A Corpus for Outbreak Detection of Diseases Prevalent in Latin America

Antonella Dellanzo
Dept. of Computer Science
FCEyN
Universidad de Buenos Aires,
Argentina
antodellanzo@gmail.com

Viviana Cotik
Dept. of Computer Science
FCEyN
Universidad de Buenos Aires,
Argentina
vcotik@dc.uba.ar

José Ochoa-Luna
Dept. of Computer Science
Universidad Católica San Pablo,
Arequipa, Peru
jeochoa@ucsp.edu.pe

Abstract

In this paper we present an annotated corpus which can be used for training and testing algorithms to automatically extract information about diseases outbreaks from news and health reports. We also propose initial approaches to extract information from it. The corpus has been constructed with two main tasks in mind. The first one, to extract entities about outbreaks such as disease, host, location among others. The second one, to retrieve relations among entities, for instance, in such geographic location fifteen cases of a given disease were reported. Overall, our goal is to offer resources and tools to perform an automated analysis so as to support early detection of disease outbreaks and therefore diminish their spreading.

1 Introduction

In recent years, several contagious diseases have aroused. Many of them, such as dengue, Guillain-Barré, Zika and microcephaly, occur in tropical regions of the world as in Latin America. There are also other diseases such as Chagas, also referred to as American trypanosomiasis, which have existed for many decades, but not many resources are used to treat them, probably because they do not appear in densely populated areas and their usual hosts belong to populations of scarce socioeconomic resources. On the other hand, there are endemic diseases, such as hantavirus, which appear recurrently in some regions of Argentina and Chile, among other countries. Finally, we are facing nowadays the COVID-19 pandemics.

Those illnesses have many issues in common: 1) they are dangerous in terms of number of losses of human life or lifelong consequences to those who carried them, 2) their economic impact can be very important (as in loss of tourism incomes with hantavirus, and loss in almost all economic areas with the isolation imposed by different governments due to the COVID-19 pandemics), and 3) when an outbreak occurs there is few available data that can support for data-driven decision making policies.

In this work we attack the third issue, we argue that information including number of cases, host and location of the host, among others, is crucial in order to understand the propagation of the illness and diminish its spreading. When available, the information usually is of poor quality (incomplete, incorrect, inconsistent, not publicly available, and it lacks the timeliness attribute, among others).

Thus, our goal is to contribute with automated tools and resources that allow to alert and inform about possible outbreak diseases and some illnesses prevalent in Latin America. To do so, we have constructed an annotated corpus based on ProMED-mail (Carrion and Madoff, 2017), a reporting system dedicated to the dissemination of information on epidemics of infectious diseases. We have also implemented an initial baseline, whose results will allow us to answer the following questions a) are there in the articles any mention to a disease?, b) in case there is: to which?, c) how many cases are reported in the article, d) in which geographic location is the disease located?, e) if there are causes or possible causes mentioned, which are they? and f) what is the date of the report?.

In the paper, we also emphasize the process followed to build the corpus which can serve as guide for future developments. In addition, we also present results on information extraction techniques applied over the corpus (namely, a named entity recognition (NER) rule-based technique). Although preliminary, the algorithm shows promising results that can be further extended to analyze news and social media.

The rest of the paper is organized as follows. Section 2 presents previous work in digital surveillance

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 543–551
Online, November 19-20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

in public health. In Section 3 we describe the creation of an annotated corpus. Section 4 describes the implemented methods for doing named entity recognition. Section 5 shows preliminary results obtained from the baseline method proposed. Finally, Section 6 describes the conclusions reached so far and the future directions we are planning.

