Combining Information about Epidemic Threats from Multiple Sources

Roman Yangarber*, Clive Best†, Peter von Etter*, Flavio Fuart†, David Horby†, Ralf Steinberger†

∗ Department of Computer Science
University of Helsinki, Finland
first.last@cs.helsinki.fi

† European Commission – Joint Research Centre
Ispra, Italy
first.last@jrc.it

Abstract

This paper describes an on-going effort to combine Information Retrieval (IR) and Information Extraction (IE) technologies, to leverage the benefits provided by both approaches to add value for the end-user, as compared with IR or IE in isolation. The main aim of the combined system is to pool together information from multiple sources to improve the quality of results. On one hand, multiple mentions of the same event or related events should be presented in a coherent fashion. On the other hand, grouping related events should improve the system's confidence in the discovered facts. We describe our approach and the results achieved in the project to date.

1 Introduction

The ability to obtain timely medical information from digital sources is essential for surveillance of diseases and epidemics. It directly impacts the work of health authorities and epidemiologists throughout the world. In this paper we present preliminary results from a project that aims to build a system for monitoring disease epidemics by analyzing textual information, mostly in the form of news, available on the Web. The system is built on top of two major components—MedISys, based mostly on IR technology, and PULS, the information extraction system.

We describe the system responsible for the IR component in section 2; in section 3 we describe the medical IE system, together with the heuristics it implements; section 4 describes how the two systems are integrated. In section 4.2 we present some quantitative measures of performance of the combined system. In conclusion, we discuss directions of on-going work.

2 Medical Information System: MedISys

The *Medical Information System*, MedISys, is an automatic tool that gathers reports concerning Public Health from many Internet sources world-wide in multiple languages, classifies them according to hundreds of categories, detects trends across categories and languages, and notifies users. The publicly accessible MedISys site *http://medusa.jrc.it/* presents a quantitative summary of latest reports on a variety of diseases and disease sub-types (e.g., respiratory infections), on bioterrorism-related issues, toxins, bacteria (e.g., anthrax), viral hemorrhagic fevers (e.g., Ebola), viruses, medicines, water contaminations, animal diseases, Public Health organisations, etc. At a second, password-restricted site, EU staff and national Public Health officials get access to an even larger variety of subject classes, such as news on additional diseases, on nuclear or chemical contamination, etc. Furthermore, users of the restricted site can see pay-for newswires, get access to mapping tools, and subscribe to automatically generated daily reports on various themes.

The development of MedISys was initiated by the European Commissions (EC) Directorate General Health and Consumer Affairs (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, bioterrorism, large-scale chemical incidents, etc.

MedISys is an automatic alternative to an otherwise time-consuming and tedious manual process. Typically, employees of national Public Health organisations look through their national press to identify reports on disease outbreaks and other Public Health issues and summarise the situation or scan the documents. The usage of MedISys saves these users time, and additionally gives them access to more news reports in more languages.

MedISys currently monitors news articles from about 1100 news portals around the world in 32 languages, from commercial news providers including 25 news agencies, LexisNexis, and from about 150 specialised Public Health sites. The system categorises all documents according to about 200 classes of predefined health threats. It uses statistical procedures to detect a sudden increase of articles in any of the classes, and visualises the trends graphically. Users can access documents and the automatically derived meta-information via Web pages, RSS feeds, through daily email alerts and summary reports, and via automatically generated SMS messages.

MedISys is part of the Europe Media Monitor (EMM) product family, developed at the EC's Joint Research Centre (JRC), which also includes NewsBrief,[1] a live news aggregation system, and NewsExplorer,[2] a news summary and analysis system [5]. The following sections cover the functionality of MedISys in more detail.

2.1 Document Gathering and Format Standardisation

MedISys ingests all EMM documents, i.e. the newswires provided by major news agencies, plus the approximately 35,000 articles per day, in 32 languages found on about 1100 news portals and 150 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 good geographical coverage. Additionally, individual users can request the inclusion of further news sources, such as all local newspapers of their country, but these user-specific sources are processed separately in order to guarantee the balance of news sources and their types across languages.

