Any Language Early Detection of Epidemic Diseases from Web News Streams

Romain Brixtel, Gaël Lejeune, Antoine Doucet, Nadine Lucas
Normandy University – Unicaen
GREYC, CNRS UMR 6072
F-14032 Caen Cedex
France
Email: first.last@unicaen.fr

***Abstract*—In this paper, we introduce a multilingual epidemiological news surveillance system. Its main contribution is its ability to extract epidemic events in any language, hence succeeding where state-of-the-art in surveillance systems usually fails : the objective of reactivity. Most systems indeed focus on a selected list of languages, deemed *important*. However, evidence shows that events are first described in the local language, and translated to other languages later, if and only if they contained important information. Hence, while systems handling only a sample of human languages may indeed succeed at extracting epidemic events, they will only do so after someone else detected the importance of the news, and made the decision to translate it. Thus, with events first described in other languages, such automated systems, that may only detect events that were already detected by humans, are essentially irrelevant for *early* detection.**

To overcome this weakness of the state-of-the-art in terms of reactivity, we designed a system that can detect epidemiological events in any language, without requiring any translation, be it automated or human-written. The solution presented in this paper relies on properties that may be called language universals. First, we observe and exploit properties of the news genre that remain unchanged, whatever the writing language. Second, we handle language variations, such as declensions, by processing text at the character-level, rather than at the word level. This additionally allows to handle various writing systems in a similar fashion.

We present experiments with 5 languages, steoreotypical of different language families and writing systems : English, Chinese, Greek, Polish and Russian. Our system, $DAnIEL$, achieves an average F-measure score around 85%, slightly below top-performing systems for the languages that such systems are able to handle. However, its performance is superior for morphologically-rich languages. And it performs of course infinitely better for the languages that other systems are not able to handle : The richest system in the state-of-the-art handles around 10 languages, while there exists about 6,000 languages in the world, 300 of which are spoken by more than one million people. The $DAnIEL$ system is able to process each of them.

I. Introduction

The Web provides many news sources in an increasing variety of languages. Information Extraction (IE) aims at extracting structured views from free text and particularly from news wires. News feeds provide instant information collected from a large number of sources. The European Media Monitor, for instance, daily collects about 40,000 news reports written in 43 languages[1]. A health authority will want to monitor information with emphasis on disease outbreaks [1].

This paper focuses on epidemiological Event Extraction from the Web, a subdomain of IE whose goal is to detect and extract disease outburst and spreading events from health-related news to send alerts to health authorities [2]. More precisely, multilingual IE with light resources is tested in order to quickly detect news with a particular style denoting concern over some disease. Tapping a wealth of information sources makes it theoretically possible to quickly detect important epidemic events over the world [3]. Until now, several approaches have been reported for epidemic surveillance on the Web [4], such as full human analysis [5], automatic keyword analysis [6] and web mining [7]. Human analysis is expected to be more precise although at a great cost, while keyword analysis is deemed cheaper but lacking precision.

To perform global epidemic surveillance, researchers are facing a challenging problem: the need to build efficient systems for multiple languages at a reasonable cost. The classic IE architecture is built for a given language first, with components for each linguistic layer operating at sentence level (morphology, syntax, semantics). It has proved its high efficiency for applications in a number of important languages [4], [8], but unfortunately most of the components involved in classical IE chains need to be rebuilt for each new language [9]. At a time when a greater variety of languages is observed on the Web, this coverage problem remains unsolved.

The approach advocated here is sharply different from the classical view of language as primarily defined by vocabulary. It uses discourse features that are common in news worldwide. It is designed to be as media-dependent as possible and as language-independent as possible. It relies on established text-genre properties to perform analysis of news discourse taking advantage of collective style, more specifically of repetition patterns at certain positions in text [10]. Though the rationale is different, the method is technically similar to relation discovery in open information extraction on the Web [11]. First, it also uses light crawled resources. Second, its algorithmic basis allows quick processing of large collections of documents.

The paper is organised as follows. In Section II, an overview of the multilingual approaches in IE is provided along with proposals to overcome shortcomings in early detection of diseases. In Section III, we introduce our system called D*ata* An*alysis for* I*nformation* E*xtraction in any* L*anguage* (DAnIEL), a genre-based IE system designed to manage

[1] http://emm.newsbrief.eu/overview.html

multilingual news. Section IV introduces the evaluation corpus that we collected for the experiments. In Section V the results are shown and elaborated upon. Finally, the efficiency of such a light approach for filtering huge multilingual news feeds is discussed and future directions are sketched in Section VI.

II. State of the Art

The use of the generic IE chain [12] as a model requires numerous diverse components for each language. Components corresponding to a new language must be gathered or constructed. Two systems that rely primarily on English, PULS[2] [6] and BIOCASTER[3] [1], [13], are well-known examples of classic IE systems with good results in English. A major disadvantage arises, however, for the end-user wishing to process a genuine multilingual corpus such as a news feed. For most languages, the necessary efficient components will lack [14]. In recent years, machine learning was successfully used to fill gaps in situations when one can find sufficient training data in a language with enough properties common to the new one [11].

However, in epidemic surveillance, there is a need to cover very scarce resource languages or even dialects without training data. In a multilingual setting, state-of-the-art systems are indeed limited by the cumulative process of their language-by-language approach. The detection and appropriate analysis of the very first news relating to an epidemic event is crucial, but it may occur in any language: usually the first language of description is that of the (remote) place where the event was located.

This is why a new hypothesis from recent studies on media rhetorical devices [10] was put to trial: that alarming news show a specific and unusual pattern of repetition. Interesting findings have be heralded in the past, concerning the distribution of proper names in breaking news [15]. The contrast with "ordinary news" has also been used to extract outburst events [13]. The underlying idea is called either pragmatics, or is altogether implicit when no specific knowledge backs the findings. As, in our system, explicit knowledge is used, it relies on specific style properties of news discourse.

III. Our system: DAnIEL

The DAnIEL system presents a full implementation of a discourse-level IE approach. It operates at the discourse-level, because it exploits the global structure of news in a newswire, that is, it exploits information ordering as defined by Lucas [10], as opposed to the usual analysis of linguistic layers at the sentence-level (morphology, syntax and semantics). Entries in the system are text news, with their title and text-body. We will hereby remind the readers of the main features of the DAnIEL system which is extensively motivated in previous works [16], [17]. Character-based refers to the fact that text is handled as sequences of characters rather than as sequences of words, in order to consider all types of languages, including those where the definition and delimitation of words is difficult. The descriptors that we use are not key words but strings of text, exploited if and only if they are repeated in pre-defined specific locations of text. Special interest has been put on describing the overall system as well as evaluating each of its parts. The aim of the process is to extract epidemic events from news feeds, and express them in the reduced form of disease-location pairs (i.e., what disease occurs where).

Figure 1 describes the mains steps of the decision process to detect if a document describes an epidemic event. The DAnIEL processing pipeline is composed of three steps: news article segmentation (Section III-A), event detection (Section III-C1) and event localization (Section III-C2) using a small knowledge base (Section III-B3) and substring patterns called *motifs* (Section III-B).

