#)).
- Carrion, M. and Madoff, L. C. (2017). “ProMED-mail: 22 years of digital surveillance of emerging infectious diseases”. *International health*, 9(3), pp. 177–183 (cited on p. [45](#)).
- Caselli, T., Sprugnoli, R., Speranza, M., and Monachini, M. (2014). “EVENTI: Evaluation of Events and Temporal Information at Evalita 2014”. *EVENTI: Evaluation of Events and Temporal Information at Evalita 2014*, pp. 27–34 (cited on p. [17](#)).

- Chambers, N. and Jurafsky, D. (2011). “Template-based information extraction without the templates”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 976–986 (cited on p. 33).
- Chan, E. H., Brewer, T. F., Madoff, L. C., Pollack, M. P., Sonricker, A. L., Keller, M., Freifeld, C. C., Blench, M., Mawudeku, A., and Brownstein, J. S. (2010). “Global capacity for emerging infectious disease detection”. *Proceedings of the National Academy of Sciences*, 107(50), pp. 21701–21706 (cited on p. 2).
- Chau, E. C., Lin, L. H., and Smith, N. A. (2020). “Parsing with multilingual BERT, a small corpus, and a small treebank”. *arXiv preprint arXiv:2009.14124* (cited on p. 6).
- Chau, M. T., Esteves, D., and Lehmann, J. (2019). “Open-domain Event Extraction and Embedding for Natural Gas Market Prediction”. *arXiv preprint arXiv:1912.11334* (cited on p. 17).
- Chawla, N. V. (2009). “Data mining for imbalanced datasets: An overview”. *Data mining and knowledge discovery handbook*, pp. 875–886 (cited on p. 56).
- Chen, C. and Ng, V. (2012). “Joint modeling for chinese event extraction with rich linguistic features”. In: *Proceedings of COLING 2012*, pp. 529–544 (cited on p. 26).
- Chen, X., Wang, S., Fu, B., Long, M., and Wang, J. (2019a). “Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning” (cited on p. 86).
- Chen, Y. (2015). “Convolutional neural network for sentence classification”. MA thesis. University of Waterloo (cited on p. 108).
- Chen, Y., Liu, S., Zhang, X., Liu, K., and Zhao, J. (2017). “Automatically labeled data generation for large scale event extraction”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 409–419 (cited on pp. 24, 34).
- Chen, Y., Xu, L., Liu, K., Zeng, D., and Zhao, J. (2015). “Event extraction via dynamic multi-pooling convolutional neural networks”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Nat-*

- ural Language Processing (Volume 1: Long Papers)*, pp. 167–176 (cited on pp. [23](#), [24](#), [28](#), [31](#), [34](#), [107](#), [108](#)).
- Chen, Y., Yang, H., Liu, K., Zhao, J., and Jia, Y. (2018). “Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1267–1276 (cited on p. [23](#)).
- Chen, Y., Chen, T., Ebner, S., White, A. S., and Van Durme, B. (2019b). “Reading the manual: Event extraction as definition comprehension”. *arXiv preprint arXiv:1912.01586* (cited on p. [35](#)).
- Chen, Z. and Ji, H. (2009). “Language specific issue and feature exploration in Chinese event extraction”. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 209–212 (cited on p. [24](#)).
- Chiron, G., Doucet, A., Coustaty, M., Visani, M., and Moreux, J.-P. (2017). “Impact of OCR errors on the use of digital libraries: towards a better access to information”. In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, pp. 1–4 (cited on pp. [96](#), [97](#)).
- Cohen, J. (1960). “A coefficient of agreement for nominal scales”. *Educational and psychological measurement*, 20(1), pp. 37–46 (cited on p. [48](#)).
- Cohen, K. B., Verspoor, K., Johnson, H. L., Roeder, C., Ogren, P., Baumgartner Jr, W. A., White, E., and Hunter, L. (2009). “High-precision biological event extraction with a concept recognizer”. In: *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 50–58 (cited on p. [22](#)).
- Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., et al. (2008). “BioCaster: detecting public health rumors with a Web-based text mining system”. *Bioinformatics*, 24(24), pp. 2940–2941 (cited on pp. [6](#), [43](#)).
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). “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*. Association for Computational Linguistics, pp. 8440–8451 (cited on pp. [4](#), [65](#), [67](#), [75](#), [78](#), [79](#)).
- Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). “Supervised learning of universal sentence representations from natural language inference data”. *arXiv preprint arXiv:1705.02364* (cited on p. [117](#)).
- Cortes, C. and Vapnik, V. (1995). “Support-vector networks”. *Machine learning*, 20(3), pp. 273–297 (cited on p. [54](#)).
- Dehouck, M. and Denis, P. (2019). “Phylogenetic Multi-Lingual Dependency Parsing”. In: *NAACL 2019-Annual Conference of the North American Chapter of the Association for Computational Linguistics* (cited on p. [95](#)).
- Deng, S., Zhang, N., Kang, J., Zhang, Y., Zhang, W., and Chen, H. (2020). “Meta-learning with dynamic-memory-based prototypical network for few-shot event detection”. In: *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 151–159 (cited on p. [33](#)).
- Deschacht, K., De Belder, J., and Moens, M.-F. (2012). “The latent words language model”. *Computer Speech & Language*, 26(5), pp. 384–409 (cited on p. [23](#)).
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. *arXiv preprint arXiv:1810.04805* (cited on pp. [4](#), [23](#), [28](#), [30](#), [55](#), [75](#), [78](#), [117](#), [118](#)).
- Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., and Weischedel, R. M. (2004). “The automatic content extraction (ace) program-tasks, data, and evaluation.” In: *Lrec*. Vol. 2. 1. Lisbon, pp. 837–840 (cited on pp. [12](#), [13](#), [41](#)).
- Dórea, F. C. and Revie, C. W. (2021). “Data-driven surveillance: Effective collection, integration and interpretation of data to support decision-making”. *Frontiers in Veterinary Science*, 8, p. 225 (cited on p. [36](#)).

- Du, M., Pivovarov, L., and Yangarber, R. (2016). “PULS: natural language processing for business intelligence”. In: *Workshop on Human Language Technology and Intelligent Applications*. Go to Print Publisher (cited on p. 38).
- Du, X. and Cardie, C. (2020). “Event extraction by answering (almost) natural questions”. *arXiv preprint arXiv:2004.13625* (cited on p. 35).
- Duan, J., Ding, X., and Liu, T. (2018). “Learning sentence representations over tree structures for target-dependent classification”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 551–560 (cited on p. 29).
- Edmunds, A. and Morris, A. (2000). “The problem of information overload in business organisations: a review of the literature”. *International journal of information management*, 20(1), pp. 17–28 (cited on p. 5).
- Ellis, J., Getman, J., and Strassel, S. M. (2014). “Overview of linguistic resources for the tac kbp 2014 evaluations: Planning, execution, and results”. In: *Proceedings of TAC KBP 2014 Workshop, National Institute of Standards and Technology*, pp. 17–18 (cited on pp. 15, 41).
- Erk, K. (2009). “Representing words as regions in vector space”. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pp. 57–65 (cited on p. 28).
- Ferguson, J., Lockard, C., Weld, D. S., and Hajishirzi, H. (2018). “Semi-supervised event extraction with paraphrase clusters”. *arXiv preprint arXiv:1808.08622* (cited on pp. 19, 32).
- Finkenstaedt, T. and Wolff, D. (1973). *Ordered profusion: Studies in dictionaries and the English lexicon*. Vol. 13. Annales Universitatis Saraviensis. C. Winter (cited on p. 44).
- Freifeld, C. C., Mandl, K. D., Reis, B. Y., and Brownstein, J. S. (2008). “HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports”. *Journal of the American Medical Informatics Association*, 15(2), pp. 150–157 (cited on pp. 36, 39).