Where available, EMM (and thus also MedISys) collects RSS feeds. RSS stands for "Really Simple Syndication" and is an XML format with standardised tags used widely for the dissemination of news and other documents. For all other source sites, scraper software firstly looks for links on pre-defined web pages and downloads the pages linked to. As news pages do not only contain the news article, but also menus, related news, advertising, information about other sections of the newspaper, and other non-news-related information, the main news article is extracted from each web page in a three-step process:

1. clean the HTML by removing Java script, non-standard tags and unnecessary tags,

2. convert the HTML code to XHTML, which includes repairing incorrect HTML code, and

3. convert XHTML to RSS format, using an XSLT transformation that needs to be produced manually and separately for each news site.

For details, see [1]. The result is a standardised document format in UTF-8 encoding that allows common processing of all texts. Information about the document's language, source country, download time and place are preserved as meta-data.

2.2 Document routing and classification

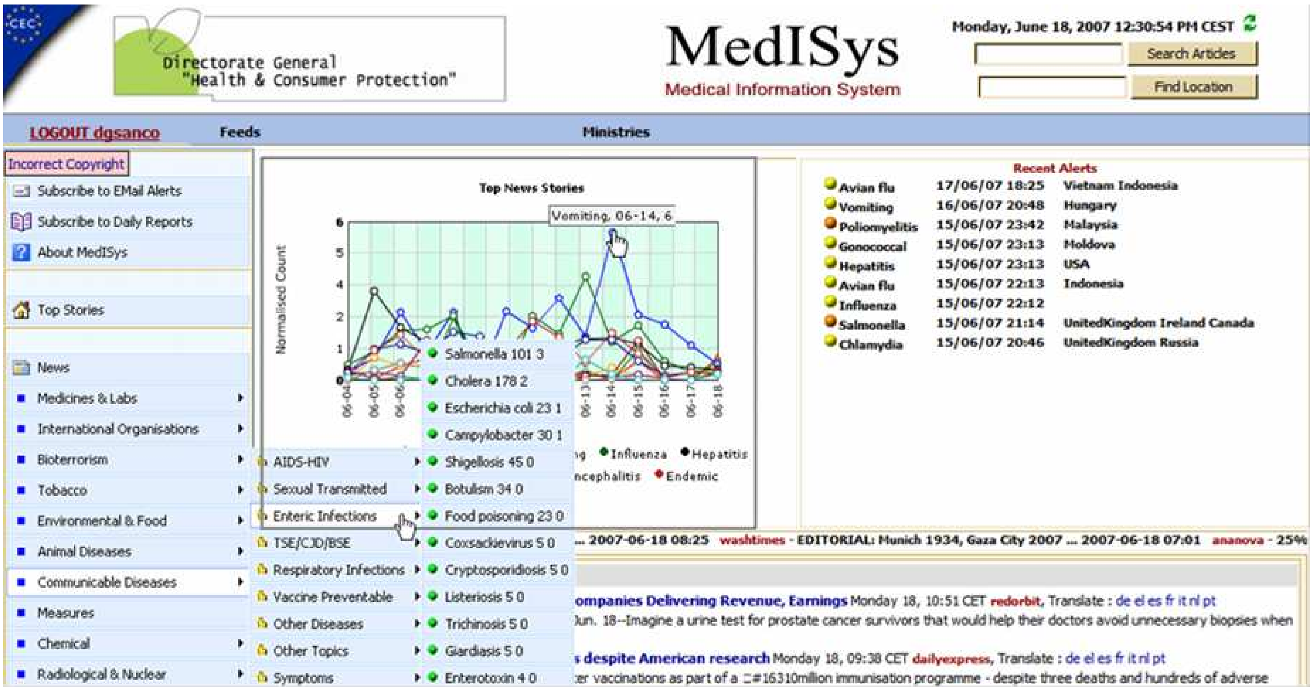
EMM allows the selection of articles about any subject using either Boolean combinations of search words or lists of search words with positive or negative weights, and the setting of an acceptance threshold. It is possible to require that search words occur within a certain proximity (number of words) and to use wild cards (single letter and word-final Kleene star). In EMM, each such subject definition is called an *alert*. EMM alerts are multilingual, i.e., search-word combinations may mix languages. In addition to the generic alerts pre-defined by the EMM team, 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 hundreds or thousands of alert definitions in real time. Information about the alerts found in each article is added to the RSS file. EMM NewsBrief has approximately 600 different alert definitions, including one for each country of the world (consisting mainly of the country name, and the name of the major city or cities). More fine-grained geo-coding and disambiguation are carried out downstream in the EMM NewsExplorer application, see [4].