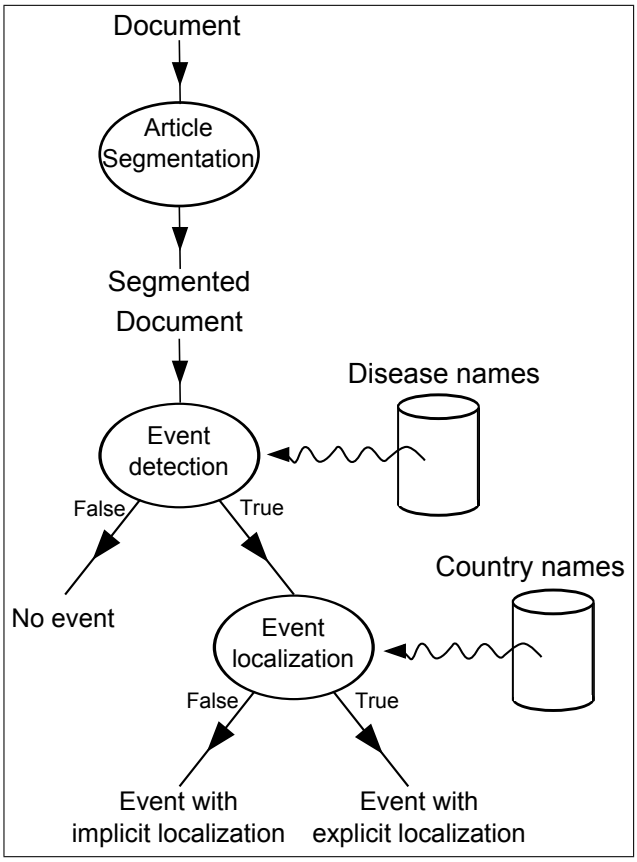


Figure 1. Overview of the $DAnIEL$ process.

A. Article segmentation

The main algorithm relies on the type of article being processed. Because our approach is style-driven, a clear understanding of text construction is crucial. The key positions are the beginning and the end of a text. For analysing press articles, the system relies on the title and beginning (a.k.a., the topical head) and checks which elements are repeated at key positions in the text. To adapt the system to the variable lengths of text, we defined the areas where to seek repeated text strings in an article, depending on its length: short, medium or long. The corresponding rules are fully described in Table I. Repetition may be looked for in : *Head* (title and first paragraph), *Tail* (last two paragraphs) and *Body* (full article except the *Head*).

For medium and long articles, the system extracts the substrings repeated in Head plus Body and Head plus Tail. For

[2]http://medusa.jrc.it/medisys/helsinkiedition/en/home.html
[3]http://born.nii.ac.jp/

Article type (example)	#paragraphs	Segments
Short (dispatches, breaking news)	3 and less	All paragraphs
Medium (updates, event evolution)	4 to 10	Head and Body
Long (analysis, less current events)	more than 10	Head and Tail

Table I. ARTICLE SEGMENTATION WITH RESPECT TO THEIR NUMBER OF PARAGRAPHS

short articles, repeated substrings are considered irrespective of their position.

B. Extraction of motifs

To find text string repetitions in the aforementioned article segments, character level analysis is performed by computing non-gapped character strings as described by Ukkonen [18]. Usually exploited in bioinformatics, where gigabytes of data are processed, this algorithm allows fast access to relevant patterns. This section formally defines motif extraction from text, before demonstrating it with a sample document from our evaluation corpus.

1) Definition of motifs: Motifs are substrings patterns of text with the following characteristics :

- they are repeated: motifs occur twice or more;
- they are maximal: motifs cannot be expanded to the left (*left maximality*) nor to the right (*right maximality*) without lowering the frequency.

For example, the motifs found in the string `HATTIVATTIAA` are `T`, `A` and `ATTI`. `TT` is not a maximal pattern because it always occurs inside an occurrence of `ATTI`. In other words, its right-context is always `I` and its left-context `A`. All the motifs in a set of strings can be efficiently enumerated using an augmented suffix array [19].

Given two strings $\mathcal{S}_0 =$ `HATTIV` and $\mathcal{S}_1 =$ `ATTIAA`, Table II shows the augmented suffix array of $\mathcal{S} = \mathcal{S}_0.\$_1.\mathcal{S}_1.\$_0$, where $\$_0$ and $\$_1$ are lexicographically lower than any character in the alphabet Σ and $\$_0 < \$_1$.

The augmented suffix array consists in the list of suffixes sorted lexicographically of $\mathcal{S}$ (SA), together with the Longest Common Prefix (LCP) between each two suffixes in SA ($LCP_i = lcp(\mathcal{S}[SA_i]...\mathcal{S}[n-1], \mathcal{S}[SA_{i+1}]...\mathcal{S}[n-1])$ and $LCP_{n-1} = 0$, n the size of $\mathcal{S}$).

i	LCP_i	SA_i	$\mathcal{S}[SA_i]...\mathcal{S}[n]$
0	0	13	$\$_0$
1	0	6	$\$_1$ATTIAA$\$_0$
2	1	12	A$\$_0$
3	1	11	AA$\$_0$
4	4	7	ATTIAA$\$_0$
5	0	1	ATTIV$\$_1$ATTIAA$\$_0$
6	0	0	HATTIV$\$_1$ATTIAA$\$_0$
7	1	10	IAA$\$_0$
8	0	4	IV$\$_1$ATTIAA$\$_0$
9	2	9	TIAA$\$_0$
10	1	3	TIV$\$_1$ATTIAA$\$_0$
11	3	8	TTIAA$\$_0$
12	0	2	TTIV$\$_1$ATTIAA$\$_0$
13	0	5	V$\$_1$ATTIAA$\$_0$

Table II. AUGMENTED SUFFIX ARRAY OF $\mathcal{S} =$ HATTIV$\$_1$ATTIAA$\$_0$

The LCP allows for the detection of repetitions. The substring `ATTI` occurs for example in $\mathcal{S}$ at the offsets $(1, 13)$, according to LCP_4 in Table II. The process enumerates all the repeated substrings by reading through LCP:

- if $LCP_i < LCP_{i+1}$: *open* a potential motif occurring at the offset SA_{i+1};
- if $LCP_i > LCP_{i+1}$: *close* motifs previously created;
- if $LCP_i = LCP_{i+1}$: *valid* motifs with the offset SA_{i+1} where it occurs in $\mathcal{S}$.

The maximal criterion is checked when a motif is closed during the enumeration process. Two different potential motifs are equivalent if the last character of these motifs occurs at the same positions. For example, `TTI` is equivalent to `ATTI` because the last characters of these two motifs occur at the offsets $(4, 10)$ (these two substrings are said to be in a relation of *occurrence-equivalence* [18]). In that case, `ATTI` is kept as a *maximal* motif, because it is the longest of its equivalents. The others motifs `A` and `T` are maximal because their contexts differ in different occurrences. All repetitions across different strings are detected at the end of the enumeration by mapping the offsets in $\mathcal{S}$ with those in $\mathcal{S}_0$ and $\mathcal{S}_1$. This way, any repetition detected in $\mathcal{S}$ can be located in any of the strings $\mathcal{S}_i$. SA and LCP are constructed in time-complexity $O(n)$ [19], while the enumeration process is done in $O(k)$, with k defined as the number of motifs and $k < n$ [18][4]. The extraction of motifs is hence very efficient.

