Multilingual Epidemic Event Extraction : From simple Classification methods to Open Information Extraction (OIE) and Ontology

Sihem Sahnoun
Sorbonne University France
sahnounsihem@yahoo.com

Gaël Lejeune
Sorbonne University France
gael.lejeune@sorbonne-universite.fr

Abstract

There is an incredible amount of information available in the form of textual documents due to the growth of information sources. In order to get the information into an actionable way, it is common to use information extraction and more specifically the event extraction, it became crucial in various domains even in public health. In this paper, We address the problem of the epidemic event extraction in potentially any language, so that we tested different corpora on an existed multilingual system for tele-epidemiology: the Data Analysis for Information Extraction in any Language (DANIEL) system. We focused on the influence of the number of documents on the performance of the system, on average results show that it is able to achieve a precision and recall around 82% but when we resorted to the evaluation by event by checking whether it has been really detected or not, results are not satisfactory according to this paper's evaluation. Our idea is to propose a system that uses an ontology which includes information in different languages and covers specific epidemiological concepts, it is also based on the multilingual open information extraction for the relation extraction step to reduce the expert intervention and to restrict the content for each text. We describe a methodology of five main stages: Preprocessing, relation extraction, named entity recognition (NER), event recognition and the matching between the information extracted and the ontology.

1 Introduction

Infectious diseases are responsible for the morbidity and mortality and like we see today with the Covid-19 pandemic, surveillance provides us with information to improve our knowledge of their epidemiology (space-time dynamics, evolution of clinical and microbiological characteristics), in order to identify an appropriate control and prevention measures. The growth of digital data sources provide an avenue for data-driven surveillance, referred to as Epidemic Intelligence. The purpose of epidemic intelligence is to detect, analyze and monitor potential health threats over time (Nash and Geng, 2020). It requires the development of tools dedicated to the collection and processing of unstructured textual data published on the Web. Information Extraction (IE) (Martinez-Rodriguez et al., 2020) is one of the areas of active research in artificial intelligence and made it possible to analyze data from web sources. Different tasks of the IE systems have been suggested such as named entity recognition (NER) to identify real-world objects such as names of people, locations, names of diseases, etc. The relation extraction , which aims to find a semantic relation between two entities in a text (Elloumi et al., 2012). The event extraction is also an important task in the field of IE to detect, from the text, the occurrence of events with specific types, and to extract arguments (i.e. typed participants or attributes) that are associated with an event (Hettiarachchi et al., 2021). In 2007, the open information extraction has appeared and has introduced a new extraction paradigm unlike the traditional IE methods, relations are automatically detected instead of specifying target relations in advance and it enables a fast extraction over huge datasets (Niklaus et al., 2018). In this paper, we applied some experiments on the DANIEL system in order to evaluate its performance especially in the multilingual event extraction in the epidemiological field and we have exploited corpora from several diverse language families namely, English, Greek, French, Spanish, Portuguese, Russian, Polish, and Chinese. On the latter we proposed a new approach which reduces the expert intervention by using a multilingual OIE systems for a relation extraction, an automatic NER, and an ontology applied for any epidemiological event. The remainder of this paper

Proceedings of Recent Advances in Natural Language Processing, pages 1227–1233
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_138

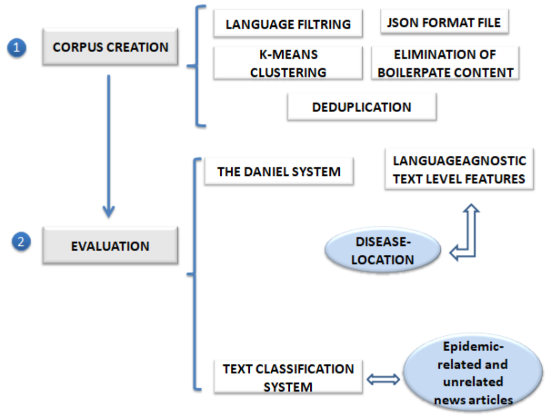
1
CORPUS CREATION
LANGUAGE FILTRING
JSON FORMAT FILE
K-MEANS CLUSTERING
ELIMINATION OF BOILERPATE CONTENT
DEDUPLICATION
2
EVALUATION
THE DANIEL SYSTEM
LANGUAGEAGNOSTIC TEXT LEVEL FEATURES
DISEASE-LOCATION
TEXT CLASSIFICATION SYSTEM
Epidemic-related and unrelated news articles