2 Related Works

Digital surveillance has been an important topic in public health. A large number of the papers published in Public Health informatics is about the epidemiological surveillance based on the new data generated in the current digital era (Thiebaut and Cossin, 2019). Research studies show that social media may be valuable tools in the disease surveillance toolkit used to detect disease outbreaks due to could be faster than traditional methods and to enhance outbreak response (Charles-Smith et al., 2015). Thus, social media constitutes a source of information for the surveillance of various public health outcomes on a real-time basis (Thiebaut and Cossin, 2019). For instance, we see that Twitter data has been used to aid in public health efforts concerned with surveillance, event detection, pharmacovigilance, forecasting, disease tracking and geographic identification, demonstrating positive results (Edo-Osagie et al., 2020). Public health surveillance is therefore a natural application for artificial intelligence techniques, the use of web-based data requires Natural Language Processing approaches to extract the information.

In this context, when national surveillance data are lacking, informal disease surveillance systems provide an opportunity to understand epidemiological trends (Desai et al., 2019). In this paper we have focused on building a corpus from ProMED (Carrion and Madoff, 2017). ProMED (The Program for Monitoring Emerging Diseases) is an internet-based reporting system for emerging infectious. Regions and countries could benefit from complementing their undiagnosed disease surveillance systems with ProMED-mail tool (Rolland et al., 2020). It is worth noting the effectiveness of ProMED as an epidemiological data source by focusing on coronaviruses (Bonilla-Aldana et al., 2020).

While ProMED is a reliable source of information, it is not currently equipped to provide detailed epidemiological data. For example, ProMED often does not report case or death counts beyond what is included in the text of a post. In order to extract this information from ProMED posts in a systematic way further analysis is required (Carrion and Madoff, 2017). One important work in this direction has been the Platform for Automated extraction of Disease Information from the web (PADI-web) (Arsevska et al., 2018). This tool generates epidemiological information on diseases, locations, dates, hosts and number of cases for outbreaks mentioned in news and social media articles. To do so, it combines Information extraction based on rule-based systems and data mining techniques. There is also a machine learning algorithm trained to better classify or identify information on diseases. This proposal is the most related to ours, however, unlike PADI-web, we focus on building a corpus from proMED-mail articles on diseases prevalent in Latin America.

3 Creating the Corpus

In this section we present the process involved to construct the corpus based on proMED-mail news articles (NPA). First, we present the dataset. Then, we explain the data cleansing processed followed, the annotation schema and criteria developed, a summary of the annotation guidelines and how we have been performing the annotation process. After, we show the statistics of our dataset and we explain the results of our inter annotator agreement. Finally, we present the data statements of our data (Bender and Friedman, 2018).

3.1 Corpus

In order to construct the corpus we downloaded articles from ProMED-mail, a reporting system dedicated to the rapid dissemination of information on epidemics of infectious diseases, among others (Arsevska et al., 2018).[1] The articles published on ProMED-mail have been edited based on journalistic notes from different media by an interdisciplinary staff.

We retrieved 811 articles written in Spanish and focused on reported issues in Latin America that mention the appearance of certain pathologies (measles, hantavirus, Guillain-barre, zika, microcephaly and Chagas). The retrieved articles were written between 11/5/2018 and 10/5/2019. Articles are formed by a title, a date and the main text. They also contain metadata, that had to be removed, as we explain in the next section.

[1] ProMED-mailhttps://promedmail.org/about-promed/

3.2 Data cleansing

We removed ProMED-Mail articles metadata with the use of regular expressions (regex). We keep only the article title, date and article note. The three components constitute what we call the article or NPA. An example of a ProMed-Mail article with its metadata is shown bellow.

Article title
Date
****
Un comunicado de ProMED-mail *(a communication of ProMED-mail)*
http://www.promedmail.org
... *(7 lines of metadata in this case: date, source, etc.)*
[Editado por *$name*] *([edited by $name])*
Article note
Comunicado por: *$name <$mailaddress>*
(Communicated by: $name <$mail address>)
– ProMED-ESP
................................jt

Freeling (Padró and Stanilovsky, 2012) was chosen as the language analyzer in order to do some natural language processing (NLP) tasks such as tokenization, sentence splitting and as an aid to our named entity (NE) recognition algorithm. We had to perform some normalization over the text to meet some criteria required by Freeling. The main normalization tasks are listed next.