2) Examples of motifs: An example from a document in Polish is given in Figure 2 to highlight the value of the process described hereabove. This document deals with a case of dengue spreading in Thailand. We will focus on two sentence extracted from this document, $\mathcal{S}_0$ and $\mathcal{S}_1$:

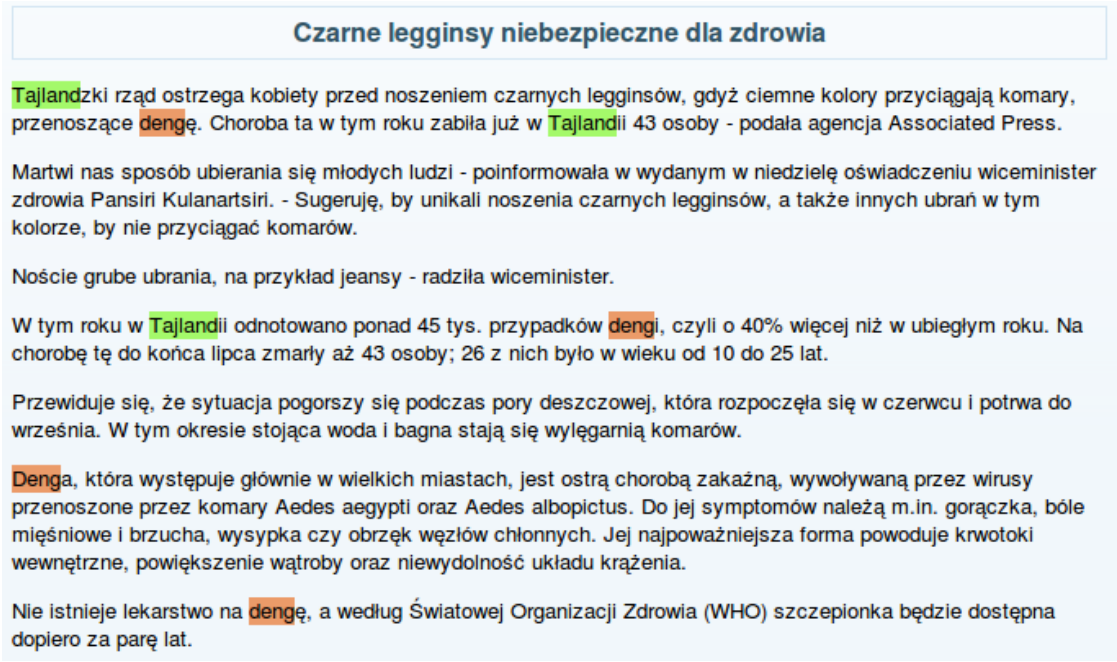
$\mathcal{S}_0$: Tajlandzki rząd ostrzega kobiety przed noszeniem czarnych legginsów, gdyż ciemne kolory przyciągają komary, przenoszące dengę.
[Thai government warns women against wearing black leggings, because dark colors attract mosquitoes carrying dengue.]

$\mathcal{S}_1$: W tym roku w Tajlandii odnotowano ponad 45 tys. przypadków dengi, czyli o 40% więcej niż w ubiegłym roku.
[This year, in Thailand, there were more than 45,000 cases of dengue fever, up 40% from last year.]

A word-based repetition detection would fail to find similarities between *dengę* and *dengi*, as well as between *Tajlandzki* and *Tajlandii*. The motif detection focuses on the detection of subpatterns of diseases names, here on the detection of the roots : *deng∼* and *Tajland∼*. Table III shows a selected sample of the augmented suffix array of the two text fragments $\mathcal{S}_0$ and $\mathcal{S}_1$.

For instance, a repetition of length 4 (LCP_{71}) is detected at the offsets (120, 186): `deng`. Another repetition, `Tajland`, is detected at the offsets (0, 140). The maximal criterion consists in verifying that those substrings are strictly included in another at each offset where they occur. From the sentences $\mathcal{S}_0$

[4]The code for computing these motifs in a set of strings is provided in PYTHON at http://code.google.com/p/py-rstr-max/



Czarne legginsy niebezpieczne dla zdrowia

Tajlandzki rząd ostrzega kobiety przed noszeniem czarnych legginsów, gdyż ciemne kolory przyciągają komary, przenoszące dengę. Choroba ta w tym roku zabiła już w Tajlandii 43 osoby - podała agencja Associated Press.

Martwi nas sposób ubierania się młodych ludzi - poinformowała w wydanym w niedzielę oświadczeniu wiceminister zdrowia Pansiri Kulanartsiri. - Sugeruję, by unikali noszenia czarnych legginsów, a także innych ubrań w tym kolorze, by nie przyciągać komarów.

Noście grube ubrania, na przykład jeansy - radziła wiceminister.

W tym roku w Tajlandii odnotowano ponad 45 tys. przypadków dengi, czyli o 40% więcej niż w ubiegłym roku. Na chorobę tę do końca lipca zmarły aż 43 osoby; 26 z nich było w wieku od 10 do 25 lat.

Przewiduje się, że sytuacja pogorszy się podczas pory deszczowej, która rozpoczęła się w czerwcu i potrwa do września. W tym okresie stojąca woda i bagna stają się wylęgarnią komarów.

Denga, która występuje głównie w wielkich miastach, jest ostrą chorobą zakaźną, wywoływaną przez wirusy przenoszone przez komary Aedes aegypti oraz Aedes albopictus. Do jej symptomów należą m.in. gorączka, bóle mięśniowe i brzucha, wysypka czy obrzęk węzłów chłonnych. Jej najpoważniejsza forma powoduje krwotoki wewnętrzne, powiększenie wątroby oraz niewydolność układu krążenia.

Nie istnieje lekarstwo na dengę, a według Światowej Organizacji Zdrowia (WHO) szczepionka będzie dostępna dopiero za parę lat.

Figure 2. Example of a relevant document in Polish: repetition of disease name and explicit location.

i	LCP_i	SA_i	$\mathcal{S}[SA_i]...\mathcal{S}[n]$
...	...	...	...
7	1	192	_czyli_o_40%_wię[...]$$_0$
8	5	185	_dengi,_czyli_o_[...]$$_0$
9	1	119	_dengę.$$_1$1W_tym_roku_w_Ta[...]$$_0$
10	1	68	_gdyż_ciemne_kolory[...]$$_1$1W_tym_roku_w_Ta[...]$$_0$
...	...	...	...
44	0	168	5_tys._przypadk[...]$$_0$
45	7	140	Tajlandii_odnot[...]$$_0$
46	0	0	Tajlandzki_rząd[...]$$_1$W_tym_roku_w_Ta[...]$$_0$
47	0	127	W_tym_roku_w_Ta[...]$$_0$
...	...	...	...
70	1	14	d_ostrzega_kobi[...]$$_1$W_tym_roku_w_Ta[...]$$_0$
71	4	186	dengi,_czyli_o[...]$$_0$
72	1	120	dengę.$$_1$W_tym_roku_w_Ta[...]$$_0$
73	1	146	dii_odnotowano_[...]$$_0$
...	...	...	...

Table III. SAMPLE OF THE AUGMENTED SUFFIX ARRAY OF $\mathcal{S} = \mathcal{S}_0\$_1\mathcal{S}_1\$_0$. WHITE SPACES ARE REPLACED BY THE SYMBOL "_".

and $\mathcal{S}_1$, the longest motifs are : `Tajland`, `ym_roku`, `_przy`, `_deng`, `y_prz`, `_prze` and `_prz`, where "_" represents a white space. `_deng` is actually extracted rather than `deng` because the left context of `deng` is always a white space.

3) Construction of the knowledge base: DAnIEL uses implicit knowledge on the news genre. Only a few rules are used here. First, the system relies on the rule that information is located in important places, called positions. Second, in journalistic style, writers use the most common disease names, known by most of their readers. Last, important information is repeated. Similar observations were stated in different studies based on pragmatics or statistical studies (estimation of positive adaptation), notably on proper names in the work of Church [15].