The medical alerts in MedISys differ from the generic EMM NewsBrief alerts. In addition to the country-based alerts, MedISys employs hundreds of health-specific alert definitions. MedISys alerts are organised into a hierarchy of classes, such as Communicable Diseases, Medicines and Labs, Organisations, Bioterrorism, Tobacco, Environmental & Food, Radiological & Nuclear, Chemical, etc., each containing finer sub-groups. Figure 1 shows the entry page of MedISys with part of its menu structure exposed (on the left and bottom-left), and a trend visualisation graph (upper-middle box).

[1] http://press.jrc.it

[2] http://press.jrc.it/NewsExplorer



Recent Alerts

Avian flu	17/06/07 18:25	Vietnam Indonesia
Vomiting	16/06/07 20:48	Hungary
Poliomyelitis	15/06/07 23:42	Malaysia
Gonococcal	15/06/07 23:13	Moldova
Hepatitis	15/06/07 23:13	USA
Avian flu	15/06/07 22:13	Indonesia
Influenza	15/06/07 22:12	
Salmonella	15/06/07 21:14	UnitedKingdom Ireland Canada
Chlamydia	15/06/07 20:46	UnitedKingdom Russia

... 2007-06-18 08:25 washtimes - EDITORIAL: Munich 1934, Gaza City 2007 ... 2007-06-18 07:01 ananova - 25%

ompanies Delivering Revenue, Earnings Monday 18, 10:51 CET redorbit, Translate : de el es fr it nl pt
Jun. 18--Imagine a urine test for prostate cancer survivors that would help their doctors avoid unnecessary biopsies when

s despite American research Monday 18, 09:38 CET dailyexpress, Translate : de el es fr it nl pt
er vaccinations as part of a □#16310million immunisation programme - despite three deaths and hundreds of adverse

Fig. 1: *Medisys main page (restricted site).*

2.3 Multilingual multi-document trend detection

The alert definitions in MedISys are multilingual, so that the mention of a disease or symptom in the news in any of the languages can be identified. The MedISys software keeps a running count of all disease alerts for any country of the world, i.e., it maintains a count of all documents mentioning both a certain country and a given disease over a fixed time window—a period of two weeks. An alerting function detects a sudden increase in the number of reports for a given disease and country by comparing the statistics for the last 24 hours with the two-week daily rolling average. It uses the Poisson distribution, which is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate.

Figure 2 shows how the intersection of country alerts with disease alerts in combination with the trend analysis can be used to alert users to a potential health threat. This screenshot, from the public MedISys site (*http://medusa.jrc.it/*), shows increasing reporting on dengue fever related to several South-East Asian countries (highlighted on map).

2.4 Users and usage of MedISys

Customers of MedISys can use the Web interface to view the latest trends and access articles about diseases and countries. However, they can also opt to receive instant email reports, or daily summaries regarding pre-selected diseases or countries, for their own choice of languages. Specific registered users can also be granted access to the JRCs *Rapid News Service*—RNS, which additionally allows 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. MedISys displays the title and the first few words of each article, plus a link to the URL containing the full text.

