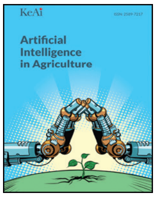
Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Identifying associations between epidemiological entities in news data for animal disease surveillance

Sarah Valentin [a,b,c], Renaud Lancelot [a,b], Mathieu Roche [a,c,*]

[a] CIRAD, UMR ASTRE, UMR TETIS, F-34398 Montpellier, France
[b] ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France
[c] TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France

ARTICLE INFO

Article history:
Received 21 March 2021
Received in revised form 6 June 2021
Accepted 29 July 2021
Available online 31 July 2021

Keywords:
Animal disease surveillance
Text mining
Event extraction

ABSTRACT

Event-based surveillance systems are at the crossroads of human and animal (and plant and ecosystem) health, epidemiology, statistics, and informatics. Thus, their deployment faces many challenges specific to each domain and their intersections, such as relations among automation, artificial intelligence, and expertise. In this context, our work pertins to the extraction of epidemiological events in textual data (i.e. news) by unsupervised methods. We define the event extraction task as detecting pairs of epidemiological entities (e.g. a disease name and location). The quality of the ranked lists of pairs was evaluated using specific ranking evaluation metrics. We used a publicly available annotated corpus of 438 documents (i.e. news articles) related to animal disease events. The statistical approach was able to detect event-related pairs of epidemiological features with a good trade-off between precision and recall. Our results showed that using a window of words outperformed document-based and sentence-based approaches, while reducing the probability of detecting false pairs. Our results indicated that Mutual Information was less adapted than the Dice coefficient for ranking pairs of features in the event extraction framework. We believe that Mutual Information would be more relevant for rare pair detection (i.e. weak signals), but requires higher manual curation to avoid false positive extraction pairs. Moreover, generalising the country-level spatial features enabled better discrimination (i.e. ranking) of relevant disease-location pairs for event extraction.

1. Introduction

Increases in the emergence or re-emergence of animal and human infectious diseases have been evident in many parts of the world for several years. Beyond the well-known role of human and animal mobility in the spread of pathogens, climate change and biodiversity loss are likely to exacerbate the global disease burden (Keesing et al., 2010; Ostfeld, 2009). National and international institutions are currently experimenting with a global paradox—reconciling trade extension with the control of the risk to public and animal health.

The growing availability of digital data represents an unprecedented source of real-time disease information (Paolotti et al., 2014). Online news, social media and electronic health records are among the so-called informal sources that have proven to be valuable sources of disease information (Soto et al., 2008; Wilson and Brownstein, 2009). Through the epidemic intelligence (EI) concept, their mainstreaming into surveillance systems has been a paradigm shift for disease surveillance and control. Driven by the International Health Regulations (IHR) (WHO, 2005), EI integrates two components in a single surveillance system: indicator-based surveillance (collection of structured data through traditional surveillance systems) and event-based surveillance (collection of unstructured data from informal sources) (Paquet et al., 2006). Combining these two components has proven to enhance surveillance systems' performance by increasing outbreak detection timeliness and number (Arsevska et al., 2018; Bahk et al., 2015; Dion et al., 2015; Barboza et al., 2013).

Informal sources cover a diverse spectrum, but they all share the information in textual format. Peculiarities of textual data include linguistic ambiguities, redundant and noisy information, a lack of normalisation, etc. Besides, daily amounts of such information can rapidly overwhelm surveillance systems, including moderation steps performed by experts. Event-based surveillance (EBS) systems thus

[*] Corresponding author at: CIRAD, UMR TETIS, F-34398 Montpellier, France.
E-mail address: mathieu.roche@cirad.fr (M. Roche).

https://doi.org/10.1016/j.aiia.2021.07.003

increasingly marshal text-mining and NLP (Natural Language Processing) methods to alleviate the amount of manual curation of the continuous flow of free text (Hartley et al., 2010; Drury and Roche, 2019).

In this setting, the Platform for Automated extraction of Animal Disease from the Web (PADI-web[1]) is an open-access EBS system dedicated to the detection of new and emerging animal infectious disease events (Arsevska et al., 2018). It was developed to meet the needs of the French Epidemic Intelligence System (FEIS, or *Veille sanitaire internationale* in French) via online news monitoring. FEIS has been involved in activities of the French Platform for Animal Health Surveillance (ESA Platform) since 2013. FEIS aims to identify, monitor and analyse reports of animal health hazards (including zoonotic diseases) threatening France as a whole by monitoring official and unofficial information sources. PADI-web monitors Google News in real-time and automatically retrieves animal disease related news articles, classifies them and extracts epidemiological entities (Valentin et al., 2020b). The classification module of PADI-web is based on a supervised machine learning approach (e.g. Random Forest, Support Vector Machine, and Multilayer Perceptron) to filter relevant news with an overall accuracy of 0.94 (Valentin et al., 2020a).

This paper tackles issues related to knowledge-based systems in agriculture (i.e. livestock farming and epidemiological surveillance) using an unsupervised machine learning approach. Many applications based on unsupervised approaches exist, e.g. clustering, anomaly detection, association extraction. Our work focuses on association mining. This task consists of identifying sets of items that often occur together in datasets (e.g. databases, textual data, etc.). We propose an unsupervised method for pattern recognition (i.e. co-occurrence of epidemiological features) applied to animal disease surveillance (i.e. PADI-web data) to extract events in news.

The motivation of this work is to use simple and effective statistical measures to extract epidemiological events that are easy to analyse by experts. The measures studied and extended in this paper use weak knowledge based on the number of examples in textual data without the need to determine counter-examples. The main objective is to determine which parameters that we need to integrate. We focus on 2 main parameters associated with our statistical methods and unsupervised approaches: (i) what textual context to use (i.e. document, sentence, and word) for extracting pairs of elements to define an epidemiological event for animal disease surveillance and (ii) which pairs of entities and generalisation to apply for event extraction. This type of results could be relevant insights for integration in embedding approach architectures.

Section 2 presents the related work on event-based surveillance systems, entity extraction, entity normalisation and linking and event extraction applied to animal disease surveillance. Section 3 presents our global process in order to extract relevant events. Sections 4 and 5 discuss the results obtained with different strategies.

2. Related work

2.1. Event-based surveillance systems

The development of EBS systems aims at meeting the challenges posed by the integration of unstructured data in the formalised EI process. Below we present the EBS systems that encompass the animal health threat in their scope.

EBS systems were pioneered in 1994 by the International Society for Infectious Diseases (ISID), through the Program for Monitoring Emerging Diseases (ProMED). ProMED is a human-curated system that relies on an extensive network of experts worldwide who detect and share reports on disease outbreaks using a common platform (Carrion and Madoff, 2017). Moderators validate the information.

BioCaster and the Platform for Automated extraction of animal Disease Information from the web (PADI-web) rely on fully automated pipelines. BioCaster was a public health surveillance system supported by the University of Tokyo from 2006, with a priority focus on the Asia-Pacific region (Kawazoe et al., 2008). BioCaster is no longer operational, but it is included in our review because it relied on a unique and well-documented ontology-based approach. PADI-web was created in 2016 to monitor online animal health-related news for the French Epidemic Intelligence System (FEIS) (Arsevska et al., 2018; Valentin et al., 2020b).

Between these two extremes of pure automation and pure manual data collection and analysis, other prominent systems combine automated text-mining based steps and a dedicated team of curators to assess and verify the outputs. Semi-automated systems include HealthMap, founded by the Boston Children's Hospital in 2006, the Canadian Public Health Agency Global Public Health Intelligence Network (GPHIN), the European Union MediSys, Argus and AquaticHealth.net.

2.2. Entity extraction

2.2.1. Entity extraction approaches

Information extraction (IE) in EBS systems aims at locating specific pieces of data in natural-language documents, thereby extracting structured information from unstructured text (Mooney and Bunescu, 2005). Entity extraction, also called named entity recognition (NER), is an IE subtask that seeks to locate and classify textual elements into predefined categories, such as locations (e.g.'Lagos', 'China'), temporal expressions (e.g. 'last month', 'July 28, 1990′), organisations (e.g. 'Ministry of Health'), person names, quantities (e.g. '2′), etc.

Regarding geographical entities in online news, it is important to distinguish: (i) geographic entity extraction and resolution from (ii) identification of the event-related location. Geographic entity extraction and resolution aim at correctly extracting and identifying all locations from a text.

The dictionary-based approach involves matching terms from a document with a list of words to extract entities from texts. Geographical dictionaries are usually called gazetteers. Some dictionaries can have an ontological structure rather than a simple list of terms. Ontologies aim at modeling the relations between entities (Guarino et al., 2009). For instance, in the GeoNames ontology (i.e. gazetteer), spatial entities are structured into different hierarchical classes identified by a letter, with each of the letters corresponding to a specific category (e.g. for administrative borders). In the health domain, an ontology can represent the causality relationships between a disease and a pathogen (Chanlekha et al., 2010).

To overcome the rigidity of the dictionary-based approach, another approach consists of considering NER as a classification task, where the type of entity is the label to assign. Extraction rules can be generated by hand or automatically. The method of the latter case relies on machine learning trained on manually annotated data. Conditional random fields (CRF) is among the most prominent classifier used for NER (Lafferty et al., 2001), at the core of well-established pre-trained NER tools, including StanfordNER (Manning et al., 2014) and NLTK (Bird and Loper, 2004). This approach is designed for sequential data: CRFs predict the probability of the output sequence according to a given input sequence (Song et al., 2019).

The classification approach is particularly suitable for misspelt locations or texts short in length, such as tweets. The gazetteer lookup suffers from low precision due to irrelevant matches (Inkpen et al., 2017). While classifier-based approaches achieve good results, they are limited to the predefined categories upon which they are trained, i.e. nonspecific domain entities (dates, locations, etc.). Recently, neural network training algorithms have shown great success in several NLP tasks, including named entity recognition. These models have achieved state-of-the-art results while alleviating the burden of the amount of feature pre-processing (Chiu and Nichols, 2016). The RNN-based

[1] https://padi-web.cirad.fr/en/

algorithm from spacy package allows users to add new types of entities to NER algorithms by training its model on annotated datasets (Honnibal and Montani, 2018).

