Information retrieval for animal disease surveillance: a pattern-based approach

Sarah Valentin[1,2,3,4], **Renaud Lancelot**[3,4] and **Mathieu Roche**[1,2]
[1]UMR TETIS, CIRAD, Montpellier, France.
[2]TETIS, AgroParisTech, CIRAD, CNRS, INRAE, Univ Montpellier, Montpellier, France.
[3]UMR ASTRE, CIRAD, Montpellier, France.
[4]ASTRE, CIRAD, INRAE, Univ Montpellier, Montpellier, France.
sarah.valentin28@gmail.com

Abstract

Animal diseases-related news articles are rich in information useful for risk assessment. In this paper, we explore a method to automatically retrieve sentence-level epidemiological information. Our method is an incremental approach to create and expand patterns at both lexical and syntactic levels. Expert knowledge input are used at different steps of the approach. Distributed vector representations (word embedding) were used to expand the patterns at the lexical level, thus alleviating manual curation. We showed that expert validation was crucial to improve the precision of automatically generated patterns.

1 Introduction

The ability to rapidly recognise emerging and re-emerging animal infectious diseases is a critical global health priority. Early warning is crucial for the quick implementation of effective control strategies at global and local levels (Heymann and Rodier, 2001). In recent decades, several outbreaks have highlighted the limitations of conventional disease surveillance, which is hampered by delayed detection and latency of the communication channels (Ben Jebara and Shimshony, 2006). The growing availability of digital data represents an unprecedented source of real-time disease information. 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). Their mainstreaming into surveillance systems, via the epidemic intelligence (EI) concept, has been a game-changer for disease surveillance and control. 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, such as online news articles) (Paquet et al., 2006). Event-based surveillance (EBS) systems increasingly marshal text-mining methods to alleviate the amount of manual curation of the continuous flow of free text from informal sources (Hartley et al., 2010).

Animal-health online news articles are rich in different types of epidemiological information. For instance, news articles that report an outbreak often also describe outbreak control measures or economic impacts, share information about the outbreak source or draw attention to a given area at risk. Those elements may be of relevance to EI teams to assess risks associated with the occurrence of an outbreak.

In this paper, our objective is to value new types of epidemiological information contained in news. We describe a pattern-based method to automatically retrieve sentence-level epidemiological information. Empirically, sentence-level seems more homogeneous in terms of the topic than the whole news content. Our method is an incremental approach to create and expand patterns at both lexical and syntactic levels, with expert knowledge input. The learnt patterns can be further integrated into EBS systems.

2 Related work

2.1 Information retrieval from disease-related news

Information retrieval methods implemented in EBS systems broadly encompass three tasks: (1) document classification, (2) named entity extraction and (3) event extraction.

Most classification approaches in EBS systems focus on the binary news relevance, so as to filter out irrelevant ones (Conway et al., 2009; Doan et al., 2007; Torii et al., 2011; Valentin et al., 2020). Other classification frames assign a broad thematic

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 70–78

https://doi.org/10.18653/v1/P17

label to the news, such as "outbreak-related" or "socioeconomic" (Zhang et al., 2009). When a news piece contains several topics, a single-label classifier has to decide on a topic (i.e. a label) among the other ones, which usually decreases the classification performance (Zhang et al., 2009).

Named entities extraction (NER) focuses on the detection of both epidemiological entities (e.g. virus, symptom) and domain unspecific entities (e.g. locations, dates). Event extraction consists in identifying the news narratives, by extracting the set of epidemiological entities describing a specific outbreak-event, also called attributes. The extraction of fine-grained temporal information such as the beginning and end of an event (Chanlekha et al., 2010), or thematic attributes such as the transmission mode (Conway et al., 2010).