Figure 1: The architecture of the DANIEL System

is organized as follows. Section 2 reviews works related to epidemic surveillance and event extraction systems, Section 3 describes the dataset used in the experimental study, in Section 4 we present the results. Finally, we provide a discussion about our new proposition and a conclusion in Section 5.

2 Related Work

Many systems have been proposed in the domain of IE and which reflects the basis that were taken to build our approach. We made a study on three different approaches that take us to our perspectives for our new system.

2.1 The DANIEL System

The system proposes techniques for identifying emerging infectious disease threats in online news text, it's based on two main steps: Corpus creation and evaluation. Figure 1 illustrates the architecture of the DANIEL system (Mutuvi et al., 2020a).

Corpus creation The system uses the Program for Monitoring Emerging Diseases (PROMED) platform to create the corpus. Firstly, the source URLs where the article was originally published, together with the other meta-data such as title, description, location, date were formatted and stored in a JSON format making the corpus easily reusable and reproducible. Then a language filtering was performed to ensure that only documents belonging to the languages of interest were retained using the K-means clustering algorithm. the boilerplate content such as navigation links, headers and footers was eliminated from HTML pages, it is among the data cleaning tasks. The final pre-processing task was de-duplication which involves eliminating perfect duplicate and near-duplicate content so that only one instance of each text was preserved.

Evaluation DANIEL is a multilingual news surveillance system, it aims to extract disease-location for each text in its corresponding language. it describes an event as a disease outbreak and the place where it occurred. It avoids grammar analysis and the usage of language-specific NLP toolkits (e.g., Part-of speech tagger, dependency parser), it considers the text as a sequence of strings instead of words. The named entities presented by a list of diseases and a list of locations in JSON files in different languages. The named entity extraction depends on a ratio r which has a default value that can be fine-tuned by the user. The ratio r is depicted in the following equation:

$$r = length(substring)/length(entity_{n}ame) \quad (1)$$

This ratio is a kind of threshold for the different size of the substrings. For example (r= 0.8) means that substrings sharing 80% of the named entity.

The news articles are grouped into various categories such as politics, wellness, travel, entertainment, sports and healthy living, among others. The models classify a news article as either relevant or non-relevant, depending on whether it alerts about a disease outbreak or not as described in (Mutuvi et al., 2020a).

2.2 The BioCaster System

BioCaster (Collier et al., 2008) ingests documents through RSS feeds. An Automatic classification of the reports was performed for topical relevance using a naïve Bayes algorithm. A named entity recognition is then accomplished for relevant documents for 18 term types based on the BioCaster ontology. The BioCaster ontology includes information in eight languages focused on the epidemiological role of pathogens as well as geographical locations with their latitudes/longitudes. At this stage disease-location pairs are plotted onto a public portal called the Global Health Monitor, to gain a geographically contextualized view of an outbreak anywhere in the world in Google Maps which can be filtered by pathogen, syndrome or text type. The event extraction is based on matching entity classes, skipwords, string literals, regular expressions, entity types as well as guard lists which include verbs of infection, common victim expressions. This is done by using a Simple Rule Language (SRL) as described in (Collier et al., 2008).

