

Text Mining from the Web for Medical Intelligence

Ralf STEINBERGER ^{a,1}, Flavio FUART ^a, Erik van der GOOT ^a, Clive BEST ^a,
Peter von ETTER ^b, and Roman YANGARBER ^b

^a *European Commission—Joint Research Centre, Ispra, Italy*

^b *Department of Computer Science, University of Helsinki, Finland*

Abstract. Global medical and epidemic surveillance is an essential function of Public Health agencies, whose mandate is to protect the public from major health threats. To perform this function effectively one requires timely and accurate medical information from a wide range of sources. In this work we present a freely accessible system designed to monitor disease epidemics by analysing textual reports, mostly in the form of news gathered from the Web. The system rests on two major components—MedISys, based on Information Retrieval technology, and PULS, an Information Extraction system.

Keywords. Information Retrieval, Information Extraction, multilinguality, medical intelligence, multi-document information aggregation

Introduction

Professionals in many fields need to sieve through large volumes of information from multiple sources on a daily basis. Most European Union (EU) countries have a national organisation that continuously monitors the media for new threats to the Public Health in their country, and for the latest events involving health threats. These threats range from outbreaks of communicable diseases and terrorism cases such as the deliberate release of biological or chemical agents, to chemical or nuclear incidents. Typically, the staff of these organisations search their national and local newspapers and/or purchase electronic news from commercial providers such as Factiva or Lexis-Nexis. Until recently, relevant news articles were cut out from printed press, and compiled into an in-house newsletter, which was then discussed among specialists who had to decide on the appropriate action. As more news sources became available on-line, it became easier to find relevant articles and to compile and manage them electronically. At the same time, the number of available sources rose, and—due to increased travel and the consequent importing of infectious diseases—it became necessary to monitor the news of neighbouring countries and major travel destinations.

These and similar professional communities can benefit from text analysis software that identifies potentially relevant news items, and thereby increases the speed and efficiency of their work, which is otherwise slow and repetitive. The search func-

¹E-mail address format: `firstname.lastname@jrc.it` or `@cs.helsinki.fi`

tions of news aggregators, such as Factiva or Google News, allow users to formulate Boolean search word combinations that filter items from large collections. The European Commission's *Medical Information System*, MedISys, in addition to providing keyword-based filtering, aggregates statistics about query matches, which enables it to provide early-warning signals by spotting sudden increases in media reports about any Public Health-related issue and alerting the interested user groups.

While this functionality is in itself helpful for the communities in question, deeper text analysis can provide further advantages, beyond those provided by classic Information Retrieval (IR) and alerting. In this chapter, we describe the IR and early-warning functionality of MedISys, and how it inter-operates with the information extraction (IE) system PULS, which analyses the documents identified by MedISys, retrieves from them *events*, or structured facts about outbreaks of communicable disease, aggregates the events into a database, and highlights the extracted information in the text. Our evaluation confirms that event extraction helps to narrow down the selection of relevant articles found in the IR step (improving precision), while on the other hand missing a small number of relevant articles (lowering recall).

The next section presents related work; sections 2 and 3 describe MedISys and PULS. Section 4 describes the mechanisms currently used for aggregating information from multiple reports. Section 5 shows how the two systems are combined into one Web application. Section 6 presents evaluation results. The final section draws conclusions and points to future work.

1. Related work

Information retrieval and information extraction have been thoroughly researched over the recent decades, with abundant literature on both topics. Typically they are studied separately, with results reported at different fora, and they are considered different problem areas, since they employ quite different methods. Conceptually, IR and IE both serve a user's information need, though they do so at different levels. It is understood that in real-world settings, IR and IE may be expected to interact, for example, in a pipeline fashion. The possibilities of tighter interaction largely remain yet to be researched.

Gaizauskas and Robertson ([1]) investigated the costs and benefits of combining IR and IE, by first applying a search engine (Excite) and its summary extraction tool, and then extracting MUC-6 "management succession" events, ([2]). The MUC-6 task is to track changes in corporate management: to find the manager's post, the company, the current manager's name, the reason why the post becomes vacant, and other relevant information about the management switch. The authors conclude that using IR as a filter before IE clearly results in a speed gain (since applying IE to *all* documents returned by the search engine would not have been possible), while the cost was a loss of 7% of the relevant documents. In further experiments by Robertson and Gaizauskas ([3], precision rose by 32% up to 100%, though at the cost of losing 65% of the retrieved relevant documents.

From an application point of view, to our knowledge, there are two other systems that attempt to gather information about infectious disease outbreaks from automatically collected news articles: Global Health Monitor [4] and HealthMap [5]. The systems provide map interfaces for visualising the events found.

Global Health Monitor follows about 1500 RSS news feeds hourly, and matches words in the new articles against a taxonomy of about 4300 named entities, i.e., 50 names of infectious diseases, 243 country names, and 4000 province or city names. For place names, the taxonomy contains latitude-longitude information. The 50 disease names are organised into an ontology with the properties relating to synonyms, symptoms, associated syndromes and hosts. The Global Health Monitor processing consists of four steps: (1) relevance decision (using Naïve Bayes classification); (2) named entity recognition (disease, location, person and organisation, using Support Vector Machine classification); (3) filtering of articles containing both disease and location names in the first half of the text. Additionally, only those disease-location pairs are retained that are frequently found in a separate *reference corpus*. Step (4) then visualises the successful matches on a map. Due to the rigorous filtering in steps (1) to (3), the system retains information on 25-30 locations and on about 40 infectious diseases a day. The system currently provides text analysis for English, though the underlying ontology includes terms in several other languages, including Japanese, Thai, and Vietnamese.