None of the available tasks described here above suits the needs of our current work. Document-based classification is not precise enough to detect the variety of information contained in a single piece of news. Named entities and event extraction mainly focus on the spatio-temporal attributes of events, and partly address the potential of other types of epidemiological information. Midway between these two approaches, (Zhang and Liu, 2007) proposed a sentence-based annotation to detect outbreak-related sentences, while recognising that a news piece contains many sentences with different semantic meanings. However, as the primary goal was outbreak detection, outbreak-unrelated sentences (e.g. describing treatment or prevention) were all merged into one negative category. Based to this approach, we precisely aim to automatically retrieve epidemiological information at the sentence-level, broadening the binary outbreak-related/unrelated categories.

2.2 Pattern-based methods

In EBS systems, as well as in the biomedical literature, pattern-based approaches have been mainly used for named entities, event and relation extraction tasks (Wang et al., 2018). For instance, the EBS systems MedISys detects event from health-related news based on the Pattern-based Understanding and Learning System (PULS) (Du et al., 2016). PULS relies on a cascade of patterns applied to each sentence of the news article content to extract the event attributes. Patterns use both syntactical and semantic information of the sentence, such as:

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

This pattern matches a noun phrase (NP) of semantic type, i.e. "disease"; 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 the auxiliary verb (e.g. "has") or adverb (e.g. "so far") (Steinberger et al., 2008).

Patterns can be generated manually or automatically. Manual construction typically relies on domain expert proposals, thereby achieving high precision. Such method is time-consuming and have two major shortcoming regarding the pattern generalisation: (i) the vocabulary is limited to the expert knowledge, and (ii) the syntactic structure may be rigid. Even if expert-based patterns may achieve high precision, the problem of recall, or coverage, is critical. Thus, weakly supervised and unsupervised methods have gained some popularity due to the marked reduction in the amount of manual curation they require (Ghosh et al., 2017; Ibekwe-Sanjuan et al., 2011; Yangarber, 2003). For instance, bootstrapping methods can generate patterns automatically from a pre-classified training corpus or from seed patterns (Jones et al., 1999; Thelen and Riloff, 2002). (Ibekwe-Sanjuan et al., 2011) used the local context of seed patterns, i.e. their surrounding words, to generate variants. This method relied on the assumption that patterns occurring in the same context tend to have the same semantic meaning, i.e. the paradigm of word embedding models. To our knowledge, the pattern-based approach was not evaluated to retrieve other types of epidemiological information in news articles. In line with (Ghosh et al., 2017; Ibekwe-Sanjuan et al., 2011), we propose an incremental approach integrating both word embedding and expert knowledge to expand patterns, in the sentence retrieval context. We propose two types of pattern expansion: (i) lexical and (ii) syntactical.

3 Method

3.1 Corpus

We used a publicly available corpus of news articles, that was annotated by four epidemiologists following specific guidelines (Valentin et al., 2019). The news articles are split in sentences and each

sentence has two levels of annotation (i.e. two labels). The first annotation level aims at identifying the relevant sentences, i.e. the sentences describing a current, hypothetical (at risk) or past outbreak event. The second annotation level, called "Information type", characterises the epidemiological topic (fine-grained information). To evaluate retrieval methods on sufficient class sizes, we increased the annotated corpus (32 news articles, 486 sentences) with 56 additional news articles. We obtained a final corpus containing 1,245 sentences (from 87 news articles). From this initial corpus (1,245 sentences), 161 sentences were labelled as irrelevant. The subset of sentences for Information type classification hence consisted of 1,084 sentences (Table 1).

Table 1: Number of sentences per Information type category.

Category	No. of sentences
Protection and control measures	401
Descriptive epidemiology	310
Concern and risk factors	110
General epidemiology	109
Economic and political consequences	69
Transmission pathway	58
Distribution	27

3.2 Approach

In this section, we aim at identifying sentences from specific classes based on the patterns they contain. We opted for the pattern definition of Du and Yangarber, 2015, i.e. a pattern consists in "a place-holder for specific tokens (terms) and their surrounding context". The surrounding context may be fixed, and the token may be the variable. For instance, "**X** was detected in **Y**", where X is a disease and Y a location.