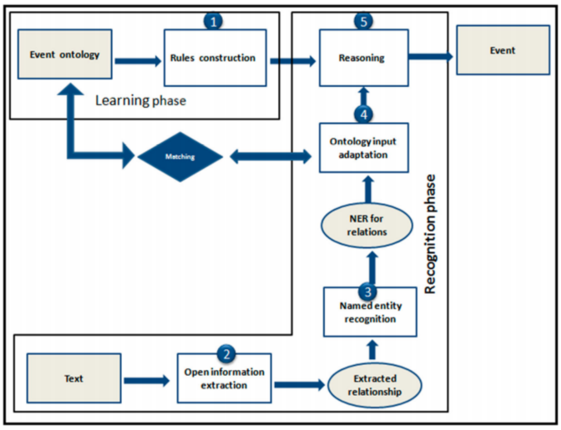

Figure 2: The Event Detection based on Open Information Extraction and Ontology System architecture

2.3 Event Detection based on Open Information Extraction and Ontology System

The approach proposes an event extraction by using an OIE system for a relation extraction without supervision, an automatic NER, and an ontology applied for any management change event (Sahnoun et al., 2020). The approach admits 2 phases as shown in Figure. 2 that depend on each other: A learning phase and a recognition phase. The learning phase consists of modeling an event by an ontology (classes, subclasses and instances in relations), and constructing a set of rules manually. The recognition phase includes the RE, the NER and an automatic reasoning between the rules and the input ontology adaptation.

The rules construction is an important step which drives to a possible event extraction. For the relation extraction the system uses OLLIE, it is an Open information extraction tool which extracts the relationship triplets that contains three textual components (Arg1, Rel, Arg2) where the first and the third stand for the pair of arguments and the second indicates the relationship. For the Named Entity Recognition step, the system can detect a person, an organization, location, etc., in any part of the triplet using the python library SPACY[1].

For the ontology input adaptation, verbs will be passed through a lemmatization layer to convert verbs to their infinitive form. An instance is every token recognized by a NE, then it will be added to the ontology and linked by relations whenever the following conditions are achieved :

- The number of named entities is greater than or equal to 2 to have a possible relationship among them.
- The lemmatized verb and the other relations between delimiters ";" should be included in the relation list of the ontology and Named Entities can be linked with these relations.

the last step is reasoning by inferring logical consequences of a set of rules to affect for each instance its role (event) as described in (Sahnoun et al., 2020).

3 Dataset and Evaluation

We tested the Daniel system on three large datasets in different languages (low- or high-resource), the first (Daniel-corpus) contains 2089 of relevant and irrelevant files (Romain et al., 2013), then we extend the dataset (BIG-corpus) (Mutuvi et al., 2020b) to include additional languages so that it covers news articles from several, diverse language families: Germanic (English, en), Hellenic (Greek, el), Romance (French, fr), Slavic (Polish, pl and Russian, ru) and Sino-tibetan (Chinese, zh). It includes 7046 irrelevant files and 1653 relevant files, and the third present the fusion of the two precedent corpora. The statistics of the dataset are presented in Table 1 and Table 2. The experimental study was carried out according to the two measurements of the true positive rate (TPR), the false positive rate (FPR), depending on the threshold r (1). The true positive rate (TPR) (2): Called also sensibility measures the likelihood of actual positive results. The false positive rate (FPR) (3):It's the probability that a positive result will be given when the true value is negative.

$$TPR = TP/TP + FN = Recall \tag{2}$$

$$FPR = FP/FP + TN \tag{3}$$

We used a ROC curve to visualize the performance of the binary classifier. It's a plot of the TPR versus the FPR for every possible classification threshold. A classifier that does a very good job separating the classes will have a ROC that hugs the upper left corner of the plot [2]. Conversely, a classifier that does a poor job will have a ROC

[1] https://www.ekino.com/articles/handson-de-quelques-taches-courantes-en-nlp

[2] https://www.dataschool.io/roc-curves-and-auc-explained/

Language	# Texts	# Paragraphs	# Char.(10^6)
Chinese (zh)	446	4428	1.14
English (en)	475	6791	1.35
Greek (el)	390	3543	2.05
Polish (pl)	352	3512	1.04
Russian (ru)	426	2891	1.56