2.2.2. Domain non-specific entities

HealthMap extracts locations using a dictionary of 2300 location-place patterns. The extraction is consolidated with heuristics to infer people's job titles and full names as well as to decipher the acronyms of organisations. Redundant locations are further filtered out based on container relationships by retaining the highest granular level of information. For instance, if 'Boston' and 'Massachusetts' are identified as locations, 'Massachusetts' is eliminated (Freifeld et al., 2008). PADI-web extracts locations by matching the text with GeoNames (Ahlers, 2013), and identifies dates using the HeidelTime rule-based system (Strotgen and Gertz, 2010). GPHIN extracts several domain-unspecific entities (e.g. person names, organisations and locations) with the classifier-based Stanford CoreNLP NER. AquaticHealth only extracts locations using the Alchemy Location Extraction application programming interface (API) developed by IBM. Users can further manually add or refine locations from a report (Lyon et al., 2013).

2.2.3. Thematic entities

In all EBS systems, thematic entity extraction is dictionary based. The lists used are either external knowledge resources (GPHIN) or manually built by domain experts (AquaticHealth.net, PADI-web, HealthMap, MedISys, BioCaster). GPHIN extracts medical entities (i.e. syndromes and disease vectors) by combining UMLS and expert heuristics. A hand-curated list of frequent false positive terms is applied to filter out irrelevant terms. All EBS systems extract at least the disease name. GPHIN, MedISys, BioCaster and PADI-web extract symptoms. PADI-web also detects the host species and the number of cases using regular expressions. Both MedISys and BioCaster use their own ontologies to extract both thematic and domain-nonspecific entities (Collier et al., 2007; Ralf et al., 2008). The multilingual BioCaster ontology (BCO) contains 18 classes encompassing both epidemiological concepts (e.g. virus, symptom) and generic concepts (e.g. locations) (Kawazoe et al., 2008; Kawazoe et al., 2006). Ontologies dedicated to the agriculture domain can be used (Drury et al., 2019).

2.3. Entity normalisation and linking

2.3.1. Domain non-specific entities

Mapping locations to an external gazetteer has several advantages, as it allows: (i) geocoding to map the detected location via latitude-longitude coordinates, (ii) inferring parent-child relationships between different granularity levels to group synonymous mentions or to synthesise local information from a global perspective (e.g. at the country level), (iii) geotagging a document to improve information retrieval from EBS databases. Both PADI-web and GPHIN detect geographical entities with the GeoNames gazetteer. In PADI-web, location mapping is merged with the entity extraction step described above, while these two phases are separate in GPHIN. Using a classification-based approach before matching with an external knowledge resource reduces geographical and non-geographical ambiguities when a noun is the same as an existing location name. For instance, the term 'More' may erroneously match the city of More in England in the PADI-web pipeline. However, in both cases, a place name has multiple gazetteer entries, thus creating geographical-geographical ambiguities. In GPHIN, such issues are resolved through heuristic rules that take where an article was published into account, but further details on the procedure are not available. PADI-web does not address this problem, and all entries are retained. In AquaticHealth.net, locations are geocoded using the Google Maps API so that reports can be presented on a Google Map on the system's website.

2.3.2. Thematic entities

Thematic entities are usually normalized to their canonical form (e.g. disease acronyms are converted into the full disease name). GPHIN provides a link between the detected entities and UMLS terminology and definitions. BioCaster Ontology provides access to term definitions, synonyms and translations in eight languages, along with a link to medical ontologies (including ICD-10, MedDRA, MeSH and SNOMED-CT) (Collier et al., 2008).

2.4. Event extraction

Event extraction methods have been extensively studied in many domains such as business and financial (Du et al., 2016), biomedical (Zhu and Zheng, 2020), and outbreak-event detection (Piskorski et al., 2011) domains. (Xiang and Wang, 2019) propose a comprehensive and synthetic survey of event extraction methods. Briefly, they include pattern-based methods, machine learning methods (supervised or semi-supervised), deep learning methods, and unsupervised methods.

In the studied EBS systems, HealthMap, PADI-web, BioCaster and MedISys include an event extraction step, all of which rely on a different approach.

2.4.1. Unsupervised approach

HealthMap event extraction is unsupervised, i.e. it does not train any event extraction models. Typically, in the unsupervised approach, the detection of triggers and arguments is based on word distributional representations. In HealthMap, the event mention and trigger detection steps are ignored. Instead, the approach relies on the document structure, based on the hypothesis that the most relevant information (i.e. event attributes) appears at the beginning of a news report. Diseases and locations are first searched in the title, then in the document headlines, and finally in the full content. If the algorithm cannot extract relevant elements from these three levels, the name of the online news source is used instead. This last step relies on the assumption that news articles that do not contain any specific location refer to a place near the publication source. Erroneous extractions are further corrected by analysts when necessary (Brownstein et al., 2008). This approach decreases the risk of false-positive extraction (i.e. extraction of locations that are not true event attribute). Two shortcomings should be noted: the spatial granularity is reduced since the attribute extraction stops at the first entity detected, and also, this approach cannot address cases of news articles containing several events.

2.4.2. Pattern-based approach

Both MediSys and BioCaster rely on pattern-based methods, which were the earliest approaches proposed for event extraction. They consist of matching text with specific event templates. Patterns are constructed manually or automatically. Manual event construction typically relies on domain expert proposals, thereby achieving high accuracy. However, manual construction is time-consuming, and expert bias can lead to a lack of recall. A weakly-supervised method or bootstrapping can automatically generate patterns from a pre-classified training corpus or seed patterns. In MediSys, event extraction is performed by the Pattern-based Understanding and Learning System (PULS) developed at the University of Helsinki. PULS relies on a cascade of patterns applied to each news article's sentence structure to extract the event attributes. For instance, the pattern in Fig. 1 uses both syntactical and semantic information of the sentence.

This pattern matches a noun phrase (NP) of semantic type (i.e. 'disease') with a verb phrase (VP) headed by the verb 'kill' (or its synonyms in the ontology) and has the adverbial phrase 'so far', etc. The square brackets indicate an optional match. If the location is omitted in the sentence, it is inferred from the surrounding context. Verb phrases are not rigid and allow the presence of modifier elements, such as an auxiliary verb (e.g. 'has') or adverb (e.g. 'so far') (Steinberger et al., 2008). PULS implements weakly-supervised learning to reduce the amount of

NP(disease) VP(kill) NP(victim)['in' NP(location)]

Fig. 1. Pattern associated with an event.

manual labour as far as possible by automatically learning new patterns via bootstrapping (Grishman et al., 2002).

BioCaster event extraction uses a simple rule language (SRL), inspired by the so-called declarative information analysis language (DIAL) (Feldman et al., 2001). SRL creates sophisticated matching patterns combining entity classes, string literals, regular expressions, entity types including verbs of infection, common victim expressions, occupation names, etc.

Several *machine-learning* and *deep learning-based methods* have been proposed for event extraction (Margineantu et al., 2010). RNN-based pre-trained models, known as word embedding models, are capable of capturing the meaning of terms depending of their context (i.e. their surrounding words) (Mikolov et al., 2013) or GloVe (Pennington et al., 2014). Such models can be integrated into event extraction pipelines, surpassing most existing methods (Yang et al., 2019). Word embedding models have yet been applied to several linguistic tasks in the disease surveillance domain, including disease taxonomy development (Ghosh et al., 2016), epidemiological feature extraction from WHO reports (Ghosh et al., 2017) and veterinary necropsy report classification (Bollig et al. (2020)). To our knowledge they are not yet implemented in any operational EBS system for animal health surveillance.

The PADI-web pipeline is based on four steps (i.e. data collection, data processing, data classification, and information extraction) summarized in Fig. 2. The extraction module aims at identifying disease, host and location which are the attributes of an event. The aim of this paper is to identify relevant associations between elements using unsupervised approaches for highlighting relevant events.

3. Event extraction approach based on an unsupervised approach

In this section, we define the event extraction task as detecting pairs of epidemiological entities from textual data (e.g. a disease name and location). Two entities form an event-related pair if they are attributes of the same event. Event-related pairs of attributes are hereafter referred to as"relevant pairs".

To address this issue, we propose association mining methods (i.e. unsupervised machine learning approaches) in order to extract entity co-occurrence in news articles. More precisely, our approach involves two steps: (i) the detection of pairs of entities based on their relative position in the news article content, and (ii) their ranking based on two state-of-the-art term association measures (pointwise mutual information and Dice coefficient). Our contribution addresses the following questions:

1. What are the best co-occurrence parameters to select relevant pairs of entities from a corpus of news articles?
2. What is the impact of two association measures for the ranking of relevant pairs of entities?
3. How can contextual aspects be integrated for the ranking measures?
4. Does the generalisation of spatial entities improve the retrieval of relevant pairs?

Below we outline the proposed statistical approach and further describe the protocol and corpus used for the evaluation.

3.1. Detection and ranking of pairs of entities

The computation of the association strength between two or more words (i.e. co-occurrence) is applied in several tasks, such as the discovery of association rules (Blanchard et al., 2005), feature extraction (Torkkola, 2003) and document summarization (Aji, 2012). Our objective is to identify the best parameters regarding entity co-occurrence and spatial hierarchy to improve the retrieval of relevant pairs, using the Dice coefficient and pointwise mutual information. In the following, pointwise mutual information will be referred to as Mutual Information (MI) for reason of simplification. MI has been used to discover and cluster words specific to events in a stream of tweets (Preotiuc-Pietro et al., 2016). Our approach is based on the same rationale. Rather than taking all the words into account, we compute the association measure only between predefined epidemiological entities (i.e. disease, host and location). Several other text-mining association metrics could be applied to our task, such as Jaccard, Cubic MI (Niwattanakul et al., 2013) or other measures such Bayes Factor, as applied in the data mining domain (Lallich et al., 2007). However, we opted to focus on Dice and MI due to their simplicity, interpretability, and highly different behaviour regarding co-occurrence counts (Roche and Prince, 2010).

