Multilingual event extraction for epidemic detection

Gaël Lejeune[a,b], Romain Brixtel[a,c], Antoine Doucet[a,d], Nadine Lucas[a]

[a] *Groupe de Recherche en Informatique, Image et Instrumentation, University of Caen Lower-Normandy, boulevard Maréchal Juin, 14032 Caen, France*
[b] *Laboratoire d'Informatique de Nantes Atlantique, University of Nantes, 2 rue de la Houssinière, 44322 Nantes, France*
[c] *Department of Organizational Behavior, Faculty of Business and Economics, Quartier Dorigny, University of Lausanne, 1015 Switzerland*
[d] *Laboratoire Informatique, Image et Interaction, University of La Rochelle, avenue Michel Crépeau, 17042 La Rochelle, France*

Abstract

Objective. This paper presents a multilingual news surveillance system applied to tele-epidemiology. It has been shown that multilingual approaches improve timeliness in detection of epidemic events across the globe, eliminating the wait for local news to be translated into major languages. We present here a system to extract epidemic events in potentially any language, provided a WIKIPEDIA seed for common disease names exists.

Methods. The DANIEL system presented herein relies on properties that are common to news writing (the journalistic *genre*), the most useful being repetition and saliency. WIKIPEDIA is used to screen common disease names to be matched with repeated characters strings. Language variations, such as declensions, are handled by processing text at the character-level, rather than at the word level. This additionally makes it possible to handle various writing systems in a similar fashion.

Material. As no multilingual ground truth existed to evaluate the DANIEL system, we built a multilingual corpus from the Web, and collected annotations from native speakers of Chinese, English, Greek, Polish and Russian, with no connection or interest in the DANIEL system. This data set is available online freely, and can be used for the evaluation of other event extraction systems.

Preprint submitted to Journal of Artificial Intelligence in Medicine *June 1, 2015*

Results. Experiments for 5 languages out of 17 tested are detailed in this paper: Chinese, English, Greek, Polish and Russian. The Daniel system achieves an average F-measure of 82% in these 5 languages. It reaches 87% on BEcorpus, the state-of-the-art corpus in English, slightly below top-performing systems, which are tailored with numerous language-specific resources. The consistent performance of Daniel on multiple languages is an important contribution to the reactivity and the coverage of epidemiological event detection systems.

Conclusions. Most event extraction systems rely on extensive resources that are language-specific. While their sophistication induces excellent results (over 90% precision and recall), it restricts their coverage in terms of languages and geographic areas. In contrast, in order to detect epidemic events in any language, the Daniel system only requires a list of a few hundreds of disease names and locations, which can actually be acquired automatically. The system can perform consistently well on any language, with precision and recall around 82% on average, according to this paper's evaluation. Daniel's character-based approach is especially interesting for morphologically-rich and low-resourced languages. The lack of resources to be exploited and the state of the art string matching algorithms imply that Daniel can process thousands of documents per minute on a simple laptop. In the context of epidemic surveillance, reactivity and geographic coverage are of primary importance, since no one knows where the next event will strike, and therefore in what vernacular language it will first be reported. By being able to process any language, the Daniel system offers unique coverage for poorly endowed languages, and can complete state of the art techniques for major languages.

Keywords: early event detection, poorly endowed languages, multilingual information access, tele-epidemiology, epidemic surveillance

1. Introduction

Information extraction (IE) aims at extracting structured views from text and particularly from newswires that provide instant information from a large number of sources. The European Media Monitor for instance collects about

40,000 news reports written in 43 languages everyday[1]. This information context provides a new opportunity for health authorities, needing to monitor information, placing emphasis on disease outbreaks and spreadings [1].

However, natural language processing historically puts a very strong emphasis on vocabulary and on differences between languages, to the extent computational models heavily rely on the constitution of lexical resources. Special effort has been exerted to collect specialized medical lexica. Therefore, although web news is available in a large number of languages and dialects, the standard pipeline in IE is designed for texts in standard English, with the need to add lexicon and special components (lemmatizer, parser) each time a new language is added. Meanwhile, disease outbreaks ignore national frontiers and when considering epidemiological event extraction (EE), one has to detect diseases from health-related news in many languages to send alerts to health authorities as quickly as possible [2].

Keller has compared existing systems [3] stressing their complementarity. In the same way, the D*ata* An*alysis for* I*nformation* E*xtraction in any* L*anguage* (Daniel) system fulfills part of the needs but not all. The strong points advocated here are quick access to new languages, very light programming needed and timeliness in IE [4]. It is also important to get a leveraged epidemiological EE, so as to detect events from multilingual sources both at the same pace and with similar reliability. Since no multilingual corpus was available for comparison with existing systems, a news corpus has been collected and made available for further tests. The Daniel system is a text-genre based EE system designed to manage multilingual news with a large geographical coverage. Multilingual IE with light resources was tested, in order to quickly detect news denoting concern about some disease. Here, the standard approach to text as a bag of words is replaced by a spatial vision of text. Three characteristics are combined to avoid the chore of constituting heavy resources for all languages. A strong hypothesis assumes the constraints of information and dissemination are common to all news writers, and that journalistic genre implies a common use of titles, headers, bodies and feet, whatever the language. The common structure in news is the rhetorical "spatial" basis for the proposed model. Information is found at a specific place. A similar notion is sometimes used in academic literature

[1] European Media Monitor: `http://emm.newsbrief.eu/overview.html` (Accessed: 20 April 2015)

analysis [5]. The second characteristic is the implicit use of discourse "time", a.k.a. narrative line in news, with some typical repetitions along the way. The third characteristic is the use of the news date, linking the event to a given time window in conjunction with a geographical location and a disease. Since the system fills a gap in epidemiological monitoring, experiments were conducted on a multilingual corpus of 17 languages. It was manually annotated for 5 of them (Chinese, English, Greek, Polish and Russian). Precision and recall are computed for document wise and event wise detection. The question is how to compare a light resource system aiming at a wide coverage, while everyone is deeply involved in enriching resources and improving results for a very few number of languages. Whenever possible, results are compared with existing systems, or on common corpora.

The present paper is organized as follows. In Section 2, an overview of the multilingual approaches in IE is provided along with proposals to overcome shortcomings in early detection of diseases. In Section 3, we introduce the Daniel system, a text-genre based EE system designed to manage multilingual news. Section 4 introduces the evaluation corpus that we collected for the experiments. In Section 5 the results are presented and discussed. Finally, the efficiency of such a light approach for filtering huge multilingual news feeds is discussed and future directions are sketched in Section 6.

2. Background

IE approaches rely mostly on the use of the generic IE chain [6]. Two systems that rely primarily on English, Puls[2] [7] and Biocaster[3] [8], are well-known examples of classic IE systems specializing in epidemiological EE with good results in English and a few other languages. HealthMap[4] [9] is another well-known example, with the additional feature that it incorporates information manually compiled by human experts. The IE processing chain involves numerous components for each language. Extending the coverage for such a system requires components corresponding to a new language to be gathered. For most languages, the necessary efficient components are lacking [2]. In recent years, machine learning has been used successfully to

[2] http://puls.cs.helsinki.fi/static/index.html (Accessed: 20 April 2015)

[3] https://sites.google.com/site/nhcollier/projects/biocaster (Accessed: 20 April 2015)

[4] http://www.healthmap.org/ (Accessed: 20 April 2015)

fill gaps in new languages that have a sufficient number of common properties with a mainstream language [10].

However, in epidemic surveillance, there is a need to cover poorly endowed languages [11] or even dialects without training data. In a multilingual setting, state-of-the-art systems are limited by the cumulative process of their language-by-language approach. A multilingual goal hardly can be fulfilled with classical monolingual approaches. This is particularly the case for highly inflected languages [12]. Despite the sequential aspect of the classic IE chain, a decomposition problem arises: a high marginal cost is needed for analyzing any new language. The detection and appropriate analysis of the very first news report relating to an epidemic event is crucial for timeliness [13], but it may occur in any language: usually the first language of description is that of the (remote) place where the event was located [11].

For these reasons, a recent assumption from studies on media rhetorical devices [14] was put to trial: expository news shows specific patterns of repetition (the main content is given first, then detailed). Interesting findings have been 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 [8]. The underlying idea is referred to as pragmatics, or is altogether implicit when no specific knowledge backs the findings. Since explicit knowledge is used in our system, it exploits style properties identified in news discourse. Lejeune *et al.* [16] introduced genre and discourse properties for EE. Liao *et al.* also advocated text level inference to improve EE though with a monolingual constraint [17, 18]. The approach presented here relies on journalists writing principles: repetition of important terms at *salient positions*, *clarity of style* and exploitation of the notion of a *model reader* (each piece of information does not have to be written explicitly since journalists make the assumption that readers can fill in the blanks). This approach leverages the unique role of structure and rhetorical principles commonly used by journalists (the *inverted-pyramid style* by Piskorski *et al.* [19]).