HealthMap monitors articles from the Google News aggregator and emails from the collaborative information-sharing portal ProMED-Mail,² and extracts information about infectious diseases and locations. After a *manual* moderation process, the results are stored in a database and visually presented on a map. Diseases and locations are identified in the text if words in the text exactly match the entities in the HealthMap taxonomy, which contains about 2300 location and 1100 disease names. Some disambiguation heuristics are applied to reduce redundancy (e.g., if the words “diarrhoea” and “shigellosis” are found, only the more specific entity “shigellosis” will be retained). HealthMap identifies between 20 and 30 disease outbreaks per day. More recent articles and those disease-location combinations reported in multiple news items and from different sources are highlighted on the map. The system developers point out the importance of using more news feeds, as their current results are focused toward the North-American continent. HealthMap currently displays articles in English, French, Spanish and Russian, ([5] describes English processing only).

The system presented in this chapter covers a large number of sources, a wide range of languages (currently, 43) and health-related topics (epidemic, nuclear, chemical and radiological incidents, bio-terrorism, etc.) Its special emphasis is on aggregation of the information collected from multiple sources and languages, and across time, and using the aggregation to provide additional functionality—for example, urgent warnings about unexpected spikes in levels of activity in a given area.

2. Information Retrieval in MedISys

The *Medical Information System*, MedISys, automatically gathers reports concerning Public Health in various languages from many Internet sources world-wide, classifies them according to hundreds of categories, detects trends across categories and languages, and notifies users. MedISys provides access at three levels: (1) free public access, (2) restricted access for Public Health professionals outside the European Commission (EC), and (3) full access inside the EC. The public MedISys site³ presents a quantitative sum-

²<http://www.promedmail.org>

³<http://medusa.jrc.it/>

mary of latest reports on a variety of diseases and disease types (e.g., respiratory infections), on bio-terrorism-related issues, toxins, bacteria (e.g., anthrax), viral hemorrhagic fevers (e.g., Ebola), viruses, medicines, water contaminations, animal diseases, Public Health organisations, and more. The restricted access site for non-Commission users offers more subject categories (e.g., covering nuclear and chemical contamination) and allows users to subscribe to daily, automatically-generated summary reports on various themes. The most complete functionality and coverage is at the Commission-internal site, which additionally monitors a number of copyright-protected news sources, such as Lexis-Nexis and about twenty news-wires.

MedISys aims to save users time, give them access to more news reports in more languages, and issue automatic alerts. An important feature of MedISys is that early-warning statistics are calculated considering the information aggregated from the news articles across all languages. MedISys is thus able to alert users of relevant events that may not yet be in the news of their country or language.

The development of MedISys was initiated by the European Commission's (EC) Directorate General *Health and Consumer Protection* (DG SANCO) for the purpose of supporting national and international Public Health institutions in their work on monitoring health-related issues of public concern, such as outbreaks of communicable diseases, bio-terrorism, and large-scale chemical incidents.⁴ The following sections cover the functionality of MedISys in more detail.

2.1. Collection and standardisation of Web documents

MedISys currently monitors an average of 50,000 news articles per day from about 1400 news portals around the world in 43 languages, from commercial news providers including 20 news agencies, Lexis-Nexis, and from about 150 specialised Public Health sites. The monitored sources were selected strategically with the aim of covering all major European news portals, plus key news sites from around the world, in order to achieve wide geographical coverage. Individual users can request the inclusion of additional news sources, such as local newspapers of their country, but these user-specific sources are normally processed separately in order to guarantee the balance of news sources and their types across languages.

Where available, MedISys collects RSS feeds. RSS ("Really Simple Syndication") is an XML format with standardised tags used widely for the dissemination of news and other documents. For other sources, scraper software looks for links on pre-defined Web pages, typically those pages that list the most recently published articles. The scraper automatically generates an RSS feed from these pages by means of specialised transformations. These transformations site-specific; they are currently produced and maintained manually, one separate transformation for each news site.

The grabber decides which of the articles in the RSS feed are new, by comparing the titles from earlier requests, and downloads the new articles. Since news pages contain

⁴MedISys users include supra-national organisations, such as the *European Commission*, the *World Health Organisation* (WHO) and the *European Centre for Disease Control* (ECDC), as well as national authorities, including the French *Institut de Veille Sanitaire* (INVS), the Spanish *Instituto de Salud Carlos III*, the Canadian *Global Public Health Intelligence Network* (GPHIN), the US *CDC*. MedISys is part of the *Europe Media Monitor* (EMM) product family, [6], developed at the EC's *Joint Research Centre* (JRC), which also includes the live news aggregation system *NewsBrief*, the news summary and analysis system *NewsExplorer* [7] and the exploratory tool set *EMM-Labs*. See <http://emm.jrc.it/overview.html> for an overview of EMM applications.

not only the news article proper, but also a great deal of irrelevant information, the main news article is extracted from each web page using a (patent-pending) text extraction process. During this process the documents are transformed into Unicode. The result is a standardised document format that allows common processing of all texts. Information about the document's language, source country, download time and source site are preserved as meta-data. The grabber software also checks whether the new text has a unique signature for this particular source. This avoids propagation of duplicate texts through the system.

2.2. Filtering and classification of the documents

MedISys allows the selection of articles about any subject, via Boolean combinations of search words or lists of search words, with positive or negative weights, and the setting of an acceptance threshold. The user may require that search words occur within a certain proximity (number of words), and may use wild cards. Using wildcards is crucial when dealing with highly-inflected languages. Each such subject definition is called an *alert*. Alerts are multilingual, which means that search word combinations mix languages. In addition to the generic alerts pre-defined by the JRC's team of developers, the specialist users may create their own subject-specific alert definitions. Users are responsible for the accuracy and completeness of their own alerts. A dedicated algorithm was developed at the JRC that allows the system to scan incoming articles for thousands of alert definitions in real time. Information about the alerts found in each article is added to the RSS file.