- in our notes dates are sometimes written using a slash (eg. *10/03/2018*) and sometimes using a dash (eg. *10-03-2018*). We used regex in order to transform dates written with dashes to dates written with slashes,
- some articles use comma and others use dot as a thousand symbol separator. We used regex in order transform all thousand symbol separators to dots,
- in the articles, some countries are written by their acronym instead of by their entire name (eg. *EE.UU.* -USA- for *Estados Unidos* -United States-). We replaced acronyms by their complete name so that Freeling would recognize them as a location when we use its NER functionality.

3.3 Annotation schema and criteria

Named entities as well as binary and ternary relations between named entities have been annotated. To do so, we are using the brat rapid annotation tool (Stenetorp et al., 2012).

In the remainder of this subsection, we describe entities, concepts and relations that have been annotated.

The following entities and characteristics are being annotated:

disease (DS): entity corresponding to a pathological finding or diagnosis, eg. zika and Chagas,

date (DT): mention to a date, eg. 28 de febrero, 2019 (02/28/19), and 30 de abril (April 30),

location (LOC): geographic location, eg. MÉXICO: (JAL), Capiatá and Montevideo,

number of cases (NoC): mentions to the number of cases of a disease,

origin (OR): entities corresponding to the cause of a disease, eg. in *The transmission of the disease occurred through the consumption of contaminated Açai*, *consumption of contaminated Açai* is annotated as origin.

transmission form (TF): refers to the form of transmission of a disease. It is only annotated when there is an origin in the same sentence. eg. in *the virus is transmitted by a mosquito bite*, *virus* is annotated as DS, *mosquito* as OR and *bite* as TF.

host (Hst): mention to the person or animal who contracted the disease, eg. dog, infant, pregnant woman, immigrant.

Other annotated concepts (called modifiers) have also been considered:

negation (NT) and uncertainty terms (UT): terms that indicate negation or lack of certainty, eg. *negative*, *absence*, *no*, *nonexistence* (for negations), *suspected cases*, *may cause*, *under observation* (for uncertainties).

past terms (PT): terms that mention a fact that happened in the past, eg. in the past, in 2014, back in 2019,

conditional (COND): terms corresponding to a fact that may happen in the future, eg. in *if dengue persists, then actions should be taken in order to (...)*, *if* is a conditional modifier.

The binary relations that have been annotated are:

disease occurs in (DsOccIn): relation among a disease and the geographic location where it occurs, eg. in *eleven cases of chagas in the state of Mérida*, *chagas* would be annotated as the disease and *Mérida* as a location. Also, a *DsOcIn* relation would be annotated among chagas and Mérida,

origin occurs in (OrOccIn): relation among cause of infection (origin) and geographic location, eg. in *an oral outbreak of the disease in the neighborhood of Belén is under investigation*, *oral outbreak* would be annotated as origin and *neighborhood of Belén* as a location. Also, a *OrOccIn* relation would be annotated among both entities,

cases of disease (NoCDs): relation among number of cases of a disease eg. a total of 391 cases of Zika were detected. Here, *391* would be annotated as number of cases and *Zika* as disease. Also, a *NoCDis* relation would be annotated among 391 and Zika,

cases in location (NoCLoc): relation among number of cases in a location, eg. in *there are already 100 patients with Guillain-Barré syndrome in Peru*, *100* would be annotated as NoC and *Peru* as a location. A *NoCHt* relation would be annotated among both entities.

cases of host (NoCHt): relation among number of cases affecting a particular host, eg. in *(...) a group of 32 newborns with microcephaly in Brazil*, *32* would be annotated as NoC and *newborns* as host. A *NoCHt* relation would be annotated among both entities.