- Frisoni, G., Moro, G., and Carbonaro, A. (2021). “A Survey on Event Extraction for Natural Language Understanding: Riding the Biomedical Literature Wave”. *IEEE Access*, 9, pp. 160721–160757 (cited on pp. 12, 17).
- Fu, J., Liu, P., and Neubig, G. (2020a). “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 (cited on pp. 69, 70).
- Fu, J., Liu, P., and Zhang, Q. (2020b). “Rethinking generalization of neural models: A named entity recognition case study”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 7732–7739 (cited on p. 70).
- Gage, P. (1994). “A new algorithm for data compression”. *C Users Journal*, 12(2), pp. 23–38 (cited on p. 77).
- Garneau, N., Leboeuf, J.-S., and Lamontagne, L. (2018). “Predicting and interpreting embeddings for out of vocabulary words in downstream tasks”. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 331–333 (cited on p. 76).
- Ghaeini, R., Fern, X. Z., Huang, L., and Tadepalli, P. (2018). “Event nugget detection with forward-backward recurrent neural networks”. *arXiv preprint arXiv:1802.05672* (cited on pp. 29, 31, 34).
- Giguet, E. and Lucas, N. (2004). “La détection automatique des citations et des locuteurs dans les textes informatifs”. *Le discours rapporté dans tous ses états: Question de frontières*, pp. 410–418 (cited on p. 60).
- Greene, D., O’Callaghan, D., and Cunningham, P. (2014). “How many topics? stability analysis for topic models”. In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer, pp. 498–513 (cited on p. 97).
- Grishman, R. (2010). “The impact of task and corpus on event extraction systems”. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)* (cited on p. 14).

- Grishman, R. and Sundheim, B. M. (1996). “Message understanding conference-6: A brief history”. In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics* (cited on p. 12).
- Grishman, R., Westbrook, D., and Meyers, A. (2005). “Nyu’s english ace 2005 system description”. *ACE*, 5 (cited on pp. 12, 24, 31).
- Grootendorst, M. (2022). “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”. *arXiv preprint arXiv:2203.05794* (cited on p. 90).
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). “Domain-specific language model pretraining for biomedical natural language processing”. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1), pp. 1–23 (cited on p. 75).
- Hammerton, J. (2003). “Named entity recognition with long short-term memory”. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pp. 172–175 (cited on p. 107).
- Harris, Z. S. (1954). “Distributional structure”. *Word*, 10(2-3), pp. 146–162 (cited on p. 28).
- Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Brownstein, J., et al. (2010). “The landscape of international event-based bio-surveillance”. *Emerging Health Threats Journal*, 3(1), p. 7096 (cited on pp. 36, 37).
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Identity mappings in deep residual networks”. In: *European conference on computer vision*. Springer, pp. 630–645 (cited on p. 114).
- Hearst, M. A. (1992). “Automatic acquisition of hyponyms from large text corpora”. In: *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics* (cited on p. 20).
- Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2020). “A survey on recent approaches for natural language processing in low-resource scenarios”. *arXiv preprint arXiv:2010.12309* (cited on p. 5).

- Heymann, D. L., Rodier, G. R., et al. (2001). “Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases”. *The Lancet infectious diseases*, 1(5), pp. 345–353 (cited on p. 2).
- Hochreiter, S. and Schmidhuber, J. (1997). “Long short-term memory”. *Neural computation*, 9(8), pp. 1735–1780 (cited on p. 111).
- Hong, J., Kim, T., Lim, H., and Choo, J. (2021). “AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain” (cited on p. 76).
- Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., and Zhu, Q. (2011). “Using cross-entity inference to improve event extraction”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 1127–1136 (cited on pp. 24, 31).
- Hong, Y., Zhou, W., Zhang, J., Zhou, G., and Zhu, Q. (2018). “Self-regulation: Employing a generative adversarial network to improve event detection”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 515–526 (cited on p. 23).
- Hossin, M. and Sulaiman, M. N. (2015). “A review on evaluation metrics for data classification evaluations”. *International journal of data mining & knowledge management process*, 5(2), p. 1 (cited on p. 56).
- Howard, J. and Ruder, S. (2018). “Universal language model fine-tuning for text classification”. *arXiv preprint arXiv:1801.06146* (cited on p. 23).
- Huang, K., Altsosaar, J., and Ranganath, R. (2019). “Clinicalbert: Modeling clinical notes and predicting hospital readmission”. *arXiv preprint arXiv:1904.05342* (cited on p. 75).
- Huang, L., Ji, H., Cho, K., Dagan, I., Riedel, S., and Voss, C. (2018). “Zero-Shot Transfer Learning for Event Extraction”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2160–2170 (cited on pp. 24, 33).



- Huang, L., Ji, H., Cho, K., and Voss, C. R. (2017). “Zero-shot transfer learning for event extraction”. *arXiv preprint arXiv:1707.01066* (cited on p. 34).
- Huang, R. and Riloff, E. (2011). “Peeling back the layers: detecting event role fillers in secondary contexts”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1137–1147 (cited on pp. 27, 31).
- (2012a). “Bootstrapped training of event extraction classifiers”. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 286–295 (cited on p. 31).
- (2012b). “Modeling textual cohesion for event extraction”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 26. 1, pp. 1664–1670 (cited on p. 31).
- Huff, A. G., Breit, N., Allen, T., Whiting, K., and Kiley, C. (2016). “Evaluation and verification of the global rapid identification of threats system for infectious diseases in textual data sources”. *Interdisciplinary perspectives on infectious diseases*, 2016 (cited on p. 36).
- Hunter, L., Lu, Z., Firby, J., Baumgartner, W. A., Johnson, H. L., Ogren, P. V., and Cohen, K. B. (2008). “OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression”. *BMC bioinformatics*, 9(1), pp. 1–11 (cited on p. 22).
- Idubor, O. I., Kobayashi, M., Ndegwa, L., Okeyo, M., Galgalo, T., Kalani, R., Githii, S., Hunsperger, E., Balajee, A., Verani, J. R., et al. (2020). “Improving detection and response to respiratory events—Kenya, April 2016–April 2020”. *Morbidity and Mortality Weekly Report*, 69(18), p. 540 (cited on p. 36).
- Intxaurreondo, A., Agirre, E., De Lacalle, O. L., and Surdeanu, M. (2015). “Diamonds in the Rough: Event Extraction from Imperfect Microblog Data.” In: *HLT-NAACL*, pp. 641–650 (cited on p. 12).
- Jebara, K. B. and Shimshony, A. (2006). “International monitoring and surveillance of animal diseases using official and unofficial sources”. *Veterinaria italiana*, 42(4), pp. 431–441 (cited on pp. 2, 36).