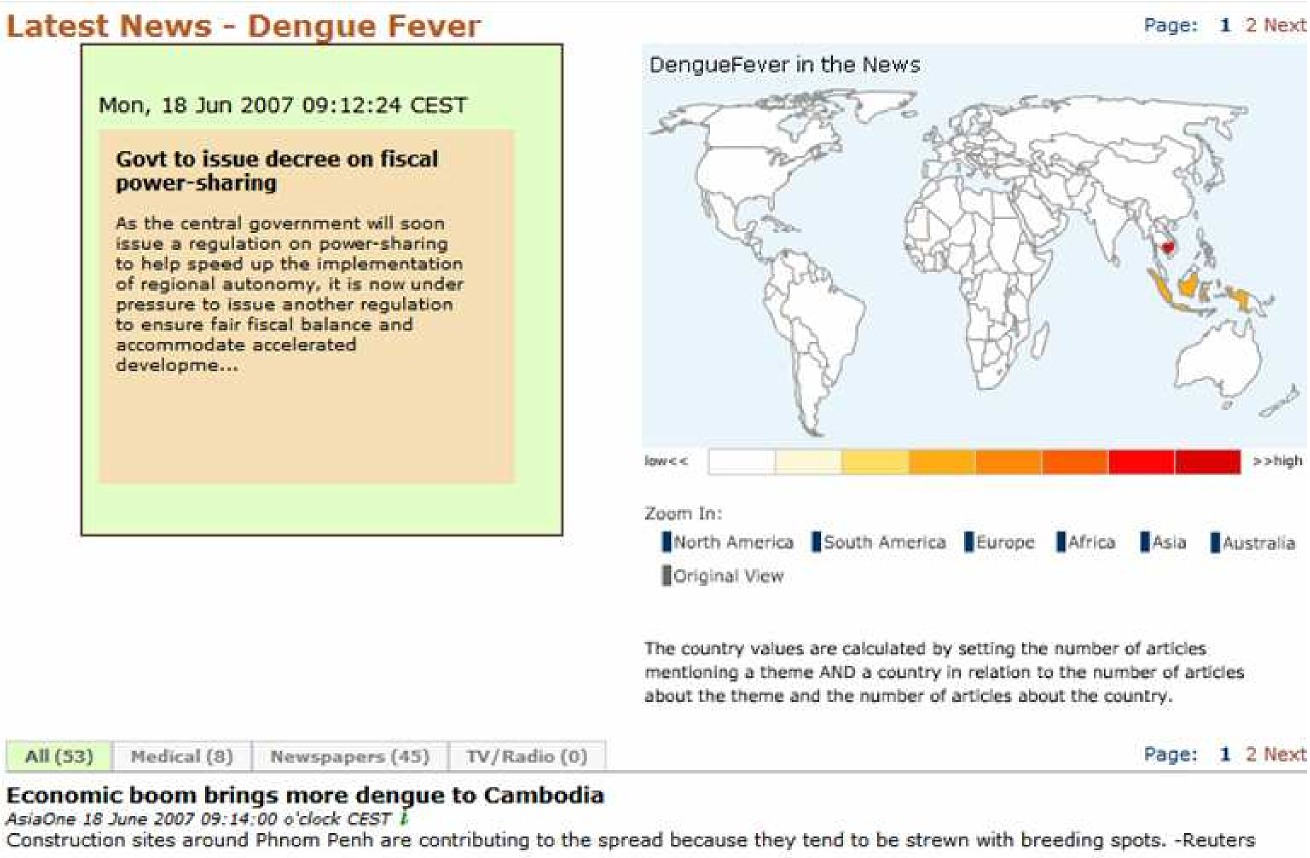
MedISys users include the European Commission, the World Health Organisation (WHO), the Canadian Global Public Health Intelligence Network (GPHIN), the European Centre for Disease Control (ECDC) and the US CDC, the French Institut de Veille Sanitaire (INVS), the Spanish Instituto de Salud Carlos III, and other national authorities.

3 Extracting Facts about Epidemics

MedISys has proved to be a useful and an effective tool, with thousands of users accessing it daily. In considering possible extensions that would add further value, a natural choice falls on IE technology:

- IE could deliver information concerning specific incidents of the diseases tracked by MedISys, whereas IR is able to return entire matched documents (along with an indication which *alerts* fired within the document).

- IE could boost precision, since keyword-based queries may trigger on documents which are off-



Latest News - Dengue Fever

Page: 1 2 Next

Mon, 18 Jun 2007 09:12:24 CEST

Govt to issue decree on fiscal power-sharing

As the central government will soon issue a regulation on power-sharing to help speed up the implementation of regional autonomy, it is now under pressure to issue another regulation to ensure fair fiscal balance and accommodate accelerated developme...

DengueFever in the News

Zoom In:

North America | South America | Europe | Africa | Asia | Australia | Original View

The country values are calculated by setting the number of articles mentioning a theme AND a country in relation to the number of articles about the theme and the number of articles about the country.

All (53) | Medical (8) | Newspapers (45) | TV/Radio (0)

Page: 1 2 Next

Economic boom brings more dengue to Cambodia

AsiaOne 18 June 2007 09:14:00 o'clock CEST

Construction sites around Phnom Penh are contributing to the spread because they tend to be strewn with breeding spots. -Reuters

Fig. 2: *Dengue alerts with geographic distribution.*

topic but happen to mention the *alerts* in unrelated contexts. Pattern matching in IE provides the mechanism that assure that the keywords appear in relevant contexts only.