DAnIEL uses only light lexical resources automatically collected from Wikipedia with light human moderation to pinpoint information that can be used to fill databases. The extracted lexicon contains common disease names and geographical locations (countries). The lexicon needed with such genre-based system is quite small: roughly hundreds of items *versus* tens of thousands in state-of-the-art systems based on linguistic knowledge [20]. Indeed, Web-extracted disease names allow to deal quickly with new languages, even without the assistance of a native speaker.

C. Use of the knowledge base

In practice the lexicons of disease names and locations is used in a very direct way. Observe, that an interesting text string is defined by 3 repetitions: two in the document (in adequate positions), and one in the lexicon. Hence, motif extraction is performed on articles combined with the external knowledge. For example, let $\mathcal{S}_2$ and $\mathcal{S}_3$ be two strings to be analysed according to $\mathcal{S}_0$ and $\mathcal{S}_1$:

$\mathcal{S}_2$: Tajlanda *[Thailand]*

$\mathcal{S}_3$: denga *[denge]*

With $\mathcal{S}_0$, $\mathcal{S}_1$ (segments of a document) and $\mathcal{S}_2$, $\mathcal{S}_3$ (external knowledge base), the augmented suffix array allows to detect repetition between selected parts of a document and any resources a system might need. Table IV shows a sample of this augmented suffix array.

It is interesting to note that the addition of the lexicons allows for sharper extraction. For example, a new detected motif is `deng`, when with the document alone, the extracted motif was instead `_deng`. Indeed, by processing the string "$\mathcal{S}_0\$_3\mathcal{S}_1\$_2$`denge`$\$_1$`Tajlandia`$\$_0$", the left context of the substring `deng` is no longer systematically the character "_" but also, for one of those occurrences, "$\$_2$". So, `deng` is now a maximal motif occurring twice in the selected parts of the document and once in the disease name lexicon (as "*denge*").

1) Event detection: DAnIEL filters out motifs according to article segmentation rules as described in Table I, and to the list of disease names as explained in Section III-B3. It keeps motifs that are substrings found in two different sub-units, typically Head and Tail, and matching with at least one disease name. This comes from the genre-related rules saying 1) that an important topic in news should be highlighted, 2) that common names should be used to catch the reader's attention and 3) that the topic should be repeated.

More formally, let $\mathcal{S}_0$ and $\mathcal{S}_1$ be the Head and the Tail of a long article and $\mathcal{S}_2$ ... $\mathcal{S}_{n+1}$ the n entries in a diseases knowledge base. The process enumerates repetitions on $\mathcal{S}_0$ ... $\mathcal{S}_{n+1}$ (section III-B) and keeps motifs that occurs in $\mathcal{S}_0$, $\mathcal{S}_1$ and any of the $\mathcal{S}_i$, for $1 < i \leq n+1$. A heuristic ratio is used to check if a motif matches an entry: for a motif m occurring in key positions and in an entry $\mathcal{S}_i$ in the list of diseases: $\frac{len(m)}{len(\mathcal{S}_i)} \geq \theta$, with $len(m)$ and $len(\mathcal{S}_i)$ the number of characters in m and $\mathcal{S}_i$. In the previous example, the process tests whether $\frac{len(\text{deng})}{len(\text{denga})} = \frac{4}{5} \geq \theta$. The choice of the value of θ is discussed in Section V-C3. This technique proves especially useful for morphologically rich languages, as it bypasses the need for a morphological analyzer. If DAnIEL finds no motif that matches its knowledge base using the θ threshold, it assumes that the document contains no event and is therefore irrelevant.

2) Event localization: An event is minimally defined as a relation between a disease name and a location. Once again, journalists' fairly strict writing principles help DAnIEL localize events without sentence-level extraction patterns. When talking about an epidemic, location of the event can be an important topic of the news. The locations are found in the same way as disease names, using repetitions and with the help of a list of countries and capitals extracted from Wikipedia.

When a journalist does not mention explicitly any location in the document, it means that this information relates to the issuing place. Hence, when no location is found using repetition rules and the list of geographical names, the location of the event is assumed to be the country of issue of the source (i.e., that of the newspaper or the news agency).

IV. Corpus

No shared corpora is available for epidemic surveillance. We hence collected a corpus in various languages from the Web. News corpora for Chinese, English and Russian were collected from Google News' health category. As this category existed neither for Polish nor Greek, corresponding documents were collected from major newspapers' health categories[5].

Restricting ourselves to documents from health category only induces surprisingly shallow filtering, as our analysis showed that only 8% documents contained epidemic events. However, this strategy allowed to collect a significant number of relevant documents at a reasonable cost.

For measuring precision and recall of the tasks of document filtering and event characterization, native speakers of each corresponding language[6] annotated sets of about 500 documents covering a 3-month period spanning from November 2011 to January 2012.

The characteristics of the evaluation corpus are shown in Table V. We can observe that the length of documents (in paragraph or characters) may vary a lot from one document to another. Annotators had to judge whether these documents were relevant for informing health authorities about infectious diseases. If a document was judged relevant, the annotator was further requested to provide the disease name and location of the event. The full annotation guidelines are available online[7]. In addition, the full corpus and the corresponding annotations are freely available to the community at the same address.

V. Results and evaluation

This section shows the performance of the repetition rule at key positions to select relevant press articles. Hence, DAnIEL is first demonstrated through examples, then evaluated quantitatively against annotators' judgements on the evaluation corpus. The system, written in Python, processes 2,000 documents in less than 15 seconds (2.4Ghz dual core processor, 2Gb RAM), which is compatible with on-line surveillance. According to our own experiments with state-of-the-art systems performing linguistic analysis, this is about 10 times faster.

A. Output examples

Figure 3 exhibits an example of the repetition phenomenon in a relevant press article. The disease name "tuberculosis" is repeated at key positions of the article: Head and Body. Only the longest common substring of the disease list found at key positions are highlighted. This is why the capitalized form "Tuberculosis" (last paragraph) is not highlighted. One can see that the seldom used abbreviation "TB" is not the only term used in the document, confirming our hypotheses on the principles of news writing.

No location is repeated in the article, hence the event is implicitly located with respect to the source[8], "India". This is a good showcase of the "implicit location" rule, used every time no location is repeated in the text.

Figure 2, already mentioned in Section III-B, shows the application of DAnIEL's principles in a morphologically rich

[5]"Gazeta", "Gazeta polska", "Dziennik zwiazkowy", etc. for Polish. "ΕΘΝΟΣ", "Το Βήμα", "ΕΞΠΡΕΣ", etc. for Greek.

[6]Nine professional translators who were not otherwise related to DAnIEL

[7]https://daniel.greyc.fr/

[8]http://www.dnaindia.com

i	LCP_i	SA_i	$\mathcal{S}[SA_i]...\mathcal{S}[n]$
...	...	...	...
46	0	168	5_tys._przypadków[...]$_2$denga$_1$Tajlandia$_0$
47	8	239	Tajlandia$_0$
48	7	140	Tajlandii_odnot[...]$_2$denga$_1$Tajlandia$_0$
49	0	0	Tajlandzki_rząd[...]$_3$W_tym_roku_w_Ta[...]$_2$denga$_1$Tajlandia$_0$
50	0	127	W_tym_roku_w_Ta[...]$_2$denga$_1$Tajlandia$_0$
...	...	...	...
77	1	14	d_ostrzega_kobi[...]$_3$W_tym_roku_w_Ta[...]$_2$denga$_1$Tajlandia$_0$
78	4	233	denga$_1$Tajlandia$_0$
79	4	186	dengi,_czyli_o_[...]$_2$denga$_1$Tajlandia$_0$
80	1	120	denge.$_3$W_tym_roku_w_Ta[...]$_2$denga$_1$Tajlandia$_0$
81	2	245	dia$_0$
...	...	...	...