3. The Daniel system

The Daniel system presents an implementation of a discourse-level EE approach. It operates at discourse-level by exploiting the global structure of news in a newswire. It harnesses information ordering as defined by Lucas [14], as opposed to the usual analysis at sentence-level (morphology, syntax and semantics). Entries in the system are news texts, including their title and text-body, and the name of the source when available. The only structural information needed are the positions of the head and the body of the news (from metadata such as RDF/microformat, or extracted with a boilerplate removal tool). The main features of the DANIEL system are that it is character-based and that it uses positions of occurrences [20]. Character-based refers to the fact that text is handled as a sequence of characters rather than as a sequence of words, in order to consider all types of languages (even if the definition and delimitation of words are difficult). The descriptors used are not key words but strings of text, exploited if and only if they are repeated in pre-defined salient zones in text. The aim of the process is to extract epidemic events from news feed, and express them in the reduced form of disease-location pairs (*i.e.* what disease occurs where).

Figure 1 describes the steps of the process to detect whether a document describes an epidemic event. The DANIEL processing pipeline is composed of three steps: news article segmentation (Section 3.1), event detection (Subsection 3.3.1), event localization (Subsection 3.3.2) using a small knowledge base (Subsection 3.2.3) and substring patterns (*motifs*, Section 3.2).

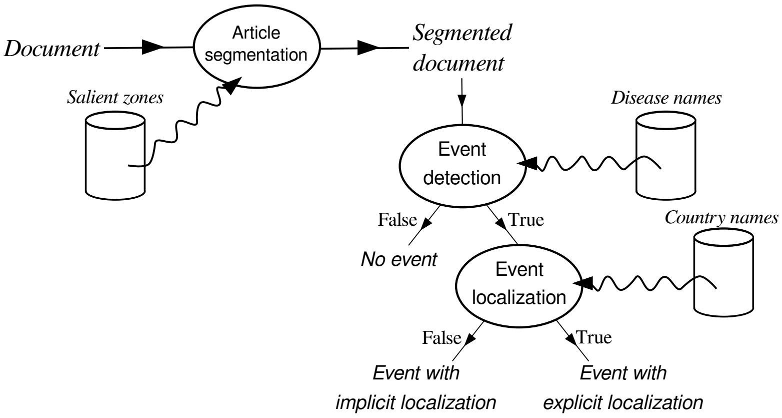


Figure 1: Overview of the DANIEL process.

3.1. *Text segmentation and salient zones*

The main algorithm exploits properties of the news genre. In a genre-driven approach, a clear understanding of text construction is crucial. The beginning and the end of a news text make up its *salient zones*. The system zooms in on the title and beginning (the topical head) of a text, and ceases elements that are repeated in these *salient positions*. The length of the text will determine the interactive relationship between *salient zones*. Rules reflecting these relationships are described in Table 1. These rules are simple to implement and largely language-independent. Salient zones -*Head* (title and first paragraph) *Tail* (last two paragraphs), and *Body* (the whole article except the *Head*)- combine for effect. The system will thus extract the substrings found in both Head and Body when confronting medium articles, and in Head and Tail in long articles. For short articles, repeated substrings are exploited irrespective of their position (the beginning overlaps the end, so the whole text is considered *salient*).

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 1: Article segmentation with respect to their number of paragraphs

3.2. 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 [21]. 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, followed by a demonstration using a sample document from our evaluation corpus.

3.2.1. Definition of motifs

Motifs are substring patterns of text with the following characteristics: they are *repeated* (motifs occur twice or more) and they are *maximal* (motifs cannot be expanded to the left (*left maximality*) or to the right (*right maximality*) without lowering the frequency). Following the example of Ukkonen [21], the motifs found in the string `HATTIVATTIAA` are `T`, `A` and `ATTI`. However, `TT` is not a motif because it always occurs inside each 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 [22] (also called Enhanced Suffix Array).

Given two strings $\mathcal{S}_0 = \texttt{HATTIV}$ and $\mathcal{S}_1 = \texttt{ATTIAA}$, Table 2 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 length of 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$\$_0$

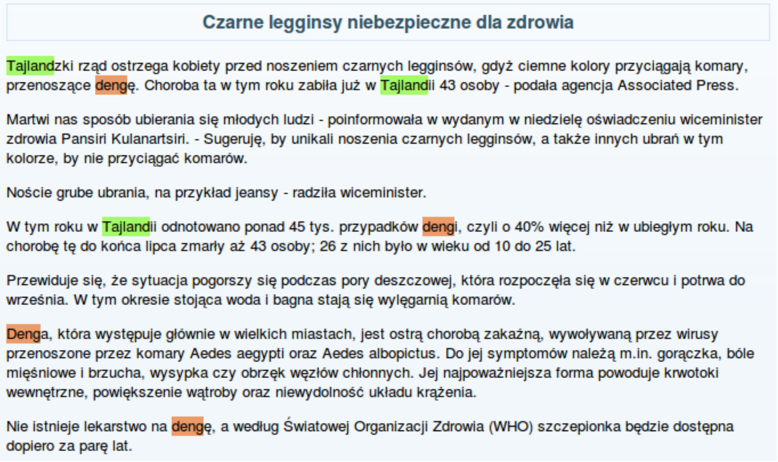
Table 2: Augmented suffix array of $\mathcal{S} = \texttt{HATTIV}\$_1\texttt{ATTIAA}\$_0$

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 2. 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}.

The maximal criterion is met 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 offset. For example, TTI is equivalent to ATTI because the last characters of these two motifs occur at the offsets $(4, 10)$ (these substrings are in a relation of *occurrence-equivalence* according to Ukkonen [21]). In this case, ATTI is held 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 $O(n)$ as described by Kärkkäinen and Sanders [22], while the enumeration process is done in $O(k)$, with k defined as the number of motifs and $k < n$ [21][5].

3.2.2. Examples of motifs



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: Relevant document (Polish) with disease name repeated and explicit location.

An example from a news article in Polish is given in Figure 2 to highlight the value of the process described here above. This document deals with a case of dengue in Thailand. We will focus on two sentences 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

[5] Code in PYTHON: `http://code.google.com/p/py-rstr-max/` (Accessed: 20 April 2015)

detection focuses on the detection of subpatterns of diseases names, here on the detection of the roots : *deng*∼ and *Tajland*∼. Table 3 shows a selected sample of the augmented suffix array of the two text fragments $\mathcal{S}_0$ and $\mathcal{S}_1$.

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.$\$1_1$W_tym_roku_w_Ta[...]$\$_0$
10	1	68	_gdyż_ciemne_kolory[...]$\$1_1$W_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 3: Sample of the augmented suffix array of $\mathcal{S} = \mathcal{S}_0\$_1\mathcal{S}_1\$_0$ (spaces replaced by "_").

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 these substrings are strictly included in another at each offset where they occur. `_deng` is actually extracted rather than `deng` because the left context of `deng` is always a white space.

3.2.3. Construction of the knowledge base

Daniel relies on implicit knowledge on the news genre, which allows it to use only light lexical resources automatically collected from Wikipedia with light human moderation to pinpoint relevant information. To integrate a new language, the adequate lexicon of disease names and geographical locations (countries) are the only resources needed. Those are built through a crawl of the Wikipedia using the following procedure:

1. Crawl the Wikipedia English list of infectious diseases[6] then fetch each outgoing link, for instance, the "smallpox"[7] page.

[6] `http://en.wikipedia.org/wiki/List_of_infectious_diseases` (Accessed: 20 April 2015)

[7] `http://en.wikipedia.org/wiki/Smallpox` (Accessed: 20 April 2015)

2. For each English WIKIPEDIA disease page, capture the interlingual outgoing links and the corresponding (language code; disease name) pairs. For instance, on the smallpox page, one of these interlingual outgoing links is `http://hu.wikipedia.org/wiki/Fekete_himlő` (Accessed: 20 April 2015), where `hu` is the language code, and `Fekete himlő` is the disease name.
3. Finally, for each language, construct the disease lexicon from the collected pairs. For instance, to build the Hungarian disease lexicon, we need to collect all the pairs corresponding to the language code `hu`.

The exact same procedure is used to build the lexicon of locations, with the exception that the initial WIKIPEDIA page is the list of sovereign states[8]. Finally, the extracted lexicon contains disease names and geographical locations (countries). The lexicon needed with our genre-based system is small: hundreds of items *versus* tens of thousands in state-of-the-art systems based on linguistic knowledge [23]. The Web-extracted disease names make it possible to deal quickly with new languages, even without the assistance of a native speaker.

3.3. Use of the knowledge base

In practice the lexica of disease names and locations is used in a very direct way. An interesting text substring is defined by at least 3 occurrences: two in the document (in salient positions) and one in the lexicon. Hence, motif extraction is performed on articles combined with the external knowledge. Let $\mathcal{S}_2$ and $\mathcal{S}_3$ items of a lexicon to be analysed according to $\mathcal{S}_0$ and $\mathcal{S}_1$:

$\mathcal{S}_2$: Tajlanda *[Thailand]*
$\mathcal{S}_3$: denga *[denge]*

With $\mathcal{S}_0$, $\mathcal{S}_1$ (two segments of a document) and $\mathcal{S}_2$, $\mathcal{S}_3$ (two items in an external knowledge base), the augmented suffix array makes it possible to detect repetition between selected parts of a document and any resources a system might need. Table 4 shows a sample of this augmented suffix array.