PULS, the *Pattern-based Understanding and Learning System*, is developed at the University of Helsinki to extract factual information from epidemiological reports.[3] Previously, PULS had been applied to two dedicated sources of epidemiological reports—ProMED-Mail,[4] and WHO epidemic and pandemic alert and response.[5]

We next briefly describe the key functionality provided by PULS for epidemics-related texts.

3.1 Medical Information Extraction

For each document, the IE system extracts a set of *incidents* reported in the text. An incident is an event involving some communicable disease, described in plain text. An incident is described by a set of attributes: the name of the disease, the location and country of the incident, the date of the incident, and descriptive information about the victims—their type (people, animals, etc.), their number, whether they survived, etc. The system also identifies events in which the disease is reported as *unknown*, or undiagnosed, which are especially crucial for surveillance.

For example, for the sentence "*Five people were reported to have contracted Ebola in Uganda last week*," the system will assign the underlined values to the corresponding attributes, and create a record in a relational database. 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.

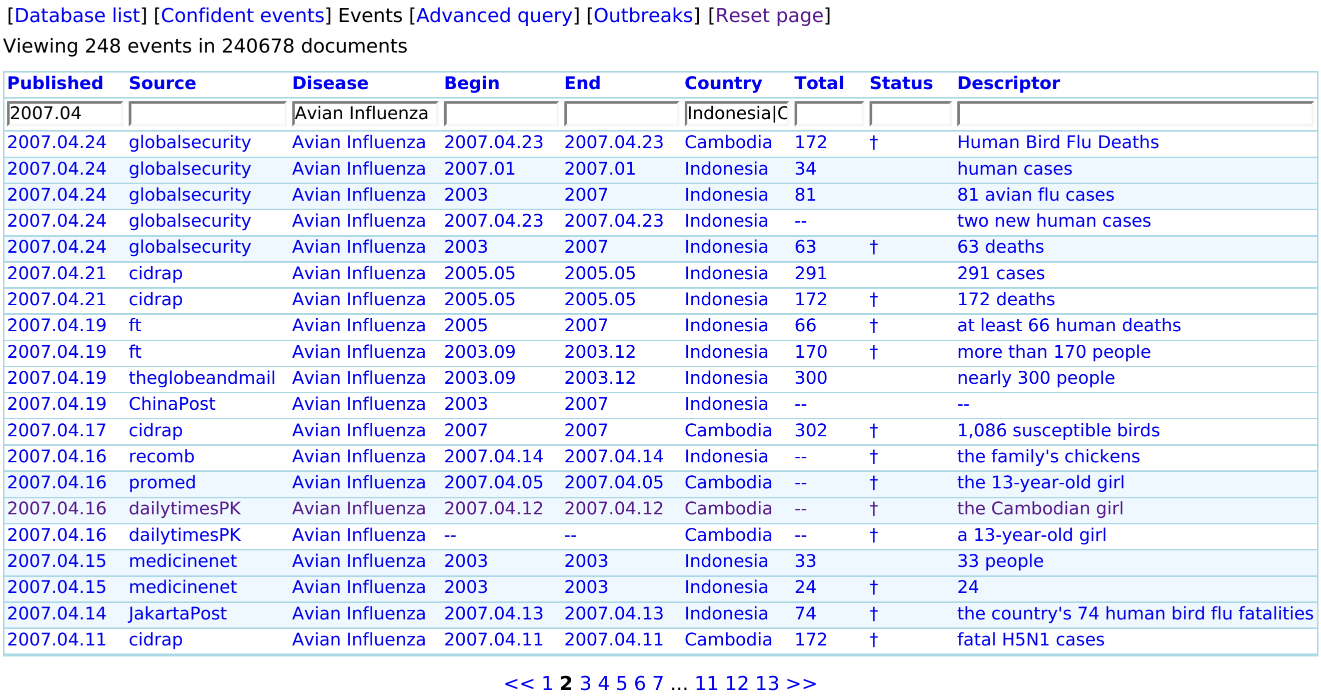
Figure 3 presents a view of the database, as it appears on the Web site. This collection of rows was returned in response to a user "query", which is specified by constraints on some of the attribute columns. This table was constrained by publication date (April 2007), disease (avian influenza) and country (Indonesia or Cambodia). The constraints are typed into the text boxes below the column names. (Rows are ordered by publication date, by default.) Blue rows in the table correspond to confident events (defined below in section 4.1), and white rows are non-confident.

