DAnIEL: Language Independent Character-Based News Surveillance

Gaël Lejeune, Romain Brixtel, Antoine Doucet, Nadine Lucas

DAnIEL: Language Independent Character-Based News Surveillance

Gaël Lejeune†, Romain Brixtel†, Antoine Doucet†, Nadine Lucas†

†GREYC, University of Caen Lower-Normandy
Boulevard du Maréchal Juin BP5186 – 14032 Caen Cedex, France
firstname.lastname@unicaen.fr

Abstract. This study aims at developing a news surveillance system able to address multilingual web corpora. As an example of a domain where multilingual capacity is crucial, we focus on Epidemic Surveillance. This task necessitates worldwide coverage of news in order to detect new events as quickly as possible, anywhere, whatever the language it is first reported in. In this study, text-genre is used rather than sentence analysis. The news-genre properties allow us to assess the thematic relevance of news, filtered with the help of a specialised lexicon that is automatically collected on Wikipedia. Afterwards, a more detailed analysis of text specific properties is applied to relevant documents to better characterize the epidemic event (i.e., which disease spreads where?). Results from 400 documents in each language demonstrate the interest of this multilingual approach with light resources. DAnIEL achieves an F_1-measure score around 85%. Two issues are addressed: the first is morphology rich languages, e.g. Greek, Polish and Russian as compared to English. The second is event location detection as related to disease detection. This system provides a reliable alternative to the generic IE architecture that is constrained by the lack of numerous components in many languages.

1 Introduction

Information Extraction (IE) aims at extracting structured views from free text and particularly from newswires. The Web provides many news sources in a variety of languages, and for instance the European Media Monitor collects about 40,000 news reports in 43 languages each day[1].

This paper focuses on multilingual IE with light resources and uses as application the epidemiological Event Extraction from the Web, a subdomain of IE whose goal is to detect and extract health-related events from news to send alerts to health authorities [1]. Tapping a wealth of information sources makes it theoretically possible to quickly detect important epidemic events over the world [2]. A health authority will want to monitor information with emphasis on disease outbreaks [3]. Until now, several approaches have been reported for

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

epidemic surveillance on the Web [4] from full human analysis [5], keyword analysis [6] and web mining [7]. Human analysis is supposed to be more precise but has a great cost; keyword analysis is cheaper but lacks 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 at sentence level (morphology, syntax, semantics). It has proved its high efficiency for applications in some important languages [8,4]. But 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, the coverage problem is still unsolved.

The approach advocated here 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 on repetition patterns at certain places in text [10]. Though the rationale is different, technically the method is similar to relation discovery in open information extraction on the Web [11]. It also uses light crawled resources. Furthermore, its algorithmic basis permits a quick processing of large collections of documents.

The paper is organised as follows. In Section 2, we provide an overview of the multilingual approaches in IE. In Section 3 we present a system called D*ata* An*alysis for* I*nformation* E*xtraction in any* L*anguage* (DAnIEL), a genre-based IE system designed for managing multilingual news. In Section 4, we describe the corpus collected for this experiment. In Section 5 we show results and we elaborate on some of the results obtained on this corpus. Lastly, we conclude with a few additional remarks in Section 6.

2 Related work

Use of the generic IE chain [12] as a model requires numerous and diverse components for each language. Components corresponding to a new language must be gathered or constructed. Two systems relying primarily on English, PULS[2] [6] and BIOCASTER[3] [3] are used as 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 news feed. For most languages, efficient components will be lacking [13]. In recent years, machine learning was successfully used to fill gaps when one can find sufficient training data in a language which has enough common properties with the new one [11].

However, in epidemic surveillance, there is need to cover even very scarce resource languages 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. The detection and appropriate analysis of the very first news relating to an epidemic event is crucial, but it may occur in any language:

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

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. It relies on what can be called either pragmatics, or genre properties related to news discourse.

3 Our system: DAnIEL

The DAnIEL system presents a full implementation of a discourse-level IE approach. It operates at text-level, because it exploits the global structure of news in a newswire, that is information ordering as defined in [10], as opposed to the usual analysis of sentence-level linguistic layers (morphology, syntax and semantics). Entries in the system are text news, with their title and text-body. The details of the model are not justified here, but the main points as far as implementation are concerned are defined. Character-based refers to the fact texts are 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 sequences that are extracted are not key words but machine-tractable strings that are linked to their order of appearance in text, paragraph after paragraph. A special interest has been put on describing the overall system as well as evaluating each part of it. 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 in what country). Time is also important but will not be explained here for lack of space.