Table IV. SAMPLE OF THE AUGMENTED SUFFIX ARRAY OF 2 SEGMENTS $\mathcal{S}_0$ AND $\mathcal{S}_1$ OF A POLISH DOCUMENT AND EXTERNAL RESOURCES $\mathcal{S}_2$ AND $\mathcal{S}_3$

	Chinese	English	Greek	Polish	Russian	Cumulated corpus
#documents (relevant)	446 (16)	475 (31)	390 (26)	352 (30)	426 (41)	2089 (144)
#paragraphs	4428	6791	3543	3512	2891	21165
avg.±std.	9.9±10.5	14.29±7.23	9.08±7.78	9.97±6.95	6.78±6.11	10.13±8.3
#characters (10^6)	1.14	1.35	2.05	1.04	1.56	7.17
avg.±std.	2568±2796	2858±1611	5264±5489	2971±2188	3680±5895	3432±4085

Table V. CHARACTERISTICS OF THE CORPUS

language, namely Polish. The disease name is repeated with different forms but still detected. The location is detected with the repetition rule, a sample case of "explicit location".

B. Global results

In this study the three main measures used for evaluation are Recall, Precision and F-measure. These measures are defined as follows:

- **Recall (R):** Number of relevant items retrieved by the system (True positives[Tp]) divided by total number of relevant items (True positives + False Negatives[Fn]): $\frac{Tp}{Tp+Fn}$;
- **Precision (P):** Number of relevant items retrieved by the system (True positives) divided by total number of retrieved items (True positives + False positives[Fp]): $\frac{Tp}{Tp+Fp}$;
- **F-measure (F):** Harmonic mean of recall and precision. This measure can be tuned (β parameter) to give a better weight to recall or precision: $(1+\beta)\frac{P.R}{(\beta.P)+R}$.

Following common practice in the field, F-measure is computed with $\beta = 1$ (F_1-measure) and $\beta = 2$ (F_2-measure), reflecting different emphasis on recall and precision (the higher β, the more the recall value is emphasized).

The performance of DAnIEL is detailed in Table VI. It shows that DAnIEL's performance is globally better in terms of recall than it is in terms of precision. DAnIEL achieved very good recall results for three languages of different families: Chinese, Greek and Polish. This is a very interesting result because Greek is a morphologically rich language whereas Chinese has poor morphology but still causes problems for machine translation. In Russian and Polish the system performance was worse, mainly because of precision. With the default θ value, DAnIEL obtained a F_1 score of 0.8 for the cumulated corpus. Tuning the best ratio for F_1-measure in each language permitted DAnIEL to increase precision to 0.74, with a slightly better recall (0.93). This result is somehow surprising as the small lexicon size was expected to impair recall more than precision. Indeed it is an important question for a system that relies on small resources: the system should not miss too many events, particularly for epidemic surveillance, where recall usually matters more than precision. Interestingly, the default ratio ($\theta = 0.8$) with its greater recall achieves a very good F_2-measure of 0.87. It is compatible with recall-oriented needs since it shows that DAnIEL can perform well without tuning. Table VII shows the extent to which DAnIEL misses events and the reasons for such errors.

	Chinese	English	Greek	Polish	Russian
Relevant documents	16	35	27	30	41
Lack in lexicon	0	1	0	1	3
No repetition	0	1	1	1	1
Wrong matching	0	2	0	0	2
Silence	0	4	0	2	6

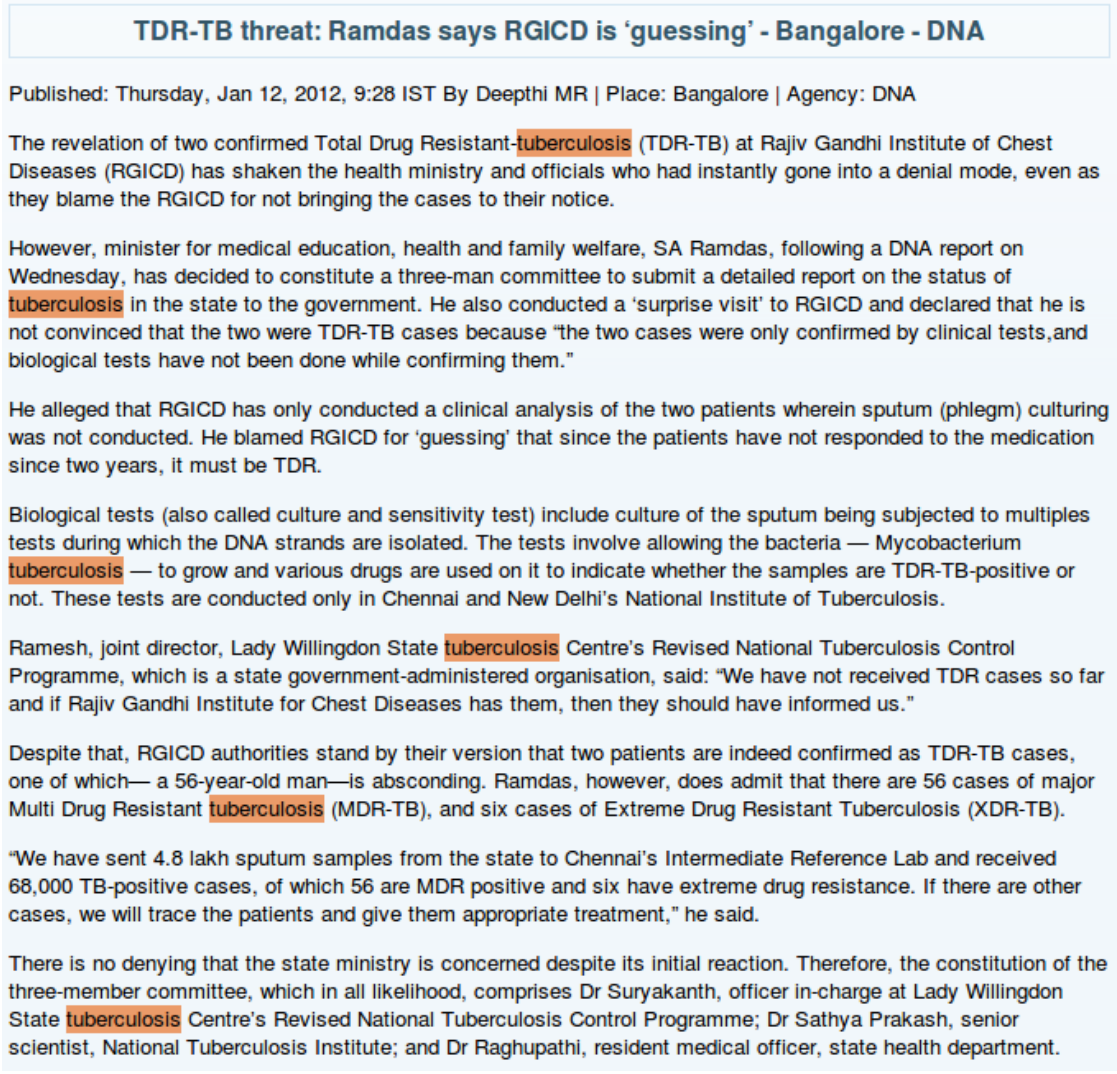
Table VII. DOCUMENT FILTERING: ERRORS IMPAIRING RECALL

Errors due to the size of the lexicon are quite rare (5 in total) and the repetition phenomenon is trustworthy: only four relevant documents were missed because no repetition matching with the disease name was found. Another issue stemmed from string recognition, as some diseases were referred to by names too short to be detected by DAnIEL.