There are about 200 alert definitions for Public Health-related subjects in MedISys. The alerts are organised into a hierarchy of classes, such as Communicable Diseases, Symptoms, Medicines, Organisations, Bio-terrorism, Tobacco, Environmental & Food, Radiological & Nuclear, Chemical, etc., each containing finer sub-groups. On average, 3–4% of the 50,000 news items gathered daily satisfy at least one MedISys alert.

In addition to the subject alerts, there is one alert for each country of the world, including the name of the country and a major city. More fine-grained geo-coding and disambiguation are carried out downstream in the EMM NewsExplorer application, see [8]. Figure 1 shows the page on *Leptospirosis*, which is a specific entry of the group *Enteric Infections* in the main section *Diseases*.

2.3. Detection of early-warning trends across languages and news sources

The alert definitions in MedISys are multilingual, so that the mention of a disease or symptom can be identified in multiple languages. MedISys keeps a running count of all disease alerts for each country, i.e., it maintains a count of all documents mentioning a given disease *and* country, over a time window of two weeks. An alerting function detects a sudden increase in the number of reports for a given category and country, by comparing the statistics for the last 24 hours with the two-week rolling average. The more articles there are for a given category-country combination compared to the expected number of articles (i.e., the two-week average), the higher the alert level. Figure 2 shows a MedISys graph with the combinations having the highest alert level at a given time. The colour codes red (leftmost bars), yellow (middle bars) and blue (rightmost bars) indicate the alert levels high, medium and low.

The alert levels are calculated assuming a normal distribution of articles per category over time. Alert levels are high (or medium), if the number of articles found is at

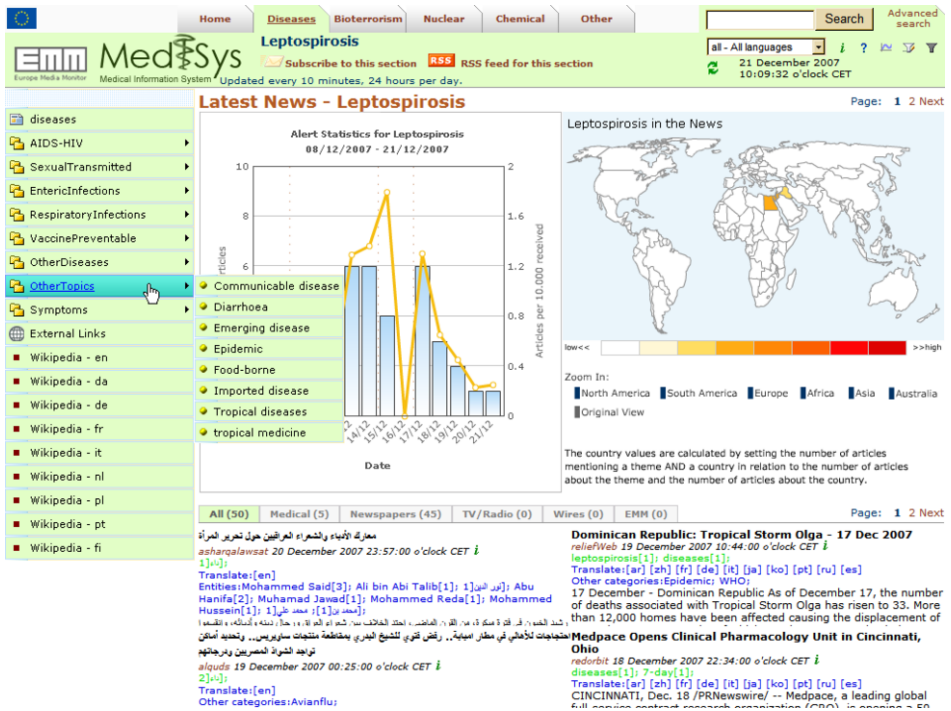


Figure 1. A page on *Leptospirosis* on the restricted MedISys site, with part of the category hierarchy exposed on the left. The bar chart in the middle column shows the absolute and relative number of articles falling into this category. The countries mentioned in articles about leptospirosis are highlighted on the map on the right (mostly Egypt and Iraq). The most recently retrieved mixed-language articles falling into this category are listed in the lower section.

least three times (or twice) the standard deviation. As the total number of articles varies throughout the week (fewer articles on Sunday and Monday), a correction factor is applied to normalise the expected frequencies according to the day of the week. The thin line in the bar chart in Figure 1 shows the relative number of articles for a given alert compared to the total number of articles that day.

2.4. Distribution of the extracted information to the MedISys users

The Web interface of MedISys can be used to view the latest trends and to access articles about diseases and countries. For each page, RSS feeds are available for integration of the results into downstream user applications. Users can also opt to receive instant email reports, or daily summaries regarding a pre-selected disease or country, for their own choice of languages. Users can subscribe to summary reports containing information on groups of alerts (e.g., *Avian Diseases*, including *avian flu*, *duck virus* and others). Registered users can also obtain access to the JRC’s *Rapid News Service*, RNS, which allows them to filter news from selected sources or countries, and which provides functionality to quickly edit and publish newsletters and to distribute them via email or to mobile phones (SMS). MedISys displays the title and the first few words of each article, plus a link to the URL where the full news text was originally found.