- Jebara, T. (2001). “Discriminative, generative and imitative learning”. PhD thesis. PhD thesis, Media laboratory, MIT (cited on p. 54).
- (2012). *Machine learning: discriminative and generative*. Vol. 755. Springer Science & Business Media (cited on p. 54).
- Jeong, J., Lee, S., and Kwak, N. (2020). “Self-Training using Selection Network for Semi-supervised Learning.” In: *ICPRAM*, pp. 23–32 (cited on p. 90).
- Ji, H. (2009). “Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning”. In: *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, pp. 27–35 (cited on p. 31).
- Ji, H. and Grishman, R. (2008). “Refining event extraction through cross-document inference”. In: *Proceedings of ACL-08: Hlt*, pp. 254–262 (cited on pp. 24, 31).
- Ji, H. and Voss, C. (2021). “Low-Resource Event Extraction via Share-and-Transfer and Remaining Challenges”. *Computational Analysis of Storylines: Making Sense of Events*, p. 163 (cited on p. 23).
- Jiangde, Y., Xinfeng, X., and Xiaozhong, F. (2007). “Chinese text event information extraction based on hidden Markov model [J]”. *Microelectronics and Computer*, 24(10), pp. 92–94+ (cited on p. 27).
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). “MIMIC-III, a freely accessible critical care database”. *Scientific data*, 3(1), pp. 1–9 (cited on p. 75).
- Joshi, A., Karimi, S., Sparks, R., Paris, C., and Macintyre, C. R. (2019). “Survey of Text-based Epidemic Intelligence: A Computational Linguistics Perspective”. *ACM Computing Surveys (CSUR)*, 52(6), pp. 1–19 (cited on p. 52).
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). “The state and fate of linguistic diversity and inclusion in the NLP world”. *arXiv preprint arXiv:2004.09095* (cited on p. 6).

- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). “Fast-Text.zip: Compressing text classification models”. *arXiv preprint arXiv:1612.03651* (cited on p. 55).
- Jung, J., Im, J. H., Ko, Y.-J., Huh, K., Yoon, C.-g., Rhee, C., Kim, Y.-E., Go, D.-S., Kim, A., Jung, Y., et al. (2019). “Complementing conventional infectious disease surveillance with national health insurance claims data in the Republic of Korea”. *Scientific reports*, 9(1), pp. 1–9 (cited on p. 2).
- Kalyan, K. S., Rajasekharan, A., and Sangeetha, S. (2021). “Ammus: A survey of transformer-based pretrained models in natural language processing”. *arXiv preprint arXiv:2108.05542* (cited on p. 117).
- Karisani, P. and Karisani, N. (2021). “Semi-supervised text classification via self-pretraining”. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 40–48 (cited on p. 87).
- Kayi, E. S., Yadav, K., and Choi, H.-A. (2013). “Topic modeling based classification of clinical reports”. In: *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pp. 67–73 (cited on p. 90).
- Kim, J.-D., Ohta, T., and Tsujii, J. (2008). “Corpus annotation for mining biomedical events from literature”. *BMC bioinformatics*, 9(1), pp. 1–25 (cited on p. 16).
- Kim, Y. (2014). “Convolutional Neural Networks for Sentence Classification”. *CoRR*, abs/1408.5882 (cited on p. 107).
- Kingma, D. P. and Ba, J. (2014). “Adam: A method for stochastic optimization”. *arXiv preprint arXiv:1412.6980* (cited on p. 82).
- Kirasich, K., Smith, T., and Sadler, B. (2018). “Random forest vs logistic regression: binary classification for heterogeneous datasets”. *SMU Data Science Review*, 1(3), p. 9 (cited on p. 54).
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). “Overcoming catastrophic

- forgetting in neural networks”. *Proceedings of the national academy of sciences*, 114(13), pp. 3521–3526 (cited on p. 86).
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “Imagenet classification with deep convolutional neural networks”. *Advances in neural information processing systems*, 25, pp. 1097–1105 (cited on p. 108).
- Lai, V., Van Nguyen, M., Kaufman, H., and Nguyen, T. H. (2021a). “Event Extraction from Historical Texts: A New Dataset for Black Rebellions”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2390–2400 (cited on p. 31).
- Lai, V. D., Derroncourt, F., and Nguyen, T. H. (2020a). “Exploiting the matching information in the support set for few shot event classification”. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 233–245 (cited on p. 33).
- (2020b). “Extensively matching for few-shot learning event detection”. *arXiv preprint arXiv:2006.10093* (cited on pp. 28, 33).
- Lai, Z., Wang, C., Oliveira, L. C., Dugger, B. N., Cheung, S.-C., and Chuah, C.-N. (2021b). “Joint Semi-supervised and Active Learning for Segmentation of Gigapixel Pathology Images with Cost-Effective Labeling”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 591–600 (cited on pp. 87, 90).
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). “Neural architectures for named entity recognition”. *arXiv preprint arXiv:1603.01360* (cited on p. 107).
- Landis, J. R. and Koch, G. G. (1977). “The measurement of observer agreement for categorical data”. *biometrics*, pp. 159–174 (cited on p. 48).
- Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. (2020). “From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers”. *arXiv preprint arXiv:2005.00633* (cited on pp. 61, 75).
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). “Deep learning”. *nature*, 521(7553), pp. 436–444 (cited on pp. 5, 56).

- Lee, D. and Seung, H. S. (2000). “Algorithms for non-negative matrix factorization”. *Advances in neural information processing systems*, 13 (cited on pp. [96](#), [97](#)).
- Lee, D. D. and Seung, H. S. (1999). “Learning the parts of objects by non-negative matrix factorization”. *Nature*, 401(6755), pp. 788–791 (cited on pp. [96](#), [97](#)).
- Lee, H.-G., Park, S.-Y., Rim, H.-C., Lee, D.-G., and Chun, H.-W. (2015). “A maximum entropy-based bio-molecular event extraction model that considers event generation”. *Journal of Information Processing Systems*, 11(2), pp. 248–265 (cited on p. [27](#)).
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. *Bioinformatics*, 36(4), pp. 1234–1240 (cited on p. [75](#)).
- Lehnert, W. (1990). “Symbolic/subsymbolic sentence analysis: Exploiting the best of two worlds”. *Advances in connectionist and neural computation theory*, 1, pp. 135–164 (cited on p. [21](#)).
- Lei, S., Zhang, H., Wang, K., and Su, Z. (2018). “How training data affect the accuracy and robustness of neural networks for image classification” (cited on p. [60](#)).
- Lei, T., Barzilay, R., and Jaakkola, T. (2015). “Molding cnns for text: non-linear, non-consecutive convolutions”. *arXiv preprint arXiv:1508.04112* (cited on p. [29](#)).
- Lejeune, G. (2013). “Veille épidémiologique multilingue: une approche parcimonieuse au grain caractère fondée sur le genre textuel”. PhD thesis (cited on pp. [53](#), [60](#)).
- Lejeune, G., Brixstel, R., Doucet, A., and Lucas, N. (2015). “Multilingual event extraction for epidemic detection”. *Artificial intelligence in medicine*, 65(2), pp. 131–143 (cited on pp. [5](#), [6](#), [8](#), [41](#), [42](#), [46](#), [53](#), [54](#), [65](#)).
- Lejeune, G., Brixstel, R., Lecluze, C., Doucet, A., and Lucas, N. (2013). “Added-value of automatic multilingual text analysis for epidemic surveillance”. In: *Artificial Intelligence in Medicine (AIME)*, pp. 284–294 (cited on p. [4](#)).
- Lejeune, G., Doucet, A., Yangarber, R., and Lucas, N. (2010). “Filtering news for epidemic surveillance: towards processing more languages with fewer resources”. In: *CLIA/COLING*, pp. 3–10 (cited on p. [60](#)).