Note that the addition of the lexica allows for sharper extraction. In the example, the detected motif is `deng`, when with the document alone, the extracted motif was `_deng`. In 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 "`_`" but

[8] `http://en.wikipedia.org/wiki/List_of_sovereign_states` (Accessed: 20 April 2015)

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	dengę.$\$_3$W_tym_roku_w_Ta[...]$\$_2$denga$\$_1$Tajlandia$\$_0$
81	2	245	dia$\$_0$
...	...	...	...

Table 4: 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 $\mathcal{S}_3$.

“$\$_2$” as well. So, `deng` is a motif occurring twice in the selected parts of the document and once in the disease name lexicon (“*denge*”).

3.3.1. Event detection

Daniel filters out motifs in response to article segmentation rules as described in Table 1, and to the list of disease names as explained in Section 3.2.3. It selects motifs that are substrings found in two different sub-units, typically head and tail, and matching at least one disease name. This comes from the genre-related rules stating that :

1. an important topic in news should be highlighted;
2. common names should be used to catch the reader's attention;
3. the topic should be repeated.

More formally, let $\mathcal{S}_0$ and $\mathcal{S}_1$ be the head and the tail of an article (*i.e.* the salient zones $\mathcal{Z}$) and $\mathcal{S}_2$ ... $\mathcal{S}_{n+1}$ the n entries in a diseases knowledge base $\mathcal{K}$ (Algorithm 1). The process enumerates repetitions on $\mathcal{S}_0$ ... $\mathcal{S}_{n+1}$ (section 3.2) and selects motifs that occur in $\mathcal{S}_0$, $\mathcal{S}_1$ and any of the $\mathcal{S}_{i\in[2,n+1]}$. A heuristic ratio is used to verify if a motif matches an entry: $len(m)/len(\mathcal{S}_i) \geq \theta$-disease (Algorithm 1, line 9), with m a motif occurring in salient zones and in an entry $\mathcal{S}_i$ of the diseases base, $len(m)$ and $len(\mathcal{S}_i)$ are the number of characters of m and $\mathcal{S}_i$. In the previous example, the process tests whether $len(\text{deng})/len(\text{denga}) = 4/5 \geq \theta$-disease. The value of θ-disease is discussed in subsection 5.3.3. This technique proves especially useful for morphologically rich languages, as it bypasses the need for a morphological analyzer. If no motif matches the knowledge base using the θ-disease threshold, it assumes that the document contains no event and is therefore irrelevant. If several items fill this criterion, the longest is selected.

Algorithm 1: `isRelevant`

```
Input: $\mathcal{Z}$, a list of salient zones $z$ of a document
Input: $\theta$-disease, a matching threshold, $\theta$-disease $\in ]0, 1]$
Input: $\mathcal{K}$, a knowledge base (a list of items $k$)
Data: $rstr(s_0, ..., s_{n-1})$, maximal repeats in strings $s_0, s_1, ..., s_{n-1}$
Data: $len(s)$, the length of a string $s$
Output: a diagnostic, $True$ if a document is relevant, $False$ otherwise
begin
    $\mathcal{R} \leftarrow rstr(\mathcal{Z} + \mathcal{K})$ // maximal repeats in salient zones $\mathcal{Z}$ and knowledge base $\mathcal{K}$;
    foreach $r \in \mathcal{R}$ do
        $match_z \leftarrow \{z \in \mathcal{Z} \mid z$ contains $r\}$ ;
        if $match_z = \mathcal{Z}$ then // if a repeat occurs in each salient zone of $\mathcal{Z}$
            $match_k \leftarrow \{k \in \mathcal{K} \mid k$ contains $r\}$ ;
            foreach $k \in match_k$ do
                /* if a repeat overlaps an item in $\mathcal{K}$                */
                if $\frac{len(r)}{len(k)} \geq \theta\text{-}disease$ then return $True$;
    return $False$;
end
```

3.3.2. Event localization

An event is minimally defined as a disease-location pair. Again, journalistic style characteristics are used in Daniel to localize events without sentence-level extraction patterns. The locations are found in the same way as disease names (Algorithm 1), using repetitions in the same salient zones $\mathcal{Z}$ as for the event detection process (as described in Table 1, Section 3.1). The motifs selected are those occurring in those zones and in a knowledge base $\mathcal{K}$ containing a list of country names extracted from Wikipedia. The matching parameter θ-location is used as an alternative to θ-disease. The impact of the value of θ-location is detailed in subsection 5.3.6. As in the previous subsection, if several locations fill the criterion according to θ-location, the longest is selected.

When no location is explicitly mentioned, the event described in the document is linked to the issuing place. Hence, the location of the event is assumed to be the country of the source (*i.e.* that of the newspaper or the news agency). This is referred to as the *implicit location rule.*

4. Corpus

To the best of our knowledge, there is no available corpus for the evaluation of multilingual epidemic surveillance. The only corpus available online, BECORPUS[9], is exclusively built with relevant documents (200), making it unsuitable for evaluating the precision of document filtering. The corpus consists of a list of uniform resource locators (URLs) of Web pages compiled in 2009, and of which 102 source documents were still available in 2014[10]. All the reports are written in English. We used this corpus to evaluate event extraction as described in Section 5.3.7.

We built a multilingual corpus with documents in Chinese, English, Greek, Polish and Russian taken from the Web. News corpora in Chinese, English and Russian were collected from the health category in GOOGLE NEWS. Since this category does not exist in Polish or in Greek, documents were collected from health categories in major newspapers[11].

Surprisingly, limiting our corpus to documents found in health categories caused low filtering power: only 8% of the resulting documents referred to epidemic events. Nonetheless, this strategy allowed us to collect a significant number of relevant documents at a reasonable cost. For measuring precision and recall of document filtering, event detection and event localization, a set of about 500 documents has been annotated for each language. Native speakers of each language[12] annotated documents covering the same 3-month period (November 2011 to January 2012). Evaluation corpus characteristics are shown in Table 5.

The length of documents (in paragraph or characters) vary a lot from one to another. Annotators had to judge whether each document was 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 guidelines, the corpus and corresponding annotations are available on the DANIEL Web site[13].

[9] `https://code.google.com/p/becorpus/` (Accessed: 20 April 2015)

[10] List available on our corpus page: `https://daniel.greyc.fr/corpus.php` (Accessed: 20 April 2015)

[11] "Gazeta", "Gazeta polska", etc. for Polish. "To Βήμα", "ΕΞΠΡΕΣ", etc. for Greek.

[12] Nine professional translators who were not otherwise related to DANIEL

[13] `https://daniel.greyc.fr/corpus.php` (Accessed: 20 April 2015)

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

Table 5: Characteristics of the corpus

5. Results and evaluation

This section shows the performance of the repetition rule in salient zones to select relevant press articles. Daniel is first demonstrated through examples, then evaluated quantitatively against annotators' judgements on the evaluation corpus. The system processes 2,000 documents in less than 15 seconds [14], which is compatible with on-line surveillance.

5.1. Output examples

Figure 3 exhibits an example of the repetition phenomenon in a relevant press article. The term "tuberculosis" is repeated at *salient positions* (*i.e.* occurs in salient zones): head and body. The longest common substrings between the disease list and salient zones are highlighted. This is why the capitalized form "Tuberculosis" (last paragraph) is not highlighted. The abbreviation "TB" is not the sole term used in the document, confirming our assumption on news writing: explicit terms are used to ease the transmission of the main topic. No location is repeated in the article, hence the event is implicitly located with respect to the source[15], "India".

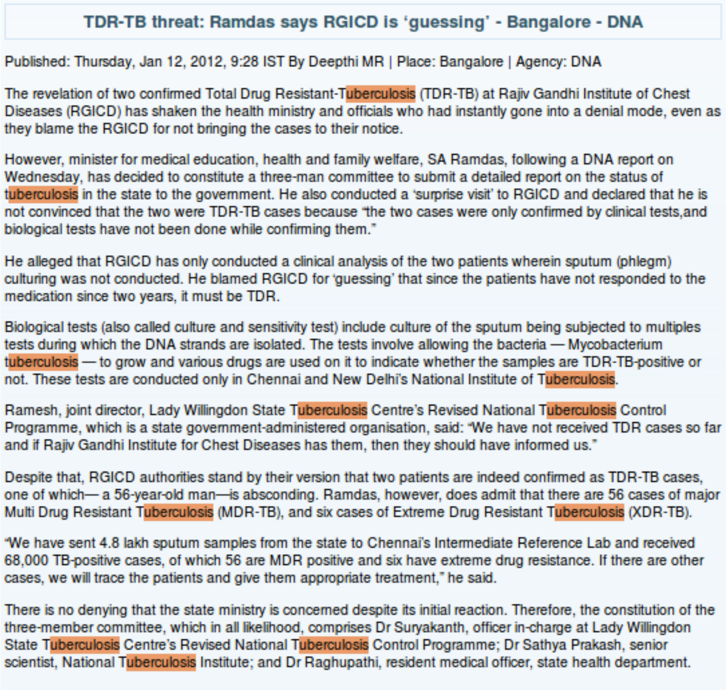
Figure 2, mentioned in Section 3.2, shows the application of Daniel's principles in Polish, a morphologically rich language. The disease name is repeated with different forms, but still detected. The location is detected with the repetition rule, a sample case of *explicit location*.