Today's Alert Statistics for Asia

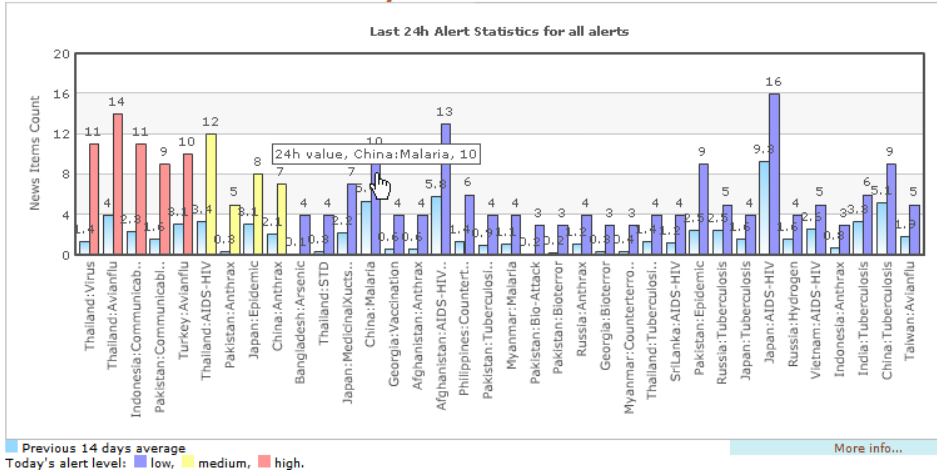


Figure 2. Alert statistics showing which category-country combinations are most active at the moment, compared to the average number of articles for the same combination. For each combination, the lower left bar indicates the expected number of articles, while the higher right bar shows the number found in the last 24 hours.

2.5. Implementation details and performance

MedISys currently processes about 50,000 media reports per day in real time. MedISys and EMM [6] are implemented in Java on Microsoft servers. MedISys shares tasks and machines with EMM, but running standalone, it would require three servers (dual processor multi-core with 4 GB of memory each). The system is scalable and could cope with processing many more feeds and news articles, for example, by processing the different languages distributed over several machines. The processes are fast and light-weight so that news monitoring and alerting happen in real-time. The slowest part of the process is downloading the articles from the Web (response time at the source). Computationally, the heaviest process currently is the real-time clustering of news articles (performed every ten minutes). The categorisation and article filtering system (see Section 2.2) matches about 30,000 patterns (multi-word terms and their combinations) against the incoming articles in a few hundred milliseconds for an average article, to categorise the articles according to about 750 categories.

3. Extraction of epidemic events in PULS

MedISys has proven to be useful and effective for finding and categorising documents from a large number of Web sources. To make the retrieved information even more useful for the end-user, it is natural to consider methodologies for deeper analysis of the texts, in particular, information extraction (IE) technology. After MedISys identifies documents where the alerts fire, IE can deliver more detailed information about the specific incidents of the diseases reported in those documents.

IE helps to boost precision, since keyword-based queries may trigger on documents which are off-topic but happen to mention the *alerts* in unrelated contexts. Pattern match-

[Database list] [Confident events] Events [Advanced query] [Outbreaks] [Reset page]

Viewing 248 events in 240678 documents

Published	Source	Disease	Begin	End	Country	Total	Status	Descriptor
2007.04		Avian Influenza			Indonesia			
2007.04.24	globalsecurity	Avian Influenza	2007.04.23	2007.04.23	Cambodia	172	†	Human Bird Flu Deaths
2007.04.24	globalsecurity	Avian Influenza	2007.01	2007.01	Indonesia	34		human cases
2007.04.24	globalsecurity	Avian Influenza	2003	2007	Indonesia	81		81 avian flu cases
2007.04.24	globalsecurity	Avian Influenza	2007.04.23	2007.04.23	Indonesia	--		two new human cases
2007.04.24	globalsecurity	Avian Influenza	2003	2007	Indonesia	63	†	63 deaths
2007.04.21	cidrap	Avian Influenza	2005.05	2005.05	Indonesia	291		291 cases
2007.04.21	cidrap	Avian Influenza	2005.05	2005.05	Indonesia	172	†	172 deaths
2007.04.19	ft	Avian Influenza	2005	2007	Indonesia	66	†	at least 66 human deaths
2007.04.19	ft	Avian Influenza	2003.09	2003.12	Indonesia	170	†	more than 170 people
2007.04.19	theglobeandmail	Avian Influenza	2003.09	2003.12	Indonesia	300		nearly 300 people
2007.04.19	ChinaPost	Avian Influenza	2003	2007	Indonesia	--		--
2007.04.17	cidrap	Avian Influenza	2007	2007	Cambodia	302	†	1,086 susceptible birds
2007.04.16	recomb	Avian Influenza	2007.04.14	2007.04.14	Indonesia	--	†	the family's chickens
2007.04.16	promed	Avian Influenza	2007.04.05	2007.04.05	Cambodia	--	†	the 13-year-old girl
2007.04.16	dailytimesPK	Avian Influenza	2007.04.12	2007.04.12	Cambodia	--	†	the Cambodian girl
2007.04.16	dailytimesPK	Avian Influenza	--	--	Cambodia	--	†	a 13-year-old girl
2007.04.15	medicinenet	Avian Influenza	2003	2003	Indonesia	33		33 people
2007.04.15	medicinenet	Avian Influenza	2003	2003	Indonesia	24	†	24
2007.04.14	JakartaPost	Avian Influenza	2007.04.13	2007.04.13	Indonesia	74	†	the country's 74 human bird flu fatalities
2007.04.11	cidrap	Avian Influenza	2007.04.11	2007.04.11	Cambodia	172	†	fatal H5N1 cases

<< 1 2 3 4 5 6 7 ... 11 12 13 >>

Figure 3. A table view of the extracted incidents.

ing in IE provides the mechanism that assure that the keywords appear in relevant contexts only. This is of value to users who are interested in *specific* scenarios involving diseases—outbreaks and epidemics, vaccination campaigns, etc.—as opposed to users who wish to monitor documents that mention the diseases in a broader context.

PULS, the *Pattern-based Understanding and Learning System*, is developed at the University of Helsinki to extract factual information from plain text. PULS has been adapted to analyse texts in the epidemiological domain, for processing documents that trigger MedISys alerts.⁵ Earlier, PULS's medical event detection had been applied to two sources dedicated to epidemiological reports—ProMED-Mail and WHO epidemic and pandemic alert and response.⁶ We next describe the processing of epidemics-related texts in PULS.