We evaluated a semi-automated and incremental process. In this approach, an expert is at the core of the process (Figure 1, steps 1 and 3). Indeed, we believe that expert knowledge is particularly suitable for the retrieval of fine-grained classes. Our objective is to use a minimal set of sentences (referred to as seed sentences) to identify patterns specific to the class (seed patterns) and expand the patterns at lexical and syntactical levels. All steps are detailed in the following subsections.

The pattern extracted and expanded after steps 1, 2, 3 and 4 are hereafter referred to as P_S (seed patterns), P_{E1} (first expansion), P_{E2} (second expansion), and P_{E3} (third expansion).

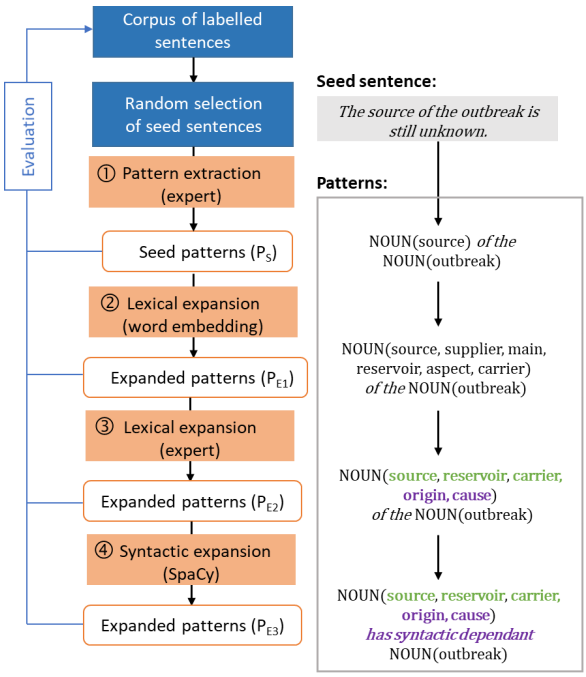
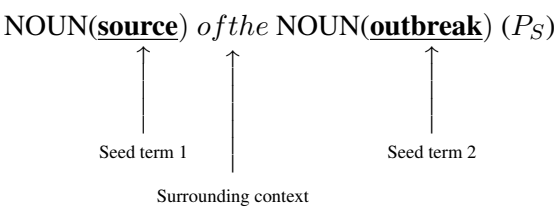


Figure 1: Incremental pattern expansion.
In the pattern box, seed terms are indicated between parentheses, and preceded by their POS. Terms validated by the expert are shown in green, and violet terms correspond to terms added by the expert. For readability, only the expansion of seed term 'source' is represented. The expansion steps (in orange) are detailed hereafter. Evaluation is done after each step.

Manual extraction of patterns (step 1) This first step aims at extracting an initial set of patterns based on a minimal subset of sentences (seed sentences). We relied on the expert to read each seed sentence and identified one or several patterns specific to the sentence class (seed patterns). For instance, based on the seed sentence from Figure 1, the identified pattern is:



Where **source** and **outbreak** are the seed terms with *of* and *the* being linking words. Linking words usually consists of prepositions (e.g. "of", "through"), adverbs or auxiliary verbs. Seed terms

include all terms present in the seed pattern. In our approach, seed terms include nouns, verbs (and their preposition), and adjectives.To match the patterns with sentences, we used the seed term lemmas labelled with their part-of-speech (POS) to avoid possible ambiguities between nouns and verbs. For instance, the seed term NOUN(cause) matches both nouns "cause" and "causes", but does not match conjugated forms of the verb "to cause", such as "causes" or "causes".

Several patterns appeared to involve disease or host as seed term, e.g. in "Investigators look for **swine fever links**". All disease names (including acronyms) and host were thus replaced in the text by the word "disease" and "host" respectively, so that they could be represented by NOUN(disease) and NOUN(host).