Mutual Information (MI - Eq. (1)) measures the relative difference between observed word co-occurrences, and their expected co-occurrence assuming independence (Church and Hanks, 1989). MI is defined as the probability that two words co-occur in the same context (the context concept is discussed below), divided by the product of the probabilities of each word occurrence in a corpus.

$$MI = log_2 \times \frac{P_{xy}}{P_x \times P_y} \tag{1}$$

where P_x is the probability of occurrence of x, P_y is the probability of occurrence of y, and P_{xy} is the probability of co-occurrence of x and y (joint probability). *Mutual Information* is sensitive to rare and specific co-occurrences (Roche et al., 2004).

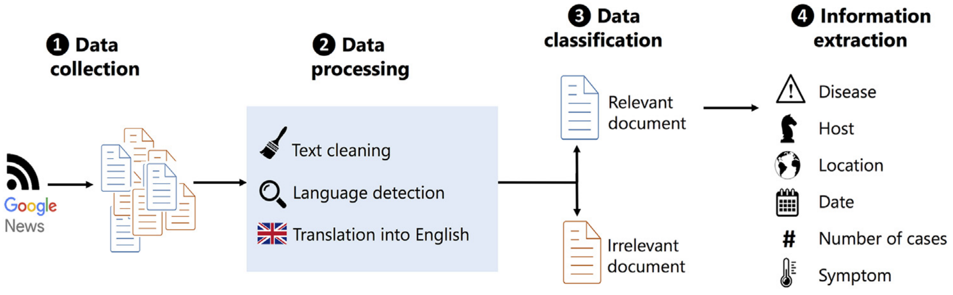


Fig. 2. PADI-Web pipeline.

Dice coefficient (Eq. (2)) is also based on the joint probability, divided by the sum of the individual occurrence probabilities. Dice is less sensitive to low-count co-occurrences (Smadja et al., 1996).

$$Dice = 2 \times \frac{P_{xy}}{P_x + P_y} \tag{2}$$

In both Eqs. (1) and (2), $P_x = \frac{N_x}{N}$, $P_y = \frac{N_y}{N}$ and $P_{xy} = \frac{N_{xy}}{N}$, where N_x is the number of occurrences of x, N_y is the number of occurrences of y and N_{xy} is the number of co-occurrences of x and y. Moreover, as both metrics are used in a ranking purpose while the *log* function is a strictly increasing function, we can simplify Eqs. (1) and (2) as:

$$MI = \frac{N_{xy}}{N_x \times N_y} \tag{3}$$

$$Dice = \frac{N_{xy}}{N_x + N_y} \tag{4}$$

The results of both metrics heavily depend on the context chosen to compute the co-occurrence between two words. In our approach, this context controls the detection of pairs of features. In this paper, we propose three definitions of co-occurrence contexts, hereafter referred to as”levels”:

1. At the document-level: N_{xy} is the number of documents in which x and y co-occur;
2. At the sentence-level: N_{xy} is the number of sentences in which x and y co-occur;
3. At the word-level: N_{xy} is the number of times that x and y co-occur in a w word window.

Word and sentence levels rely on two parameters, i.e. the window size and the window side. The window size corresponds to the number of words (or sentences) separating two entities. The window side can be positive (y appears after x), negative (y appears before x) or bidirectional (y appears before or after x). For both disease-location and disease-host pairs, disease entities are considered as ”pivot”. Thus, a positive (resp. negative) window of w words corresponds to searching for another entity within the w words on the right (resp. on the left) of the disease feature. A bilateral window consists of searching for an entity on the right or left of a disease feature in a sliding window of w words.

We illustrate the influence of the window parameters on pair detection with an example extracted from a news article.[2] Location features are in bold while disease features are in italic (Fig. 3).

When setting the word window size at 15 words and the sentence window size at 1 sentence, [3] the disease-location pairs are:

- At the document level: {*African swine fever*, **Więckowice**}, {*African swine fever*, **Poznań**}, {*African swine fever*, Poland}, {*African swine fever*, **Germany**};
- At a word level, right side, window of 15 words: {*African swine fever*, **Więckowice**}, {*African swine fever*, **Poznań}**;
- At a word level, left side, window of 15 words: no co-occurence;
- At a word level, both sides, window of 15 words: {*African swine fever*, **Więckowice**}, {*African swine fever*, **Poznań**};
- At the sentence level: {*African swine fever*, **Więckowice**}, {African swine fever, Poznań}, {*African swine fever*, **Poland**}, {*African swine fever*, **Germany**}.

3.2. Spatial generalisation

As illustrated in the previous example, spatial information can be provided at different granularity levels (e.g. city and administrative level). These levels generate different pairs of entities while representing the same location. Thus, we evaluated the impact of generalising the spatial entities on different granularity levels. More precisely, based on the GeoNames hierarchy, we converted the location entities into lower granular levels (e.g. converting ”Allier” into ”France”), hereafter referred to as ”generalisation”. We evaluated three generalisation levels:

- Level 0: No generalisation. This level corresponds to raw location values without applying any generalisation. It includes spatial entities with heterogeneous granularity levels (e.g. cities, villages, countries, etc.)
- Level 1: Administrative generalisation. This level corresponds to the conversion of spatial features into their first administrative level. This conversion is applied if the initial spatial granularity is higher than the first administrative level. This level thus still includes heterogeneous granularity levels, such as administrative regions and countries.
- Level 2: Country generalisation. This level corresponds to the conversion of spatial features into their country. This last level only contains countries and supranational entities (e.g. Asia and the European Union).

We illustrate the impact of generalisation on co-occurrence weights with the previous example (Fig. 4):

At level 1, all locations with a lower granularity than the first administrative level (i.e. **Więckowice** and **Poznań**) are converted into their administrative level (**Greater Poland**). At level 2, all locations are converted into their country level, which increases the joint probability of the pair {*African swine fever*, **Poland**}:

- Level 0: {*African swine fever*, **Więckowice**}: $N_{xy} = 1$, {*African swine fever*, **Poznań**}: $N_{xy} = 1$, {*African swine fever*, **Poland**}: $N_{xy} = 1$, {*African swine fever*, **Germany**}: $N_{xy} = 1$;
- Level 1: {*African swine fever*, **Greater Poland**}: $N_{xy} = 2$, {*African swine fever*, **Poland**}: $N_{xy} = 1$, {*African swine fever*, **Germany**}: $N_{xy} = 1$;
- Level 2: {*African swine fever*, **Poland**}: $N_{xy} = 3$, {*African swine fever*, **Germany**}: $N_{xy} = 1$.

The combination of association measures (Eqs. (3, 4)), co-occurrence contexts and spatial generalisation provides a mixed measure to evaluate both the detection quality and the ranking of relevant pairs:

- The window parameters control pair detection;
- The association measure (MI or Dice) controls the ranking of the detected pairs;
- For disease-location pairs, the spatial generalisation level jointly contributes to the detection of a pair and its ranking.

In the following section, we describe the evaluation protocol and the corpus used for the experiments.

4. Experiments

To evaluate the proposed approach, we first annotated a corpus of news articles with events (Section 4.1). We further used the list of annotated events as a gold-standard to automatically determine the relevance of the retrieved pairs of entities (Section 4.2). The quality of the

[2] https://www.theguardian.com/environment/2020/apr/08/african-swine-fever-outbreak-reported-in-western-poland

[3] These parameters were chosen as an example, but a range of values are evaluated in Section 4.3.

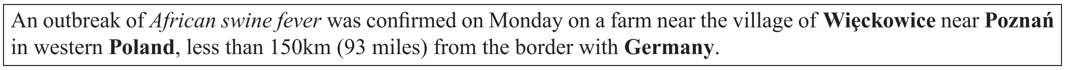

An outbreak of *African swine fever* was confirmed on Monday on a farm near the village of **Więckowice** near **Poznań** in western **Poland**, less than 150km (93 miles) from the border with **Germany**.

Fig. 3. A news article content extract (The Guardian, 8 April 2020).

ranked lists of pairs was evaluated using specific ranking evaluation metrics (Section 4.3).

4.1. Event corpus

We used a publicly available annotated corpus of 438 documents (i.e. news articles) related to animal disease events (either describing a recent outbreak or providing complementary insight regarding control measures, economic impacts, etc.) (Rabatel et al., 2019). This corpus was initially designed for training and evaluating the PADI-web information extraction module. The corpus contains information about the news article itself (publication date, title, content, URL, etc.), as well as epidemiological features (locations, diseases, hosts, dates and symptoms), which were first automatically identified by data mining and rule-based approaches. A veterinary epidemiologist and a computer scientist subsequently labelled each candidate as correct or incorrect. For each document and type of feature (i.e. disease, host, date and location), only candidates manually labelled as correct in the corpus were retained for analysis (including the geographical-geographical disambiguation of locations).

An epidemiologist read each of the 438 documents to detect all disease events contained within them. To ensure consistent and reproducible annotation, events found in the documents were compared to a gold-standard database, i.e. the Emergency Prevention System for Priority Animal and Plant Pests and Disease (EMPRES-i) database. EMPRES-i is a publicly available animal disease information system created by the Food and Agriculture Organization of the United Nations (FAO) (Martin et al., 2007). Among other sources, EMPRES-i stores the official notifications from the World Animal Health Organization (OIE). Each detected event was labelled using the unique EMPRES-i identifier. When the epidemiologist could not link an event to an official one, she created a new event identifier and manually recorded the epidemiological features (location, date, disease and host). The final corpus annotated with the event identifiers is hereafter referred to as the *event corpus*.

The number of news articles containing at least one event represented 53% of the corpus ($n = 229/438$). Among them, 52% ($n = 127/229$) reported several events, with a median number of 3 events (Table 1). One news article contained a maximum number of 208 events due to the reporting of 200 avian influenza outbreaks in Taiwan on 28 January 2015.

Overall, 771 events were detected in the corpus. Among them, 70% ($n = 541/771$) were reported in a single news article. The events present in several news articles were reported in up to 11 news articles (median number of 3 news articles).