The news discourse model implemented through repetition rules at special positions efficiently selects relevant press articles on epidemiological events. Figure 4 shows how frequent disease name repetition behaves in relevant articles (dotted line) and how rare it is in irrelevant ones (continuous line). This shows how this simple rule truly helps to filter documents out: 97% of irrelevant and only 0.7% of relevant articles contained no repetition.

C. Detailed evaluation

This section evaluates the performance of DAnIEL's processing steps and compares its results with two baselines.



TDR-TB threat: Ramdas says RGICD is 'guessing' - Bangalore - DNA

Published: Thursday, Jan 12, 2012, 9:28 IST By Deepthi MR | Place: Bangalore | Agency: DNA

The revelation of two confirmed Total Drug Resistant-tuberculosis (TDR-TB) at Rajiv Gandhi Institute of Chest Diseases (RGICD) has shaken the health ministry and officials who had instantly gone into a denial mode, even as they blame the RGICD for not bringing the cases to their notice.

However, minister for medical education, health and family welfare, SA Ramdas, following a DNA report on Wednesday, has decided to constitute a three-man committee to submit a detailed report on the status of tuberculosis in the state to the government. He also conducted a 'surprise visit' to RGICD and declared that he is not convinced that the two were TDR-TB cases because "the two cases were only confirmed by clinical tests,and biological tests have not been done while confirming them."

He alleged that RGICD has only conducted a clinical analysis of the two patients wherein sputum (phlegm) culturing was not conducted. He blamed RGICD for 'guessing' that since the patients have not responded to the medication since two years, it must be TDR.

Biological tests (also called culture and sensitivity test) include culture of the sputum being subjected to multiples tests during which the DNA strands are isolated. The tests involve allowing the bacteria — Mycobacterium tuberculosis — to grow and various drugs are used on it to indicate whether the samples are TDR-TB-positive or not. These tests are conducted only in Chennai and New Delhi's National Institute of Tuberculosis.

Ramesh, joint director, Lady Willingdon State tuberculosis Centre's Revised National Tuberculosis Control Programme, which is a state government-administered organisation, said: "We have not received TDR cases so far and if Rajiv Gandhi Institute for Chest Diseases has them, then they should have informed us."

Despite that, RGICD authorities stand by their version that two patients are indeed confirmed as TDR-TB cases, one of which— a 56-year-old man—is absconding. Ramdas, however, does admit that there are 56 cases of major Multi Drug Resistant tuberculosis (MDR-TB), and six cases of Extreme Drug Resistant Tuberculosis (XDR-TB).

"We have sent 4.8 lakh sputum samples from the state to Chennai's Intermediate Reference Lab and received 68,000 TB-positive cases, of which 56 are MDR positive and six have extreme drug resistance. If there are other cases, we will trace the patients and give them appropriate treatment," he said.

There is no denying that the state ministry is concerned despite its initial reaction. Therefore, the constitution of the three-member committee, which in all likelihood, comprises Dr Suryakanth, officer in-charge at Lady Willingdon State tuberculosis Centre's Revised National Tuberculosis Control Programme; Dr Sathya Prakash, senior scientist, National Tuberculosis Institute; and Dr Raghupathi, resident medical officer, state health department.

Figure 3. Example of a relevant document in English: repetition of disease name and use of implicit location rule.

	Chinese	English	Greek	Polish	Russian	cumulated corpora	
θ	0.75	0.8	0.75	0.8	0.85	$\theta = 0.8$ (default)	best θ for each language
Precision	**0.84**	0.70	0.68	0.65	0.81	0.72	0.74
Recall	**1.0**	0.89	0.96	0.93	0.85	0.91	0.93
F_1-measure	**0.91**	0.78	0.80	0.77	0.83	0.80	0.82
F_2-measure	**0.96**	0.84	0.89	0.86	0.84	0.87	0.88

Table VI. DOCUMENT FILTERING: PRECISION, RECALL, F_1 AND F_2-MEASURE FOR BEST θ

1) Segmentation filtering: The news segmentation described in Section III-A is intended to filter out uninteresting motifs. Table VIII shows the impact of this filtering on the total number of motifs. The point of segmentation filtering is to reduce the noise produced by the system without impairing recall too much. The filtering rate is higher in Chinese since this language contains much more characters and there is hence less very frequent n-grams that are found all over the documents (including at key positions). These frequent n-grams are much more common in other languages, for instance "_th" in English.

2) Filtering relevant documents: In order to evaluate the different features of our system, Table IX shows the performance of two baselines B1 and B2: B1 assumes an epidemic event whenever a disease name is present in the document while B2 does only so if the disease name is repeated. B1 highlights the problems one can encounter with morphologically rich languages because of the exact matching needed for the disease name. B2 shows the improvement in precision with the use of repetitions. To focus exclusively on repetition, both baselines use $\theta = 1$ and neither takes positions into account.

3) Evaluating the threshold $\frac{len(motif)}{len(entry)} \geq \theta$*:* This section describes the results of empirical experiments to determine the

	Chinese	English	Greek	Polish	Russian
#medium and large documents	429	467	358	348	401
#motifs without segmentation (avg.±std.)	676.98±317.27	5583.48±1952.33	4698.41±2800.28	7520.75±2496.48	5638.76±2032.96
#motifs with segmentation (avg.±std.)	8.23±11.89	1386.72±237.18	249.30±47.08	1811.02±383.38	1135.70±172.64
Filtering rate	82.25	4.02	18.84	4.15	4.96

Table VIII. ASSESSMENT OF FILTERING IMPACT, NUMBER OF MOTIFS FOR MEDIUM AND LONG ARTICLES

		Chinese	English	Greek	Polish	Russian	Cumulated corpora
Baseline 1 (B1)	Precision	0.50	0.42	0.47	0.50	0.48	0.47
	Recall	1.0	0.9	**0.96**	**0.91**	**0.87**	**0.93**
	F_1-measure	0.66	0.57	0.63	**0.65**	0.62	0.6
	F_2-measure	0.8	0.7	0.76	**0.76**	0.72	0.73
Baseline 2 (B2)	Precision	**0.8**	**0.64**	**0.61**	**0.53**	**0.67**	**0.62**
	Recall	1.0	**0.89**	0.92	0.5	0.75	0.81
	F_1-measure	**0.88**	**0.74**	**0.73**	0.51	**0.70**	**0.73**
	F_2-measure	**0.94**	**0.81**	**0.82**	0.51	**0.73**	**0.77**