Automatic lexical expansion (step 2) At this step, we aim at expanding extracted patterns (P_s) at the lexical level. We focus on the seed terms, by automatically generating closed terms. We used a property of word embedding models, whereby words are represented as vectors. The vector values correspond to a densely distributed representation of the word. The values are learned according to the context in which the word appears, based on the assumption that words that frequently appear in the same context (i.e. surrounded by the same words) tend to have the same meaning (Goldberg, 2017). Words with common contexts have close vectors in the produced vector space. Several pre-trained word embeddings are publicly available, but training a word embedding model on corpus specific to the target domain has been shown to improve performances (Pyysalo et al., 2015). We thus decided to train a word embedding model on a dataset of animal-disease related news articles. We extracted the news from PADI-web database, published from February 20, 2014 to December 14, 2018. PADI-web is an open-source EBS system dedicated to animal health monitoring (Arsevska et al., 2018). We obtained a training set of 35,577 news articles. The training set length was 33,417,501 words, corresponding to a vocabulary of 464,536 single terms. The corpus was tokenized and lemmatized using the NLTK library (Bird and Loper, 2004). We trained the Word2Vec model, developed by Tomas Mikolov in 2013, which is one of the most popular techniques to learn word embedding (Mikolov et al., 2013). We used the continuous bag-of-words algorithm (CBOW) and we set the dimension for the trained vectors to 300, which is the dimension used in various studies (Mikolov et al., 2017, 2013; Pennington et al., 2014). We used the default parameter for the window size (5 words)— while setting the minimum word frequency at 10. In Word2Vec, the metric used to calculate the distance between two vectors is the standard cosine similarity (Leeuwenberg et al., 2016). The closer the cosine similarity between two word vectors is to one, the more similar the words are according to the model. Thus, for each word, cosine similarity can be used to rank terms in decreasing order of similarity.

For each seed term, we retrieved the K closest terms based on word embedding cosine similarity. For instance, the five closest terms for "source" are "supplier", "main", "reservoir", "aspect" and "carrier". The pattern is expanded based on the list of seed term synonyms. Finally, each **seed term** generates a set of K+1 variants (the seed term plus its close terms). Thus, if a pattern contains two seed terms generating K+1 variants each, we obtain $K+1^{K+1}$ combinations. We set K at 15, as a trade-off between the amount of input information and limited manual curation.

Manual lexical expansion (step 3) At the previous step, the K closest terms provided by the word embedding model were considered as seed term synonyms by default. At this step, we use expert knowledge to validate this assumption and enhance the lexical expansion at different steps:

1. Manual validation of the list of variants (K closest terms) generated automatically; terms judged as irrelevant or not specific enough regarding the sentence category are removed.

2. Adding a new term, if not present in the initial list of variants;

3. Merging seed terms (and their corresponding variants), when they are considered as synonyms.

The variants "supplier", "main", and "aspect" were considered as irrelevant by the expert and removed. The expert further added the variants "origin" and "cause" (Figure 1).

Structure expansion (step 4) The final step of manual curation consists of modifying the rigid contextual surrounding to improve the generalisation of patterns. A common approach is to use

wildcards, i.e. symbols representing optional or specific characters and words in pattern matching. For instance, the pattern NOUN(source) of the NOUN(infection) could be replaced by:

NOUN(**source**) (W)? NOUN(**infection**)

The symbol (W)? indicates that NOUN(**source**) and NOUN(**infection**) can be separated by zero or more words. A major shortcoming of using wildcards is that syntactic information is not taken into account: the previous pattern only matches the source (or its variants) followed by the term infection (or its variants) but is not able to detect "the infection's source", for instance.

Thus, we proposed to expand the pattern structure based on the syntactic dependence between terms. More precisely, we modified the pattern structure when two (or more) seed terms were immediately syntactic dependent, i.e. connected by a single arc in the dependency tree (e.g. the subject of a verb, the adjective of a noun, etc.). In the previous example, the new pattern is:

NOUN(**source**) *has immediate syntactic dependant* NOUN(**infection**) *(P_{E3})*

This new pattern is now able to match both "the source of the infection" and "the infection's source". This approach has two advantages regarding pattern generalisation. First, as shown in the previous example, it increases the recall, as it does not rely on a specific sequence of terms such as wildcards. Second, the immediate syntactic relation a more fine-grained understanding of the meaning, thus avoiding irrelevant matches.