Table 1
Descriptive statistics of the number of articles ($N_{article}$) per event and number of events (N_{event}) per articles in the event corpus.

	Min	Median	Mean	Max
n_{event} per article:				
Articles with $N_{event} >= 1$ (n = 229)	1	2.0	5.1	208
Articles with $N_{event} >= 2$ (n = 127)	2	3.0	8.4	208
$N_{article}$ per event:				
Events with $N_{article} >= 1$ ($n = 771$)	1	1.0	1.5	11
Events with $N_{article} >= 2$ ($n = 230$)	2	3	2.8	11

In the following experiments, we selected only news articles containing at least one event (a corpus of 229 documents). Even if still modest in size, our corpus is highly specialized regarding both its domain (i.e. animal health) and its nature (i.e. online news articles).

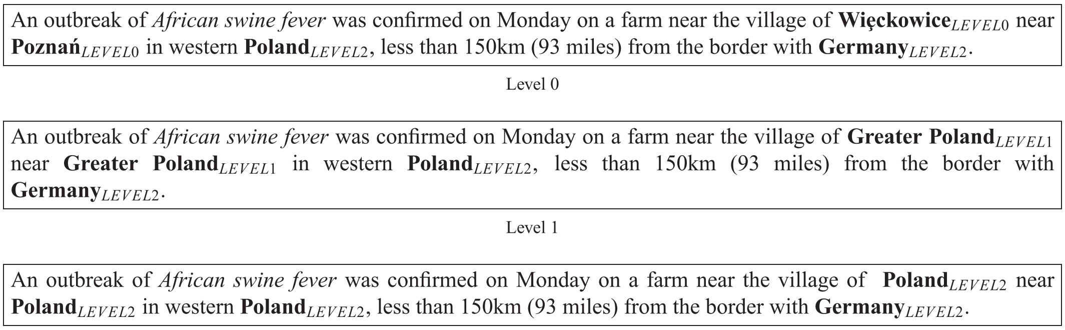
4.2. Relevant pairs

From the annotated event corpus, we can map each publication date $date_i$ with the set of its corresponding events (i.e. the events annotated in the news articles published on $date_i$). Sets including the publication dates and their epidemiological attributes are used as gold-standard lists to evaluate the relevance of extracted pairs of features. We created a gold-standard list specific to each type of pairs, as follows:

1. We aggregated news articles from the event corpus by publication date;
2. For each distinct date $date_i$, we extracted all events $event_i$ labelled in the set of news articles published on $date_i$ (gold-standard list);
3. For each event $event_j$, we retrieved its disease ($disease_j$), host ($host_j$) and country ($country_j$);
4. The gold-standard lists include all of the formed $date_i$, $disease_j$, $country_j$ and $date_i$, $disease_j$ and $host_j$ sets.

Identifying the extracted events for each publication date was required to avoid false positive matches between retrieved pairs and the gold-standard lists. As the event corpus covers a 2-year period, a pair of features extracted at $date_j$ could erroneously correspond to a pair corresponding to a different event.

The disease-location and disease-host gold-standard lists contained 248 and 228 sets, respectively. Each retrieved pair extracted from a set of articles at date $date_i$ was relevant if it matched at least one pair from the gold-standard list corresponding to date $date_i$ (Fig. 5). To match the gold-standard terms (disease names, species names and


An outbreak of *African swine fever* was confirmed on Monday on a farm near the village of **Więckowice**$_{LEVEL0}$ near **Poznań**$_{LEVEL0}$ in western **Poland**$_{LEVEL2}$, less than 150km (93 miles) from the border with **Germany**$_{LEVEL2}$.

Level 0

An outbreak of *African swine fever* was confirmed on Monday on a farm near the village of **Greater Poland**$_{LEVEL1}$ near **Greater Poland**$_{LEVEL1}$ in western **Poland**$_{LEVEL2}$, less than 150km (93 miles) from the border with **Germany**$_{LEVEL2}$.

Level 1

An outbreak of *African swine fever* was confirmed on Monday on a farm near the village of **Poland**$_{LEVEL2}$ near **Poland**$_{LEVEL2}$ in western **Poland**$_{LEVEL2}$, less than 150km (93 miles) from the border with **Germany**$_{LEVEL2}$.

Level 2

Fig. 4. Generalisation levels of spatial entities. The level of each location (based on the GeoNames hierarchy) is shown.

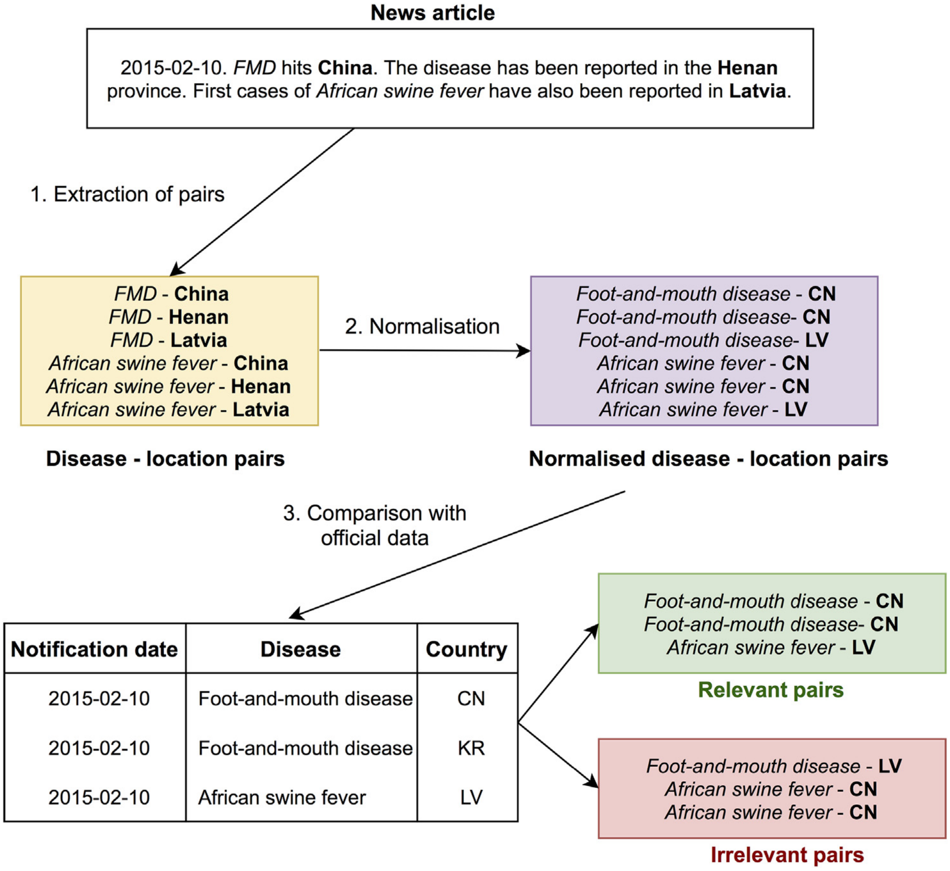

News article
2015-02-10. FMD hits China. The disease has been reported in the Henan province. First cases of African swine fever have also been reported in Latvia.
1. Extraction of pairs
FMD - China
FMD - Henan
FMD - Latvia
African swine fever - China
African swine fever - Henan
African swine fever - Latvia
Disease - location pairs
2. Normalisation
Foot-and-mouth disease - CN
Foot-and-mouth disease- CN
Foot-and-mouth disease- LV
African swine fever - CN
African swine fever - CN
African swine fever - LV
Normalised disease - location pairs
3. Comparison with official data
Notification date
Disease
Country
2015-02-10
Foot-and-mouth disease
CN
2015-02-10
Foot-and-mouth disease
KR
2015-02-10
African swine fever
LV
Foot-and-mouth disease - CN
Foot-and-mouth disease- CN
African swine fever - LV
Relevant pairs
Foot-and-mouth disease - LV
African swine fever - CN
African swine fever - CN
Irrelevant pairs

Fig. 5. Steps to evaluate the relevance of the disease-location pairs extracted from a news article. After extraction (1), disease and location features are normalized (2). A pair present in the gold-standard list is considered relevant (3).

country codes from the EMPRES-i database), diseases and hosts were normalized to their canonical form using a manually built dictionary, and locations were normalized to their country code.

Note that the normalisation of locations differs from the spatial generalisation described in Section 3.2. Normalisation aims at matching a pair with the gold-standard list features. Locations from the same country are not aggregated and are considered as two distinct values in the pair extraction step. In the example from Fig. 5, "Henan" and"China" are considered as two distinct values, even though they are normalized to the same country code.

4.3. Evaluation

4.3.1. Pair extraction and ranking

We extracted all the disease-host and disease-location pairs using the co-occurrence parameters described in Section 3.1. The word window size ranged from 1 to 200 words on each side (left, right, and both). This window was chosen by (Piskorski et al., 2011) for an event extraction task, assuming that most relevant information would be present in the first 200 words. The sentence window size ranged from 0 to 20 sentences per side. We ranked the retrieved pairs in decreasing order based on their *Mutual Information* or Dice values. We evaluated the quality of the ranked list according to the parameters and association measures' ability to assign a better rank to relevant pairs than to irrelevant ones. The ranking was evaluated in terms of normalized precision (P_{norm}), normalized recall (R_{norm}) and F-measure (F_{norm}). R_{norm} and P_{norm} are based on the difference between the sum of ranks of R relevant pairs obtained by a ranking function, and the sum of ranks of an ideal list, where all relevant pairs are retrieved before all the irrelevant pairs (Kishida, 2005; Salton and Lesk, 1968):

$$R_{norm} = 1 - \frac{1}{R*(N-R)} \times \sum_{i=1}^{R} r_i - \sum_{i=1}^{R} i \tag{5}$$

$$P_{norm} = 1 - \frac{1}{log\,(C(N,R))} \times \sum_{i=1}^{R} log\,(r_i) - \sum_{i=1}^{R} log\,(i) \tag{6}$$

where N is the total number of pairs, r_i is the rank of the i^{th} relevant pair in the ordered list, and $C(N,R) = \frac{N!}{R! \times (N-R)!}$.