The system description is split in five subsections. DAnIEL uses a small knowledge base (Section 3.1) and its processing pipeline contains four steps: news article segmentation (Section 3.2), motifs extraction (Section 3.3), event detection (Section 3.4) and event localization (Section 3.5)

3.1 Knowledge base

DAnIEL uses implicit knowledge on news writing and reading. Information is displayed carefully in news. The rules that are useful here are that information is displayed at important places, called positions. In journalistic style, writers use common disease names, because all newspaper readers know them. Media style rules also say they will appear before more specialised words in a piece of news. Last, important information is repeated, probably twice. Some similar observations are stated in different studies based on pragmatics or statistical studies (estimation of positive adaptation), notably on proper names [14].

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 lexicon contains disease common names and some geographical names (countries). The lexicon needed with text-genre-based IE is quite small: roughly hundreds of items instead of tens of thousands in IE systems [15]. Indeed, Web-extracted disease names prove useful for dealing quickly with new languages, even without the assistance of a native speaker.

3.2 Article segmentation

The main algorithm relies on the type of article being processed. The segmentation is important: as the approach is style-driven, having good judgement about the way the text is constructed is crucial. Key positions are the beginning and the end of text. For analysing press articles, the system relies on the title and beginning (the topical head) and checks which elements are repeated at key positions in the text. Because the hypothesis needed to be tested first, a coarse simplification was made to handle text length. Table 1 shows the three types of text according to length and the text windows corresponding to the text highlighted positions. Repetition are looked for in : Head (title plus the first paragraph), Tail (last two paragraphs) and Body (news article minus the Head).

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

Table 1. Article segmentation with respect to their number of paragraphs

For medium and long articles, the system extracts the substrings repeated in Head plus Body and Head plus Tail. For short articles, repeated substrings are considered irrespective of their position.

3.3 Motifs extraction

The system checks repetitions at given positions in text, mainly beginning and end of text or text sub-units (paragraphs). To achieve that, character level analysis is allowed by computing non-gapped character strings as described by Ukkonen under the name motifs [16]. The main ideas are given in this section to enumerate those motifs in one or more text. Those motifs are substrings patterns of text with the following characteristics :

repeated: motifs occur twice or more;
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 each occurrence of `ATTI`. In other words its right-context (the characters on the right of all the occurrences of `TT`) is always `I` and its left-context `A`. All of these motifs in a set of strings are enumerated using an augmented suffix array [17].

For two strings $\mathcal{S}_0 =$ `HATTIV` and $\mathcal{S}_1 =$ `ATTIAA`, both string in Σ^*, Table 2 shows the augmented suffix array of $\mathcal{S} = \mathcal{S}_0.\$_1.\mathcal{S}_1.\$_0$. $\$_0$ and $\$_1$ are lexicographically lower than any character in Σ and $\$_0 < \$_1$. Augmented suffix array consists in the list of suffixes sorted lexicographically of $\mathcal{S}$ (SA) and the Longest Common Prefix (LCP) between suffixes two at a time consecutively 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 $\mathcal{S}$ = HATTIV$\$_1$ATTIAA$\$_0$

The LCP allows the detection of repetitions, for example, the substring ATTI occurrs at the offsets $(1, 13)$ in $\mathcal{S}$ according to LCP_4 in Table 2. The process enumerates all the repeated substrings by reading through LCP.

- $LCP_i < LCP_{i+1}$: *open* a potential motif occurring at the offset SA_{i+1}
- $LCP_i > LCP_{i+1}$: *close* motifs previously created
- $LCP_i = LCP_{i+1}$: *sustain* 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 $(4, 10)$. 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 are different according to their 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 $\mathcal{S}_1$. SA and LCP are constructed in $O(n)$ time [17], the enumeration process is done in $O(k)$ time, with k defined as the number of motifs and $k < n$ [16][4].

3.4 Event detection

DAnIEL filters out motifs according to article segmentation rules as described in Table 1, and to the list of disease names as explained in Section 3.1. It keeps motifs that are substrings found in two different sub-units, typically Head and Tail, and matching at least with one disease name. This comes from the genre-related rules saying that an important topic is highlighted in news, that common names are used to catch the reader's attention and that the topic is repeated.

[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/

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 3.3) and keeps motifs that occurs in $\mathcal{S}_0$, $\mathcal{S}_1$ and any $\mathcal{S}_{1<i\leq n+1}$. 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 the number of characters in m and $\mathcal{S}_i$. The value of θ is discussed in section 5.2. This proves especially useful for morphologically rich languages; the need for a morphological analyzer is thus avoided. If DAnIEL finds no motif that matches its knowledge base using the θ threshold, it considers the document contains no event and thus is not relevant.