Table 1: Statistics for the DANIEL-corpus : number of texts, paragraphs and characters

Language	# Texts	# 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

Table 2: Statistics for the BIG-corpus (number of texts, sentences and tokens)

curve that is close to the diagonal line. The purpose of AUC, which stands for Area Under the Curve. A very poor classifier has an AUC of around 0.5 and a perfect classifier has an AUC close to 1. After Evaluating the Daniel system and by taking into account the influence of the number of documents on the performance of the system, we resorted to the evaluation by event, each disease-location pair (e.g. flu in Spain) is considered as a unique event, regardless of the number of documents in which it has been reported, and then we check whether it has been detected or not.

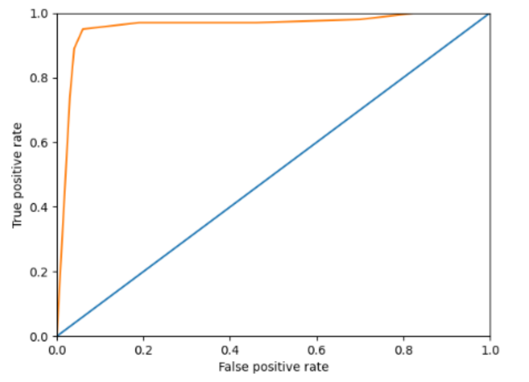

Figure 3: ROC curve for the DANIEL corpus (2089 files) (Romain et al., 2013)

4 Results

4.1 Evaluation By Documents

Figure 3 depicts the tracing of the ROC curve for the first corpus (Daniel-corpus) for 2089 files, the system represents its performance with a threshold of 0.8: The TPR has for percentage of 89% while we find only 4% of FPR that explains that the system has detected a significant number of events in the relevant documents. The Daniel-corpus contains texts in five different languages: Greek, English, Chinese, Russian and Polish as it shows in Figure 5, For the five languages, the ROC curve hugs the upper left corner of the plot for each curve (i.e. the system detects well the events for each language). Figure 4 demonstrate that the ROC curve is close to the diagonal for the BIG-corpus, we have 39% of TPR and 10% of FPR for a threshold of 0.8 results are not satisfactory for this corpus, so when we have increased the number of documents the TPR decreased because there are new names of diseases and locations that are not detected and there are new languages have been added. We merged the two corpora in order to obtain a more representative corpus, the results of the subsequent corpus are shown in figures 6 and 7. The results have been slightly improved for Spanish, Portuguese, Greek, Russian, Polish and Chinese the system gives a high-performance but this not the case for English and French. For a threshold of 0.8 we observe a TPR of 42% and a 9% FPR.

4.2 Evaluation By Event

Table 3 shows the results of the evaluation by event for the DANIEL-corpus, demonstrating that there

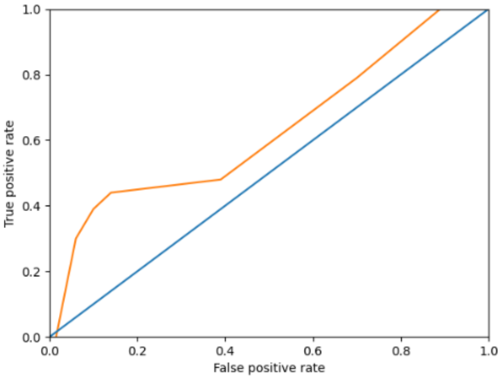

Figure 4: ROC curve for the BIG-corpus (8699 files) (Mutuvi et al., 2020b)