Graphically, R_{norm} corresponds to the area under curve (AUC) of the receiver operating characteristics (ROC) curve, or AUC. Fig. 6 provides an example of how ROC curves work regarding ranking evaluation.

Let R_1 and R_2 being the two ranked lists of pairs P_i:

- $R_1 = \mathbf{P_2}, \mathbf{P_1}, P_4, \mathbf{P_6}, P_5, P_3$
- $R_2 = P_3, P_4, \mathbf{P_6}, P_5, \mathbf{P_1}, \mathbf{P_2}$

For each relevant pair (in bold), the curve increases one unit in the Y-axis direction. For each irrelevant pair, the curve increases one unit in the X-axis direction. Consequently, the AUC of the best ranking function (here, R_1) is greater than that of a function giving a poorer ranking (here, R_2).

The normalized F-measure F_{norm} is the harmonic mean of R_{norm} and P_{norm} (Eq. (7)).

$$F_{norm} = 2 \times \frac{R_{norm} \times P_{norm}}{R_{norm} + P_{norm}} \tag{7}$$

We also evaluated the quality of the 5 first pairs retrieved by calculating the precision at k ($P @ k$), recall at k ($R @ k$), and the F-measure ($F @ k$), with k=5. We chose this threshold because it provides the

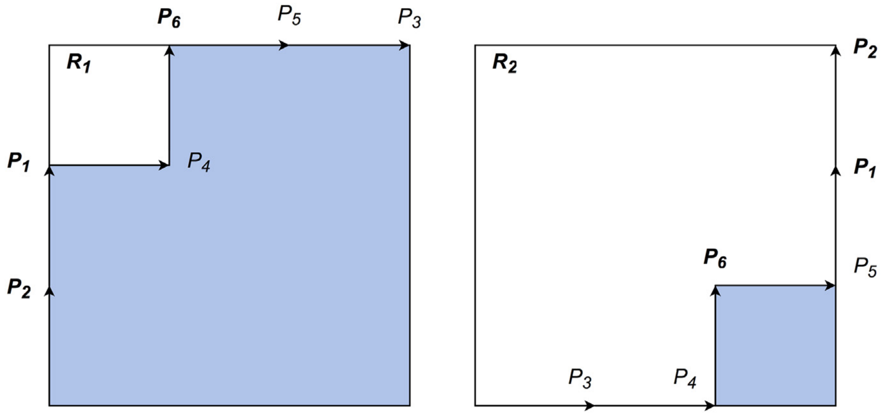


Fig. 6. ROC curves obtained by two different rankings, R_1 and R_2. The pairs in bold correspond to the relevant pairs, and the blue areas correspond to the AUC.

Table 2
Performance of MI and Dice to retrieve and rank relevant disease-host pairs at document level, based on P_{norm}, R_{norm} and F_{norm}.

	Mutual information			Dice		
	R_{norm}	P_{norm}	F_{norm}	R_{norm}	P_{norm}	F_{norm}
Document level	0.79	0.80	0.80	0.83	0.85	0.84
Sentence level	0.78	0.81	0.80	0.84	0.88	0.86
Word level	0.82	0.85	0.87	**0.90**	**0.92**	**0.91**

For sentence-level and word-level windows, the performance corresponds to the best values among the range of window sizes and sides.

local ranking quality, and 95% of the sets of relevant pairs had 1 to 5 elements (pairs).

4.4. Results

4.4.1. Disease - host pairs

Table 2 summarises the best results obtained among all the window parameters evaluated and the document-level performance. The word-level window outperformed document-level and sentence-level windows in terms of normalized precision and recall. The highest precision and recall values were obtained with *Dice* using a window of 26 words on the right side (R_{norm}=0.90, P_{norm}=0.92). The performance obtained with *Dice* values exceeded that with the *MI* values.

The maximum recall at 5 (R@5) ranged from 0.89 to 0.92, while the precision at 5 ($P@5$) reached a maximum value of 0.88 (Table 3). The word-level obtained the best recall-precision balance ($F@5 = 0.88$).

Fig. 7 shows the normalized F-measure (F_{norm}) among the word window sizes and sides. The horizontal lines correspond to the F_{norm} values obtained at the document level. At a given window size and side, *Dice* systematically outperformed *MI*. For both metrics, we achieved better F_{norm} values using a right or bilateral window, clearly outperforming the left-side windows. For all curves, the slope rapidly increased when the word distances increased from 1 to 100. The *MI* performance decreased with window sizes of more than 100 words. *Dice* exhibited a different behaviour, where the performance remained stable among all the window sizes when the values peaked (100 to 200 words).

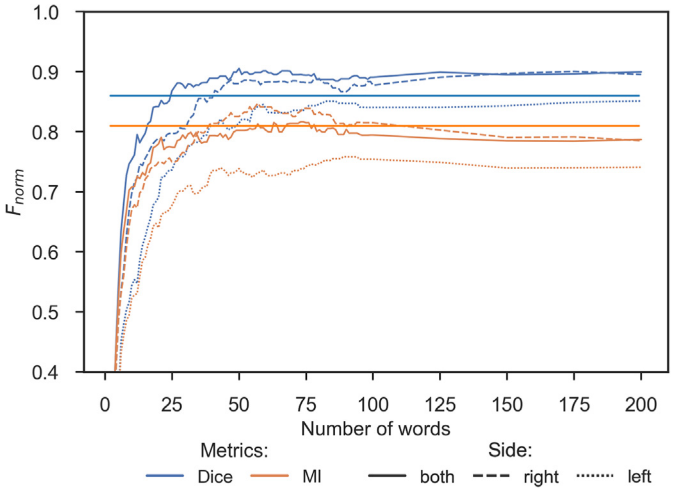


Fig. 7. Performance of MI and Dice to retrieve and rank relevant disease-host pairs in terms of F_{norm}, depending on the window parameters used for the co-occurence count.
For the left side, distances were converted into their positive values. Horizontal lines correspond to the F_{norm} values obtained at the document-level for MI (orange line) and Dice (blue line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Performance of MI and Dice to retrieve and rank relevant disease-host pairs, based on $P@5$, $R@5$ and $F@5$.

	Mutual information			Dice		
	$R@5$	$P@5$	$F@5$	$R@5$	$P@5$	$F@5$
Document level	0.91	0.81	0.86	**0.92**	0.81	0.86
Sentence level	0.89	0.85	0.87	0.90	**0.85**	0.87
Word level	0.90	0.84	0.87	0.91	0.84	**0.88**

For sentence-level and word-level windows, the performance corresponds to the best values among the range of window sizes and sides.

In contrast to the global ranking, Dice and MI obtained similar performance in retrieving the first five pairs (Fig. 8). The F-measure behaviour was identical to the global ranking, with right and bilateral sides obtaining the best results while remaining stable among the window sizes.

4.4.2. Disease-location pairs

The maximal normalized F-measure values for disease-host pairs ranged from 0.62 (document-level, MI) to 0.88 (word-level, Dice) (Table 4). At the document level, the generalisation at the administrative level (level 1) slightly improved performance. The second level (country level) improved the *Dice* ranking's recall and precision at the word level (improving the F-measure from 0.81 to 0.88). However, it decreased the MI ranking performance (at the word level, the F-measure decreased by 0.11).

Fig. 9 highlights the different behaviours of *Dice* and *MI* regarding the generalisation level. Without generalisation (level 0), the best F-measures were obtained for both metrics with a window of 25 words

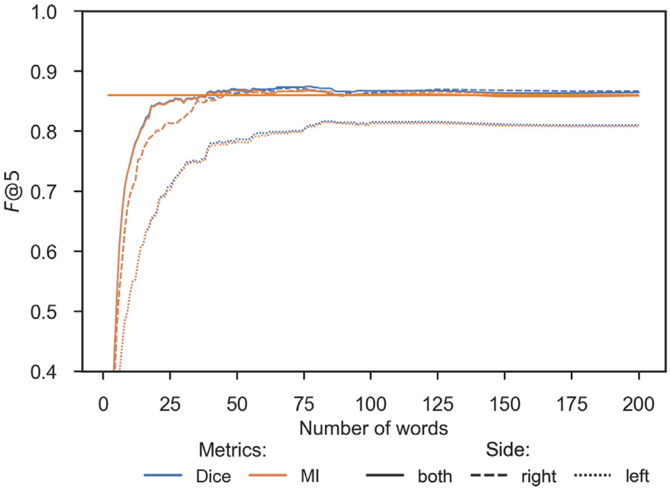


Fig. 8. Performance of MI and Dice to retrieve and rank relevant disease-host pairs in terms of *F* @ 5, depending on the window parameters used for the co-occurence count.
For the left side, distances were converted into their positive values. Horizontal lines correspond to the P@5 values obtained at the document level for MI (orange line) and Dice (blue line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Performance of MI and Dice based on P_{norm}, R_{norm} and F_{norm} to retrieve and rank relevant disease-location pairs at the document level according to the spatial generalisation level.

		Mutual information			Dice		
	Generalisation	R_{norm}	P_{norm}	F_{norm}	R_{norm}	P_{norm}	F_{norm}
Document-level	Level 0	0.73	0.73	0.73	0.75	0.76	0.76
	Level 1	0.76	0.76	0.76	0.77	0.80	0.78
	Level 2	0.64	0.60	0.62	0.81	0.82	0.81
Sentence-level	Level 0	0.73	0.74	0.74	0.75	0.75	0.75
	Level 1	0.73	0.74	0.74	0.75	0.76	0.75
	Level 2	0.63	0.66	0.64	0.62	0.66	0.64
Word-level	Level 0	0.72	0.78	0.75	0.78	0.84	0.81
	Level 1	0.71	0.74	0.73	0.77	0.82	0.80
	Level 2	0.68	0.73	0.71	**0.88**	**0.88**	**0.88**

Level 0: no generalisation, level 1: first generalisation level, level 2: second generalisation level.