3.1. Event extraction in the medical domain

For each document, the IE system extracts a set of *incidents* reported in the text. An incident is a structured representation of an event⁷ involving some communicable disease, described in the text in natural language. An incident consists of a set of attributes: the location and country of the incident, the name of the disease, the date of the incident, and descriptive information about the victims—their type (people, animals, etc.), their number, whether they survived, etc. The incident may cover a single occurrence “80 chickens died on the farm on Wednesday,” or larger time interval, as in “Two people in the region have contracted the disease since the beginning of the year.” Text may also contain ‘periodic’ incidents: “according to authorities, 330 people die of malaria in

⁵<http://doremi.cs.helsinki.fi/jrc>

⁶<http://www.who.int/csr/don/en/>

⁷The term *event* is used differently in the medical and the computational communities. To the medics, an event denotes the *entire course* of an epidemic episode, from its inception to completion. In the computational literature on IE, event denotes a single, atomic “factoid”, that may be isolated, or may belong to a group of factoids that together describe the entire course of an epidemic. In the rest of this chapter, the computational reading is intended.

Uganda daily” (these are not currently handled by the system). The system also identifies events in which the disease is *unknown*, or undiagnosed, which are especially important for surveillance. For example, the sentence:

The deadly Ebola outbreak has so far killed 16 people in Gabon”

will trigger the creation of an incident—a record in a relational database—and assign the underlined values to the corresponding attributes. Each record extracted from the document is permanently stored, together with links to the exact offsets in the text where its attributes were found within the document.

Figure 3 presents a view of the database, as it appears on the Web site. This collection of rows was returned in response to a user query, which is specified by constraints on some of the attribute columns. Here, the constraints are on publication date (April 2007), disease (avian influenza) and country (Indonesia or Cambodia). The constraints are entered into the text boxes below the column names. (The table is ordered by publication date by default.) Blue rows in the table correspond to confident events (defined below in section 5), and white rows are less confident.

For detailed information about the design principles behind PULS, see, e.g., [9,10]. The system relies on several kinds of domain-independent and domain-specific *knowledge bases*. An example of domain-independent knowledge is the location hierarchy, containing names of countries, states or provinces, cities, etc. An example of a domain-specific knowledge base is the medical ontology, containing names of diseases, viruses, drugs, etc., organised in a conceptual hierarchy. The ontology currently contains 2,400 disease terms; 400 vectors (organisms that transmit disease, like rats, mosquitoes, etc.); 1,500 political entities—countries, their top-level divisions and name variants; over 70,000 location names (towns, cities, provinces).

PULS uses pattern matching to extract events; the system contains a domain-specific pattern base—a cascade of finite-state patterns, which map information from its syntactic representation in the sentence to its semantic representation in the database records. For example, the above sentence about Ebola will be matched by a pattern like:

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

The pattern first matches a noun phrase (NP) of semantic type *disease*; “Ebola” is a descendant of the *disease* node in the ontology. Then it matches a verb phrase (VP) headed by the verb *kill* (or its synonyms in the ontology). The verb phrase also subsumes modifier elements, such as the auxiliary verb *has*, the adverbial phrase *so far*, etc. The square brackets indicate that the locative prepositional phrase is optional; in case the location is omitted in the sentence, it is inferred from the surrounding context.

Populating the knowledge bases requires a significant investment of time and manual labour. PULS employs weakly-supervised learning to reduce the amount of manual labour as far as possible, by bootstrapping the knowledge bases from large, un-annotated document collections, [11,12].

Various views may be used to present the relational data. Especially important are views that aggregate the information according to the user’s criteria. Examples of views the PULS system provides include geographic maps, as in Figure 4 and charts or histograms, as in Figure 5.

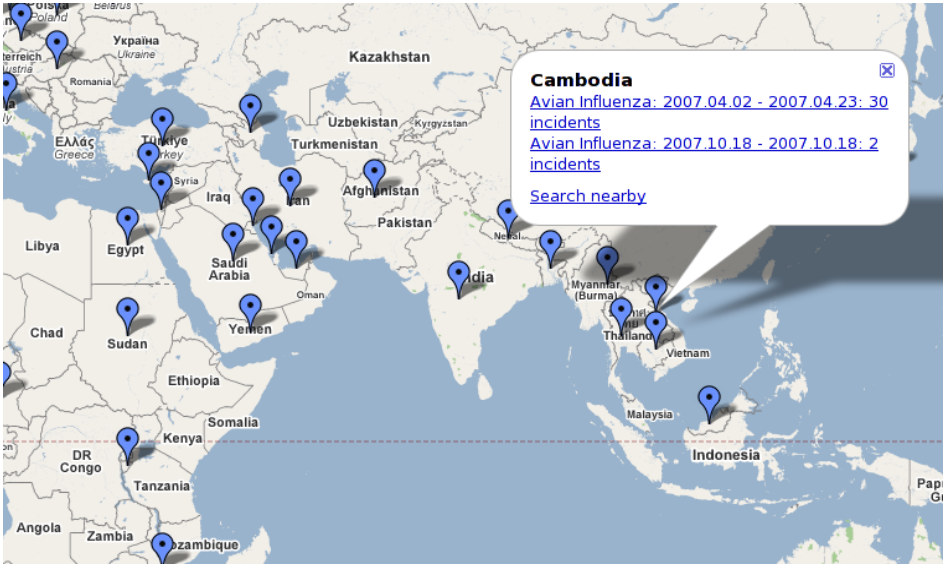


Figure 4. A geographic map of bird flu incidents

4. Cross-document aggregation

Besides the accuracy of the MedISys filtering and categorisation, an important issue for users is multiple reporting: due to the high number of independent news sources, MedISys captures many reports that readers of one or a few news sources would miss, but the flip-side of the coin is multiple reporting. This causes extra work for the users and makes monitoring daily news a time-consuming task. The solution to this problem lies in the aggregation of reports into larger units. MedISys and PULS use different approaches to aggregation, which are not currently integrated.