date of disease (DtDs): date in which a disease occurred eg. *so far in 2013, 4 persons died (...) from chagas*. Here, *2013* would be annotated as date and *chagas* as disease. Also, a *DtDs* relation would be annotated among 2013 and chagas,

cause of disease (OrDs): cause of a disease (eg. zika (cause -and also disease-) in pregnant women is the cause of Guillain-Barré syndrome in their babies. In this case zika would be annotated as cause (origin entity) and as disease and Guillain-Barré as disease. Also a *OrDs* relation would be annotated among zika and Buillain-Barré,

negates disease, location, number of cases or cause (NegDs, NegLoc, NegNoC, NegOr, NegTf): relations among a negation term and: a disease, a location, a number of cases, a cause or a transmission form, eg. *10 of the cases didn't present microcephaly*. Here, *microcephaly* would be annotated as disease and *didn't* as a negation. Also, a *NegDs* relation would be annotated among microcephaly and didn't.

speculates disease, origin or transmission form (UcDs, UcOr, UcTf, UcNoC): relations among a speculation term and: a disease, an origin, a number of cases or a transmission form, eg. in *the transmission form is suspected to be through ingestion or orally (...)*, *suspected* would be annotated as uncertainty, and *ingestion* and *orally* as origin. Also, an *UcOr* relation would be annotated among suspected and ingestion, and another among suspected and orally,

occurs to (OccTo): relation among a disease and a host (who suffered from the disease), eg. in *chagas prevalence in donors (blood donors) has decreased*, *chagas* would be annotated as disease and *donors* as host. Also, an *OccTo* relation would be annotated among chagas and donors,

transmission cause (TfOr): relation among a transmission form and a cause, eg. *there was an oral transmission through food contaminated by chipo feces*. Here, *oral* would be annotated as transmission form and*food contaminated by chipo feces* as origin. Also, a *TfOr* relation would be annotated among them,

temporal conditional or past (TempCond, TempNt): relations among a conditional or past term and a disease, or among a past term and number of cases, eg. in *last week the appearance of cases of chagas disease has been reported (...)*, *last week* would be annotated as past and *chagas* as disease. Also, a *TempNt* relation would be annotated between *last week* and *chagas*. Finally, the following ternary relations have been annotated:

speculates relation (UcOrDs): relation among an uncertainty term and the cause of a disease, eg. in *a significant number of virus strains (OR) circulate in the country and all of them can cause hantavirus pulmonary syndrome*, *virus strains* would be annotated as OR, *can cause* as uncertainty and *hantavirus pulmonary syndrome* as a disease. Also *can cause* should be related to *virus strains* and also to *hantavirus pulmonary syndrome* with a UcNoCDs relation,

speculates number of cases of disease (UcNoCDs): allows relations among an UT, NoC and a disease, eg. in *283 cases of people are suspected of having zika (...)*, *283* would be annotated as NoC, *zika* as Ds and *suspected* as Uncertainty. A ternary *SUcNoCDs* relation would be annotated among 283, zika and suspected.

3.4 Annotation guidelines

Next, we present a summary of the annotation guidelines:

- **diseases** will be annotated as such, only if they appear in the same sentence as NoC or other complementary information.
- **hosts** are annotated only if they have a special characteristic (eg. pregnant, infant and dog). If a person without further description of its characteristics contracted a disease, *person* is not annotated as host. Hosts are the carriers of the disease (not those of a secondary virus that causes the disease).
- **negations and speculations** that are not related to a NE or to a RE have not been annotated.
- the **largest possible term** has to be annotated in the case there is one entity embedded in a bigger one, eg. ”mosquito aedes aegipty” (*aedes aegipty mosquito*) should be annotated rather than *mosquito*.
- **relations between sentences** have not been annotated.
- **Other:**
  - terms corresponding to a NE, that are misspelled must be annotated.
  - only NPAs, whose title refers to a Latin American country will be annotated. Nevertheless, if the note mentions other countries outside Latin America, these countries or cities will be annotated as locations.