(on both sides), and the scores slightly decreased with the largest window sizes. At the country level, the MI F-measure remained below 0.70 while that of *Dice* ranged from 0.80 to 0.88. The Dice ranking reached maximum values between 100 and 125 words (both sides) and remained increased for all window sizes.

The ranking quality at 5 was sensitive to the word windows regarding both the level and generalisation (Fig. 10, Table 5), with the best F-measures reached within a 50-word window.

5. Discussion

5.1. Behaviour of statistical measures

The statistical approaches used in this paper (i.e. *Dice* and *MI*) detected event-related pairs of epidemiological features with a good trade-off between precision and recall. Our results showed that using a window of words outperformed document-based and sentence-based methods while reducing the probability of detecting false pairs.

Our results indicated that Mutual Information was less adapted than the *Dice* coefficient for ranking pairs of features in the event extraction framework. This was especially true when generalising spatial features and increased occurrence counts. Besides, Dice ranking was found more resistant to larger word windows, in line with the findings of (Bouma, 2009), who proposed to add a normalisation factor to the *MI* formula to address low-count issues. We believe that *MI* would be more relevant for rare pair detection (i.e. weak signals) but requires higher manual curation to avoid false-positive extraction pairs.

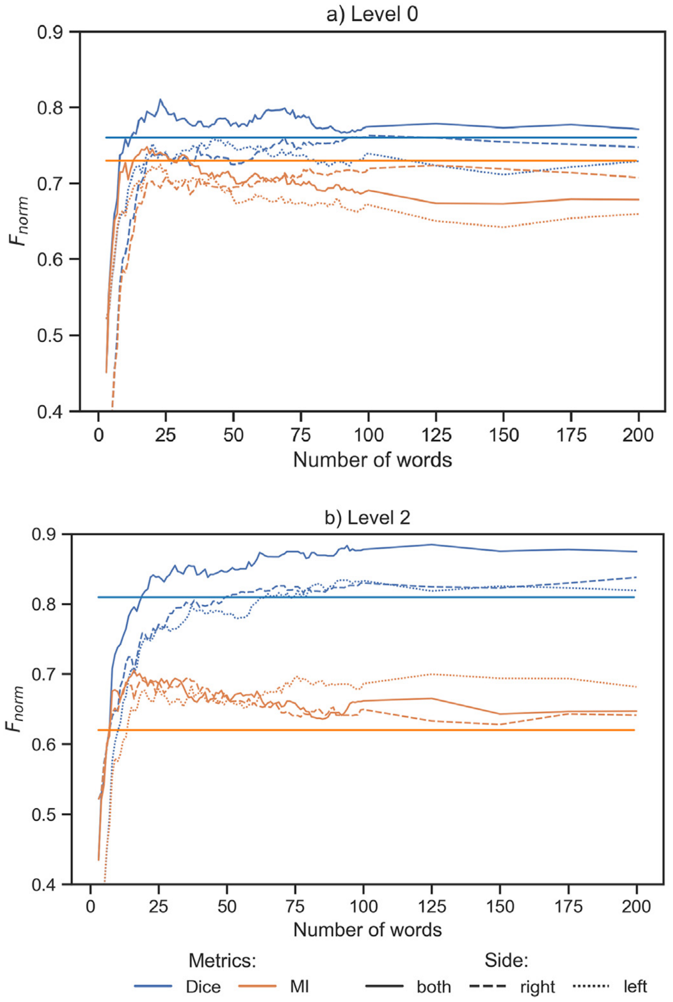


Fig. 9. Performance of MI and Dice to retrieve and rank relevant disease-location pairs in terms of F_{norm} depending on the window parameters used for the co-occurrence count and two spatial generalisation levels.
For left side, distances are converted to their positive value. Horizontal lines correspond to the normalized F-measure values at document-level. Level 0: no generalisation, level 2: generalisation at the country level.

MI tends to extract rare and specific co-occurrences that is highlighted in several studies (Thanopoulos et al., 2002) (Agnihotri et al., 2016). Some variants have been proposed, they consist in introducing factors to the numerator to empirically correct the bias of *MI* that extracts low frequency events (Roche and Prince, 2010; Role and Nadif, 2011). *Dice* coefficient does not favor numerator as these variants of *MI* do but denominator value is less important for computing the sum of both elements.

5.2. Specificities of the corpus

The use of the BERT (Bidirectional Encoder Representations from Transformers) model for event extraction could represent an attractive future work as discussed in Section 6. But the use of pre-trained language models has some limitations with respect to the specific domain addressed in this paper. For example, we studied the use of embedding methods like the Word2vec model that consists in a 2-layer neural network. We used a pretrained corpus (i.e. Google News corpus with 3

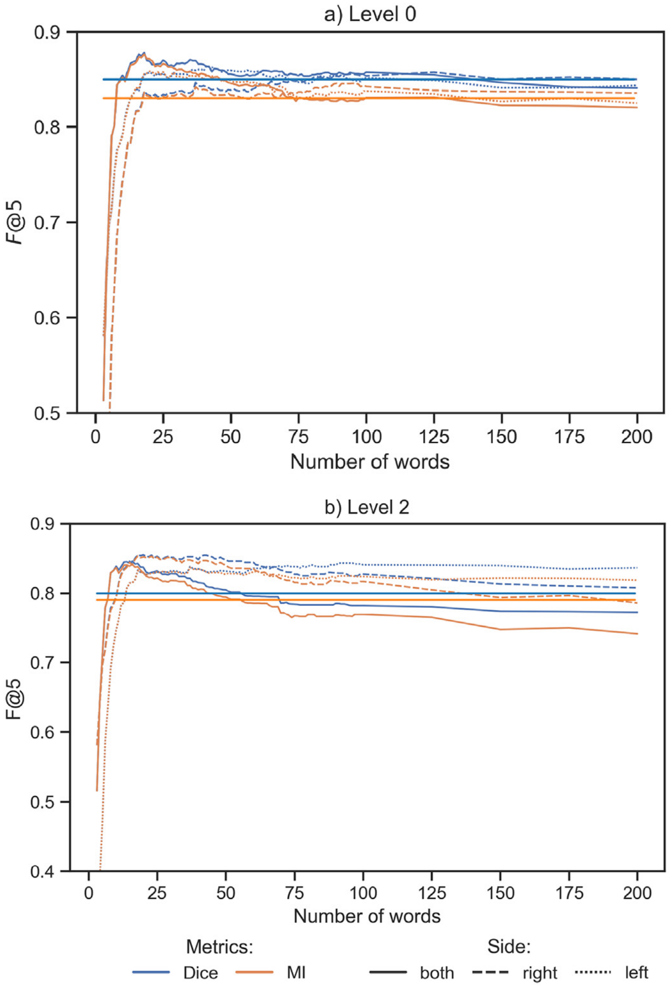


Fig. 10. Performance of MI and Dice to retrieve and rank relevant disease-location pairs in terms of *F*@5 depending on the window parameters used for the co-occurrence count and two spatial generalisation levels.
For the left side, distances are converted into their positive value. Horizontal lines represent the *F* @ 5 values at the document level. Level 0: no generalisation, level 2: generalisation at the country level.

Table 5
Performance of MI and Dice based on *P* @ 5, *R* @ 5 and *F* @ 5 to retrieve and rank relevant disease-location pairs at the document level according to the level of spatial generalisation.

		Mutual information			Dice		
	Generalisation	*R* @ 5	*P* @ 5	*F* @ 5	*R* @ 5	*P* @ 5	*F* @ 5
Document-level	Level 0	0.89	0.78	0.83	0.95	0.77	0.85
	Level 1	0.90	0.77	0.83	0.95	0.77	0.85
	Level 2	0.95	0.67	0.79	**0.98**	0.68	0.80
Sentence-level	Level 0	0.90	0.85	0.87	0.89	0.86	0.87
	Level 1	0.90	0.85	**0.88**	0.90	0.85	**0.88**
	Level 2	0.93	0.76	0.84	0.93	0.76	0.84
Word-level	Level 0	0.91	0.85	0.88	0.91	0.85	**0.88**
	Level 1	0.91	0.80	0.84	0.92	0.80	0.85
	Level 2	0.84	**0.87**	0.85	0.85	0.86	0.86

Level 0: no generalisation, level 1: first generalisation level, level 2: second generalisation level.

billion words) for a classification task (e.g. classification of sentences based on 6 classes: Descriptive epidemiology, Protection and control measures, Concern and risk factors, Transmission pathway, Economic and political consequences and Distribution) using supervised machine learning techniques. The pre-trained Word2vec models (CBOW) did not improve the results with the best classification algorithm (i.e. multilayer perceptron): the accuracy was 0.69 with pretrained models and 0.74 using a specialized corpus (i.e. PADI-web data with 33 million words) for the learning step. Some details of the experiments conducted are given in (Valentin, 2020). These results highlight the limitation of transfer learning approaches. Even with this limitation we plan to investigate other methods that consists in generating labelled data by using the information about the role (Yang et al., 2019).

5.3. Disease-location detection and retrieval

The results obtained for retrieving disease-location pairs without any generalisation suggested that relevant spatial features tended to occur within a small window around the disease feature (25 words, bilateral window). Beyond this window range, the global ranking performance decreased. However, global ranking with the *Dice* coefficient after country-level generalisation exhibited a different behaviour, i.e. remaining stable and close to its maximum value throughout the window size range. Event-related spatial features are provided at different granularity levels. Spatial generalisation allowed us to aggregate related locations in single features, thus increasing event-related pairs' weights. Moreover, spatial generalisation overcomes possible location extraction and disambiguation errors. For instance, news articles often refer to where the sample analyses are performed, thus citing a laboratory location. If the laboratory-based city is extracted as an event-related candidate, using the city feature itself would generate a false alarm. This issue may be overcome by converting the location value into its country, which would lead to lower spatial precision. In the epidemic intelligence framework, the country-level is acceptable for signal analysis, but this approach may not relevant if fine-grained location extraction is needed.

Several gold-standard disease-location pairs were not detected due to a linkage between GeoNames and the Global Administrative Unit Layers (GAUL) used by the EMPRES-i database. In the latter, Taiwan and Hong Kong are considered distinct countries. On the contrary, the GeoNames hierarchy considers them as administrative units of China. (Claes et al., 2014) used a manual procedure to link both databases.