5.2. Global results

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

[14] Program in Python, using a 2.4Ghz dual core processor with 2Gb RAM

[15] `http://www.dnaindia.com` (Accessed: 20 April 2015)



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: Relevant document (English) with disease name repeated and *implicit location.*

- **Recall (R):** Number of relevant items retrieved by the system (True positives T_p) divided by total number of relevant items (True positives + False Negatives: $T_p + F_n$): $R = \frac{T_p}{T_p+F_n}$
- **Precision (P):** Number of relevant items retrieved by the system (True positives T_p) divided by total number of retrieved items (True positives + False positives: $T_p + F_n$): $P = \frac{T_p}{T_p+F_p}$
- **F-measure (F_β):** Harmonic mean of recall and precision. This measure can be tuned (β parameter) to add weight to recall or precision: $F_\beta = (1+\beta)\frac{P.R}{(\beta.P)+R}$

In harmony with common field practice, the F-measure is computed with $\beta = 1$ (F_1) and $\beta = 2$ (F_2), the higher β, the more the recall is emphasized. The items considered in the following experiments are documents. Hence, this evaluation is referred to as document wise evaluation (event wise evaluation is discussed in Section 5.3.6).

The performance of the Daniel system is detailed in Table 6. We can see that the performance is globally better in terms of recall than in terms

	Chinese	English	Greek	Polish	Russian	Cumulated corpora	
θ-disease	$[0.68, 1.0]$	0.82	$[0.90, 0.92]$	0.76	$[0.82, 0.86]$	0.80	best combination
Precision	**0.84**	0.70	0.70	0.65	0.76	0.72	0.74
Recall	**1.0**	0.89	0.96	0.87	0.90	0.91	0.93
F_1	**0.91**	0.78	0.81	0.77	0.82	0.80	0.82
F_2	**0.96**	0.84	0.90	0.86	0.86	0.87	0.88

Table 6: Document filtering – precision, recall, F_1 and F_2-measure for θ-disease values achieving the best F_1-measure score (0.80 being the default value)

of precision. Good recall results are achieved for three languages of different families: Chinese, Greek and Polish. This is a significant result because Greek is a morphologically rich language whereas Chinese has poor morphology but still causes problems for machine translation. In Polish the system performance was less satisfying due to lack of precision.

With the default θ-disease value (0.80), a F_1 score of 0.80 for the cumulated corpus. Tuning the best ratio θ-disease for F_1-measure in each language increased the 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. 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 θ-disease value 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 7 shows the extent to which Daniel misses events and the reasons for such errors.

	Chinese	English	Greek	Polish	Russian	Cumulated Corpora
#relevant documents	16	35	27	30	41	149
Lack in lexicon	0	1	0	1	3	5
No repetition	0	1	1	1	1	4
Wrong matching	0	2	0	0	2	4
Silence	0	4	1	2	6	13

Table 7: Errors impairing recall for the filtering task (with θ-disease = 0.80)

Errors due to the size of the lexicon are rare (5). The repetition phenomenon is trustworthy: only four relevant documents were missed because no repetition matching any disease name in the knowledge base 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 salient 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 out irrelevant documents: 97% of irrelevant as opposed to only 0.7% of relevant articles contained no repetition.

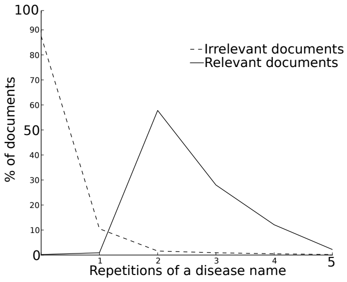


Figure 4: Repetitions of disease name in relevant and irrelevant articles

5.3. Detailed evaluation

This section first evaluates the performance of Daniel's processing steps and compares results to three baselines. The influence of the parameters θ-disease and θ-location is evaluated, and then the question of using alternative resources is tackled. Finally, an event-based evaluation is proposed for our multilingual reference corpus as well as for a corpus from the state-of-the-art BEcorpus.

5.3.1. Segmentation filtering

The news segmentation described in Section 3.1 is intended to filter out uninteresting motifs. Table 8 shows the impact of this filtering. The point of segmentation filtering is to reduce the noise produced by the system without significantly impairing recall. The filtering rate is lower in Chinese since the alphabet size is much higher (around 3,000 items). Hence, the motif distribution is sparser, and repetitions are less frequent. Frequent n-grams are much more common in other languages (*i.e.* "`_th`" in English).

5.3.2. Filtering relevant documents

In order to evaluate the different features of our system, Table 9 shows the performance of three baselines B1, B2 and B3. B1 assumes an epidemic event

	#documents	#motifs (avg.)		Filtering rate
		without segmentation	with segmentation	
Chinese	415	271.72	120.70	2.62
English	396	1101.45	114.67	9.60
Greek	159	1242.81	148.33	8.67
Polish	192	1128.12	129.05	8.74
Russian	90	1311.07	159.72	8.20

Table 8: Assessment of filtering impact, number of motifs for medium and long articles

whenever a disease name is present in the document while B2 does so only if the disease name is repeated. Finally, B3 combines the repetition criteria to the position of repetition. B1 highlights the problems with morphologically rich languages because of the exact matching required for the disease name. B2 shows the improvement in precision obtained with the use of repetitions. The additional constraint of position used in B3 leads to even better precision while hindering recall. All three baselines use θ-disease = 1.

		Chinese	English	Greek	Polish	Russian	Cumulated corpora
Baseline 1 (B1) *presence*	P	0.47	0.30	0.41	0.39	0.59	0.41
	R	1.00	1.00	0.96	0.90	0.88	0.94
	F_1	0.64	0.47	0.57	0.54	0.71	0.57
	F_2	0.82	0.69	0.76	0.71	0.80	0.74
Baseline 2 (B2) *repetition*	P	0.76	0.44	0.57	0.50	0.76	0.57
	R	1.00	0.91	0.92	0.60	0.78	0.83
	F_1	0.86	0.59	0.69	0.55	0.77	0.68
	F_2	0.94	0.75	0.80	0.58	0.78	0.76
Baseline 3 (B3) *repetition* & *position*	P	0.80	0.63	0.74	0.63	0.76	0.71
	R	1.00	0.71	0.93	0.33	0.76	0.72
	F_1	0.89	0.67	0.82	0.43	0.76	0.71
	F_2	0.95	0.69	0.88	0.37	0.76	0.72

Table 9: Evaluation of three baselines – Precision (P), Recall (R), F_1 and F_2-measure

5.3.3. Evaluating the overlap between knowledge base and documents

This section describes the determination of the appropriate string matching ratio between motifs extracted and knowledge base entries for the five languages. For instance, a small θ-disease offers a perfect recall with high noise (many irrelevant documents are selected). The aim of the following experiments is to find the value allowing for the best trade-off between recall and precision. Figure 5 and 6 shows that in Chinese, English and Greek, an increase in the value of θ-disease 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.

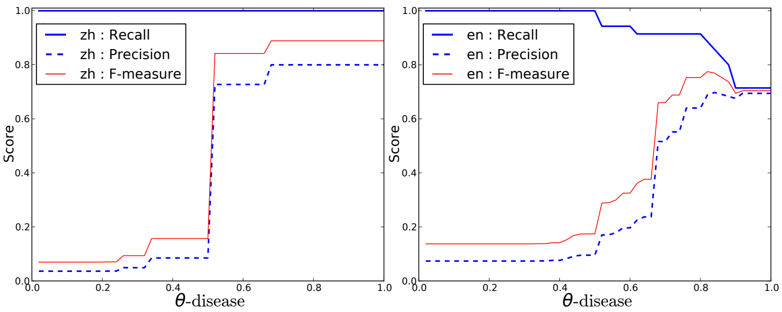


Figure 5: Evaluation according to θ-disease (Chinese and English)

Conversely, in Figure 6 performance drops for Polish (respectively, Russian) when θ-disease is greater than 0.80 (respectively, 0.85). The choice of θ-disease matters more for these two languages, due to their rich morphology. The same experiment was performed with a same θ-disease value for the cumulated corpora. The left graph of figure 7 shows that θ-disease $= 0.80$ is a good empirical value for processing the five different languages simultaneously. Table 6 contains the optimal value of θ-disease for each language and the scores obtained with θ-disease uniformly set to 0.80.