The analysis, including pattern-matching and generation of syntactic dependencies, was done using spaCy, i.e. a free open-source library for NLP in Python (Honnibal and Montani, 2017). We chose spaCy because it provides a pattern-matching function that readily allows pattern creation and enrichment. Syntactic relations were based on the spaCy dependency parser.

3.3 Evaluation

The pattern-based approach is particularly relevant for under-represented classes containing highly precise information, which may be hard to identify by supervised methods. In this context, we evaluated the pattern-based approach on two Information type classes, i.e. Concern and risk factors (CRF) and Transmission pathway (TP). As seed sentences, we extracted 15% of an initial set of sentences belonging to the same class. The number of seed sentences was 16 (among 110) and 10 (among 69), respectively.

As shown in Figure 1, we evaluated the pattern performances after each expansion step in terms of recall, precision and F-measure. Here, as we focused on two categories, we evaluated the pattern approach as a binary classification—a sentence is classified as TP (resp. CRF) if it matches at least one TP (resp. CRF) patterns (positive sentence). Otherwise it is considered a negative sentence. We excluded seed sentences from the evaluation to avoid artificial positive matches. The testing dataset hence respectively consisted of 59 positive and 1,025 negative sentences for TP, and 94 positive and 990 negative sentences for CRF categories.

4 Results and discussion

4.1 Pattern-based retrieval

At step 1, the expert extracted 9 and 12 patterns from TP and CRF seed sentences, respectively (Table 2). No identifiable pattern could be found in two sentences (one in each category). In the CRF category, three sentences had the same pattern. After step 3 (manual lexical expansion), the number of terms represented 27% (65/240) and 68% (113/165) of the terms generated automatically in TP and CRF classes, respectively. For the same final number of patterns (7), the number of term variants in the CRF class was twofold higher in the CRF class than in the TP class.

Tables 3 and 4 show the performances of the pattern-based approach for the retrieval of TP and CRF sentences. The first extracted patterns (P_S) retrieved only 7% (4/59) of sentences and 47% (44/94) of CRF sentences.

In both classes, the precision decreased after the automatic lexical expansion (step 2) and increased after the manual expansion (step 3).

Manual and automatic pattern expansion did not impact the TP and CRF recall in similar ways. In the TP class, lexical expansions obtained mitigate improvement in recall, reaching a maximum value of 0.36 after step 3. The syntactic expansion increased the number of retrieved sentences by 179% (28 to 78 sentences), thus obtaining the highest recall of all steps (0.68).

In the CRF class, the automatic lexical expan-

Table 2: Numbers of patterns and terms at the different pattern expansion steps for both Transmission pathway (TP) and Concern and risk factor (CRF) categories. P_S, P_{E1}, P_{E2} and P_{E3} correspond to the seed patterns, and the 1st, 2nd and 3rd expansions, respectively.

	TP				CRF			
	P_S	P_{E1}	P_{E2}	P_{E3}	P_S	P_{E1}	P_{E2}	P_{E3}
No. of patterns	9	9	7	7	12	12	8	7
No. of terms (seeds and variants)	16	240	65	65	11	165	113	113

sion reached a recall of 0.80. Manual lexical and syntactic expansion did not improve the recall but contributed to increasing the precision from 0.33 to 0.53. The syntactic expansion did not impact the number sentences retrieved (+0.7%).

These results were consistent with the characteristics of the patterns extracted from both TP and CRF classes at the first step. In the CRF category, the seed terms mostly consisted of a noun such as threat, risk, or fear (14 out of 16 seed sentences). The recall stability among the lexical and syntactic expansion steps confirmed that CRF sentences were homogeneous regarding their syntactic structure and vocabulary. On the contrary, the extracted TP patterns were more complex. Seed terms mostly consisted of a verbal-linguistic structure such as "could have been brought by". Such syntactic structures were not generalizable, as highlighted by the poor recall at the first step.