Manual analysis of irrelevant retrieved pairs showed that most of them were due to multiple events, which no statistical approach succeeded in separating.

5.4. Disease-host detection and retrieval

The detection and ranking of disease-host pairs achieved better results than the retrieval of disease-location pairs. The variation in the word window had less impact than for the disease-location pairs. This finding indicates that the precision also remained high when the highest recall value was achieved. This result was expected because thematic features are much less prone to ambiguity than spatial or temporal features. When additional events were present in a news article, they were often summarized in a few sentences containing only the disease and location, thereby reducing the probability of creating false disease-host pairs. Several pairs were detected because irrelevant host terms were extracted from two disease variants, i.e. “small ruminant plague” and “cattle plague”. As these two expressions were not in the dictionary, they were not recognised as diseases, which led to erroneous extraction of ”small ruminant” and ”cattle” as hosts. Both cases were corrected by adding the new variants to the dictionary and reconducting the experiments. Animal disease expressions are often composed of several terms containing epidemiological entities such as hosts and symptoms, so the information extraction performance (and, in this particular case, the coverage of dictionary-based methods) is critical for event extraction. We observed that we did not achieve a recall of 1 when using the broader co-occurrence detection level (i.e. document level). After manual investigation, we discovered that 5 news articles did not contain any reference to a host. Four out of 5 news articles were about the foot-and-mouth disease in endemic areas, and the last article referred to a suspected case of African swine fever. Such cases were rare in the studied corpus. Still, it should be noted that the host attribute is not necessarily communicated in news content, either because the disease is yet well known (e.g. endemic disease in an area) or because the event is only a suspicion. The host is thus implied because this information is secondarily compared to spatial features. This behaviour may bias models that determine news articles' relevance based on the presence or absence of a host's name.

6. Conclusion and future work

This paper highlights that the detection of relevant epidemiological entity association based on unsupervised approaches depends on the textual context, i.e. the window used to retrieve the features in the news. Restraining the context to a fixed window of words achieved better results than retrieving all pairs occurring in a document. Association measures, such as Dice and Mutual Information, could also compute the co-occurrence strength in a simple and interpretable way. Besides, generalising the country-level spatial features enabled better discrimination (i.e. ranking) of relevant disease-location pairs.

A comprehensive comparison of the state-of-the-art association metrics could provide an extensive overview of their performance as further work.

Before applying extraction methods, expansion of text content could be proposed using word embedding architectures like the BERT model. BERT produces word representations that are dynamically informed by the words around them (i.e. context dependent). This model achieved new state-of-the-art results for several NLP tasks (Piskorski et al., 2020; Torregrossa et al., 2021; Trieu et al., 2020).

Moreover, we would like to identify weak signals (i.e. weak pairs of epidemiological information). This concept should be precisely defined in event-based surveillance to formalise specific research questions and assess adapted methods. Rare occurrences of event attributes, such as disease-location pairs, are potential weak signals. They can also signal an event that has not yet been confirmed at the national or international level. In this context, a weak signal could be defined as a temporal anomaly of the frequency of a term or an association of terms compared to a baseline. Detection of weak signals by event-based systems could consist of implementing alerts based on terms-weighted metrics which take the temporal dimension into account.

Acknowledgements

This work was partially funded by the French General Directorate for Food (DGAL), the French Agricultural Research Centre for International Development (CIRAD) and the SONGES Project (FEDER and Occitanie). This work was also supported by the French National Research Agency under the Investments for the Future Program, referred to as ANR-16-CONV-0004.

This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD010. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

References

Agnihotri, D., Verma, K., Tripathi, P., 2016. Computing symmetrical strength of n-grams: a two pass filtering approach in automatic classification of text documents. SpringerPlus 5.

Ahlers, D., 2013. Assessment of the Accuracy of GeoNames Gazetteer Data., in: Proceedings of the 7th Workshop on Geographic Information Retrieval. ACM, New York, NY, USA, pp. 74–81.

Aji, S., 2012. Document summarization using positive pointwise mutual information. Int. J. Comp. Sci. Inform. Technol. 4, 47–55. https://doi.org/10.5121/ijcsit.2012.4204.

Arsevska, E., Valentin, S., Rabatel, J., de Goër de Hervé, J., Falala, S., Lancelot, R., Roche, M., 2018. Web monitoring of emerging animal infectious diseases integrated in the French animal health epidemic intelligence system. PLoS One 13, e0199960. https://doi.org/10.1371/journal.pone.0199960.

Bahk, C.Y., Scales, D.A., Mekaru, S.R., Brownstein, J.S., Freifeld, C.C., 2015. Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting. BMC Infect. Dis. 15. https://doi.org/10.1186/s12879-015-0885-0.

Barboza, P., Vaillant, L., Mawudeku, A., Nelson, N.P., Hartley, D.M., Madoff, L.C., Linge, J.P., Collier, N., Brownstein, J.S., Yangarber, R., Astagneau, P., on behalf of the Early Alerting, Reporting Project of the Global Health Security Initiative, 2013. Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of a/H5N1 influenza events. PLoS One 8, e57252. https://doi.org/10.1371/journal.pone.0057252.

Bird, S., Loper, E., 2004. NLTK: The natural language toolkit. Proceedings of the ACL Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Barcelona, Spain, pp. 214–217.

Blanchard, J., Guillet, F., Gras, R., Briand, H., 2005. Using Information-Theoretic Measures to Assess Association Rule Interestingness, in: 5th IEEE International Conference on Data Mining ICDM'05. IEEE Computer Society, United States, pp. 66–73 https://doi.org/10.1109/ICDM.2005.149.

Bollig, N., Clarke, L., Elsmo, E., Craven, M., 2020. Machine learning for syndromic surveillance using veterinary necropsy reports. PLoS One 15, e0228105. https://doi.org/10.1371/journal.pone.0228105 publisher: Public Library of Science.

Bouma, G., 2009. Normalized (Pointwise) mutual information in collocation extraction. Proceedings of German Society for Computational Linguistics & Language Technology Conference, pp. 31–40.

Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D., 2008. Surveillance sans Frontieres: internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med. 5, e151.

Carrion, M., Madoff, L.C., 2017. ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. Int. Health 9, 177–183. https://doi.org/10.1093/inthealth/ihx014.

Chanlekha, H., Kawazoe, A., Collier, N., 2010. A framework for enhancing spatial and temporal granularity in report-based health surveillance systems. BMC Med. Inform. Decision Making 10, 1.

Chiu, J.P., Nichols, E., 2016. Named entity recognition with bidirectional LSTM-CNNs. Trans. Assoc. Comput. Linguistics 4, 357–370. https://doi.org/10.1162/tacl_a_00104.

Church, K.W., Hanks, P., 1989. Word association norms, mutual information, and lexicography. Proceedings of the 27th Annual Meeting on Association for Computational Linguistics -, Association for Computational Linguistics, Vancouver, British Columbia, Canada, pp. 76–83 https://doi.org/10.3115/981623.981633.

Claes, F., Kuznetsov, D., Liechti, R., Von Dobschuetz, S., Dinh Truong, B., Gleizes, A., Conversa, D., Colonna, A., Demaio, E., Ramazzotto, S., Larfaoui, F., Pinto, J., Le Mercier, P., Xenarios, I., Dauphin, G., 2014. The EMPRES-i genetic module: a novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses. Database 2014. https://doi.org/10.1093/database/bau008 bau008–bau008.

Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R.A., Takeuchi, K., Kawtrakul, A., 2007. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. Lang. Resour. Eval. 40, 405.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., Taniguchi, K., 2008. BioCaster: detecting public health rumors with a web-based text mining system. Bioinformatics 24, 2940–2941. https://doi.org/10.1093/bioinformatics/btn534.

Dion, M., AbdelMalik, P., Mawudeku, A., 2015. Big data and the global public health intelligence network (GPHIN). Can. Commun. Dis. Rep. 41, 209–214.

Drury, B., Roche, M., 2019. A survey of the applications of text mining for agriculture. Comput. Electron. Agric. 163, 104864. https://doi.org/10.1016/j.compag.2019.104864.

Drury, B., Fernandes, R., Moura, M.F., de Andrade Lopes, A., 2019. A survey of semantic web technology for agriculture. Inform. Proc. Agric. 6, 487–501. https://doi.org/10.1016/j.inpa.2019.02.001.

Du, M., Pivovarova, L., Yangarber, R., 2016. PULS: natural language processing for business intelligence. Proceedings of the 2016 Workshop on Human Language Technology and Intelligent Applications, New York, United States.

Feldman, R., Aumann, Y., Liberzon, Y., Ankori, K., Schler, J., Rosenfeld, B., 2001. A domain independent environment for creating information extraction modules. Proceedings of the Tenth International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA, pp. 586–588 https://doi.org/10.1145/502585.502699.

Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S., 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports. J. Am. Med. Inform. Assoc. 15, 150–157. https://doi.org/10.1197/jamia.M2544.

Ghosh, S., Chakraborty, P., Cohn, E., Brownstein, J.S., Ramakrishnan, N., 2016. Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2vec Approach arXiv:1603.00106 [cs, stat].

Ghosh, S., Chakraborty, P., Lewis, B.L., Majumder, M.S., Cohn, E., Brownstein, J.S., Marathe, M.V., Ramakrishnan, N., 2017. Guided Deep List: Automating the Generation of Epidemiological Line Lists from Open Sources arXiv:1702.06663 [cs].

Grishman, R., Huttunen, S., Yangarber, R., 2002. Information extraction for enhanced access to disease outbreak reports. J. Biomed. Inform. 35, 236–246. https://doi.org/10.1016/S1532-0464(03)00013-3.

Guarino, N., Oberle, D., Staab, S., 2009. What is an ontology? In: Staab, S., Studer, R. (Eds.), Handbook on Ontologies. International Handbooks on Information Systems, Springer, Berlin, Heidelberg, pp. 1–17 https://doi.org/10.1007/978-3-540-92673-3_0

Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Brownstein, J., Thinus, G., Lightfoot, N., 2010. The landscape of international event-based biosurveillance. Emerging Health Threats J. 3. https://doi.org/10.3402/ehtj.v3i0.7096.