4.1. Clustering in MedISys

MedISys presents news *clusters* to the users, grouping similar news reports arriving within at most 8 hours of each other. The short time window means that clusters normally contain articles published within the same day. If reporting continues steadily, articles from different days will be grouped into the same cluster. The similarity measure for the news articles is based on cosine similarity on a simple vector-space representation of the first 200 word tokens of each article. This means that not only multiple reports of the same story, but also similar reports about different cases for the same disease may be grouped together. This method also allows users to discard entire groups of *non-relevant* articles (e.g., discussions about vaccination campaigns) at once.

4.2. Grouping disease events into outbreaks in PULS

PULS provides another means of aggregating multiple individual facts into larger units. PULS goes beyond the traditional IE paradigm in two respects. First, in a typical IE system, documents are processed separately and independently; facts found in one document

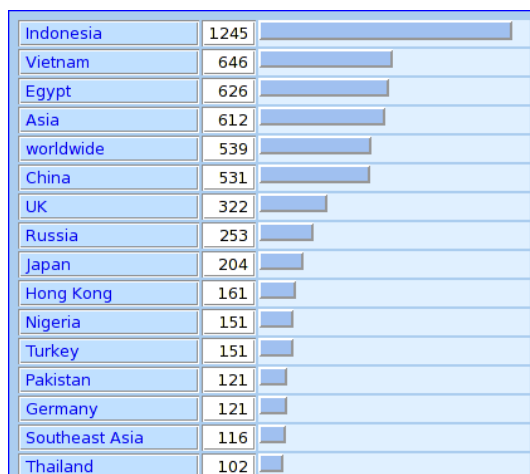


Figure 5. A chart of bird flu reports (over a time period specified by a user's query)

do not interact with information found in other documents. Second, for each attribute in an extracted incident, the IE system stores only one value in the database record—the locally best guess for that attribute.

1. After PULS extracts information from each document locally, it attempts to globally unify the extracted facts into groups, which we call *outbreaks*. An outbreak is a set of related incidents. Currently, incidents are related by simple heuristics: they must share the same disease name and the same country, and occur reasonably 'close' in time. Closeness is determined by a time window, currently fixed at 15 days.⁸ A chain of incidents, any pair of which are separated by no more than the time window, are aggregated into the same group. Thus, an outbreak is a kind of a 'bin' containing related incidents, and provides an added level of abstraction above the 'low-level' facts/incidents.

2. When PULS stores a record in the database, for each attribute, in general, rather than storing a single value, PULS stores a distribution over a set of possible values. For example, the sample text (in the first paragraph of this section) might read instead "*Five more people died last week.*" PULS will then try to fill in the missing attributes (i.e., the disease name, location) by searching for entities of the corresponding semantic type elsewhere in the discourse. In general, for a given attribute of an event, the document will contain several possible candidate entities, and each candidate will have a corresponding score—measuring how well it fits the event. The score depends on certain features of the candidate value. These features include whether the value is mentioned inside the *trigger*—the piece of text that triggered some pattern from the pattern base; whether it appears in the same sentence as the trigger; whether it appears before or after the sentence containing the trigger; whether this value is the unique value of its type in the sentence that contains the trigger (e.g., the sentence mentions only a single country, or disease); whether the value is unique in the entire document; etc.

Using a set of candidate values rather than a single candidate is helpful in two ways. First, it allows us to compute the *confidence* of an incident, which is used in cross-

⁸The time window could be made more sensitive, e.g., dependent on the disease type.

document aggregation (in section 5). Second, it allows us to explore methods for recovery from locally-best but incorrect guesses by using global information.⁹

The next section shows how these features are used in the combined system.

5. Integration

This section describes the integration between MedISys and PULS, and tries to demonstrate that even in this early state, the whole is greater than the sum of its parts.

A special RSS tunnel has been set up between MedISys and PULS. At present, PULS processes only English-language documents. MedISys forwards documents which it categorises as relevant to the medical domain through the tunnel to PULS. Currently, the documents arrive as plain text. This is done in addition to the normal processing on the MedISys side, where running averages are monitored for all alerts, etc. A document batch is sent every 10 minutes, with documents newly discovered on the Web.

On the PULS side, the IE system analyses all documents received from MedISys, and returns information that it extracted from the received documents back through the tunnel—in structured form (also at 10 minute intervals). This communication is asynchronous, while both sites are operating in real-time.

When PULS receives documents from MedISys, it performs the following steps:

First, the IE system analyses the documents, extracts incidents, and stores them in the database (at <http://doremi.cs.helsinki.fi/jrc>). Second, PULS uses document-local heuristics to compute the *confidence* of the attributes in the extracted incidents.

The confidence of an attribute is computed from the set of candidate values for that attribute, based on their scores, which are in turn based on the features, as explained in Section 4.2. If the score of the best value exceeds a certain threshold, the attribute is considered *confident*. Some of the attributes of an incident are considered to be more important than others: here, in the case of epidemic events, these *principal* attributes are the disease name, location and date. If all principal attributes of an incident are confident, the entire incident is considered confident as well.¹⁰

Third, the system aggregates the extracted incidents into *outbreaks*, across multiple documents and sources. The aggregation process requires that at least one of the incidents in each outbreak chain must be confident (that is, chains composed entirely of non-confident incidents are discarded).

Finally, PULS returns a batch of recent incidents to MedISys, for displaying on its pages. The goal is to return a set of incidents with high confidence and low redundancy—a complete yet manageably-sized set for the user to explore. The batch is restricted to documents published within the last 10 days; from this period, PULS returns the most recent 50 incidents, filtering out duplicates: if multiple incidents of the same disease in the same location are reported, PULS returns only the most recent one.¹¹

⁹This line of current research is not covered in the present chapter.