3.5 Event localization

An event is minimally defined as a relation between a disease name, highlighted by its position and a place name. 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 explicit place names are found in the same way disease names are found, with the help of a reduced list 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 (as seen in Section 3.4) and the list of geographical names, the location of the event is assumed to be the country of issue of the source by default (i.e., the newspaper or news agency country).

4 Corpus

Since the method that is tested considers full news, including title and text-body, no shared corpus was available. A corpus was to be collected in various languages from the Web. News corpora for 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]. We resorted to such pre-filtered sources because they are available. Even in health categories, however, only 8% of documents contained epidemic events. This strategy thus permitted to collect a significant number of relevant documents at a reasonable cost.

For measuring precision and recall on document filtering and event characterization, native speakers of each language[6] annotated sets of about 500 documents covering a 8 week period from November 2011 to January 2012.

The characteristics of the evaluation corpus are shown in Table 3. The main characteristic is the fact that the length (in paragraph or characters) may vary a

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

[6] Eight professional translators who were never involved in DAnIEL

	English	Greek	Polish	Russian
#documents (relevant)	475 (31)	390 (26)	390 (30)	426 (40)
#paragraphs	10419	4216	4986	3565
avg.±std.	21.9±8.89	13.8±10.22	12.82±9.34	8.37±8.33
#characters (10^6)	1.60	2.09	1.19	1.64
avg.±std.	3372±1796	5382±5001	3059±2032	3871±5902

Table 3. Characteristics of the corpus

lot from one document to another. Annotators had to judge if these documents were relevant for informing health authorities about infectious diseases. If they judged them relevant, they had to further give the disease and location. This annotated corpus and the full annotation guidelines are available online[7]. The corpus is freely available for the community for further experiments.

5 Results and evaluation

This section first highlights the efficiency of the repetition rule at key positions to select relevant press articles. Then DAnIEL is evaluated against annotators' judgements on the evaluation corpus. The program to run the experiments, written in PYTHON, processes 2000 documents in less than 15 seconds (2.4Ghz dual core processor, 2Gb RAM), which is compatible with on-line surveillance.

5.1 Global results

It is difficult to measure recall and precision when large amount of documents are processed, here the ground truth is the set of documents independent annotators found relevant. The F-measure is calculated with $\beta = 1$.

	English	Greek	Polish	Russian	cumulated corpora
θ	0.85	0.75	0.8	0.85	best θ per language
Precision	0.77	0.76	0.73	**0.85**	0.78
Recall	0.97	**1.0**	0.85	0.88	0.93
F_1-measure	**0.86**	**0.86**	0.79	**0.86**	0.85

Table 4. Document filtering: precision, recall, F_1-measure for best θ

Table 4 shows that recall is slightly better than precision. In this table, a different θ ratio was used for each language. Tuning the best ratio for each language permitted DAnIEL to achieve 0.78 in precision with a 0.93 recall, for the cumulated corpus. This is unexpected, because it was feared that choosing to use a small lexicon would 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. Table 5 shows the extent to which DAnIEL generates silence and the reasons for errors.

[7] `https://lejeuneg.users.greyc.fr/daniel/`

	English	Greek	Polish	Russian
#of relevant documents	31	26	30	40
Lack in lexicon	0	0	2	3
No repetition	1	0	1	1
Wrong matching	0	0	2	0
Silence	1	0	5	4

Table 5. Document filtering, errors impairing recall

Errors due to the size of the lexicon are quite rare (5 are missed) and the repetition phenomenon is trustworthy: only three relevant documents were missed because no repetition matching with the disease name was found. More errors came from string recognition, because some diseases are referred by short names (in number of characters) and DAnIEL was unable to detect whether a disease was involved.

The news discourse model implemented through repetition rules at special positions efficiently selects relevant press articles on epidemiologic events. Figure 1 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 filter documents out: 97% of irrelevant and only 0.7% of relevant articles contained no repetition.

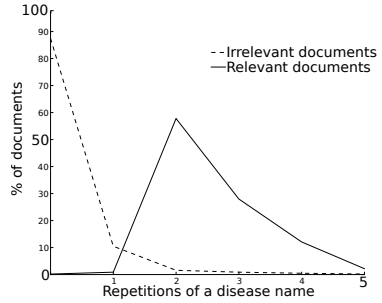

Fig. 1. Repetitions of disease name in relevant and irrelevant articles

5.2 Detailled evaluation

Segmentation filtering. The news segmentation described in Section 3.2 is intended to filter out uninteresting motifs. Table 6 shows the impact of this filtering in the total number of motifs.