Honnibal, M., Montani, I., 2018. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., Ghazi, D., 2017. Location detection and disambiguation from twitter messages. J. Intell. Inf. Syst. 49, 237–253. https://doi.org/10.1007/s10844-017-0458-3.

Kawazoe, A., Jin, L., Shigematsu, M., Barrero, R., Taniguchi, K., Collier, N., 2006. The development of a schema for the annotation of terms in the Biocaster disease detecting/tracking system. KR-MED.

Kawazoe, A., Chanlekha, H., Shigematsu, M., Collier, N., 2008. Structuring an event ontology for disease outbreak detection. BMC Bioinform. 9, S8. https://doi.org/10.1186/1471-2105-9-S3-S8.

Keesing, F., Belden, L.K., Daszak, P., Dobson, A., Harvell, C.D., Holt, R.D., Hudson, P., Jolles, A., Jones, K.E., Mitchell, C.E., Myers, S.S., Bogich, T., Ostfeld, R.S., 2010. Impacts of biodiversity on the emergence and transmission of infectious diseases. Nature 468, 647–652. https://doi.org/10.1038/nature09575 number: 7324 Publisher: Nature Publishing Group.

Kishida, K., 2005. Property of average precision and its generalization: an examination of evaluation indicator for information retrieval experiments. NII Techn. Rep. 2005, 1–19.

Lafferty, J., McCallum, A., Pereira, F.C.N., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proc. 18th International Conf. on Machine Learning, pp. 282–289.

Lallich, S., Teytaud, O., Prudhomme, E., 2007. Association rule interestingness: measure and statistical validation, in: Kacprzyk, J., Guillet, F.J., Hamilton, H.J. (Eds.), Quality Measures in Data Mining. Springer Berlin Heidelberg, Berlin, Heidelberg. volume 43, pp. 251–275. doi:https://doi.org/10.1007/978-3-540-44918-8_11. series Title: Studies in Computational Intelligence.

Lyon, A., Grossel, G., Burgman, M., Nunn, M., 2013. Using internet intelligence to manage biosecurity risks: a case study for aquatic animal health. Divers. Distrib. 19, 640–650.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D., 2014. The stanford CoreNLP natural language processing toolkit. Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60.

Margineantu, D., Wong, W.K., Dash, D., 2010. Machine learning algorithms for event detection: a special issue of machine learning. Mach. Learn. 79, 257–259. https://doi.org/10.1007/s10994-010-5184-9.

Martin, V., Von Dobschuetz, S., Lemenach, A., Rass, N., Schoustra, W., DeSimone, L., 2007. Early warning, database, and information systems for avian influenza surveillance. J. Wildl. Dis. 43, S71.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed Representations of Words and Phrases and their Compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Curran Associates Inc., USA, pp. 3111–3119.

Mooney, R.J., Bunescu, R., 2005. Mining knowledge from text using information extraction. ACM SIGKDD 7, 3–10. https://doi.org/10.1145/1089815.1089817.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S., 2013. Using of Jaccard coefficient for keywords similarity. Hong Kong 5.

Ostfeld, R.S., 2009. Biodiversity loss and the rise of zoonotic pathogens. Clin. Microbiol. Infect. 15, 40–43. https://doi.org/10.1111/j.1469-0691.2008.02691.x.

Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J., Gomes, G., Koppeschaar, C., Rehn, M., Smallenburg, R., Turbelin, C., Van Noort, S., Vespignani, A., 2014. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clin. Microbiol. Infect. 20, 17–21. https://doi.org/10.1111/1469-0691.12477.

Paquet, C., Coulombier, D., Kaiser, R., Ciotti, M., 2006. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Eurosurveillance 11, 5–6. https://doi.org/10.2807/esm.11.12.00665-en.

Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543 https://doi.org/10.3115/v1/D14-1162.

Piskorski, J., Tanev, H., Atkinson, M., van der Goot, E., Zavarella, V., 2011. Online news event extraction for global crisis surveillance. In: Nguyen, N.T. (Ed.), Transactions on Computational Collective Intelligence V. Springer Berlin Heidelberg, Berlin, Heidelberg. 6910, pp. 182–212. https://doi.org/10.1007/978-3-642-24016-4_10.

Piskorski, J., Haneczok, J., Jacquet, G., 2020. New benchmark corpus and models for fine-grained event classification: To BERT or not to BERT? Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 6663–6678 https://doi.org/10.18653/v1/2020.coling-main.584

Preotiuc-Pietro, D., Srijith, P.K., Hepple, M., Cohn, T., 2016. Studying the temporal dynamics of word co-occurrences: an application to event detection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portoroz, Slovenia , pp. 4380–4387 URL. https://www.aclweb.org/anthology/L16-1694.

Rabatel, J., Arsevska, E., Roche, M., 2019. PADI-web corpus: Labeled textual data in animal health domain. Data in Brief 22, 643–646. https://doi.org/10.1016/j.dib.2018.12.063.

Ralf, S., Flavio, F., Erik, V.D.G., Clive, B., Peter, V.E., Roman, Y., 2008. Text Mining from the Web for Medical Intelligence. NATO Science for Peace and Security Series, D: Information and Communication Security. , pp. 295–310 https://doi.org/10.3233/978-1-58603-898-4-295.

Roche, M., Prince, V., 2010. A web-mining approach to disambiguate biomedical acronym expansions. Informatica 34, 12. http://www.informatica.si/index.php/informatica/article/view/296.

Roche, M., Azé, J., Kodratoff, Y., Sebag, M., 2004. Learning interestingness measures in terminology extraction. A ROC-based approach. ROCAI, pp. 81–88.

Role, F., Nadif, M., 2011. Handling the impact of low frequency events on co-occurrence based measures of word similarity - a case study of pointwise mutual information. Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011), pp. 218–223.

Salton, G., Lesk, M.E., 1968. Computer evaluation of indexing and text processing. J. Assoc. Comput. Mach. 15, 8–36.

Smadja, F., Hatzivassiloglou, V., McKeown, K.R., 1996. Translating collocations for bilingual lexicons: a statistical approach. Comput. Linguistics 2 URL. http://aclweb.org/anthology/J/J96/J96-1001.

Song, S., Zhang, N., Huang, H., 2019. Named entity recognition based on conditional random fields. Clust. Comput. 22, 1–12. https://doi.org/10.1007/s10586-017-1146-3.

Soto, G., Araujo-Castillo, R.V., Neyra, J., Fernandez, M., Leturia, C., Mundaca, C.C., Blazes, D.L., 2008. Challenges in the implementation of an electronic surveillance system in a resource-limited setting: Alerta, in Peru. BMC Proceedings, BioMed Central, p. S4.

Steinberger, R., Fuart, F., Goot, E., Best, C., Etter, P., Yangarber, R., 2008. Text mining from the web for medical intelligence. Mining Massive Data Sets for Security. IOS Press URL. https://www.researchgate.net/profile/Erik_Van_der_Goot/publication/252032768_Text_Mining_from_the_Web_for_Medical_Intelligence/links/54e46a9d0cf2dbf6069671a0.pdf.

Strotgen, J., Gertz, M., 2010. HeidelTime: high quality rule-based extraction and normalization of temporal expressions. Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 321–324.

Thanopoulos, A., Fakotakis, N., Kokkinakis, G., 2002. Comparative evaluation of collocation extraction metrics. Proceedings of the Third International Conference on Language Resources and Evaluation LREC, pp. 620–625.

Torkkola, K., 2003. Feature extraction by non parametric mutual information maximization. J. Mach. Learn. Res. 3, 1415–1438 Publisher: JMLR.org.

Torregrossa, F., Allesiardo, R., Claveau, V., Kooli, N., Gravier, G., 2021. A survey on training and evaluation of word embeddings. Int. J. Data Sci. Analytics, 1–19 https://doi.org/10.1007/s41060-021-00242-8.

Trieu, H.L., Tran, T.T., Duong, K.N.A., Nguyen, A., Miwa, M., Ananiadou, S., 2020. DeepEventMine: end-to-end neural nested event extraction from biomedical texts. Bioinformatics 36, 4910–4917. https://doi.org/10.1093/bioinformatics/btaa540.

Valentin, S., 2020. Extraction and Combination of Epidemiological Information from Informal Sources for Animal Infectious Diseases Surveillance. Ph.D. thesis. University of Montpellier, France.

Valentin, S., Arsevska, E., Mercier, A., Falala, S., Rabatel, J., Lancelot, R., Roche, M., 2020a. PADI-web: an event-based surveillance system for detecting, classifying and processing online news. Post-Proceedings of 8th Language & Technology Conference, LTC 2017, November 17-19, 2017. Springer, LNCS, Poznań, Poland.

Valentin, S., Lancelot, R., Roche, M., 2020b. Automated processing of multilingual online news for the monitoring of animal infectious diseases. Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020), European Language Resources Association, Marseille, France , pp. 33–36 URL. https://www.aclweb.org/anthology/2020.multilingualbio-1.6.

WHO (Ed.), 2005. International Health Regulation (2005), 3rd ed. WHO Press, Geneva.

Wilson, K., Brownstein, J.S., 2009. Early detection of disease outbreaks using the internet. Can. Med. Assoc. J. 180, 829–831. https://doi.org/10.1503/cmaj.090215.

Xiang, W., Wang, B., 2019. A survey of event extraction from text. IEEE Access 7, 173111–173137. https://doi.org/10.1109/ACCESS.2019.2956831.

Yang, S., Feng, D., Qiao, L., Kan, Z., Li, D., 2019. Exploring pre-trained language models for event extraction and generation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp. 5284–5294 https://doi.org/10.18653/v1/P19-1522.

Zhu, L., Zheng, H., 2020. Biomedical event extraction with a novel combination strategy based on hybrid deep neural networks. BMC Bioinform. 21, 47. https://doi.org/10.1186/s12859-020-3376-2.