Table IX. EVALUATION OF TWO BASELINES: PRECISION, RECALL, F_1-MEASURE AND F_2-MEASURE

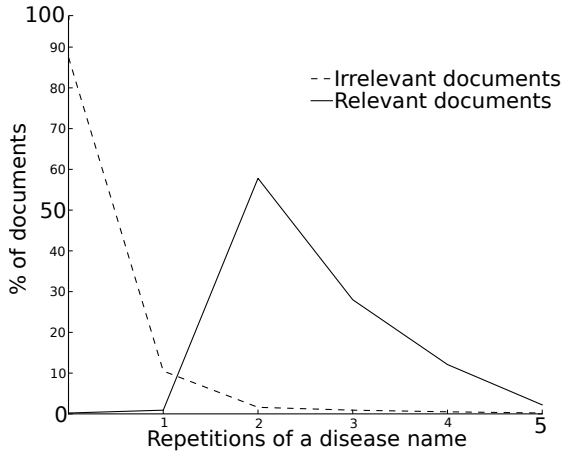


Figure 4. Repetitions of disease name in relevant and irrelevant articles

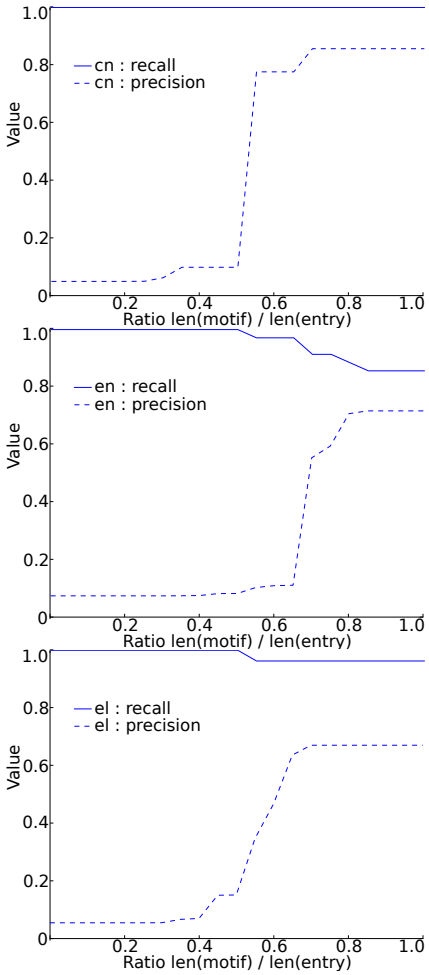


Figure 5. Recall and precision according to θ (Chinese, English and Greek)

appropriate string matching ratio (θ) between motifs extracted and knowledge base entries for the five languages of this experiment. For instance, a small θ offers a perfect recall with high noise (many wrong events are extracted). The following experiments were performed with the aim to find the value allowing the best trade-off between recall and precision.

Figure 5 shows that in Chinese, English and Greek, an increase in the value of θ causes an increase in precision with little impact on recall. This result was expected for Chinese and English but not for Greek which has richer morphology.

In the contrary, Figure 6 shows that for Polish (respectively, Russian), performance drops quickly when θ is greater than 0.8 (respectively, 0.85). As could be expected, the choice of θ matters more for these two languages, due to their rich morphology.

The same experiment was performed with a same θ value for the whole corpus. Figure 7 shows that $\theta = \frac{4}{5}$ is a good empirical value for processing the five different languages simultaneously. The results are summarized in Table VI, containing the optimal value of θ for each language and the scores obtained with θ uniformly set to $\frac{4}{5}$.

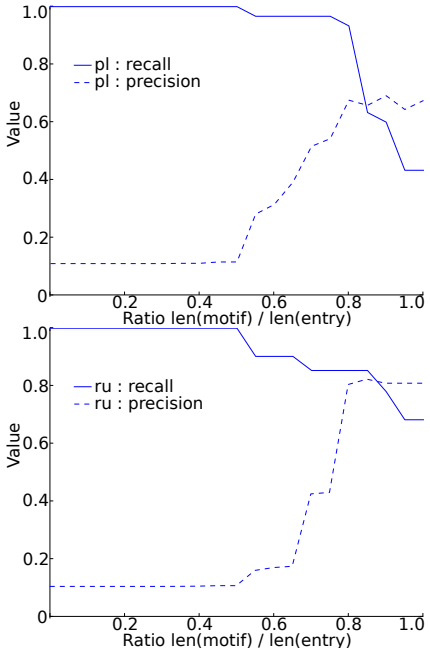


Figure 6. Recall and precision according to θ (Polish and Russian)

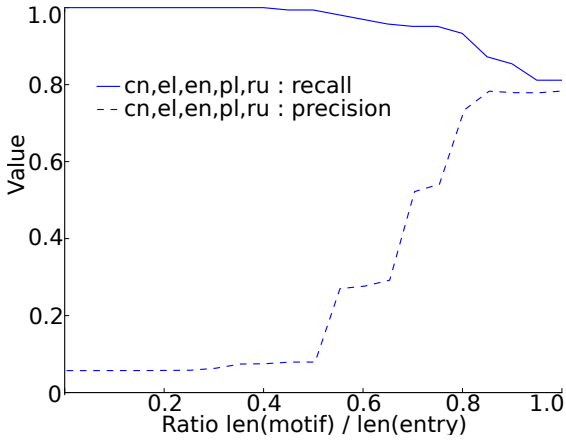


Figure 7. Recall and precision according to θ (all languages)

4) Event localization: Table X exhibits the performance of the localization algorithm on our annotated corpus. This experiment compares the location given by DAnIEL and the location given by the annotators. The implicit location rule has been applied to the majority of the detected events (99 over 134) and achieved a very good performance with 85% precision. Two errors came from a source to which the wrong country had been assigned. The explicit location rule performed slightly worse with 79% precision. Most of the errors could actually be called "partial", since the detected location was correct, but incomplete. For instance, events concerning the whole Europe were incorrectly located in Poland only.

5) Level of evaluation unit: document or event?: An evaluation per document is not necessarily adequate, when one considers a typical use case [21]. It is for instance possible to detect 99 documents describing the same event (e.g., flu in Spain) but miss an event because it was contained in only one document (e.g., Ebola in Congo). A document-level evaluation would rank this as 99% recall, which is intuitively wrong, as only one out of two events was detected. To evaluate how DAnIEL performs with respect to events rather than documents, further event-based annotations were compiled. Each disease-location pair (flu in Spain for instance) was considered as a unique epidemiological event regardless of the number of documents it had been reported in over a 3-month time window.

	Unique events	Detected	Missed
Chinese	5	5	0 (0%)
English	15	14	1 (6,6%)
Greek	17	17	0 (0%)
Polish	28	26	2 (7,1%)
Russian	23	21	2 (8,6%)
Total	62	59	3 (4,8%)

Table XI. EVALUATION BY UNIQUE EVENT

Table XI shows the results of the evaluation by event, demonstrating that only few full-fledged events (3 out of 62) were missed. The system takes advantage of the fact that it has coverage in more than one language which gives it additional opportunities to detect events [22]. For instance, an event missed in Polish had been detected in Russian. Note that the total number of unique events in Table XI is not the sum of unique events in reports since a single epidemiological event can be reported in several languages.

This final experiment highlights the importance of increasing the coverage by processing more languages rather than optimizing a system in a small number of languages. A greater coverage limits both the time needed to detect an event and the risk to miss it.

VI. CONCLUSION

The principles of a genre-based information extraction system called DAnIEL have been tested with success on Chinese, English, Greek, Polish and Russian news data. The system relies on very light, easy-to-get resources, and is intended to help health authorities gather precious information about ongoing infectious diseases spreading all around the world. In order to be multilingual, it uses genre-related features and relies on text style, in particular, relying on carefully selected types of string repetitions, rather than on sentence-level words or patterns specific to one or few languages.