4.2 Error analysis

Our results indicated that both lexical and syntactic features were crucial for improving the retrieval quality by the pattern-based approach. Yet the relative importance of each expansion step depended on semantic and syntactic specificity of each category. The manual curation by an expert allowed filtering out of irrelevant terms automatically generated by the word embedding model, thus improving the precision. However, the final precision (after step 4) did not exceed 0.53, indicating that some patterns were ambiguous and not sufficiently class-specific. This constraint was offset by the fact that we aimed to minimize the number of false-negative instances.

To understand what impacted the final recall, we manually evaluated the sentences not detected by the pattern-based approach (false negative sentences).

We found that 13/19 and 11/20 sentences were based on identifiable patterns that were not captured as seed patterns in step 1, in TP and CRF classes, respectively. For instance, four TP sentences referred to ongoing investigations about the outbreak's cause. Indeed, the term "investigators" was present in one of the seed sentences but was not identified by the expert as specific to the TP category.

In the second group of sentences (5/19 and 9/20 sentences), no specific patterns could be identified in, for instance:
"The minister said in one case a man bought meat from Ukraine and gave the water he washed the meat in to his pigs, which got sick and died."

This type of sentence underlines the limitations of the pattern-based approach. The classification decision is based only on matching with predefined terms (pattern seed terms), which are sometimes not sufficient to capture the whole semantics of the sentences. We hypothesise that supervised classification that takes all the sentence terms into account may perform better in such cases.

Eventually, one last TP sentence was not detected because it contained the pronoun "it" in reference to the disease, which did not match the list of expanded terms. This classic NLP problem is known as a noun phrase coreference resolution (Ng and Cardie, 2002). It highlights one limitation of the sentence-based approach in which references to entities are not inferred from other sentences. Further work could be focused on coreference resolution at the corpus level.

As a lexical expansion, we relied on terms automatically generated by the word embedding models. While a lot of retrieved terms were highly relevant, i.e. semantically close to the seed term, a substantial number of them were irrelevant, as shown by the drop in the number of variants after expert validation (step 3). A well-known alternative to retrieve term variants is the use of an external lex-

Table 3: Performances of the patterns for TP sentence retrieval in terms of precision, recall, and F-measure. The variation in the percentage corresponds to the change from the previous step.

	P_S	P_{E1}	P_{E2}	P_{E3}
Sentences retrieved (nb)	7	32 (+357%)	28 (-12,5%)	78 (+179%)
Precision	0.57 (4/7)	0.25 (10/32)	0.75 (21/28)	0.50 (39/78)
Recall	0.07 (4/59)	0.17 (10/59)	0.36 (21/59)	0.68 (40/59)
F-measure	0.12	0.20	0.49	0.58

Table 4: Performances of the patterns for CRF sentence retrieval in terms of precision, recall, and F-measure. The variation in the percentage corresponds to the change from the previous step.

	P_S	P_{E1}	P_{E2}	P_{E3}
Sentences retrieved (nb)	70	227 (+224%)	138 (-39%)	139 (+0.7%)
Precision	0.63 (44/70)	0.33 (75/227)	0.52 (72/138)	0.53 (74/139)
Recall	0.47 (44/94)	0.80 (75/94)	0.77 (72/94)	0.79 (74/94)
F-measure	0.53	0.47	0.62	0.63

ical database such as case WordNet (Luhn, 1958). However, as highlighted by Ibekwe-Sanjuan et al., 2011, WordNet is not domain-specific and may fail to provide appropriate words. For instance, the seed term "source" has "beginning", "root" and "informant" among its synonyms in WordNet. Word embedding models are also prone to generate irrelevant terms, i.e. not only retrieving synonyms but also antonyms, derived forms, hyper and hyponyms (Nooralahzadeh et al., 2018). Besides, there is no consensus regarding the number of terms to retrieve (K). Ghosh et al., 2017 used the top-5 terms to expand medical expression patterns. Relying on the same K for all seed terms unavoidably overlooks some relevant terms, or retrieves irrelevant ones, as the number of variants varies between terms. An alternative could be to set a minimum threshold for the cosine similarity value (Rekabsaz et al., 2017). But determining whether the similarity score obtained from word embedding is indicative of term synonymy is still an open question. Leeuwenberg et al., 2016 showed that cosine similarity alone is a bad indicator to determine if two words are synonymous. They proposed a new measure, i.e. relative cosine similarity, which calculates similarity relative to other cosine-similar words in the corpus.