Filtering relevant documents. In order to ponder the different features of our system, Table 7 shows the performance of two baselines: B1 relies on the presence of a disease name in the document and B2 relies on the repetition of the disease name. B1 highlights the problems one can encounter with morphologically rich

	English	Greek	Polish	Russian
#documents	396	159	192	90
#motifs without segmentation (avg.)	1101.45	1242.81	1128.12	1311.07
#motifs with segmentation (avg.)	114.67	143.33	129.05	159.72
Filtering rate	9.60	8.67	8.74	8.20

Table 6. Assessment of filtering impact, number of motifs for medium and long articles

languages because of the exact matching needed for the disease name. B2 shows the improvement in precision with the use of repetitions. Both baselines ignore the position criterion (Section 3.2) with $\theta = 1$ (Section 3.4).

		English	Greek	Polish	Russian	All
Baseline 1 (B1)	Precision	0.17	0.69	0.33	0.59	0.43
	Recall	**1.0**	**0.62**	**0.79**	**0.61**	**0.68**
	F_1-measure	0.29	0.65	0.47	0.60	0.53
Baseline 2 (B2)	Precision	**0.33**	**0.76**	**0.48**	**0.74**	**0.63**
	Recall	0.97	0.45	0.56	**0.61**	0.60
	F_1-measure	0.49	0.57	0.52	0.67	0.61

Table 7. Evaluation of two baselines: precision, recall and F_1-measure

Evaluating the threshold $\frac{len(motif)}{len(entry)} \geq \theta$. Figure 3 shows the results of empirical experiments to determine the appropriate string matching ratio between motifs extracted and knowledge base entry. $\theta = \frac{4}{5}$ is a good empirical value for processing the four different languages simultaneously and it might be optimized individually for each language (for example, 0.85 in Russian (Figure 2)).

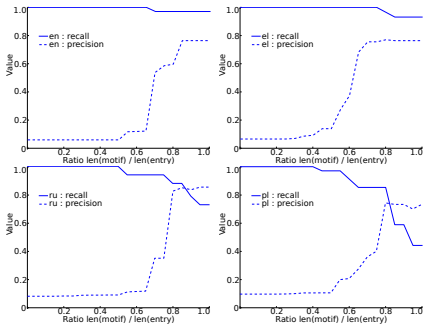

Fig. 2. Recall and precision according to θ (English, Greek, Russian and Polish)

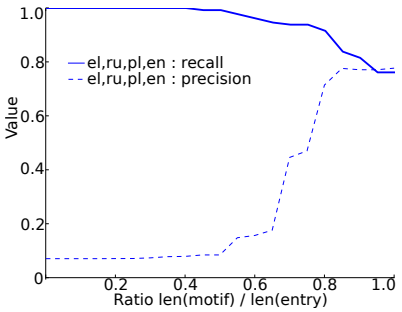

Fig. 3. Recall and precision according to θ (all languages)

Event localization. A large corpus (2000 documents for each language) was processed by DAnIEL. Then, a subcorpus of relevant documents without explicit location was extracted. Those documents have been checked to assess if linking the events they describe to the source location is acceptable (Table 8).

	English	Greek	Polish	Russian
# documents DAnIEL found relevant	93	188	213	230
# relevants documents without explicit location	46	33	35	51
Location = source	78.3%	81.8%	82.9%	78.4%
Location ≠ source	21.7%	18.2%	17.1%	21.6%
Overall errors	12.2%	3%	2.8%	4.8%

Table 8. Performance of the implicit location rule

In this corpus, roughly 70% of epidemic events contained an explicit location. Therefore results obtained show that the "implicit location" rule is efficient. For instance in Russian, among the 22% of documents where no location is explicitly mentioned, 78% are accurately localized with this simple rule. That leaves only 4.8% of all events incorrectly localized in Russian news.

Level of evaluation unit: document or event? Evaluation per document is not necessarily adequate, when one considers a typical use case [18]. One can 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). This should not be valued at 0.99 recall for the end-user. To evaluate how DAnIEL performs with respect to events rather than documents, further event-based annotations are made. Each disease-location pair (flu in Spain for instance) was considered as a unique epidemiological event regardless of the number of documents it has been reported in, over a 8 week time window.

	Unique events	Detected	Missed
English	15	14	1 (6,6%)
Greek	17	17	0 (0%)
Polish	28	26	2 (7,1%)
Russian	23	21	2 (8,6%)
Total	57	54	3 (5,2%)

Table 9. Evaluation by unique event

Table 9 shows results of this experiment, demonstrating that only few full-fledged events (3 among 57) were missed. The system takes advantage of the fact that it has coverage in more than one language to detect events [19]. For instance, an event missed in Polish had been detected in Russian. Note that the total number of unique events in Table 9 is not the sum of unique events in reports, since a single epidemiological event can be reported in several languages.