3.5 Annotation process

The annotation was carried out by Spanish native speakers from Peru and Argentina. Some of them are computer science master students, others are linguists, and others are computer scientists that do research in NLP and with a background on annotation in different areas (from now on, *the experts*).

A document with the annotation schema and criteria was written and many meetings were held with the annotators in order to solve doubts. After having annotated a first dataset doubts and differences in criteria were reviewed and the annotation guidelines (described in Subsection 3.4) were written by the experts with more detail. After two annotation-revision iterations, the final guidelines were defined and annotations were performed (in what we call iteration 3). Now we are on working in iteration number 3.

Disagreements were solved by the experts.

3.6 Dataset Analysis

Since the title of the notes already contain a lot of information, we will work with 1) only the title and the date of the article (reduced article), and 2) the entire NPA (title, date and article note -body of the article-).

Once the annotation was performed by all the annotators, the final dataset was evaluated to know how many entities, relations and events of each type were found.

Overall 170 different newspaper articles have been annotated. Average number of sentences: 11, average number of words per article 367. Average length of titles 11 words.

The average number of sentences per article was 11, the average number of words per article was 367, and the average number of words in titles was 11.

Tables 1 and 2 show the number of annotated entities and modifiers (eg. negations) and the number of different entities and modifiers for the entire article and for the reduced article.

type	total	different
Date	245	213
Disease	1087	143
Host	283	134
Location	759	315
Number of cases	606	264
Origin	417	217
Transmission form	106	80
Negation	22	14
Past	154	118
Uncertainty	108	68

Table 1: Type and amount of entities, modifiers and other characteristics with more than five occurrences for NPA.

It may seem strange that there are 143 different diseases. That is for many reasons: some diseases that are not of interest for our study (eg. hypertension and diabetes) have also been annotated. Also, some diseases are written with many variations (eg. Guillain-Barré is written in ten different ways).

Tables 3 and 4 show for NPAs and reduced articles relations, the entities related by them, and the total number of relations and the number of different relations appearing in the annotated texts.

3.7 Inter-annotator agreement

To evaluate the consistency among the annotations performed between pairs of annotators, the inter-annotator agreement (IAA) was calculated using the Cohen’s Kappa coefficient (κ) (Cohen, 1960).

type	total	different
Date	156	149
Disease	183	26
Host	25	17
Location	160	59
Origin	51	18
Transmission form	13	10
Uncertainty	7	3

Table 2: Type and amount of entities, modifiers and other characteristics with more than five occurrences for article title and date.

relation	entities	total	different
DsOccIn	DS-LOC	551	330
DtDs	DT-DS	52	42
NegDs	NT-DS	9	7
NegOr	NT-CA	13	12
NoCDs	NoC-DS	340	291
NoCHt	NoC-HT	76	71
NoCLoc	NoC-LOC	144	139
OccTo	DS-HT	192	121
OrDs	CA-DS	288	182
OrOccIn	CA-LOC	38	33
PtDs	PT-DS	103	92
PtNoC	PT-NoC	50	49
TfOr	TF-CA	69	65
UcDs	UT-DS	14	14
UcNoC	UT-NoC	55	55
UcNoCDs	UC-NoC-DS	5	4
UcOr	UT-CA	28	25
UcOrDs	UC-CA-DS	7	6
UcOrDs	UC-DS-CA	7	7

Table 3: Relations with more than five occurrences annotated among entities in NPAs.

The Cohen Kappa Score was calculated with scikit-learn library,[2] which given two arrays (one corresponding to the annotation of each annotator), returns the score for that annotation.

In our implementation a token will be considered to have the same label (type of named entity or of relation) if and only if both annotators assigned the exact set of labels to it (or none).

For a set of 27 NPA annotated in common by both students and experts, we obtained a minimum κ value of 0.16 (obtained in the first iteration of annotations), a maximum of 0.73 (obtained in the second iteration) and an average value of 0.52 throughout the 27 annotated NPAs.