- Lewis, D. D. (1998). “Naive (Bayes) at forty: The independence assumption in information retrieval”. In: *European conference on machine learning*. Springer, pp. 4–15 (cited on p. 54).
- Lex, A., Gehlenborg, N., Strobel, H., Vuillemot, R., and Pfister, H. (2014). “UpSet: visualization of intersecting sets”. *IEEE transactions on visualization and computer graphics*, 20(12), pp. 1983–1992 (cited on p. 69).
- Li, D., Huang, L., Ji, H., and Han, J. (2020a). “Biomedical event extraction based on knowledge-driven tree-LSTM”. In: *Proc. 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT2019)* (cited on pp. 24, 25, 29).
- Li, F., Peng, W., Chen, Y., Wang, Q., Pan, L., Lyu, Y., and Zhu, Y. (2020b). “Event extraction as multi-turn question answering”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 829–838 (cited on p. 35).
- Li, L., Liu, Y., and Qin, M. (2018). “Extracting biomedical events with parallel multi-pooling convolutional neural networks”. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(2), pp. 599–607 (cited on p. 29).
- Li, P., Zhou, G., Zhu, Q., and Hou, L. (2012). “Employing compositional semantics and discourse consistency in Chinese event extraction”. In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 1006–1016 (cited on pp. 24, 31).
- Li, Q., Ji, H., Hong, Y., and Li, S. (2014). “Constructing information networks using one single model”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1846–1851 (cited on pp. 25, 26, 31).
- Li, Q., Ji, H., and Huang, L. (2013). “Joint event extraction via structured prediction with global features”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 73–82 (cited on pp. 25, 26, 31, 34).
- Li, Q., Li, J., Sheng, J., Cui, S., Wu, J., Hei, Y., Peng, H., Guo, S., Wang, L., Beheshti, A., et al. (2021). “A Compact Survey on Event Extraction: Approaches and Applications”. *arXiv preprint arXiv:2107.02126* (cited on pp. 24, 25, 34).

- Li, X. and Gong, H. (2021). “Robust optimization for multilingual translation with imbalanced data”. *Advances in Neural Information Processing Systems*, 34, pp. 25086–25099 (cited on p. 6).
- Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., and Zhang, C. (2020). “BOND: Bert-Assisted Open-Domain Named Entity Recognition with Distant Supervision”. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (cited on p. 88).
- Liang, H., Sun, X., Sun, Y., and Gao, Y. (2017). “Text feature extraction based on deep learning: a review”. *EURASIP journal on wireless communications and networking*, 2017(1), pp. 1–12 (cited on pp. 55, 108).
- Liang, Z., Noriega-Atala, E., Morrison, C., and Surdeanu, M. (2022). “Low Resource Causal Event Detection from Biomedical Literature”. In: *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 252–263 (cited on p. 9).
- Liao, S. and Grishman, R. (2010). “Using document level cross-event inference to improve event extraction”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 789–797 (cited on pp. 24, 33).
- (2011). “Acquiring topic features to improve event extraction: in pre-selected and balanced collections”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 9–16 (cited on p. 33).
- Lin, B. Y., Lee, D.-H., Shen, M., Moreno, R., Huang, X., Shiralkar, P., and Ren, X. (2020). “Triggerer: Learning with entity triggers as explanations for named entity recognition”. *arXiv preprint arXiv:2004.07493* (cited on p. 14).
- Lin, C., Miller, T., Dligach, D., Bethard, S., and Savova, G. (2019). “A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction”. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 65–71 (cited on p. 107).
- Lin, H., Lu, Y., Han, X., and Sun, L. (2018). “Nugget proposal networks for Chinese event detection”. *arXiv preprint arXiv:1805.00249* (cited on p. 31).
- Linge, J. P., Steinberger, R., Fuart, F., Bucci, S., Belyaeva, J., Gemo, M., Al-Khudhairi, D., Yangarber, R., and Goot, E. van der (2010). “MedISys: medical information system”. In:

- Advanced ICTs for disaster management and threat detection: Collaborative and distributed frameworks*. IGI Global, pp. 131–142 (cited on pp. 6, 37, 43).
- Liu, B., Wei, Y., Zhang, Y., and Yang, Q. (2017). “Deep Neural Networks for High Dimension, Low Sample Size Data.” In: *IJCAI*, pp. 2287–2293 (cited on p. 5).
- Liu, J., Chen, Y., Liu, K., Bi, W., and Liu, X. (2020). “Event extraction as machine reading comprehension”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1641–1651 (cited on p. 35).
- Liu, J., Chen, Y., Liu, K., and Zhao, J. (2018a). “Event detection via gated multilingual attention mechanism”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1 (cited on pp. 31, 34).
- Liu, M., Liu, Y., Xiang, L., Chen, X., and Yang, Q. (2008). “Extracting key entities and significant events from online daily news”. In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, pp. 201–209 (cited on pp. 18, 23).
- Liu, S., Liu, K., He, S., and Zhao, J. (2016). “A probabilistic soft logic based approach to exploiting latent and global information in event classification”. In: *Thirtieth AAAI Conference on Artificial Intelligence* (cited on pp. 31, 34).
- Liu, X., Huang, H., and Zhang, Y. (2019a). “Open domain event extraction using neural latent variable models”. *arXiv preprint arXiv:1906.06947* (cited on p. 17).
- Liu, X., Luo, Z., and Huang, H. (2018b). “Jointly multiple events extraction via attention-based graph information aggregation”. *arXiv preprint arXiv:1809.09078* (cited on pp. 23, 26, 30).
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). “Roberta: A robustly optimized bert pretraining approach”. *arXiv preprint arXiv:1907.11692* (cited on p. 118).
- Liu, Z., Lin, W., Shi, Y., and Zhao, J. (2021). “A Robustly Optimized BERT Pre-training Approach with Post-training”. In: *China National Conference on Chinese Computational Linguistics*. Springer, pp. 471–484 (cited on p. 117).



- Llorens, H., Saquete, E., and Navarro, B. (2010). “TimeML events recognition and classification: learning CRF models with semantic roles”. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 725–733 (cited on p. 27).
- Lu, W. and Roth, D. (2012). “Automatic event extraction with structured preference modeling”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 835–844 (cited on pp. 24, 34).
- Lu, Z. and Nie, J.-Y. (2019). “RALIGRAPH at HASOC 2019: VGCN-BERT: Augmenting BERT with Graph Embedding for Offensive Language Detection”. In: (cited on p. 55).
- Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., and Hajishirzi, H. (2019). “A general framework for information extraction using dynamic span graphs”. *arXiv preprint arXiv:1904.03296* (cited on p. 30).
- Lucas, N. (2009). “Modélisation différentielle du texte, de la linguistique aux algorithmes”. PhD thesis. Université de Caen (cited on p. 60).
- Lyu, C. (2022). “Knowledge and Pre-trained Language Models Inside and Out: a deep-dive into datasets and external knowledge” (cited on p. 4).
- Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., et al. (1999). “Performance measures for information extraction”. In: *Proceedings of DARPA broadcast news workshop*. Herndon, VA, pp. 249–252 (cited on p. 66).
- Mawudeku, A. and Blench, M. (2005). “Global public health intelligence network (GPHIN)”. In: *Proceedings of Machine Translation Summit X: Invited papers* (cited on pp. 3, 43).
- McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). “Learned in translation: Contextualized word vectors”. *arXiv preprint arXiv:1708.00107* (cited on p. 117).
- McClosky, D., Charniak, E., and Johnson, M. (2006). “Effective self-training for parsing”. In: *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*. Citeseer, pp. 152–159 (cited on p. 88).