In Figure 7, the graph on the right-hand side illustrates the results obtained when all knowledge bases for all languages are merged. In this framework, the language of each document is unknown to the system. The results are very close to those obtained on the left-hand side of the figure, in which only the knowledge bases in the document's language are used. Interestingly, this implies that knowing the language of the document is not decisive for Daniel. This is mostly due to the fact that the languages used in this experiment are significantly different, which implies that there is little overlap between the various lexica. The potential of incorrectly matching a disease in the knowledge base of a given language with an irrelevant string in another language is indeed very unlikely, hence limiting the impact on the results.

5.3.4. Evaluation of document filtering using the ICD-10 lexicon

The Wikipedia lexica used in Daniel are easy-to-collect and multilingual. Domain ontologies could be used but few offer multilingual coverage. The international classification of diseases provided by World Health Orga-

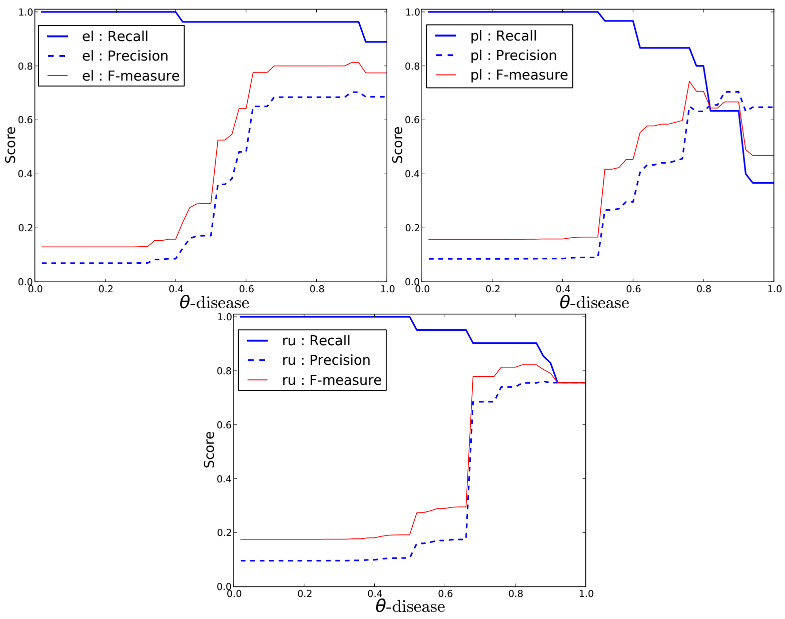


Figure 6: Evaluation according to θ-disease (Greek, Polish and Russian)

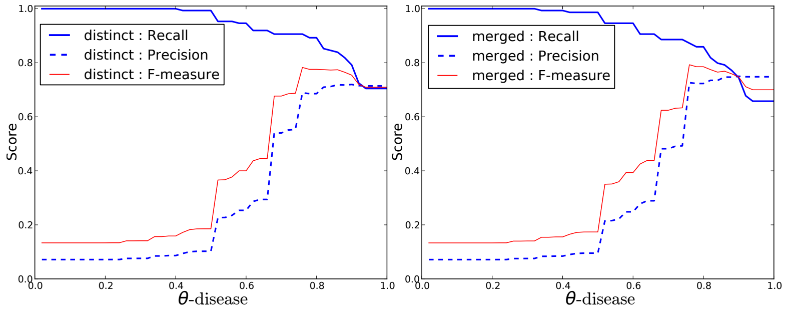


Figure 7: Experiments on the cumulated corpora, on the left distinct language resources are used whereas on the left the resources are merged

nization's (WHO) ICD-10[16] is one of them. ICD-10 covers 42 languages, several of which are available online. DANIEL has been tested with a lexicon extracted from ICD-10 (using chapters I to XV, II and IV excluded). Because the entries in ICD-10 might be complex (sometimes composed of

[16] 2010 version on the WHO's website: `http://apps.who.int/classifications/icd10/browse/2010/en` (Accessed: 20 April 2015)

	ICD-10, word split	ICD-10, word split (manual cleansing)	WIKIPEDIA
Resource size (#items)	2991	1347	147
Recall	**1.0**	0.77	0.91
Precision	0.07	0.23	**0.67**
F_1-measure	0.14	0.36	**0.77**

Table 10: Results for document filtering using ICD-10 and WIKIPEDIA (English).

a dozen of words), two different sub-lexica are exploited. The first one is composed of all the words in the entries of ICD-10. The second one was obtained by removing grammatical and vague words (45% of the English lexicon, e.g., “disease”, “sick” etc.). Performances are analyzed with regards to the document filtering task.

Experiments have been performed on the English corpus (Table 4) with θ-disease = 0.80. The results obtained with WIKIPEDIA (Table 10) are different from Table 6 since the θ-disease value is the default one. The ICD-10 lexicon induces very low precision since all the documents are tagged as relevant. After manually cleansing the ICD-10, by removing grammatical words and vague terms, precision rose from 0.07 to 0.23 which is far from the results obtained with WIKIPEDIA. It appears that ICD-10 gives no added-value to the results. Most of the terms are very specialized and seldom used in the news genre. A more thorough manual cleansing may improve the results further but this would be a costly and language-dependent procedure.

5.3.5. Event localization

Table 11 exhibits the performance of the localization algorithm. 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 (98 over 136) and achieved a good performance with 87% precision. Two errors came from a source to which the wrong country had been assigned. The explicit location rule performed worse with 79% precision. Most of the mislocations were actually partially correct, since the detected location was often a subregion of the annotated location. (e.g. events concerning the whole Europe were *incorrectly* located in Poland).

5.3.6. Evaluation by event

Evaluation can be carried out with respect to the number of documents selected, the technical unit commonly used in information retrieval, or to

	Chinese	English	Greek	Polish	Russian	Cumulated corpora
#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	12
No repetition detected	0	0	1	1	1	3
Lack in lexicon	0	0	0	0	2	2
Error in the source	0	1	0	0	1	2

Table 11: Performance of the location rules

the number of events, unit expressing the meaningful information for the task [24, 25]. For instance, it is possible to detect 99 documents describing the same epidemic event (e.g. flu in Spain in April 2012) and yet miss an event that is contained in only one document (e.g. Ebola in Congo in April 2012). A document wise evaluation would rank this case as 99% recall, which is intuitively wrong since only one out of two events is detected [17].

To evaluate how DANIEL performs with respect to events rather than documents, event-based annotations were compiled (corpus described in Section 4). Here, an event is a disease-location pair and a time period. All documents were published during the same 3-month time window. Therefore, each disease-location pair (e.g. flu in Spain) is considered as a unique event, regardless of the number of documents in which it has been reported.

	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%)
Cumulated Corpora	62	59	3 (4.8%)

Table 12: Evaluation by unique event

Table 12 shows the results of the evaluation by event, demonstrating that only a few full-fledged epidemic events (3 out of 62) were missed. The total number of unique events in the corpus (Table 12) is not the sum of unique events in each subcorpus. A single epidemiological event can be reported in several languages. The system takes advantage of its language coverage, which gives it additional opportunities to detect events [26] (e.g. an event missed in Polish documents was detected in Russian documents). This experiment highlights the importance of increasing the geographical coverage by processing more languages rather than optimizing a system in a small number of languages. A more extended coverage limits the time needed to detect an event and minimizes the risk of missing it [4].

Figure 8 exhibits heatmaps to show how the θ-location and θ-disease values affect event extraction. The lighter a zone, the better the results for a particular combination of θ-disease and θ-location values. Recall, precision and F_1-measure are computed as described in Section 5.2, based on disease-location pairs. In other words, let (d_1, l_1) be a disease-location pair of the gold standard. If d_2 and l_2 are a disease and a location in the knowledge bases, then neither (d_2, l_1) nor (d_1, l_2) are true positives.

The recall is slightly lower than in Table 12 since each distinct disease-location pair represents a class. For recall, the lighter zone (≥ 0.8) corresponds to the following combination of parameters: θ-location $\in [0.55, 1]$ and θ-disease $\in [0.6, 0.9]$. θ-location has little influence on results compared to θ-disease. Two factors lead to this. First, the implicit location rule is used for many documents (72% in the standard configuration as shown in Section 5.3.5). Second, location names include specific substrings that are less commonly found in the corpus (they have a relative invariant basis).