Figure 4 shows a MedISys document which generated an event (corresponding to the sixth line from the bottom of the table in Figure 3). The values of the attributes of the event are shown in the box on the right.

[3] http://doremi.cs.helsinki.fi/jrc

[4] http://www.promedmail.org

[5] http://www.who.int/csr/don/en/



[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|C			
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 >>

Fig. 3: *A view of extracted incidents.*

For detailed information about the design and operation principles behind the PULS system, see, e.g., [7, 2]. PULS operates by pattern matching, and 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., organized in a conceptual hierarchy. The system also contains a domain-specific pattern base—which contains patterns that map the surface-syntactic representation of the information in the sentence to the semantic representation in the database records. Populating the knowledge bases requires a significant investment of time and manual labor. PULS employs weakly-supervised methods to reduce the amount of manual labor as far as possible, by bootstrapping the knowledge bases from large, unannotated document collections, [6, 3].

3.2 Toward Cross-Document IE

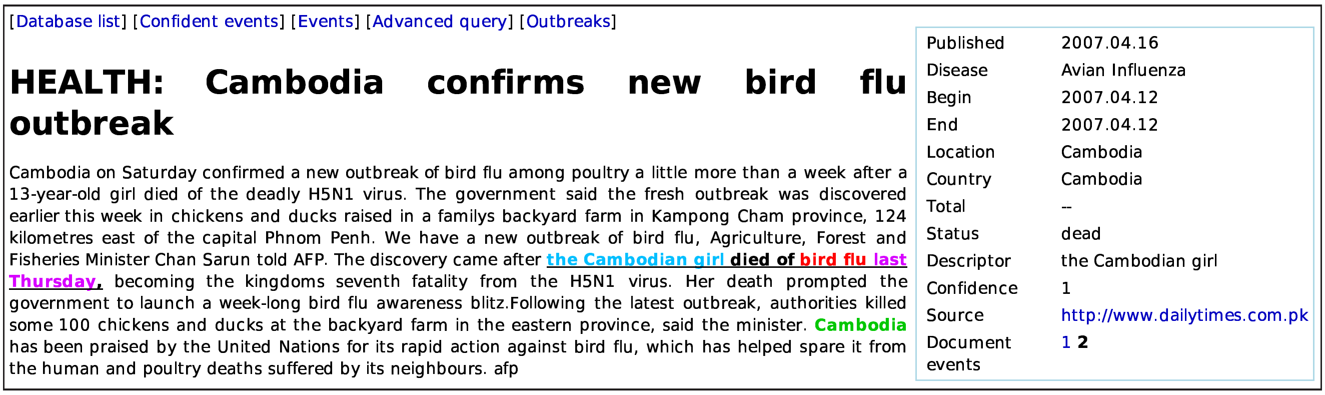
PULS goes beyond the traditional IE paradigm in two respects. First, in a traditional IE system, documents are processed separately and independently; facts found in one document do not interact with information found in other documents. Second, for each attribute in an extracted incident, traditionally, the IE system stores only one value in the database record—the value that is 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 straightforward heuristics: they must share the same disease name and the same country, and be “reasonably close” in time. Closeness is determined by a time window, currently fixed at 15 days.[6] Any chain of incidents which are separated by no more than the time window are aggregated into the same group.

An outbreak therefore serves as a kind of a “summary” of the incidents it contains, and provides an extra level of abstraction between the user and 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

[6] This could be made more flexible, e.g., dependent on the disease type.