The algorithm is based on the way news articles are rhetorically constructed. The detection of string repetitions permits to limit the number of components needed for monitoring new languages. No linguistic analysis is performed and only a small-sized external lexicon is used. Experiments showed that the system might lack in precision, but has good recall (0.97 for English, 0.92 for the whole corpus). This performance is of interest for online epidemic surveillance. DAnIEL is efficient at distinguishing irrelevant documents which makes it useful to filter large corpora, even with low-resourced languages.

With an average F_1-measure of 0.85, DAnIEL scores are below state-of-the-art systems like PULS or BIOCASTER, which are closer to 0.9 with English and a few other languages. However, the resources that these systems require (lexicon,

	Chinese	English	Greek	Polish	Russian	Global
# Events retrieved by DAnIEL	16	31	26	28	35	136
Implicit location performance	15/16	20/21	11/13	14/18	27/30	87/98 (87%)
Explicit location performance	N/A	7/10	11/13	8/10	4/5	30/38 (79%)
Area error	1	3	3	4	1	9
No repetition detected	0	0	1	1	1	1
Lack in lexicon	0	0	0	0	2	2
Error in the source	0	1	0	0	1	2

Table X. PERFORMANCE OF THE LOCATION RULES

language parser, ontologies) are far more extensive and costly. To the contrary, DAnIEL makes it possible to process new languages without any knowledge in programming. To process a new language, the only thing the user needs to provide is a list of diseases names. Providing a list of locations is not a strict requirement since the implicit location rule of DAnIEL performs reasonably well. The parameter θ can be used at its default value (0.8) with reliable results, or be tuned if expert knowledge is available (of a linguist or a native speaker).

When no classical IE system is available or training data is too scarce, a text genre-based IE system can fill the gap efficiently. In particular, the method described here is aimed to increase the coverage in number of languages rather than optimizing results with a particular language. It can save efforts to filter relevant documents to be thoroughly parsed by existing techniques with high precision over major languages. In order to help IE research, the corpora used for this experiment are available to the community with annotations detached from original urls. It will be of interest for morphologically rich languages. New corpora are currently being annotated for numerous other languages (Arabic, French, Portuguese, Spanish, Swahili…) in order to assess DAnIEL's ability for a greater multilingual coverage.

REFERENCES

[1] D. Son, H.-N. Quoc, K. Ai, and N. Collier, "Global health monitor - a web-based system for detecting and mapping infectious diseases," *International Joint Conference on Natural Language Processing*, pp. 951–956, 2008.

[2] J. Linge, R. Steinberger, T. Weber, R. Yangarber, E. van der Goot, D. Al Khudhairy, and N. Stilianakis, "Internet surveillance systems for early alerting of threats," *Eurosurveillance*, vol. 14, no. 13, 2009. [Online]. Available: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19162

[3] A. Lyon, M. Nunn, G. Grossel, and M. Burgman, "Comparison of web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap," *Transboundary and Emerging Diseases*, 2011. [Online]. Available: http://dx.doi.org/10.1111/j.1865-1682.2011.01258.x

[4] D. M. Hartley, N. P. Nelson, R. Walters, R. Arthur, R. Yangarber, L. Madoff, J. Linge, A. Mawudeku, N. Collier, J. S. Bronstein, G. Thinus, and N. Lightfoot, "The landscape of international event-based biosurveillance," *Emerging Health Threats Journal*, vol. 3, no. e3, 2010.

[5] A. R. Reilly, E. A. Iarocci, C. M. Jung, D. M. Hartley, and N. P. Nelson, "Indications and warning of pandemic influenza compared to seasonal influenza," *Advances in disease surveillance*, vol. 5, p. 190, 2008.

[6] R. Steinberger, F. Fuart, E. van der Goot, C. Best, P. von Etter, and R. Yangarber, "Text mining from the web for medical intelligence," in *Mining massive data sets for security*. OIS Press, 2008, pp. 295–310.

[7] S. Huttunen, V. Arto, P. von Etter, and R. Yangarber, "Relevance prediction in information extraction using discourse and lexical features," in *Nordic Conference on Computational Linguistics, Nodalida 2011*, 2011, pp. 114–121.

[8] H. Ji, "Challenges from information extraction to information fusion," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 507–515.

[9] M. Du, P. Von Etter, M. Kopotev, M. Novikov, N. Tarbeeva, and R. Yangarber, "Building support tools for Russian-language information extraction," in *Proceedings of the 14th international conference on Text, Speech and Dialogue*, 2011, pp. 380–387. [Online]. Available: http://dl.acm.org/citation.cfm?id=2040037.2040088

[10] N. Lucas, "Stylistic devices in the news, as related to topic recognition," in *Texts and Minds : Papers in Cognitive Poetics and Rhetoric*, ser. Łódź, Studies in language, A. Kwiatkowska, Ed. Frankfurt am Main: Peter Lang, 2012, vol. 26, pp. 301–316.

[11] O. Etzioni, A. Fader, J. Christensen, and S. Soderland, "Open information extraction: The second generation," *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pp. 3–10, 2011.

[12] J. R. Hobbs, "The generic information extraction system," in *Proceedings of the 5th conference on Message understanding*, ser. MUC5 '93. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993, pp. 87–91. [Online]. Available: http://dx.doi.org/10.3115/1072017.1072029

[13] N. Collier, "Towards cross-lingual alerting for bursty epidemic events," *Journal of Biomedical Semantics*, vol. 2, pp. 1–11, 2011.

[14] R. Steinberger, "A survey of methods to ease the development of highly multilingual text mining applications," *Language Resources and Evaluation*, pp. 1–22, 2011.

[15] K. Church, "Empirical estimates of adaptation: the chance of two noriegas is closer to $\frac{p}{2}$ than p^2," in *Proceedings of the 18th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2000, pp. 173–179.

[16] G. Lejeune, R. Brixtel, A. Doucet, and N. Lucas, "Daniel: Language independent character-based news surveillance." in *8th International Conference on NLP (JapTAL 2012)*, ser. Lecture Notes in Computer Science, H. Isahara and K. Kanzaki, Eds., vol. 7614. Springer, 2012, pp. 64–75.

[17] G. Lejeune, R. Brixtel, C. Lecluze, A. Doucet, and N. Lucas, "Added-value of automatic multilingual text analysis for epidemic surveillance," in *14th Conference on Artificial Intelligence in Medicine (AIME 2013)*, ser. Lecture Notes in Computer Science, N. Peek, R. M. Morales, and M. Peleg, Eds., vol. 7885. Springer, 2013, pp. 284–294.

[18] E. Ukkonen, "Maximal and minimal representations of gapped and non-gapped motifs of a string," *Theoretical Computer Science*, vol. 410, no. 43, pp. 4341–4349, 2009.

[19] J. Kärkkäinen, P. Sanders, and S. Burkhardt, "Linear work suffix array construction," *Journal of the ACM*, vol. 53, no. 6, pp. 918–936, 2006.

[20] N. Collier, K. Ai, L. Jin *et al.*, "A multilingual ontology for infectious disease surveillance: rationale, design and challenges," *Journal of Language Resources and Evaluation*, pp. 405–413, 2007.

[21] S. Liao and R. Grishman, "Using document level cross-event inference to improve event extraction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '10, 2010, pp. 789–797.

[22] J. Piskorski, J. Belyaeva, and M. Atkinson, "On refining real-time multilingual news event extraction through deployment of cross-lingual information fusion techniques," in *Proceedings of European Intelligence and Security Informatics Conference (EISIC).*, 2011, pp. 38–45.