- McClosky, D., Surdeanu, M., and Manning, C. D. (2011). “Event extraction as dependency parsing”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1626–1635 (cited on p. 31).
- Mejri, M. and Akaichi, J. (2017). “A Survey of Textual Event Extraction from Social Networks.” In: *LPKM* (cited on pp. 17, 20).
- Mihalcea, R. and Chklovski, T. (2003). “Open mind word expert: Creating large annotated data collections with web users’ help”. In: *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003* (cited on p. 5).
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). “Efficient estimation of word representations in vector space”. *arXiv preprint arXiv:1301.3781* (cited on p. 23).
- (2013b). “Efficient Estimation of Word Representations in Vector Space”. *Computer Science* (cited on p. 76).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013c). “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*, pp. 3111–3119 (cited on pp. 23, 108).
- Miller, T., Dligach, D., and Savova, G. (2016). “Unsupervised document classification with informed topic models”. In: *Proceedings of the 15th workshop on biomedical natural language processing*, pp. 83–91 (cited on p. 90).
- Misra, R. (2018). *News Category Dataset* (cited on p. 46).
- Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., and Strassel, S. (2015). “Event nugget annotation: Processes and issues”. In: *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pp. 66–76 (cited on pp. 15, 31).
- Munkhdalai, T., Namsrai, O.-E., and Ryu, K. H. (2015). “Self-training in significance space of support vectors for imbalanced biomedical event data”. *BMC bioinformatics*, 16(7), pp. 1–8 (cited on p. 32).

- Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., and Odeo, M. (2020). “Multilingual Epidemiological Text Classification: A Comparative Study”. In: *COLING, International Conference on Computational Linguistics* (cited on pp. [61](#), [75](#)).
- (2021). “Token-level multilingual epidemic dataset for event extraction”. In: *International Conference on Theory and Practice of Digital Libraries*. Springer, pp. 55–59 (cited on pp. [8](#), [46](#), [48](#)).
- Naughton, M., Kushmerick, N., and Carthy, J. (2006). “Event extraction from heterogeneous news sources”. In: *proceedings of the AAI workshop event extraction and synthesis*, pp. 1–6 (cited on p. [33](#)).
- Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-J., Ohta, T., Pyysalo, S., and Zweigenbaum, P. (2013). “Overview of BioNLP shared task 2013”. In: *Proceedings of the BioNLP shared task 2013 workshop*, pp. 1–7 (cited on p. [16](#)).
- Neubig, G., Dou, Z.-Y., Hu, J., Michel, P., Pruthi, D., Wang, X., and Wieting, J. (2019). “comparemt: A tool for holistic comparison of language generation systems”. *arXiv preprint arXiv:1903.07926* (cited on p. [70](#)).
- Nguyen, T. and Grishman, R. (2018). “Graph convolutional networks with argument-aware pooling for event detection”. In: *Proceedings of the AAI Conference on Artificial Intelligence*. Vol. 32. 1 (cited on pp. [23](#), [24](#), [28](#)).
- Nguyen, T. H., Cho, K., and Grishman, R. (2016). “Joint event extraction via recurrent neural networks”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 300–309 (cited on pp. [23](#), [26](#), [28](#), [29](#), [31](#), [34](#)).
- Nguyen, T. H. and Grishman, R. (2015). “Event detection and domain adaptation with convolutional neural networks”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 365–371 (cited on pp. [23](#), [24](#), [28](#), [31](#), [34](#), [107](#), [108](#)).

- Nguyen, T. H. and Grishman, R. (2016). “Modeling skip-grams for event detection with convolutional neural networks”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 886–891 (cited on pp. [23](#), [24](#), [29](#)).
- Nguyen, T. M. and Nguyen, T. H. (2019). “One for all: Neural joint modeling of entities and events”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 6851–6858 (cited on p. [30](#)).
- Njeru, I., Kareko, D., Kisangau, N., Langat, D., Liku, N., Owiso, G., Dolan, S., Rabinowitz, P., Macharia, D., Ekechi, C., et al. (2020). “Use of technology for public health surveillance reporting: opportunities, challenges and lessons learnt from Kenya”. *BMC Public Health*, 20(1), pp. 1–11 (cited on p. [2](#)).
- Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G. S., and Dean, J. (2013). “Zero-shot learning by convex combination of semantic embeddings”. *arXiv preprint arXiv:1312.5650* (cited on pp. [61](#), [75](#)).
- Ochani, R. K., Batra, S., Shaikh, A., and Asad, A. (2019). “Nipah virus-the rising epidemic: a review”. *Infez Med*, 27(2), pp. 117–27 (cited on p. [1](#)).
- Ogueji, K., Zhu, Y., and Lin, J. (2021). “Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages”. In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116–126 (cited on p. [6](#)).
- Okamoto, M. and Kikuchi, M. (2009). “Discovering volatile events in your neighborhood: Local-area topic extraction from blog entries”. In: *Asia Information Retrieval Symposium*. Springer, pp. 181–192 (cited on p. [23](#)).
- Olivier, C., Bernhard, S., and Alexander, Z. (2006). “Semi-supervised learning”. In: *IEEE Transactions on Neural Networks*. Vol. 20. 3, pp. 542–542 (cited on p. [87](#)).
- Otte, J. and Pica-Ciamarra, U. (2021). “Emerging infectious zoonotic diseases: The neglected role of food animals”. *One Health*, 13, p. 100323 (cited on p. [1](#)).
- Pan, S. J. and Yang, Q. (2009). “A survey on transfer learning”. *IEEE Transactions on knowledge and data engineering*, 22(10), pp. 1345–1359 (cited on p. [23](#)).

- Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J., Gomes, G., Koppeschaar, C., Rehn, M., Smallenburg, R., Turbelin, C., et al. (2014). “Web-based participatory surveillance of infectious diseases: the InfluenzaNet participatory surveillance experience”. *Clinical Microbiology and Infection*, 20(1), pp. 17–21 (cited on pp. 3, 35).
- Paquet, C., Coulombier, D., Kaiser, R., and Ciotti, M. (2006). “Epidemic intelligence: a new framework for strengthening disease surveillance in Europe”. *Eurosurveillance*, 11(12), pp. 5–6 (cited on p. 2).
- Patwardhan, S. and Riloff, E. (2007). “Effective information extraction with semantic affinity patterns and relevant regions”. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 717–727 (cited on p. 32).
- (2009). “A unified model of phrasal and sentential evidence for information extraction”. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 151–160 (cited on p. 27).
- Pennington, J., Socher, R., and Manning, C. D. (2014). “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543 (cited on pp. 23, 108).
- Pérez-Cruz, F. (2008). “Kullback-Leibler divergence estimation of continuous distributions”. In: *2008 IEEE international symposium on information theory*. IEEE, pp. 1666–1670 (cited on p. 89).
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018a). “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237 (cited on p. 23).
- (2018b). “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

- Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237 (cited on p. 30).
- Petroni, F., Raman, N., Nugent, T., Nourbakhsh, A., Panić, Ž., Shah, S., and Leidner, J. L. (2018). “An extensible event extraction system with cross-media event resolution”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 626–635 (cited on p. 19).
- Pires, T., Schlinger, E., and Garrette, D. (2019). “How Multilingual is Multilingual BERT?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001 (cited on p. 94).
- Piskorski, J., Tanev, H., Atkinson, M., Goot, E. v. d., and Zavarella, V. (2011). “Online news event extraction for global crisis surveillance”. In: *Transactions on computational collective intelligence V*. Springer, pp. 182–212 (cited on pp. 9, 18, 21, 53).
- Piskorski, J., Tanev, H., and Oezden Wennerberg, P. (2007). “Extracting violent events from on-line news for ontology population”. In: *International conference on business information systems*. Springer, pp. 287–300 (cited on p. 21).
- Poerner, N., Waltinger, U., and Schütze, H. (2020). “Inexpensive Domain Adaptation of Pre-trained Language Models: Case Studies on Biomedical NER and Covid-19 QA”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online (cited on p. 76).
- Poon, H. and Vanderwende, L. (2010). “Joint inference for knowledge extraction from biomedical literature”. In: *Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 813–821 (cited on p. 25).
- Powers, D. M. (2020). “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. *arXiv preprint arXiv:2010.16061* (cited on p. 56).
- Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). “TimeML: Robust specification of event and temporal expressions in text.” *New directions in question answering*, 3, pp. 28–34 (cited on p. 16).

- Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). “ISO-TimeML: An International Standard for Semantic Annotation.” In: *LREC*. Vol. 10, pp. 394–397 (cited on p. 16).
- Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). “Pre-trained models for natural language processing: A survey”. *Science China Technological Sciences*, 63(10), pp. 1872–1897 (cited on p. 51).
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). “Language models are unsupervised multitask learners”. *OpenAI blog*, 1(8), p. 9 (cited on p. 23).
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). “Squad: 100,000+ questions for machine comprehension of text”. *arXiv preprint arXiv:1606.05250* (cited on p. 118).
- Ramponi, A., Goot, R. van der, Lombardo, R., and Plank, B. (2020). “Biomedical event extraction as sequence labeling”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5357–5367 (cited on p. 34).
- Rao, C. and Gudivada, V. N. (2018). *Computational analysis and understanding of natural languages: principles, methods and applications*. Elsevier (cited on p. 51).
- Rehurek, R. and Sojka, P. (2011). “Gensim–python framework for vector space modelling”. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2) (cited on p. 97).
- Ribeiro, S., Ferret, O., and Tannier, X. (2017). “Unsupervised event clustering and aggregation from newswire and web articles”. In: *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pp. 62–67 (cited on p. 33).
- Riedel, S., Chun, H.-W., Takagi, T., and Tsujii, J. (2009). “A markov logic approach to biomedical event extraction”. In: *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 41–49 (cited on p. 25).
- Riedel, S. and McCallum, A. (2011a). “Fast and robust joint models for biomedical event extraction”. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1–12 (cited on p. 25).

- Riedel, S. and McCallum, A. (2011b). “Robust biomedical event extraction with dual decomposition and minimal domain adaptation”. In: *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 46–50 (cited on p. 25).
- Riloff, E. (1996). “Automatically generating extraction patterns from untagged text”. In: *Proceedings of the national conference on artificial intelligence*, pp. 1044–1049 (cited on p. 32).
- Riloff, E. and Shoen, J. (1995). “Automatically acquiring conceptual patterns without an annotated corpus”. In: *Third Workshop on Very Large Corpora* (cited on p. 21).
- Rish, I. et al. (2001). “An empirical study of the naive Bayes classifier”. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22, pp. 41–46 (cited on p. 54).
- Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). “A primer in bertology: What we know about how bert works”. *Transactions of the Association for Computational Linguistics*, 8, pp. 842–866 (cited on p. 117).
- Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). “Semi-supervised self-training of object detection models” (cited on pp. 87, 88).
- Roser, M. and Ritchie, H. (2021). “Burden of Disease”. *Our World in Data*. <https://ourworldindata.org/burden-of-disease> (cited on p. 1).
- Rubenstein, H. and Goodenough, J. B. (1965). “Contextual correlates of synonymy”. *Communications of the ACM*, 8(10), pp. 627–633 (cited on p. 28).
- Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. (2021). “How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 3118–3135 (cited on p. 79).
- Rutherford, G. W. (1998). “Public Health, Communicable Diseases, and Managed Care:: Will Managed Care Improve or Weaken Communicable Disease Control?” *American journal of preventive medicine*, 14(3), pp. 53–59 (cited on p. 2).



- Saha, S., Majumder, A., Hasanuzzaman, M., and Ekbal, A. (2011). “Bio-molecular event extraction using Support Vector Machine”. In: *2011 Third International Conference on Advanced Computing*. IEEE, pp. 298–303 (cited on p. 27).
- Sahlgren, M. (2008). “The distributional hypothesis”. *Italian Journal of Disability Studies*, 20, pp. 33–53 (cited on p. 28).
- Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). “Earthquake shakes twitter users: real-time event detection by social sensors”. In: *Proceedings of the 19th international conference on World wide web*, pp. 851–860 (cited on p. 27).
- Sang, E. F. and De Meulder, F. (2003). “Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition”. *arXiv preprint cs/0306050* (cited on p. 118).
- Santos, C. N. d. and Guimaraes, V. (2015). “Boosting named entity recognition with neural character embeddings”. *arXiv preprint arXiv:1505.05008* (cited on p. 107).
- Sarker, I. H. (2021). “Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions”. *SN Computer Science*, 2(6), pp. 1–20 (cited on p. 5).
- Schick, T. and Schütze, H. (2020). “Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8766–8774 (cited on p. 79).
- Schuster, M. and Nakajima, K. (2012). “Japanese and korean voice search”. In: *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5149–5152 (cited on pp. 58, 76, 79, 119).
- Schuster, M. and Paliwal, K. K. (1997). “Bidirectional recurrent neural networks”. *IEEE transactions on Signal Processing*, 45(11), pp. 2673–2681 (cited on p. 108).
- Sennrich, R., Haddow, B., and Birch, A. (2016). “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725 (cited on pp. 77, 79, 119).

- Sha, L., Qian, F., Chang, B., and Sui, Z. (2018). “Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1 (cited on pp. 14, 29).
- Sheng, J., Guo, S., Yu, B., Li, Q., Hei, Y., Wang, L., Liu, T., and Xu, H. (2021). “Case: A joint learning framework with cascade decoding for overlapping event extraction”. *arXiv preprint arXiv:2107.01583* (cited on p. 19).
- Silfverberg, M. and Rueter, J. (2015). “Can Morphological Analyzers Improve the Quality of Optical Character Recognition?” In: *Septentrio Conference Series*. 2, pp. 45–56 (cited on p. 96).
- Song, Z., Bies, A., Strassel, S., Ellis, J., Mitamura, T., Dang, H. T., Yamakawa, Y., and Holm, S. (2016). “Event nugget and event coreference annotation”. In: *Proceedings of the Fourth Workshop on Events*, pp. 37–45 (cited on p. 15).
- Song, Z., Bies, A., Strassel, S. M., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., Ma, X., et al. (2015). “From Light to Rich ERE: Annotation of Entities, Relations, and Events.” In: *EVENTS@ HLP-NAACL*, pp. 89–98 (cited on pp. 14, 15).
- Soper, E., Fujimoto, S., and Yu, Y.-Y. (2021). “BART for Post-Correction of OCR Newspaper Text”. In: *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pp. 284–290 (cited on p. 7).
- Sprugnoli, R. (2018). “Event detection and classification for the digital humanities”. PhD thesis. University of Trento (cited on p. 18).
- Sprugnoli, R. and Tonelli, S. (2017). “One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective”. *Natural language engineering*, 23(4), pp. 485–506 (cited on p. 12).
- Srivastava, A., Makhija, P., and Gupta, A. (2020). “Noisy text data: Achilles’ heel of BERT”. In: *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pp. 16–21 (cited on pp. 5, 7).