¹⁰In the PULS tables, confident attributes appear in bold, and confident incidents are highlighted.

¹¹Note that this implies that a recent event that was last reported more than 10 days ago, will not appear in the result list, while an event from several months ago may appear—if it is mentioned in a very recently published report. This is a design choice that aims for a balance between recency of *publication* vs. recency of *occurrence* of an incident: both may be important to the user. Note also that in any case *all* events are available for browsing in the PULS database.

On the MedISys side, the returned events are displayed in two views. The main MedISys page shows the five most recent events—these correspond to the most urgent news. For more detail, this box has a link to the batch of 50 most recent incidents. For the complete view, the recent list has a link to the PULS database.

6. Evaluation—Summary of Results

The public MedISys site is currently accessed by an average of 1,700 distinct users every day and over twenty Public Health authorities make use of the restricted site, the customisable Regional News Service view of the data, and the automatically sent notifications. Generally, the feedback from users has been positive. When giving feedback, users typically ask for more news sources or alert categories. Heavy MedISys users sometimes express the wish for a more thorough filtering of the news in order to reduce the number of articles that mention an alert out of context. For instance, news about a celebrity in the context of which a disease is mentioned is considered to be unwanted noise. One way to tackle this issue would be through using automatically trained classifiers that separate relevant from irrelevant news (e.g., sports, film, etc.). Another option would be to tighten the filter by tuning the binary alert definitions, using more search words and applying weights. This tuning of alert definitions is an on-going process, as alert definitions are updated all the time. The third option is to combine the Information Retrieval approach used in MedISys with a subsequent Information Extraction phase, as presented in this chapter.

6.1. Evaluation of the filtering and classification in MedISys

We evaluated disease-related alerts in MedISys by selecting 100 English-language articles.¹² The articles were selected randomly, with a maximum of ten articles taken from a given day, in order to investigate a broader variety of news articles and alerts. Only those articles were considered which triggered some MedISys alert definition.

In these 100 news articles, 156 relevant alerts were found. For these alerts, a human expert judged the accuracy of the alert, by assigning it to one of four categories:

1. The MedISys alert assigned to the news article is appropriate and the article mentioned a disease outbreak event. 63 alerts fell into this category.
2. The disease name was mentioned in the article, but in the context of vaccines, new drugs, or similar. 74 alerts fell into this category.
3. The MedISys alert was inappropriate. This category consists of cases where the disease name was mentioned, but the article was about generic issues such as politics, literature, finance or sports. In a small number of cases, person names or other words are homographic with a disease name triggered the category. For example, the term *Mobility Aids* triggered the category HIV. 19 alerts fell into this category.

¹²Note that the mandate of MedISys is broader than monitoring communicable diseases, as some users are interested in chemical or nuclear incidents, in mentions of vaccines or new medicines, etc. However, because PULS at present identifies only events describing outbreaks of diseases, the evaluation was limited to this subset of alerts. All other alerts were ignored.

4. The fourth group consists of articles mentioning a disease that should have been identified, but were not, corresponding to the *false-negative* measure (which has an impact on recall). Some examples: The word *HIV-positive* did not trigger the disease HIV, due to a tokenisation error; *Foot & Mouth Disease* was not recognised because the disease name variant with the ampersand was not part of the alert definition. 13 alerts fell into this category.

Problems in category 4, such as tokenisation errors and missing disease name variants, are easy to correct. This example shows that continuous quality control is essential, and that the performance of queries involving Boolean operators is highly dependent on the quality of the search words. Since MedISys and NewsBrief have hundreds of alert categories for a wide variety of users, it is clear that the users have to use their subject knowledge and control their own alerts. For that purpose, an alert-editing interface is available to specialist users.

The evaluation shows that the results for concrete alerts involving one or more disease names are: 137 (63+74) out of 156 alerts were relevant, which corresponds to a *precision* of 0.88. The performance for more abstract alerts which cannot so simply be defined via the occurrence or non-occurrence of specific words (e.g., 'fraud' or 'stress at the work place') is expected to be lower.

6.2. Evaluation of event detection

PULS receives on the order of 10,000 documents from MedISys each month. From 27% of these documents, PULS extracts about 6,000 incidents, on average.¹³ The remaining 73% of the documents processed by PULS yield no incidents. This is as expected, since MedISys does not explicitly select for outbreaks, but for mentions of disease names in *any* context, and many documents may mention diseases in contexts unrelated to epidemics and outbreaks.

To estimate the proportion of documents rejected by PULS that contain missed events—false negatives—we manually checked 200 MedISys documents that produced no events. Among these documents, 14% contained an event that the IE system had missed.¹⁴ As PULS filters out 73% of the incoming documents, adding back the incorrectly filtered documents (14% of all filtered), yields that about 63% of the documents that arrive to PULS contain no *epidemic* events. In this way, the IE phase helps to distinguish reports about epidemic outbreaks from documents that mention diseases in other contexts.¹⁵

We tried to estimate the accuracy of event detection and the confidence heuristic. Twenty percent of all extracted incidents are rated as confident by PULS. We selected 100 confident incidents at random, and checked their correctness manually. In this evaluation we took a conservative (strict) approach: we considered an incident to be correct

¹³In IE, it is typical for a relevant document to contain more than one incident, since often there is one or more main events, and other, related events are mentioned as part of background discussion.

¹⁴NB: this does not correspond to the false-negative rate. To compute the false-negative rate, recall at the document level, and recall at the level of events would require a more detailed evaluation, to be conducted in the future.