6 Conclusion

The principles of a genre-based information extraction system called DAnIEL have been tested with success on English, Greek, Polish and Russian news. The system relies on very light, easy to get resources, and it is intended to help health authorities get precious information about on going infectious diseases spreading all around the world. In order to be multilingual, it uses genre related features and relies on text-style, specifically 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 local analysis is used and a limited-size lexicon is enough. Experiments showed that the system might lack in precision, but has good recall (0.97 for English, 0.92 for the whole corpus). DAnIEL is efficient at distinguishing relevant and really irrelevant documents, which makes it useful to filter large corpora, even with less known languages.

With an average F_1-measure of 0.85 with appropriate tuning, DAnIEL scores are below state-of-the-art systems like PULS or BIOCASTER, which are closer to 0.9 on English and a few other languages. But the resources that these systems need (lexicon, language parser, ontologies) are much more extensive and costly.

When no classical IE system is available or training data are too scarce, a text genre-based IE system can fill the gap efficiently. It can save efforts to filter relevant documents to be thoroughly parsed by existing techniques with high precision on 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.

References

1. Linge, J., Steinberger, R., Weber, T., Yangarber, R., van der Goot, E., Al Khudhairy, D., Stilianakis, N.: Internet surveillance systems for early alerting of threats. Eurosurveillance **14**(13) (2009)
2. Lyon, A., Nunn, M., Grossel, G., Burgman, M.: Comparison of web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap. Transboundary and Emerging Diseases (2011)
3. Son, D., Quoc, H.N., Ai, K., Collier, N.: Global health monitor - a web-based system for detecting and mapping infectious diseases. International Joint Conference on Natural Language Processing (2008) 951–956
4. Hartley, D.M., Nelson, N.P., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Bronstein, J.S., Thinus, G., Lightfoot, N.: The landscape of international event-based biosurveillance. Emerging Health Threats Journal **3**(e3) (2010)
5. Reilly, A.R., Iarocci, E.A., Jung, C.M., Hartley, D.M., Nelson, N.P.: Indications and warning of pandemic influenza compared to seasonal infliuenza. Advances in disease surveillance **5** (2008) 190
6. Steinberger, R., Fuart, F., van der Goot, E., Best, C., von Etter, P., Yangarber, R.: Text mining from the web for medical intelligence. In: Mining massive data sets for security, OIS Press (2008) 295–310

7. Huttunen, S., Arto, V., von Etter, P., Yangarber, R.: Relevance prediction in information extraction using discourse and lexical features. In: Nordic Conference on Computational Linguistics, Nodalida 2011. (2011) 114–121
8. Ji, H.: Challenges from information extraction to information fusion. In: Proceedings of the 23rd International Conference on Computational Linguistics. (2010) 507–515
9. Du, M., Von Etter, P., Kopotev, M., Novikov, M., Tarbeeva, N., Yangarber, R.: Building support tools for Russian-language information extraction. In: Proceedings of the 14th international conference on Text, Speech and Dialogue. (2011) 380–387
10. Lucas, N.: Stylistic devices in the news, as related to topic recognition. In Kwiatkowska, A., ed.: Texts and Minds : Papers in Cognitive Poetics and Rhetoric. Volume 26 of Łódź, Studies in language. Peter Lang, Frankfurt am Main (2012) 301–316
11. Etzioni, O., Fader, A., Christensen, J., Soderland, S.: Open information extraction: The second generation. Proceedings of the 22nd International Joint Conference on Artificial Intelligence (2011) 3–10
12. Hobbs, J.R.: The generic information extraction system. In: Proceedings of the 5th conference on Message understanding. MUC5 '93, Stroudsburg, PA, USA, Association for Computational Linguistics (1993) 87–91
13. Steinberger, R.: A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation (2011) 1–22
14. Church, K.: 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) 173–179
15. Collier, N., Ai, K., Jin, L., et al.: A multilingual ontology for infectious disease surveillance: rationale, design and challenges. Journal of Language Resources and Evaluation (2007) 405–413
16. Ukkonen, E.: Maximal and minimal representations of gapped and non-gapped motifs of a string. Theorie in Computer Science **410**(43) (2009) 4341–4349
17. Kärkkäinen, J., Sanders, P., Burkhardt, S.: Linear work suffix array construction. Journal of the ACM **53**(6) (2006) 918–936
18. Liao, S., Grishman, R.: Using document level cross-event inference to improve event extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. ACL '10 (2010) 789–797
19. Piskorski, J., Belyaeva, J., Atkinson, M.: 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) 38–45