5 Conclusion

In this paper, we presented an incremental approach to generate patterns for sentence-based retrieval, in the context of animal disease surveillance. The role of the expert was crucial to pinpoint the irrelevant terms generated by the word embedding model. Besides, the time cost of manual curation was minimal, as it mostly consisted of validating or adding terms. This time depends directly on the threshold chosen for top K retrieved terms. The use of syntactic dependency was easy to implement, thus alleviating the cost and bias of wildcard fine-tuning.

Contrary to the supervised approach, the pattern-based method is not hampered by so-called "black box" problems and can be easily enhanced by expert knowledge. Besides, even though it was not evaluated in this study, it is suitable for multi-label sentences as a sentence can match patterns from several classes. In future work, we aim at evaluating the pattern-based method on a new and larger corpus of news. Besides, it would be interesting to compare the pattern-based results with supervised sentence-based classifiers. We also think that the advantages of each method could be synergistic. A promising perspective could thus be to evaluate how to jointly take full advantage the strength of both methods. Cui et al., 2019 proposed an interesting approach to combine both supervised learning and manually built patterns and rules. They applied heuristic-based regular expression when the prediction confidence of the supervised classifier confidence was less than a specific threshold. Such

an approach may help enhance the performance of Information retrieval in the event-based surveillance context.

6 Acknowledgments

This work was 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 supported by the French National Research Agency (ANR) under the Investments for the Future Program (ANR-16-CONV-0004).

References

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):e0199960.

M. Karim Ben Jebara and Arnon Shimshony. 2006. International monitoring and surveillance of animal diseases using official and unofficial sources. *Veterinaria Italiana*, 42(4):431–441.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Hutchatai Chanlekha, Ai Kawazoe, and Nigel Collier. 2010. A framework for enhancing spatial and temporal granularity in report-based health surveillance systems. *BMC medical informatics and decision making*, 10(1):1.

Mike Conway, Son Doan, Ai Kawazoe, and Nigel Collier. 2009. Classifying Disease Outbreak Reports Using N-grams and Semantic. *International Journal of Medical Informatics*, 78(12).

Mike Conway, Ai Kawazoe, Hutchatai Chanlekha, and Nigel Collier. 2010. Developing a Disease Outbreak Event Corpus. *Journal of Medical Internet Research*, 12(3):e43.

Menglin Cui, Ruibin Bai, Zheng Lu, Xiang Li, Uwe Aickelin, and Peiming Ge. 2019. Regular Expression Based Medical Text Classification Using Constructive Heuristic Approach. *IEEE Access*, 7:147892–147904.

Son Doan, Ai Kawazoe, and Nigel Collier. 2007. The Role of Roles in Classifying Annotated Biomedical Text. In *Biological, translational, and clinical language processing*, pages 17–24, Prague, Czech Republic. Association for Computational Linguistics.

Mian Du, Lidia Pivovarova, and Roman Yangarber. 2016. PULS: natural language processing for business intelligence. In *Proceedings of the 2016 Workshop on Human Language Technology and Intelligent Applications*, New York, United States.

Mian Du and Roman Yangarber. 2015. Acquisition of domain-specific patterns for single document summarization and information extraction. In *Proceedings of the The Second International Conference on Artificial Intelligence and Pattern Recognition*, Shenzhen, China.

Saurav Ghosh, Prithwish Chakraborty, Bryan L. Lewis, Maimuna S. Majumder, Emily Cohn, John S. Brownstein, Madhav V. Marathe, and Naren Ramakrishnan. 2017. Guided Deep List: Automating the Generation of Epidemiological Line Lists from Open Sources. *arXiv:1702.06663*.