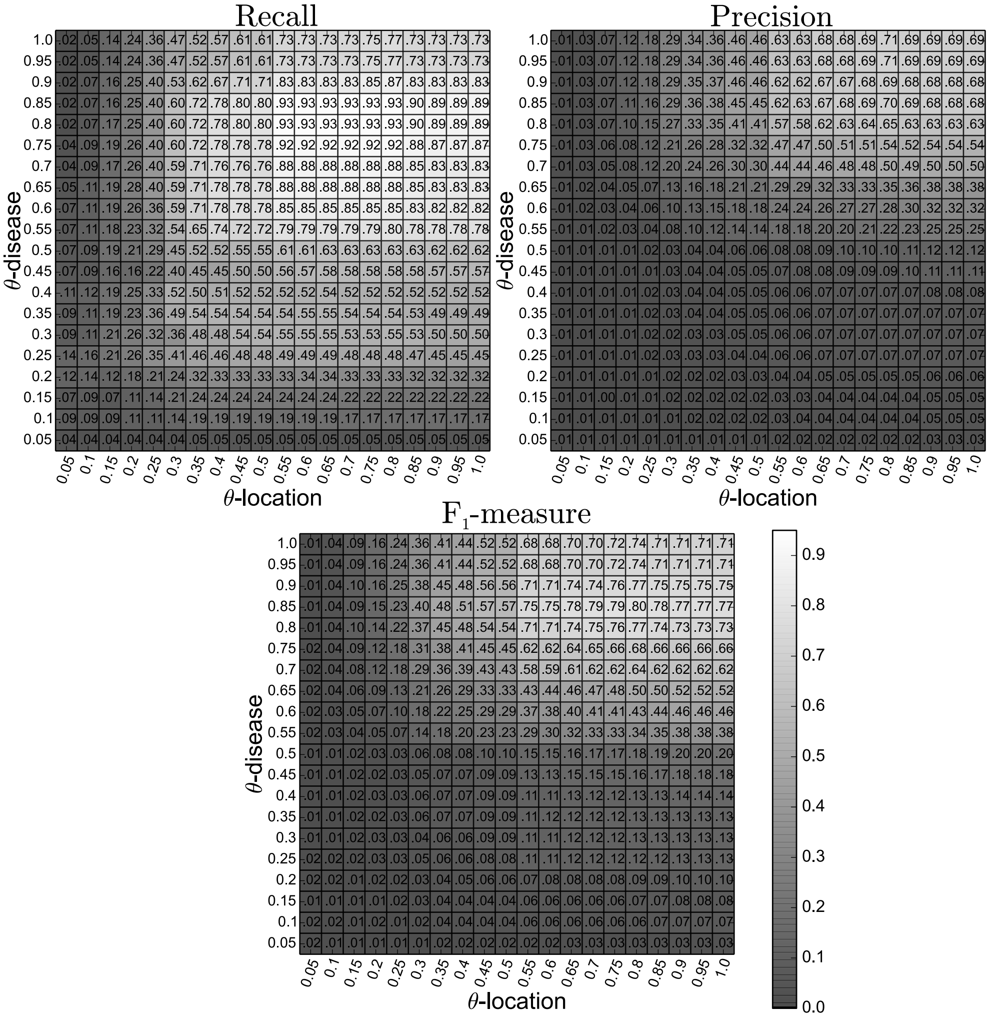
The best parameter combinations for precision are comparable to the ones for recall: θ-location $\in [0.65, 1]$ and θ-disease $\in [0.80, 0.90]$. The lighter zones cover a smaller area than in the heatmap for recall. However, few false positives represent noise since these events can easily be connected to human-validated ones (for instance *(H1N1,China)* and *(avian flu,China)*). This echoes the results shown at the document level (Table 6).

Finally, the heatmap for F_1-measure appears as a synthesis of the previous ones. The range of values of both parameters for achieving the best results (F_1-measure ≥ 0.7) are: θ-location $\in [0.55, 1]$ and θ-disease $\in [0.80, 1]$. The parameters can be adjusted in accordance with users' objectives. Still, using 0.80 for both θ-location and θ-disease achieves good results.

5.3.7. Document filtering and evaluation by event on the BEcorpus

This corpus has been released and described by Conway *et al.* in 2009 [27], and is available online[17]. The BioCaster team has used this corpus to evaluate event classification for its system [28]. It consists of 200 reports supplied with, among other things, the URL of the source and the metadata

[17] `https://code.google.com/p/becorpus/` (Accessed: 20 April 2015)



Recall

	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1.0
1.0	.02	.05	.14	.24	.36	.47	.52	.57	.61	.61	.73	.73	.73	.73	.75	.77	.73	.73	.73	.73
0.95	.02	.05	.14	.24	.36	.47	.52	.57	.61	.61	.73	.73	.73	.73	.75	.77	.73	.73	.73	.73
0.9	.02	.07	.16	.25	.40	.53	.62	.67	.71	.71	.83	.83	.83	.83	.85	.87	.83	.83	.83	.83
0.85	.02	.07	.16	.25	.40	.60	.72	.78	.80	.80	.93	.93	.93	.93	.93	.93	.90	.89	.89	.89
0.8	.02	.07	.17	.25	.40	.60	.72	.78	.80	.80	.93	.93	.93	.93	.93	.93	.90	.89	.89	.89
0.75	.04	.09	.19	.26	.40	.60	.72	.78	.78	.78	.92	.92	.92	.92	.92	.92	.88	.87	.87	.87
0.7	.04	.09	.17	.26	.40	.59	.71	.76	.76	.76	.88	.88	.88	.88	.88	.88	.85	.83	.83	.83
0.65	.05	.11	.19	.28	.40	.59	.71	.78	.78	.78	.88	.88	.88	.88	.88	.88	.85	.83	.83	.83
0.6	.07	.11	.19	.26	.36	.59	.71	.78	.78	.78	.85	.85	.85	.85	.85	.85	.83	.82	.82	.82
0.55	.07	.11	.18	.23	.32	.54	.65	.74	.72	.72	.79	.79	.79	.79	.79	.80	.78	.78	.78	.78
0.5	.07	.09	.19	.21	.29	.45	.52	.52	.55	.55	.61	.61	.63	.63	.63	.63	.63	.62	.62	.62
0.45	.07	.09	.16	.16	.22	.40	.45	.45	.50	.50	.56	.57	.58	.58	.58	.58	.58	.57	.57	.57
0.4	.11	.12	.19	.25	.33	.52	.50	.51	.52	.52	.52	.52	.54	.52	.52	.52	.52	.52	.52	.52
0.35	.09	.11	.19	.23	.36	.49	.54	.54	.54	.54	.54	.55	.55	.54	.54	.54	.53	.49	.49	.49
0.3	.09	.11	.21	.26	.32	.36	.48	.48	.54	.54	.55	.55	.55	.53	.53	.55	.53	.50	.50	.50
0.25	.14	.16	.21	.26	.35	.41	.46	.46	.48	.48	.49	.49	.49	.48	.48	.48	.47	.45	.45	.45
0.2	.12	.14	.12	.18	.21	.24	.32	.33	.33	.33	.33	.34	.34	.33	.33	.33	.32	.32	.32	.32
0.15	.07	.09	.07	.11	.14	.21	.24	.24	.24	.24	.24	.24	.24	.22	.22	.22	.22	.22	.22	.22
0.1	.09	.09	.09	.11	.11	.14	.19	.19	.19	.19	.19	.19	.19	.17	.17	.17	.17	.17	.17	.17
0.05	.04	.04	.04	.04	.04	.04	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05

Precision

	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1.0
1.0	.01	.03	.07	.12	.18	.29	.34	.36	.46	.46	.63	.63	.68	.68	.69	.71	.69	.69	.69	.69
0.95	.01	.03	.07	.12	.18	.29	.34	.36	.46	.46	.63	.63	.68	.68	.69	.71	.69	.69	.69	.69
0.9	.01	.03	.07	.12	.18	.29	.35	.37	.46	.46	.62	.62	.67	.67	.68	.69	.68	.68	.68	.68
0.85	.01	.03	.07	.11	.16	.29	.36	.38	.45	.45	.62	.63	.67	.68	.69	.70	.69	.68	.68	.68
0.8	.01	.03	.07	.10	.15	.27	.33	.35	.41	.41	.57	.58	.62	.63	.64	.65	.63	.63	.63	.63
0.75	.01	.03	.06	.08	.12	.21	.26	.28	.32	.32	.47	.47	.50	.51	.51	.54	.52	.54	.54	.54
0.7	.01	.03	.05	.08	.12	.20	.24	.26	.30	.30	.44	.44	.46	.48	.48	.50	.49	.50	.50	.50
0.65	.01	.02	.04	.05	.07	.13	.16	.18	.21	.21	.29	.29	.32	.33	.33	.35	.36	.38	.38	.38
0.6	.01	.02	.03	.04	.06	.10	.13	.15	.18	.18	.24	.24	.26	.27	.27	.28	.30	.32	.32	.32
0.55	.01	.01	.02	.03	.04	.08	.10	.12	.14	.14	.18	.18	.20	.20	.21	.22	.23	.25	.25	.25
0.5	.01	.01	.01	.01	.02	.03	.04	.04	.06	.06	.08	.08	.09	.10	.10	.10	.11	.12	.12	.12
0.45	.01	.01	.01	.01	.01	.03	.04	.04	.05	.05	.07	.08	.08	.09	.09	.09	.10	.11	.11	.11
0.4	.01	.01	.01	.01	.02	.03	.04	.04	.05	.05	.06	.06	.07	.07	.07	.07	.07	.08	.08	.08
0.35	.01	.01	.01	.01	.02	.03	.04	.04	.05	.05	.06	.06	.07	.07	.07	.07	.07	.07	.07	.07
0.3	.01	.01	.01	.01	.02	.02	.03	.03	.05	.05	.06	.06	.07	.07	.07	.07	.07	.07	.07	.07
0.25	.01	.01	.01	.01	.02	.03	.03	.03	.04	.04	.06	.06	.07	.07	.07	.07	.07	.07	.07	.07
0.2	.01	.01	.01	.01	.01	.02	.02	.02	.03	.03	.04	.04	.05	.05	.05	.05	.05	.06	.06	.06
0.15	.01	.01	.00	.01	.01	.02	.02	.02	.02	.02	.03	.03	.04	.04	.04	.04	.04	.05	.05	.05
0.1	.01	.01	.01	.01	.01	.01	.02	.02	.02	.02	.03	.04	.04	.04	.04	.04	.04	.05	.05	.05
0.05	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.02	.02	.02	.02	.02	.02	.02	.03	.03	.03