- Strassel, S. M., Graff, D., Martey, N., and Cieri, C. (2000). “Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora.” In: *LREC* (cited on p. 18).
- Styler, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., De Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., et al. (2014). “Temporal annotation in the clinical domain”. *Transactions of the association for computational linguistics*, 2, pp. 143–154 (cited on p. 16).
- Subramaniam, L. V., Roy, S., Faruque, T. A., and Negi, S. (2009). “A survey of types of text noise and techniques to handle noisy text”. In: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, pp. 115–122 (cited on p. 5).
- Sun, F., Li, F.-L., Wang, R., Chen, Q., Cheng, X., and Zhang, J. (2021). “K-AID: Enhancing Pre-trained Language Models with Domain Knowledge for Question Answering”. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 4125–4134 (cited on p. 74).
- Sun, W., Rumshisky, A., and Uzuner, O. (2013). “Evaluating temporal relations in clinical text: 2012 i2b2 challenge”. *Journal of the American Medical Informatics Association*, 20(5), pp. 806–813 (cited on p. 16).
- Sundheim, B. M. (1991). “Third message understanding evaluation and conference (muc-3): Phase 1 status report”. In: *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991* (cited on p. 33).
- Tai, K. S., Socher, R., and Manning, C. D. (2015). “Improved semantic representations from tree-structured long short-term memory networks”. *arXiv preprint arXiv:1503.00075* (cited on p. 29).
- Tai, W., Kung, H., Dong, X. L., Comiter, M., and Kuo, C.-F. (2020). “exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 1433–1439 (cited on pp. 4, 76).

- Tanev, H., Piskorski, J., and Atkinson, M. (2008). “Real-time news event extraction for global crisis monitoring”. In: *International Conference on Application of Natural Language to Information Systems*. Springer, pp. 207–218 (cited on pp. 18, 23).
- Tarvainen, A. and Valpola, H. (2017). “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results”. *arXiv preprint arXiv:1703.01780* (cited on p. 88).
- Thacker, S. B. and Berkelman, R. L. (1988). “Public health surveillance in the United States”. *Epidemiologic reviews*, 10(1), pp. 164–190 (cited on p. 2).
- Tian, L., Zhang, X., and Lau, J. H. (2021). “Rumour Detection via Zero-shot Cross-lingual Transfer Learning”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 603–618 (cited on p. 75).
- Tobius, B. S., Babirye, C., Nakatumba-Nabende, J., and Katumba, A. (2022). “A Comparison of Topic Modeling and Classification Machine Learning Algorithms on Luganda Data”. In: *3rd Workshop on African Natural Language Processing* (cited on p. 90).
- Tong, M., Xu, B., Wang, S., Cao, Y., Hou, L., Li, J., and Xie, J. (2020). “Improving event detection via open-domain trigger knowledge”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5887–5897 (cited on p. 28).
- Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Versley, Y., Candito, M., Foster, J., Rehbein, I., and Tounsi, L. (2010). “Statistical parsing of morphologically rich languages (spmrl) what, how and whither”. In: *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pp. 1–12 (cited on pp. 45, 50).
- Tuo, A., Besançon, R., Ferret, O., and Tourille, J. (2022). “Better Exploiting BERT for Few-Shot Event Detection”. In: *International Conference on Applications of Natural Language to Information Systems*. Springer, pp. 291–298 (cited on p. 33).
- Uddin, S., Khan, A., Hossain, M. E., and Moni, M. A. (2019). “Comparing different supervised machine learning algorithms for disease prediction”. *BMC medical informatics and decision making*, 19(1), pp. 1–16 (cited on p. 52).

- UNEP and ILRI, U. N. E. P.-I. L. R. I. (2020). *Preventing the next pandemic - Zoonotic diseases and how to break the chain of transmission | UNEP - UN Environment Programme*. [https://www.unep.org/resources/report/preventing-future-zoonotic-disease-outbreaks-protecting-environment-animals-and?\\_ga=2.188238440.1789729362.1660202212-1548959296.1660202212](https://www.unep.org/resources/report/preventing-future-zoonotic-disease-outbreaks-protecting-environment-animals-and?_ga=2.188238440.1789729362.1660202212-1548959296.1660202212). (Accessed on 08/11/2022) (cited on p. 1).
- Valentin, S. (2020). “Extraction and combination of epidemiological information from informal sources for animal infectious diseases surveillance”. PhD thesis. Université Montpellier (cited on pp. 7, 89, 117).
- Valentin, S., Arsevska, E., Falala, S., De Goër, J., Lancelot, R., Mercier, A., Rabatel, J., and Roche, M. (2020a). “PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases”. *Computers and Electronics in Agriculture*, 169, p. 105163 (cited on p. 36).
- Valentin, S., Arsevska, E., Mercier, A., Falala, S., Rabatel, J., Lancelot, R., and Roche, M. (2017). “PADI-web: an event-based surveillance system for detecting, classifying and processing online news”. In: *Language and Technology Conference*. Springer, pp. 87–101 (cited on p. 43).
- Valentin, S., Lancelot, R., and Roche, M. (2020b). “Automated processing of multilingual online news for the monitoring of animal infectious diseases”. In: ELRA (cited on pp. 38, 53).
- (2020c). “Information retrieval for animal disease surveillance: a pattern-based approach”. In: Association for Computational Linguistics (cited on pp. 3, 21, 38).
- (2021). “Identifying associations between epidemiological entities in news data for animal disease surveillance”. *Artificial Intelligence in Agriculture*, 5, pp. 163–174 (cited on pp. 4, 6, 37, 40).
- Van Engelen, J. E. and Hoos, H. H. (2020). “A survey on semi-supervised learning”. *Machine Learning*, 109(2), pp. 373–440 (cited on p. 87).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). “Attention is all you need”. In: *Advances in neural information processing systems*, pp. 5998–6008 (cited on pp. 4, 108, 113, 114).

- Velasco, E., Agheneza, T., Denecke, K., Kirchner, G., and Eckmanns, T. (2014). “Social media and internet-based data in global systems for public health surveillance: a systematic review”. *The Milbank Quarterly*, 92(1), pp. 7–33 (cited on p. 36).
- Venugopal, D., Chen, C., Gogate, V., and Ng, V. (2014). “Relieving the computational bottleneck: Joint inference for event extraction with high-dimensional features”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 831–843 (cited on p. 25).
- Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). “Semeval-2007 task 15: Tempeval temporal relation identification”. In: *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pp. 75–80 (cited on p. 16).
- Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). “SemEval-2010 Task 13: TempEval-2”. In: *Proceedings of the 5th international workshop on semantic evaluation*, pp. 57–62 (cited on p. 16).
- Veysel, A. P. B., Nguyen, T. H., and Dou, D. (2019). “Graph based neural networks for event factuality prediction using syntactic and semantic structures”. *arXiv preprint arXiv:1907.03227* (cited on p. 28).
- Von Etter, P., Huttunen, S., Vihavainen, A., Vuorinen, M., and Yangarber, R. (2010). “Assessment of utility in web mining for the domain of public health”. In: *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pp. 29–37 (cited on p. 54).
- Wadden, D., Wennberg, U., Luan, Y., and Hajishirzi, H. (2019). “Entity, relation, and event extraction with contextualized span representations”. *arXiv preprint arXiv:1909.03546* (cited on pp. 28, 30).
- Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). “ACE 2005 multilingual training corpus”. *Linguistic Data Consortium, Philadelphia*, 57, p. 45 (cited on p. 13).
- Wang, H., Yu, D., Sun, K., Chen, J., and Yu, D. (2019a). “Improving pre-trained multilingual models with vocabulary expansion”. *arXiv preprint arXiv:1909.12440* (cited on p. 6).