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

HEALTH: Cambodia confirms new bird flu outbreak

Cambodia on Saturday confirmed a new outbreak of bird flu among poultry a little more than a week after a 13-year-old girl died of the deadly H5N1 virus. The government said the fresh outbreak was discovered earlier this week in chickens and ducks raised in a familys backyard farm in Kampong Cham province, 124 kilometres east of the capital Phnom Penh. We have a new outbreak of bird flu, Agriculture, Forest and Fisheries Minister Chan Sarun told AFP. The discovery came after **the Cambodian girl died of bird flu last Thursday,** becoming the kingdoms seventh fatality from the H5N1 virus. Her death prompted the government to launch a week-long bird flu awareness blitz.Following the latest outbreak, authorities killed some 100 chickens and ducks at the backyard farm in the eastern province, said the minister. **Cambodia** has been praised by the United Nations for its rapid action against bird flu, which has helped spare it from the human and poultry deaths suffered by its neighbours. afp

Published	2007.04.16
Disease	Avian Influenza
Begin	2007.04.12
End	2007.04.12
Location	Cambodia
Country	Cambodia
Total	--
Status	dead
Descriptor	the Cambodian girl
Confidence	1
Source	http://www.dailytimes.com.pk
Document events	1 **2**

Fig. 4: *An epidemic event extracted from a document.*

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-document aggregation (in section 4.1). Second, it allows us to explore methods for recovery from locally-best but incorrect guesses by using global information.[7]

In the next section, we will discuss how these features of the PULS system are used in the combined, multi-source system.

4 Integration of MedISys and PULS

This section will describe the integration between MedISys and PULS, and will try to demonstrate that, even in its current, preliminary state, the integrated whole is greater than the sum of its parts.

A special RSS tunnel has been set up between MedISys and PULS. At present, PULS is able to process only English-language documents. MedISys forwards documents which it categorizes as relevant to the medical domain through the tunnel to PULS. Currently, the documents arrive as plain text, with no layout information (paragraphs, sections, etc). 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 analyzes 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, and does not affect the functioning of both sites, which are inter-operating normally in real-time.

4.1 Multi-document Aggregation

When documents are received from MedISys, PULS performs the following processing steps:

First, the IE system analyzes the documents, extracts incidents, and stores them in the local database (*doremi.cs.helsinki.fi/jrc*). Second, PULS uses 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 3.2. 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.[8]

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 prepares a batch of recent incidents to return to MedISys, for displaying on its pages. The goal is to return a set of recent incidents with high confidence and low redundancy—a complete yet manageably-sized set of news for MedISys users to explore.

The batch is restricted to documents published within the last 10 days; from this period, PULS re-

[7] This line of our current research is not covered in this paper.

[8] In the PULS tables, confident attributes are set off in bold text, and confident incidents are highlighted in blue.

turns 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.[9]

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

4.2 Performance

We now discuss some of the on-going evaluations of the currently deployed systems.

The number of documents PULS receives from MedISys is approximately 10,000 per month. From 2,700 of these, PULS extracts approximately 6,000 incidents per month, on average. (It is quite common for a relevant document to contain more than one incident.) The remaining 6,300 documents fed to PULS by MedISys produce no incidents. That is to be expected, since MedISys does not explicitly search for outbreaks, but for any *mentions* of disease names, and many documents may mention the crucial diseases in the context of new vaccines or treatments, eradication campaigns, etc.

To determine what proportion of these 6,300 documents actually do contain events—and are therefore false negatives from the perspective of PULS—we randomly selected and checked 100 MedISys documents that produced no events. Of these, 15% contained an event that the IE system missed.

From the perspective of MedISys, this roughly indicates that at least 64% of the documents fed to PULS on average contain no events. This confirms that the IE component indeed serves its purpose by helping to distinguish reports about epidemic outbreaks from other discussions concerning diseases.

About 20% of all extracted incidents are rated as confident. We tried to estimate the accuracy of the confidence heuristics. We selected 100 confident incidents at random, and checked their correctness by hand. Without employing a rigorous (e.g., MUC-style) evaluation, we consider an incident to be correct only if all of its principal attributes are correct (no partial credit). This evaluation yielded: 72% of the confident incidents are correct; in 14% of the cases, the information extraction is erroneous, i.e., PULS extracts an incident where there should be none; in 14% of the cases, the confident incident is incorrect—for at least one attribute, the top-ranked value is not correct. 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 is simpler to correct, as it usually entails some tuning of the knowledge bases. Thus, if we could correct the erroneous cases with some tuning labor, we might expect the confidence measure to be correct in just under 84% of the confident incidents.