F_1-measure

	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1.0
1.0	.01	.04	.09	.16	.24	.36	.41	.44	.52	.52	.68	.68	.70	.70	.72	.74	.71	.71	.71	.71
0.95	.01	.04	.09	.16	.24	.36	.41	.44	.52	.52	.68	.68	.70	.70	.72	.74	.71	.71	.71	.71
0.9	.01	.04	.10	.16	.25	.38	.45	.48	.56	.56	.71	.71	.74	.74	.76	.77	.75	.75	.75	.75
0.85	.01	.04	.09	.15	.23	.40	.48	.51	.57	.57	.75	.75	.78	.79	.79	.80	.78	.77	.77	.77
0.8	.01	.04	.10	.14	.22	.37	.45	.48	.54	.54	.71	.71	.74	.75	.76	.77	.74	.73	.73	.73
0.75	.02	.04	.09	.12	.18	.31	.38	.41	.45	.45	.62	.62	.64	.65	.66	.68	.66	.66	.66	.66
0.7	.02	.04	.08	.12	.18	.29	.36	.39	.43	.43	.58	.59	.61	.62	.62	.64	.62	.62	.62	.62
0.65	.02	.04	.06	.09	.13	.21	.26	.29	.33	.33	.43	.44	.46	.47	.48	.50	.50	.52	.52	.52
0.6	.02	.03	.05	.07	.10	.18	.22	.25	.29	.29	.37	.38	.40	.41	.41	.43	.44	.46	.46	.46
0.55	.02	.03	.04	.05	.07	.14	.18	.20	.23	.23	.29	.30	.32	.33	.33	.34	.35	.38	.38	.38
0.5	.01	.01	.02	.03	.03	.06	.08	.08	.10	.10	.15	.15	.16	.17	.17	.18	.19	.20	.20	.20
0.45	.01	.01	.02	.02	.03	.05	.07	.07	.09	.09	.13	.13	.15	.15	.15	.16	.17	.18	.18	.18
0.4	.01	.01	.02	.03	.03	.06	.07	.07	.09	.09	.11	.11	.13	.12	.12	.13	.13	.14	.14	.14
0.35	.01	.01	.02	.02	.03	.06	.07	.07	.09	.09	.11	.12	.12	.12	.12	.13	.13	.13	.13	.13
0.3	.01	.01	.02	.03	.03	.04	.06	.06	.09	.09	.11	.11	.12	.12	.12	.13	.13	.13	.13	.13
0.25	.02	.02	.02	.03	.03	.05	.06	.06	.08	.08	.11	.11	.12	.12	.12	.12	.12	.13	.13	.13
0.2	.02	.02	.01	.02	.02	.03	.04	.05	.06	.06	.07	.08	.08	.08	.08	.09	.09	.10	.10	.10
0.15	.01	.01	.01	.01	.02	.03	.04	.04	.04	.04	.06	.06	.06	.06	.06	.07	.07	.08	.08	.08
0.1	.02	.02	.01	.02	.01	.02	.04	.04	.04	.04	.06	.06	.06	.06	.06	.06	.07	.07	.07	.07
0.05	.02	.01	.01	.01	.01	.01	.02	.02	.02	.02	.02	.02	.03	.03	.03	.03	.03	.03	.03	.03

Figure 8: Evaluation of event detection (recall, precision and F-Measure) on the cumulated corpora (el, en, pl, ru and zh) for different combinations of θ-disease and θ-location. The lighter the squares, the better DANIEL performs with the corresponding parameters.

in the form of disease-location pair. Unfortunately, only 102 source web pages (among 200) were still available online at the time of this publication (100 in English, one in Russian and one in French). The evaluation in [27] is done using English reports and news articles, whereas DANIEL is specialized in processing news only. DANIEL has been evaluated on 102 source web pages using merged resources as described in Section 5.3.3 and θ-disease =

θ-location = 0.80, the standard configuration of the system.

First, the performance for the document filtering task is evaluated. The precision measure is inappropriate since all documents are tagged as relevant. The recall is 0.88 (90/102). This figure is comparable to the results presented in our own reference corpus (Table 6).

This figure can not be compared with BIOCASTER since to our knowledge the authors did not report this kind of evaluation. This is probably due to the fact that the corpus was designed for event-wise evaluation only. However, it is interesting to give some insights into the 12 misclassified documents (all in English). First, 6 of them concerned events that were not included in DANIEL guidelines (bacterial infections and diseases affecting animals). Second, 4 documents were misclassified because they did not fulfill the genre requirements: they were reports from the program for monitoring emerging diseases (PROMED). DANIEL is designed to process press articles whilst 23 documents are PROMED reports[18]. With the aim of transmitting information as quickly as possible, DANIEL is a good alternative. It annihilates the delay in writing reports about an epidemic issue.

Finally, an event-wise evaluation was performed. For 81% of the documents the appropriate disease-location pair was detected. For unique events, the performances are better than those obtained with our own corpus: 0.85 for recall and 0.88 for precision. The recall is lower (0.85 *vs.* 0.93) but the precision is very high (0.88 *vs.* 0.80, Figure 8). The F_1-measure increases by 0.07 with a 0.87 score. The BIOCASTER system obtained an even better score with 0.94 F_1-measure [28]. Considering the fact that we expose a simpler and more multilingual scheme, this is a very good result. The heatmaps showed that DANIEL achieves comparable results even for poorly endowed languages.

6. Discussion

6.1. Objective

The challenge in health surveillance is to ensure world coverage. The current approach is to multiply dedicated systems for each language, but resources are lacking for a very large number of them. The richest state-of-the-art system handles 10 languages, whereas there are about 6,000 languages

[18] from the human-produced reports available at `http://www.promedmail.org` (Accessed: 20 April 2015)

in the world, 300 of which are spoken by more than one million people. The principles of a *genre*-based IE system called DANIEL have been tested on 17 languages and evaluated on 5 languages: Chinese, English, Greek, Polish and Russian. The system relies on light, easy-to-obtain resources, and is intended to help health authorities gather information about on-going infectious diseases spreading throughout the world. In order to be multilingual, it uses news *genre*-related features. Carefully selected types of string repetitions are used as clues to relevance of a document. Experiments show that the system is lacking in precision, but has a good recall (0.89 for English, 0.91 for the whole corpus), an excellent result for global online epidemic surveillance.

6.2. Contribution

The DANIEL algorithm is based on the rhetorical construction of news articles, unlike state-of-the-art systems relying on extended lexicon and syntactic parsing. It focuses on *where* the useful information should be rather than on *what* it should be. The detection of string repetitions in texts might seem costly, however, cost is curbed by exploring only salient positions in the text. The longer the document, the more constrained the search space. In short news pieces, the beginning and end cover the whole text (technically there is no middle), but in longer news pieces, the middle is larger. Another savings in processing costs is the fact that an external list of disease names from WIKIPEDIA is used to filter repetition candidates.

Given these constraints, is DANIEL *truly* language-independent? Language-independence is relative and may be identified with respect to three characteristics:

1. Consistency in journalistic style;
2. Establishment of a knowledge base from easily accessible resources;
3. Determination of a parametric model of language (θ-disease, θ-location).

We have proposed a simple and effective model for comparing lexical entries that can have several forms in a document. The parameters θ take variability of prefixes/suffixes into account to find the largest root occurring in a document and in lexical resources. This parameter shows low variability in the languages processed. A single value of θ (0.80) can be used to cover different languages for both θ-disease and θ-location. DANIEL factorises the diversity of the entries in the knowledge base. This can be explained by the journalistic-genre assumption (well-known terms are used in news wires) and by the specificity of entries in the knowledge base. Medical terms have a large invariant root from their Greek and Latin origin. Place names also have a relative invariant basis in a given alphabet. It might be argued that Daniel results are not directly linked to specialized medical databases through the Unified Medical Language System, ICD or any other nomenclature. For example, in Figure 3, the detection of tuberculosis does not necessarily lead to feed databases using the proper entry (TDR-TB) in an ontology. Post-processing would be needed to achieve this goal.