Yoav Goldberg. 2017. Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

David Hartley, Noele Nelson, Ronald Walters, Ray Arthur, Roman Yangarber, Larry Madoff, Jens Linge, Abla Mawudeku, Nigel Collier, John Brownstein, Germain Thinus, and Nigel Lightfoot. 2010. The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, 3(0).

David L. Heymann and Guénaël R. Rodier. 2001. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. *Lancet Infectious Diseases*, 1(5):345–353.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. In *To appear*.

Fidelia Ibekwe-Sanjuan, Fernandez Silvia, Sanjuan Eric, and Charton Eric. 2011. Annotation of Scientific Summaries for Information Retrieval. *arXiv:1110.5722*.

Rosie Jones, Andrew Mccallum, Kamal Nigam, and Ellen Riloff. 1999. Bootstrapping for Text Learning Tasks. In *In IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, pages 52–63.

Artuur Leeuwenberg, Mihaela Vela, Jon Dehdari, and Josef van Genabith. 2016. A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. *The Prague Bulletin of Mathematical Linguistics*, 105(1):111–142.

Hans Peter Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv:1712.09405*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Vincent Ng and Claire Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Evaluation of domain-specific word embeddings using knowledge resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christophe Paquet, Daniel Coulombier, Reinhard Kaiser, and Marco Ciotti. 2006. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Eurosurveillance*, 11(12):5–6.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou. 2015. Overview of the Cancer Genetics and Pathway Curation tasks of BioNLP Shared Task 2013. *BMC Bioinformatics*, 16(10):S2.

Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. 2017. Exploration of a Threshold for Similarity Based on Uncertainty in Word Embedding. In *Advances in Information Retrieval*, volume 10193, pages 396–409. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Giselle Soto, Roger V. Araujo-Castillo, Joan Neyra, Miguel Fernandez, Carlos Leturia, Carmen C. Mundaca, and David L. Blazes. 2008. Challenges in the implementation of an electronic surveillance system in a resource-limited setting: Alerta, in Peru. In *BMC proceedings*, volume 2, page S4. BioMed Central.

Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter, and Roman Yangarber. 2008. Text Mining from the Web for Medical Intelligence. In *Mining Massive Data Sets for Security*. IOS Press.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, volume 10, pages 214–221. Association for Computational Linguistics.

Manabu Torii, Lanlan Yin, Thang Nguyen, Chand T. Mazumdar, Hongfang Liu, David M. Hartley, and Noele P. Nelson. 2011. An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. *International Journal of Medical Informatics*, 80(1):56–66.

Sarah Valentin, Elena Arsevska, Sylvain Falala, Jocelyn de Goër, Renaud Lancelot, Alizé Mercier, Julien Rabatel, and Mathieu Roche. 2020. PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169:105163.

Sarah Valentin, Valérie De Waele, Aline Vilain, Elena Arsevska, Renaud Lancelot, and Mathieu Roche. 2019. Annotation of epidemiological information in animal disease-related news articles: guidelines and manually labelled corpus. *Dataverse Cirad*. Type: dataset.

Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, and Jiawei Han. 2018. Open Information Extraction with Meta-pattern Discovery in Biomedical Literature. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 291–300, Washington DC USA. ACM.

K. Wilson and J. S. Brownstein. 2009. Early detection of disease outbreaks using the Internet. *Canadian Medical Association Journal*, 180(8):829–831.

Roman Yangarber. 2003. Counter-training in discovery of semantic patterns. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 343–350. Association for Computational Linguistics.

Yi Zhang and Bing Liu. 2007. Semantic text classification of emergent disease reports. In *Proceedings of the 11th European Conference on Principles and Pratice of Knockledge Discovery in Databases (PKDD)*, Warsaw, Poland. Springer.

Yulei Zhang, Yan Dang, Hsinchun Chen, Mark Thurmond, and Cathy Larson. 2009. Automatic online news monitoring and classification for syndromic surveillance. *Decision Support Systems*, 47(4):508–517.