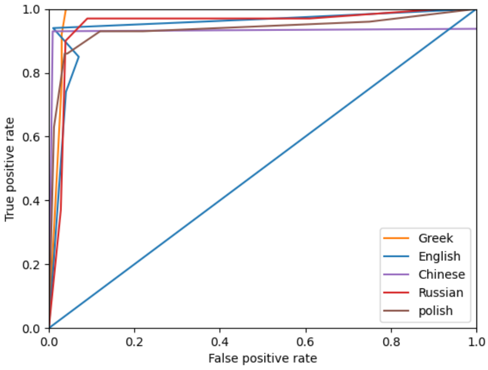

Figure 5: The ROC curves for each language of the DANIEL-corpus

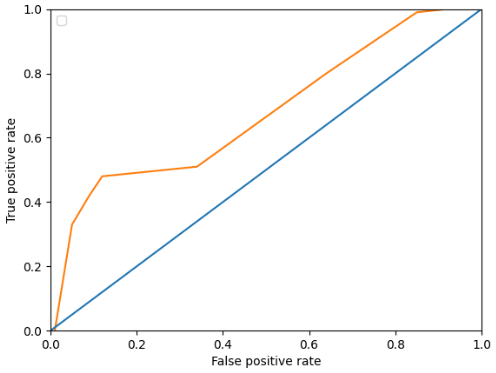

Figure 6: The ROC curve for the MERGED corpora

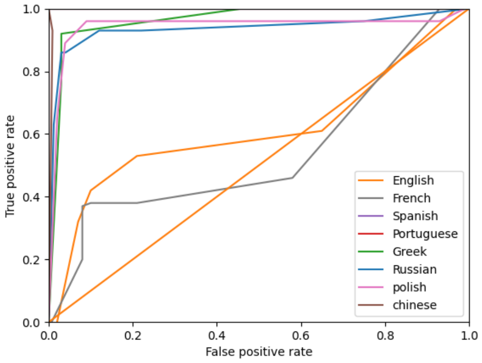

Figure 7: The ROC curve for each language for the MERGED corpora

are 66 events out of 86 are missed if no entity linking is performed. The total number of unique events in the corpus is not the sum of unique events in each subcorpus. In the cumulated corpora there are 16 events that are reported in more than one language.

	# Events	Detected	Missed
Chinese	9	1	8 (89%)
English	19	6	13 (68%)
Greek	16	5	11 (69%)
Polish	21	4	17 (81%)
Russian	21	4	17 (81%)
All	70	18	52 (74%)

Table 3: Evaluation by unique event for the DANIEL-corpus with a threshold of 0.8 (NB: this a strict scenario where no entity linking has been performed)

The results of the relevant documents for The BIG-corpus in Table 4 shows that there is 75 of events detected between 939 unique events (in all languages) so the number of events detected is not great if we consider that we have a much larger corpus. Table 5 presents an example of an event extraction in the form of a peer of disease-location, the real date that documents have mentioned the event and the date when it was detected.

5 Conclusion and Perspectives

In this paper we have focused on the study of some systems in the epidemiological field such as Daniel and BioCaster, and an event extraction system which is based on a methodology that opens

	# Events	Detected	Missed
English	328	20	308 (93%)
French	169	20	149 (88%)
Spanish	278	33	245 (88%)
Portuguese	183	2	181 (99%)
All	939	75	864 (92%)

Table 4: Evaluation by unique event for the BIG corpus for a threshold of 0.8 (NB: this is a strict scenario where no entity linking has been performed)

Event	Real date	Detection date	Diff.
Flu-India	12-01-2012	17-01-2010	5 days
Grypa-Chiny	06-01-2012	12-01-2012	6 days
H5N1-Κίνα	30-12-2011	30-12-2011	0 days

Table 5: Example of an Event Extraction

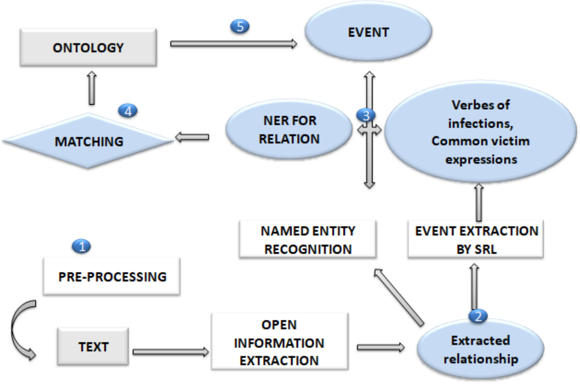

Figure 8: The Epidemiological event extraction based on open information extraction and ontology system architecture