3.8 Data statements

Data statements were proposed by Bender and Friedman (2018) to address critical issues, such as biases, when working with natural language data.

relation	entities	total	different
DsOccIn	DS-LOC	162	89
OccTo	DS-HT	26	21
OrDs	CA-DS	42	17
OrOccIn	CA-LOC	17	15
TfOr	TF-CA	6	5

Table 4: Relations with more than five occurrences annotated among entities in reduced articles.

In following paragraphs we describe the data statements of our corpus, including some annotation decisions.

We selected texts from ProMed-mail written in Spanish that mention at least one of the following terms: chagas, measles, hantavirus, Guillain Barré, zika or microcephaly and that were written between 04/01/2001 and 10/5/2019.

From those, we selected only those that talk about Spanish speaking Latin American countries. Our goal was to obtain quality news about previously mentioned illnesses and their prevalence in Latin American countries. Therefore we selected ProMed-mail as source.

The language used is the usual in newspaper articles. Nevertheless, news are shorter than they usually are. Each country has its particularities in the use of Spanish (in what regards texts, different word choice). Nevertheless, in the articles, standard Spanish is used (it can not be identified to the variant used in a particular country).

Annotator guideline developers speak different variations of Spanish, so do the annotators. Nevertheless, we evaluate that this fact did not hinder an accurate understanding of the annotation criteria or the newspapers articles. We do not have information of ProMed-mail editors' demographics.

4 Approaches for Extracting Information

In this section we present a rule-based named entity recognizer applied to reduced NPAs and to whole NPAs. We also developed a connectionist state of the art method for doing NER. However, since we still do not have a sufficient amount of annotated texts as to train the neural network, we will only show preliminary results of our rule-based method for the titles of the notes and for the complete articles.

We used Freeling to perform named entity recognition and classification (NERC) of some entities and regular expressions to detect others. After, based on an analysis of a subset of our data we de-

[2] Sklearn Metrics Cohen Kappa Score. Available at: http://https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score [Accessed June 2020].

fined a rule-based method to detect named entities.

In next sections we present a summary of our rule-based and machine learning methods we are developing.

4.1 Rule-based method

Below we describe the defined rules.

Recognition of number of cases (NoC)

To detect NoC entities we used the following heuristic:
1) we PoS-tagged the NPAs with Freeling. Tokens tagged as numbers (*Z*) or as as ordinal adjectives (*A0*) are potential NoC candidates.
2) We kept those candidates containing only numbers and eventually dots or commas ([0-9.,]+) (eg. the token *D8*, tagged as *Z*, was eliminated).
3) Candidates that met following rules were also removed: i) those that are followed by any of the words *day*, *month*, *year*, *percentage* or *%*, since the detected number probably refers to a point in time or to a percentage and not to a NoC, ii) some ordinal adjectives, like *last*,[3] and
4) finally, from the remaining candidates, we picked only those that are at a maximum distance of 7 to some of the words belonging to a list of words that might be related to number of cases (eg. *cases* and *infected*).

Recognition of locations (LOC)

The process to detect LOC entities is as follows:

1) we created a Gazetteer with data downloaded from *GeoNames*[4] (cities,[5] regions, countries -and their official language- and continents were downloaded).

2) Freeling NER and NEC modules were ran and we kept only those tokens tagged as locations (*NP00G00*).

3) Almost all articles mention a location in their title. To collect the valid ones, we filtered those tagged as such (step 2) and eliminated those that do not belong to Latin America or those that belong, but whose inhabitants native language is not Spanish or Portuguese. [6]
4) We only keep those NPA entities tagged as locations by Freeling that are related to the location found in the article note title.[7]
5) If there exists some kind of ambiguity (eg. *El Salvador* is a country and a city in Mexico), we select the location with the highest population.
6) Finally, we keep only those locations that co-occur in the same sentence as a NoC entity.