Inflections can affect several words in multiword entries. Therefore, relevant substrings between a document and this kind of units are harder to detect. For instance, “**птичь**его **грипп**а” and “**птичь**им **грипп**ом” are two inflections of avian flu, found in a relevant Russian article[19]. To tackle this problem, the motif extraction module might be shifted to a *gapped-motif* extraction module [21]. The detection of these patterns is greedy, but the complexity of their enumerations can be channeled by limiting the maximum size (in number of characters) of gapped-motif considered. The longest gapped-motifs cannot be longer than the longest entry in the knowledge base.

6.3. Conclusion

Daniel is a text *genre*-based IE system devoted to news. It is efficient at distinguishing irrelevant documents in epidemic surveillance and at filtering streams of documents with low-resourced languages. When no classical IE system is available or training data is scarce, Daniel can fill the gap efficiently. The method described increases coverage in number of languages at low cost, rather than optimizing results with a particular language. Wikipedia is used to screen some common disease names to be matched with repeated character strings. The language variations, such as declensions, are handled by processing text at the character level, rather than at the word level. This additionally allows Daniel to handle various writing systems in a similar fashion.

With an average F_1-measure of 0.85, Daniel scores are below state-of-the-art systems (Puls or Biocaster), as we confirmed with our comparative evaluation over the BEcorpus. However, the resources that these systems require (lexicon, language parser, ontologies) are far more extensive and costly to acquire. Daniel makes it possible to immediately process new languages if a list of disease names is provided. A list of locations is not a strict requirement since the implicit location rule of Daniel performs well.

[19] `https://daniel.greyc.fr/public/index.php?id=1577` (Accessed: 20 April 2015)

Daniel results have demonstrated great promise in multilingual EE at minimal marginal cost. Further research on document structure and segmentation will lead to more refined rhetorical rules. It is also possible to build a hybrid system in which Daniel will filter relevant documents from a general news feed. A language detector and filter could then direct documents in dominant languages to a classical EE system that achieves high precision. High precision with a language can provide a more precise tag to a cluster of related documents [17].

In order to advance EE research, the corpora used for these experiments are available to the community with annotations detached from original URLs. News corpora in Arabic, French, Portuguese, Spanish, Swahili etc. are being annotated to assess Daniel's quality in a wider range of languages as part of the effort to improve multilingual world coverage.

References

[1] S. Doan, Q.-H. Ngo, A. Kawazoe, N. Collier, Global Health Monitor – a Web-based System for Detecting and Mapping Infectious Diseases, in: J.-S. Chang (Ed.), Proceedings of the 3rd International Joint Conference on Natural Language Processing: Volume-II, Association for Computational Linguistics, Hyderabad, India, 2008, pp. 951–956.

[2] R. Steinberger, A survey of methods to ease the development of highly multilingual text mining applications, Language Resources and Evaluation 46 (2011) 155–176.

[3] M. Keller, M. Blench, H. Tolentino, C. Freifeld, K. Mandl, A. Mawudeku et al., Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance, Emerging Infectious Diseases 15 (2009) 689–695.

[4] G. Lejeune, R. Brixtel, C. Lecluze, A. Doucet, N. Lucas, Added-Value of Automatic Multilingual Text Analysis for Epidemic Surveillance, in: N. Peek, R. M. Morales, M. Peleg (Eds.), Proceedings of the 14th Conference on Artificial Intelligence in Medicine, Lecture Notes in Computer Science, Springer, Murcia, Spain, 2013, pp. 284–294.

[5] B. Webber, A. Joshi, Discourse Structure and Computation: Past, Present and Future, in: R. E. Banchs (Ed.), Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, Asso-

ciation for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 42–54.

[6] J. Hobbs, Generic Information Extraction System, in: B. Sundheim (Ed.), Proceedings of the 5th conference on Message Understanding Conference, Association for Computational Linguistics, Baltimore, Maryland, 1993, pp. 87–91.

[7] M. Du, P. Von Etter, M. Kopotev, M. Novikov, N. Tarbeeva, R. Yangarber, Building support tools for Russian-language information extraction, in: Habernal, I. and Matoušek, V. (Ed.), Proceedings of the 14th international conference on Text, Speech and Dialogue, Springer, Pilsen, Czech Republic, 2011, pp. 380–387.

[8] N. Collier, Towards cross-lingual alerting for bursty epidemic events, Journal of Biomedical Semantics 2 (2011) 1–11.

[9] C. C. Freifeld, K. D. Mandl, B. Y. Reis, J. S. Brownstein, HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports, Journal of the American Medical Informatics Association 15 (2008) 150–157.

[10] O. Etzioni, A. Fader, J. Christensen, S. Soderland, Open Information Extraction: The Second Generation, in: T. Walsh, NICTA, University of NSW (Eds.), Proceedings of the 22nd International Joint Conference on Artificial Intelligence, AAAI Press, Barcelona, Spain, 2011, pp. 3–10.

[11] R. Munro, Processing short message communications in low-resource languages, Ph.D. thesis, Stanford University, 2012.

[12] R. Steinberger, M. Ehrmann, J. Pajzs, M. Ebrahim, J. Steinberger, M. Turchi, Multilingual Media Monitoring and Text Analysis – Challenges for Highly Inflected Languages, in: I. Habernal, V. Matousek (Eds.), Proceedings of the 16th international conference on Text, Speech and Dialogue, Lecture Notes in Computer Science, Springer, Plzeň, Czech Republic, 2013, pp. 22–33.

[13] F.-J. Tsai, E. Tseng, C.-C. Chan, H. Tamashiro, S. Motamed, A. Rougemont, Is the reporting timeliness gap for avian flu and H1N1 outbreaks in global health surveillance systems associated with country transparency?, Globalization and Health 9 (2013) 14–21.

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

[15] K. W. Church, Empirical estimates of adaptation: the chance of two Noriegas is closer to $\frac{p}{2}$ than p^2, in: M. Kay (Ed.), Proceedings of the 18th conference on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2000, pp. 173–179.

[16] G. Lejeune, A. Doucet, R. Yangarber, N. Lucas, Filtering news for epidemic surveillance: towards processing more languages with fewer resources, in: S. Sarkar, M. Zhang, A. Lopez, R. Udupa (Eds.), Proceedings of the 4th Workshop on Cross Lingual Information Access, Association for Computational Linguistics, Beijing, China, 2010, pp. 3–10.

[17] S. Liao, R. Grishman, Using document level cross-event inference to improve event extraction, in: J. Hajič, S. Carberry, S. Clark, J. Nivre (Eds.), Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 789–797.

[18] Z. Chen, S. Tamang, A. Lee, H. Ji, A toolkit for knowledge base population, in: W.-Y. Ma, J.-Y. Nie, R. Baeza-Yates, C. Tat-Seng, W. Croft (Eds.), Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, Association for Computational Linguistics, Beijing, China, 2011, pp. 1267–1268.

[19] J. Piskorski, J. Belyaeva, M. Atkinson, Exploring the Usefulness of Cross-lingual Information Fusion for Refining Real-time News Event Extraction: A Preliminary Study, in: G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov (Eds.), Proceedings of Recent Advances in Natural Language Processing, Association for Computational Linguistics, Hissar, Bulgaria, 2011, pp. 210–217.

[20] R. Brixtel, G. Lejeune, A. Doucet, N. Lucas, Any Language Early Detection of Epidemic Diseases from Web News Streams, in: C. Yang (Ed.),

Proceedings of the 15th International Conference on Healthcare Informatics, Institute of Electrical and Electronics Engineers, Philadelphia, PA, USA, 2013, pp. 159–168.

[21] E. Ukkonen, Maximal and minimal representations of gapped and non-gapped motifs of a string, Theoretical Computer Science 410 (2009) 4341–4349.

[22] J. Kärkkäinen, P. Sanders, S. Burkhardt, Linear work suffix array construction, Journal of the Association for Computing Machinery 53 (2006) 918–936.

[23] 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 40 (2007) 405–413.

[24] S. Morse, Public health surveillance and infectious disease detection, Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science 10 (2012) 6–16.

[25] P. Barboza, L. Vaillant, A. Mawudeku, N. Nelson, D. Hartley, L. Madoff et al., Evaluation of Epidemic Intelligence Systems Integrated in the Early Alerting and Reporting Project for the Detection of A/H5N1 Influenza Events, PLoS ONE 8 (2013) e57252.

[26] J. Piskorski, J. Belayeva, M. Atkinson, On Refining Real-Time Multilingual News Event Extraction through Deployment of Cross-Lingual Information Fusion Techniques, in: N. Memon, D. Zeng (Eds.), Proceedings of the 2nd European Intelligence and Security Informatics Conference., Institute of Electrical and Electronics Engineers, Athens, Greece, 2011, pp. 38–45.

[27] M. Conway, A. Kawazoe, H. Chanlekha, N. Collier, Developing a disease outbreak event corpus, Journal of medical Internet research 12 (2010).

[28] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno et al., Biocaster: detecting public health rumors with a web-based text mining system, Bioinformatics 24 (2008) 2940–2941.