Since outbreak aggregation is our primary means of reducing redundant information in the flow of news, it is important to have an estimate of the accuracy of the outbreak calculation. We analyzed a randomly chosen set of medium-sized outbreaks, 20 outbreaks, about 10 incidents each. For each incident we tried to determine whether it was appropriately included in the outbreak. We found that 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.[10]

22.5% of the examined incidents were confident (i.e., on average, the outbreaks contained only 2–3 confident incidents).

5 Conclusion and Future Work

The public and the restricted MedISys applications are currently independent of each other, and they provide different functionality. The medical event information is only available on the public site. The two systems will soon be integrated in order to allow a single entry point and visual presentation. Registered users will receive access to more functionality and more alert definitions. Depending on the users' access rights, they may get access to newswires and to commercial sources as well.

We further plan to integrate a tool that automatically extracts terms from the comprehensive medical thesaurus MeSH (Medical Subject Headings),[11] and to allow users to select articles by browsing and drilling down in the multilingual MeSH hierarchy. This will give the user an alternative entry point to the same information.

We need to resolve some technical problems to improve the quality of the input data. One problem relates to the way MedISys extracts textual content from

[9] Note that under this arrangement, 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 decision that aims to balance the tension between recency of *publication* vs. recency of actual *occurrence* of an incident: both may be important to the user. Note also that in any case *all* events are always available in the PULS database for browsing.

[10] It was interesting to observe that aggregation is often useful even when the outbreak consists entirely of incorrectly analyzed incidents. E.g., 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 analyzed incorrectly, the error is *consistent*, and the grouping is still useful: it helps reduce the load on the user by aggregating related facts.

[11] www.nlm.nih.gov/mesh

source sites. Because the original focus of MedISys was on the keywords contained in the text, it ignored document layout information (such as headings, sub-headings, by- and date-lines, paragraph breaks, etc.), which provides important cues when detailed text analysis is required. The lack of this information is known to confuse the IE process, and needs to be addressed to improve IE accuracy.[12]

We are currently investigating methods for extending the measure of local confidence to global confidence, across multiple documents and sources.

We also plan to develop methods for MedISys to exploit the information returned from PULS in novel ways. One problem that needs to be studied is to what extent the outbreaks extracted by MedISys based on keyword frequencies agree with outbreaks extracted by PULS, and how they can be best integrated. Another path under consideration is to incorporate the PULS confidence as a criterion for the urgency of MedISys alerts. The current scheme, based on cumulative statistics, assumes that if something is newly prominent in many news sources, it is urgent or interesting news. However, in some cases, news that appears everywhere is "dated" news—it is already highly publicized. For timely surveillance, it is also interesting to detect outlier reports—those that have not yet achieved wide publicity, but in this case it is crucial that the system be certain that it was correctly identified. Here, a high score on the PULS confidence scale may serve as a complementary criterion for the urgency of an event.

Acknowledgements

This work was in part supported by the Academy of Finland. We wish to thank the anonymous reviewers for their thorough and helpful comments.

References

[1] C. Best, E. van der Goot, K. Blackler, T. Garcia, and D. Horby. Europe media monitor—system description. Technical Report 22173 EN, EUR, 2005.

[2] R. Grishman, S. Huttunen, and R. Yangarber. Information extraction for enhanced access to disease outbreak reports. *J. of Biomed. Informatics*, **35**(4), 2003.

[3] W. Lin, R. Yangarber, and R. Grishman. Bootstrapped learning of semantic classes from positive and negative examples. In *Proc. ICML Workshop*, Washington, DC, 2003.

[4] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A. Forslund, and C. Best. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of LREC-2006*, Genova, Italy, 2006.

[5] R. Steinberger, B. Pouliquen, and C. Ignat. Navigating multilingual news collections using automatically extracted information. *Journal CIT*, 13(4), 2005.

[6] R. Yangarber. Counter-training in discovery of semantic patterns. In *Proc. ACL-2003*, Sapporo, Japan, 2003.

[7] R. Yangarber, L. Jokipii, A. Rauramo, and S. Huttunen. Extracting information about outbreaks of infectious epidemics. In *Proc. HLT-EMNLP 2005*, Vancouver, Canada, 2005.

[12] Extracting document layout accurately is a highly non-trivial problem, since source sites are completely unstandardized, and in general the layout is hard to infer automatically.