¹⁵MedISys has an optional boolean filter that tries to capture outbreaks by requiring the name of the disease to occur in combination with keywords like *bedridden*, *hospital**, *deadly*, *cases*, etc. This has not yet been evaluated.

only if all of its 'principal' attributes are correct, i.e. no credit is granted for partially correct events, unlike in standard IE evaluation. The result was: 72% of the confident incidents are correct; in 14% of the cases, the information extraction is spurious, i.e., PULS extracts an incident where there should be none; in another 14% of the cases, the confident incident is incorrect—i.e., at least one of the attributes has an incorrect value (the top-ranked value is wrong). The latter category of error is difficult to correct, since it is usually due to an inherent complexity in the text. The former type of error may be simpler to correct, through further tuning of the knowledge bases.

Since outbreak aggregation is our primary means of reducing redundant information in the flow of news to the user, it is important to estimate the accuracy of the outbreak grouping. We analysed a randomly chosen set of medium-sized outbreaks: 20 outbreaks with approximately 10 incidents in each. For each incident we determined whether it was appropriately included in the outbreak. 68% of the incidents were correctly identified with their outbreaks. Three of the outbreaks (about 15%) were erroneous, i.e., based on incorrect confident incidents.¹⁶

Of all the incidents examined in this evaluation 22.5% were confident (i.e., on average, the outbreaks contained only 2–3 confident incidents).

7. Conclusion and future work

The combination of the two initially independent systems MedISys and PULS has led to a stronger application offering users complementary functionality through a unified user interface. For communicable disease outbreaks, which are covered by both systems, the combination of IR in MedISys and IE in PULS leads to additional advantages: Firstly, PULS's computationally heavier methods only need to be applied to the document collection pre-filtered by MedISys. Secondly, the medical event extraction patterns act as an additional filter to identify only disease outbreak reports. MedISys is designed to capture not only disease outbreak reports, but also other news articles mentioning diseases. For users interested specifically in disease outbreaks, PULS's event recognition helps reduce the number of reports by filtering out just under three quarters of incoming reports, of which about 14% are incorrectly filtered relevant reports.

The current status of integration can be taken further: the systems don't yet make full use of the other's information aggregation methods. The categorisation of news items by MedISys can be useful for the analysis performed by PULS, and is yet to be utilised. The taxonomies used by the systems are overlapping, but have not yet been fully integrated. These and other issues are to be tackled in future work.

While we believe that the combination of IR in MedISys and IE in PULS provides added value, it is not a universal solution. An important strength of MedISys is its multi-linguality: it monitors media reports in currently 43 languages. Developing PULS-style event extraction grammars for so many languages is not currently possible: porting the IE

¹⁶It was interesting to observe that aggregation is often useful even when the outbreak consists entirely of incorrectly analysed incidents. For example, in high-profile cases picked up by main news agencies, reports are re-circulated through multiple sites worldwide. Because the text is very similar to the original report, the IE system extracts similar incidents from all reports, and correctly groups them together. Although some attribute is always analysed incorrectly, the error is *consistent*, and the grouping is still useful: it helps reduce the load on the user by aggregating related facts.

system to a new language requires pre-existing robust lower-level linguistic components (named entity tagger, ontology, parser) for each new language, which are unlikely to be available for all the languages covered by MedISys in the near future. However, focusing on the major languages for which lower-level linguistic resources have been developed is planned for future extensions.

Acknowledgements

The development of MedISys would not have been possible without the members of the Web Mining and Intelligence team, who contributed over the years. Work on PULS was supported in part by the Academy of Finland grant 118653, National Center of Excellence “*Algodan*” (*Algorithmic Data Analysis*).

References

- [1] R. Gaizauskas and A. Robertson, “Coupling information retrieval and information extraction: A new text technology for gathering information from the web,” in *Proceedings of the 5th RIAO Computer-Assisted Information Searching on Internet*, Montreal, Canada, 1997.
- [2] Defence Advanced Research Projects Agency, “Information extraction task: scenario on management succession,” in *Proc. 6th Message Understanding Conf. (MUC-6)*. Columbia, MD: Morgan Kaufmann, 1995.
- [3] A. Robertson and R. Gaizauskas, “On the marriage of information retrieval and information extraction,” in *Information retrieval research 1997: Proceedings of the 1997 annual BCS-IRSG colloquium on IR research, Aberdeen, Scotland*, J. Furner and D. Harper, Eds. London: Springer-Verlag, 1997.
- [4] S. Doan, Q. Hung-Ngo, A. Kawazoe, and N. Collier, “Global Health Monitor—a web-based system for detecting and mapping infectious diseases,” in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2008.
- [5] C. Freifeld, K. Mandl, B. Reis, and J. Brownstein, “HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports,” *J Am Med Inform Assoc*, vol. 15, pp. 150–157, 2008.
- [6] C. Best, E. van der Goot, K. Blackler, T. Garcia, and D. Horby, “Europe Media Monitor—system description,” EUR, Tech. Rep. 22173 EN, 2005.
- [7] R. Steinberger, B. Pouliquen, and C. Ignat, “Navigating multilingual news collections using automatically extracted information,” *Journal CIT*, vol. 13, no. 4, 2005.
- [8] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouni, A. Widiger, A.-C. Forslund, and C. Best, “Geocoding multilingual texts: Recognition, disambiguation and visualisation,” in *Proceedings of LREC-2006*, Genova, Italy, 2006.
- [9] R. Yangarber, L. Jokipii, A. Rauramo, and S. Huttunen, “Extracting information about outbreaks of infectious epidemics,” in *Proc. HLT-EMNLP 2005*, Vancouver, Canada, 2005.
- [10] R. Grishman, S. Huttunen, and R. Yangarber, “Information extraction for enhanced access to disease outbreak reports,” *J. of Biomed. Informatics*, vol. 35, no. 4, 2003.
- [11] R. Yangarber, “Counter-training in discovery of semantic patterns,” in *Proc. ACL-2003*, Sapporo, Japan, 2003.
- [12] W. Lin, R. Yangarber, and R. Grishman, “Bootstrapped learning of semantic classes from positive and negative examples,” in *Proc. ICML Workshop: Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, Washington, DC, 2003.