- Wang, L. (2005). *Support vector machines: theory and applications*. Vol. 177. Springer Science & Business Media (cited on p. 54).
- Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M. (2017). “Gated self-matching networks for reading comprehension and question answering”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 189–198 (cited on p. 107).
- Wang, X., Han, X., Lin, Y., Liu, Z., and Sun, M. (2018). “Adversarial multi-lingual neural relation extraction”. In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1156–1166 (cited on pp. 61, 75).
- Wang, X., Han, X., Liu, Z., Sun, M., and Li, P. (2019b). “Adversarial training for weakly supervised event detection”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 998–1008 (cited on pp. 28, 30, 32, 107).
- Wang, X., Wang, Z., Han, X., Jiang, W., Han, R., Liu, Z., Li, J., Li, P., Lin, Y., and Zhou, J. (2020). “MAVEN: A massive general domain event detection dataset”. *arXiv preprint arXiv:2004.13590* (cited on p. 30).
- Wang, X., Wang, Z., Han, X., Liu, Z., Li, J., Li, P., Sun, M., Zhou, J., and Ren, X. (2019c). “HMEAE: Hierarchical modular event argument extraction”. 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)*, pp. 5777–5783 (cited on pp. 28, 30).
- Wang, Z., Mayhew, S., Roth, D., et al. (2019d). “Cross-lingual ability of multilingual bert: An empirical study”. *arXiv preprint arXiv:1912.07840* (cited on pp. 61, 75).
- Wei, S., Korostil, I., Nothman, J., and Hachey, B. (2017). “English event detection with translated language features”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 293–298 (cited on p. 31).
- Woodall, J. P. (2001). “Global surveillance of emerging diseases: the ProMED-mail perspective”. *Cadernos de saude publica*, 17, S147–S154 (cited on pp. 36, 37).

- Woolhouse, M. E. and Gowtage-Sequeria, S. (2005). “Host range and emerging and reemerging pathogens”. *Emerging infectious diseases*, 11(12), p. 1842 (cited on p. 1).
- World Health Organization, W. (2022). “World health statistics 2022: monitoring health for the SDGs, sustainable development goals” (cited on p. 1).
- Wright, R. E. (1995). “Logistic regression.” (cited on p. 54).
- Wu, S. and Dredze, M. (2019). “Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT”. *arXiv preprint arXiv:1904.09077* (cited on pp. 61, 75).
- (2020a). “Are all languages created equal in multilingual BERT?” *arXiv preprint arXiv:2005.09093* (cited on p. 6).
- (2020b). “Are All Languages Created Equal in Multilingual BERT?” In: *Proceedings of the 5th Workshop on Representation Learning for NLP*. Online: Association for Computational Linguistics (cited on p. 79).
- Yang, B., Wang, L., Wong, D. F., Chao, L. S., and Tu, Z. (2019a). “Assessing the ability of self-attention networks to learn word order”. *arXiv preprint arXiv:1906.00592* (cited on p. 113).
- Yang, B. and Mitchell, T. (2016). “Joint extraction of events and entities within a document context”. *arXiv preprint arXiv:1609.03632* (cited on pp. 19, 31).
- Yang, H., Chen, Y., Liu, K., Xiao, Y., and Zhao, J. (2018). “Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data”. In: *Proceedings of ACL 2018, System Demonstrations*, pp. 50–55 (cited on p. 19).
- Yang, S., Feng, D., Qiao, L., Kan, Z., and Li, D. (2019b). “Exploring pre-trained language models for event extraction and generation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5284–5294 (cited on pp. 28, 30, 108).
- Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., and Lin, J. (2019c). “End-to-end open-domain question answering with bertserini”. *arXiv preprint arXiv:1902.01718* (cited on p. 107).



- Yang, Y., Agarwal, O., Tar, C., Wallace, B. C., and Nenkova, A. (2019d). “Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction”. *arXiv preprint arXiv:1905.07791* (cited on pp. 4, 58).
- Yang, Z., Zhan, S., Hou, M., Zeng, X., and Zhu, H. (2020). “Injecting Event Knowledge into Pre-Trained Language Models for Event Extraction”, n. pag. Web (cited on p. 75).
- Yangarber, R., Grishman, R., and Tapanainen, P. (2000). “Unsupervised discovery of scenario-level patterns for information extraction”. In: *Sixth Applied Natural Language Processing Conference*, pp. 282–289 (cited on p. 32).
- Yangarber, R., Jokipii, L., Rauramo, A., and Huttunen, S. (2005). “Extracting information about outbreaks of infectious epidemics”. In: *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pp. 22–23 (cited on p. 50).
- Yarowsky, D. (1995). “Unsupervised word sense disambiguation rivaling supervised methods”. In: *33rd annual meeting of the association for computational linguistics*, pp. 189–196 (cited on p. 88).
- Yu, S. and Wu, B. (2018). “Exploiting structured news information to improve event detection via dual-level clustering”. In: *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. IEEE, pp. 873–880 (cited on p. 18).
- Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). “Relation classification via convolutional deep neural network”. In: *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pp. 2335–2344 (cited on p. 107).
- Zhang, T., Ji, H., and Sil, A. (2019a). “Joint entity and event extraction with generative adversarial imitation learning”. *Data Intelligence*, 1(2), pp. 99–120 (cited on p. 30).
- Zhang, X., Zhao, J., and LeCun, Y. (2015). “Character-level convolutional networks for text classification”. *Advances in neural information processing systems*, 28 (cited on p. 108).
- Zhang, Y. and Wallace, B. (2015). “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification”. *arXiv preprint arXiv:1510.03820* (cited on pp. 108, 110).

- Zhang, Y., Xu, G., Wang, Y., Liang, X., Wang, L., and Huang, T. (2019b). “Empower event detection with bi-directional neural language model”. *Knowledge-Based Systems*, 167, pp. 87–97 (cited on p. 29).
- Zhao, W., Zhang, J., Yang, J., He, T., Ma, H., and Li, Z. (2021). “A novel joint biomedical event extraction framework via two-level modeling of documents”. *Information Sciences*, 550, pp. 27–40 (cited on pp. 26, 30).
- Zheng, S., Hao, Y., Lu, D., Bao, H., Xu, J., Hao, H., and Xu, B. (2017). “Joint entity and relation extraction based on a hybrid neural network”. *Neurocomputing*, 257, pp. 59–66 (cited on p. 107).
- Zhu, X. and Goldberg, A. B. (2009). “Introduction to semi-supervised learning”. *Synthesis lectures on artificial intelligence and machine learning*, 3(1), pp. 1–130 (cited on pp. 87, 88).
- Zhu, X. J. (2005). “Semi-supervised learning literature survey” (cited on p. 87).