the door for a new proposition using the open information extraction and the ontology. We propose a procedure to an eventual epidemic event extraction which consists of five main stages: Pre-processing, relation extraction, named entity extraction and the matching between the information extracted and the ontology. The first task is to Retrieve articles from PROMED in different languages, then a pre-processing step based on a Data cleaning task to Eliminate the boilerplate content from the corpus.

The text then will be passed to a relation extraction using an open information extraction to restrict the content of a text into triplets of relationships so that it is in form (Arg1, verb, Arg2) and we can differentiate between the nominal part and the verbal part. In the nominal part where we can find the named entities and in the verbal part we can visualize verbal expressions in the epidemiological context. The relation extraction method aim to represent semantic relations between entities. The entities have numerous applications in building knowledge representation models that report relations between words, such as ontologies, semantic networks.. In this work, we will investigate the area of multilingual open information extraction for the Portuguese, English and other languages (Claro et al., 2019). The extracted relations will be passed to a named entity recognition layer. The system can detect the location, the date, the percentage and the name of disease... For the event extraction we can use the simple rule language (SRL) with a capability to match regular expressions and guard lists include verbs of infection, common victim expressions, occupation names. The ontology in our approach present the event which is an object that admits an existence in the space of time and depends on other objects in relation. An ontology is a set of concepts, as well as relationships between these concepts that's why the event was modeled by an ontology. The matching between the extracted information and the ontology will bring us to an eventual event extraction as depicted in Figure. 8, we can use an ontology already existed like the BioCaster ontology but it is not available online so we're going to do it ourselves.

We are looking for evaluating the results of the test obtained and compare them by another systems like BioCaster and DANIEL.

References

Daniela Barreiro Claro, Marlo Souza, Clarissa Castellã Xavier, and Leandro Oliveira. 2019. Multilingual open information extraction: Challenges and opportunities. *Information 2019*, pages 1–25.

Nigel Collier, Son Doan, Ai Kawazoe1, Reiko Matsuda Goodwin1, Yoshio Tateno Mike Conway, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, pages 2940–2941.

Samir Elloumi, Ali Jaoua, Fethi Ferjani, Nasredine Semmar, Romaric Besançon, Jihad Al-Jaam, and

Helmi Hmmami. 2012. General learning approach for event extraction: Case of management change event. pages 211–224.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2021. Embed2detect: Temporally clustered embedded words for event detection in social media. pages 1–39.

Jose L. Martinez-Rodriguez, Hogan Aidanb, and van Lopez-Arevalo. 2020. Information extraction meets the semantic web: A survey. *Semantic Web*, pages 255–335.

Stephen Mutuvi, Antoine Doucet, Gael Lejeune, and Moses Odeo. 2020a. A dataset for multilingual epidemiological event extraction. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4139—4144.

Stephen Mutuvi, Antoine Doucet Emanuela Boros, Gael Lejeune, Adam Jatowt, and Moses Odeo. 2020b. Multilingual epidemiological text classification: A comparative study. *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6172—6183.

Denis Nash and Elvin Geng. 2020. Goal-aligned, epidemic intelligence for the public health response to the covid-19 pandemic. *American Journal of Public Health*, pages 1154–1156.

Christina Niklaus, Matthias Cetto, Andre Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. *27th International Conference on Computational Linguistics*.

Brixtel Romain, Lejeune Gael, Doucet Antoine, and Lucas Nadine. 2013. Any language early detection of epidemic diseases from web news streams. *International Conference on Healthcare Informatics (ICHI)*, pages 159–168.

Sihem Sahnoun, Samir Elloumi, and Sadok Ben Yahia. 2020. Event detection based on open information extraction and ontology. *Journal of Information and Telecommunication*, pages 383–403.