Recognition of other named entities

The rest of the named entities (diseases, hosts, origin, transmission form, and negation, uncertainty, past and conditional terms) were detected by the use of regular expressions and lists of terms developed ad-hoc by us based on the subset of NPA analyzed. Dates were detected with the use of Freeling.

Dates and negation, uncertainty, past and conditional terms were recognized as NER only if they co-occurr in the same sentence and within a fixed distance to another entity to which they might be related.

4.2 Machine learning method

We are currently working on a machine learning method for NER based on the work from Akbik et al (2018).

Therefore, we are using the library provided by the authors. We initialized a set of stacked embeddings in Spanish. Then, we are going to train the sequence tagger with our annotated articles by adapting the proposed architecture. This process will be finished once we have a higher number of annotated articles.

5 Preliminary Results

In this Section we show preliminary results of our rule-based method. 20 % of our dataset, not used for doing the analysis was used to test the results. We show the usual Precision, Recall and F1 metrics for each entity and an overall average score that considers all named entities.

Therefore, the annotated files were transformed to coNLL format with the script *anntoconll.py* provided by brat. A normalization and transformation was performed and conlleval perl script[8] was ran.

[3] Others, as *first*, may refer to NoC, so they were not removed.

[4] GeoNames WebServices. Available at: http://www.geonames.org/export [Accessed April 2020].

[5] For Latin American cities, those with population over 5000 inhabitants were obtained, for other cities, those with more than 15.000.

[6] This information is taken from the information downloaded from GeoNames (step 1).

[7] Eg. cities belonging to the country, and country where the city or region belongs to.

[8] Available at: https://www.clips.uantwerpen.

Tables 5 and 6 show results for NPAs and for the reduced articles.

NE	P	R	F1
Date	0.63	0.44	0.52
Disease	0.74	0.67	0.70
Host	0.53	0.28	0.37
Location	0.31	0.61	0.41
Number of cases	0.63	0.43	0.51
Origin	0.07	0.13	0.09
Transmission form	0.03	0.15	0.05
Conditional	0.00	0.00	0.00
Negation	0.58	0.09	0.16
Past	0.07	0.17	0.10
Uncertainty	0.16	0.13	0.14
Total	0.48	0.46	0.47

Table 5: Performance of the rule-based method for NPAs.

NE	P	R	F1
Date	0.79	0.65	0.71
Disease	0.89	0.85	0.87
Host	0.64	0.64	0.64
Location	0.69	0.63	0.66
Origin	0.18	0.50	0.26
Transmission form	0.08	0.14	0.10
Negation	1.00	0.33	0.50
Uncertainty	0.00	0.00	0.00
Total	0.72	0.69	0.70

Table 6: Performance of the rule-based method for reduced NPAs. Only those entities and modifiers that appear in the title are shown.

Next, we do a brief analysis of our results. 1) For NPAs, generally those NE detected only by regex and lists of terms have the worst results (Hst, OR, TF, NT, UT, PT and COND). This is mainly due to differences with the annotation criteria and to the fact that lists of terms were not comprehensive enough. 2) Nevertheless, in both cases (NPAs and partial notes) diseases had much better results. We assume that it is because there is a reduced number of diseases we are looking for. 3) For NPAs, those NE that were based on the use of Freeling and of more elaborated rules had better results. 4) The main differences among the algorithm and the annotation criteria are between i) what is considered past, ii) the difference among OR and TF (it occurs in partial NPAs and NPAs), and iii) the dates. 5) the algorithm does not detect zika as a cause of microcephaly. This is due to the fact that zika is not in the list of possible causes (entity origin). 6) We also notice a difference among the annotation criteria and the development of the algorithm when seeing the results for LOC NE. That, together with the fact that mainly in the title, countries are usually mentioned together with the abbreviation of a city (eg. *BRASIL: (SP))* made that LOC results are not good. With our gazetteers, we don't have a way to know the expansion of the abbreviations.

Furthermore, those abbreviations are usually non-standard. 7) Normalizing dates, would enhance results of Freeling date detection. 8) Finally, uncertainty has bad results in reduced NPAs. It only appears in eight titles and there was a mismatch among those detected by the algorithm and those annotated by the annotators.

Overall, as we imagined, NER worked better for the title than for the NPAs. Those results that are low are mainly associated with the lack of normalization and the difference with the annotation criteria (eg. OR and TF). Finally, even though the results do not seem encouraging, there are entities that were recognized correctly by the algorithm but no by the annotators (eg. in *the Ministry reported that there are 24.011 suspected cases of (...))*, the algorithm correctly recognized *24.011* as NoC and *suspected cases* as UT, but the annotator did not. This shows that the annotation and annotation criteria still have to be improved.

6 Conclusion

We have just presented an initial annotated corpus based on ProMED-mail which can be used for early detection systems of outbreak diseases prevalent in Latin America. While there are several social media sources for disease surveillance, there is a lack of tools that can automatically provide detailed epidemiological data. Thus, the whole process of building the corpus was oriented to aid the extraction of useful health facts. With this corpus, it is possible to analyze medium/long articles from reliable sources such as ProMED-mail and answer queries regarding detailed and direct diseases incidences. This can be useful for understanding epidemiological trends.

We have also proposed an initial baseline rule based algorithm that automatically extract diseases related entities. The results reported are promising and we also plan to work on relation extraction. We are developing a baseline based on co-occurrence of named entities. The analysis of the preliminary results shows different criteria between the algorithm rules and the annotation criteria. This analysis will help us review and enhance our corpus. Finally, the rule-based method is very laborious

be/conll2000/chunking/conlleval.txt [Accessed July 2020].

and is not always useful to detect entities without understanding the contexts in which they appear.

Acknowledgements

This work was funded by CONCYTEC-FONDECYT under the call E041-01 [contract number 34-2018-FONDECYT-BM-IADT-SE].

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Elena Arsevska, Sarah Valentin, Julien Rabatel, Jocelyn de Goër de Hervé, Sylvain Falala, Renaud Lancelot, and Mathieu Roche. 2018. Web monitoring of emerging animal infectious diseases integrated in the french animal health epidemic intelligence system. *PloS one*, 13(8).

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

D. K. Bonilla-Aldana, Y. Holguin-Rivera, I. Cortes-Bonilla, M. C. Cardona-Trujillo, A. García-Barco, H. A. Bedoya-Arias, A. A. Rabaan, R. Sah, and A. J. Rodriguez-Morales. 2020. Coronavirus infections reported by promed, february 2000-january 2020. *Travel medicine and infectious disease*, 35(101575).

M. Carrion and L. C. Madoff. 2017. Promed-mail: 22 years of digital surveillance of emerging infectious diseases. *International health*, 9(3):177—-183.

L. E. Charles-Smith, T. L. Reynolds, M. A. Cameron, M. Conway, E. H. Lau, J. M. Olsen, J. A. Pavlin, M. Shigematsu, L. C. Streichert, K. J. Suda, and C. D. Corley. 2015. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PloS one*, 10(10).

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

A. N. Desai, A. Anyoha, L. C. Madoff, and B. Lassmann. 2019. Changing epidemiology of listeria monocytogenes outbreaks, sporadic cases, and recalls globally: A review of promed reports from 1996 to 2018. *International journal of infectious diseases*, 84:48—-53.

O. Edo-Osagie, B. De La Iglesia, I. Lake, and O. Edeghere. 2020. A scoping review of the use of twitter for public health research. *Computers in biology and medicine*, 122(103770).

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

C. Rolland, C. Lazarus, C. Giese, B. Monate, A. S. Travert, and J. Salomon. 2020. Early detection of public health emergencies of international concern through undiagnosed disease reports in promed-mail. *Emerging infectious diseases*, 26(2):336—-339.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

R. Thiebaut and S. Cossin. 2019. Artificial intelligence for surveillance in public health. *Yearbook of medical informatics*, 28(1):232—-234.