Global Public Health
Vol. 7, No. 7, August 2012, 731–749

Uncovering text mining: A survey of current work on web-based epidemic intelligence

Nigel Collier*

National Institute of Informatics, Tokyo, Japan

(*Received 20 October 2011; final version received 6 March 2012*)

Real world pandemics such as SARS 2002 as well as popular fiction like the movie Contagion graphically depict the health threat of a global pandemic and the key role of epidemic intelligence (EI). While EI relies heavily on established indicator sources a new class of methods based on event alerting from unstructured digital Internet media is rapidly becoming acknowledged within the public health community. At the heart of automated information gathering systems is a technology called *text mining*. My contribution here is to provide an overview of the role that text mining technology plays in detecting epidemics and to synthesise my existing research on the BioCaster project.

Keywords: natural language processing; BioCaster; text mining; artificial intelligence; ontologies; evaluation; web-based discovery; social media

Introduction

Epidemic intelligence (EI) is the early identification, assessment and verification of potential public health hazards (Paquet *et al*. 2006) and the timely dissemination of alerts to appropriate stakeholders. The discipline includes both indicator surveillance techniques such as sentinel networks of physicians as well as event techniques that gather data from Internet-based digital news media (Hartley *et al*. 2010) as well as official sources such as World Health Organisation (WHO) alerts. Event techniques, in particular, with their emphasis on sifting through large volumes of dynamically changing unstructured data, lie at the crossroads where public health and informatics intersect. The technological discipline that has grown from this and similar interactions is called text mining (Hearst 1999). Text mining is a relatively new human language processing technology that aims to meet the knowledge discovery needs of professionals struggling under pressure of information overload, be it from the need to find facts and opinions on the Internet or making new discoveries in literature databases like PubMed's Medline (Swanson 1986). Text mining aims to discover novel information in a timely manner from large-scale text collections by developing high performance algorithms for sourcing and converting unstructured textual data to a machine understandable format and then filtering this according to the needs of its users. In later stages, text mining systems perform domain analysis (e.g., to determine topical details or identify aberrations from past norms) and

*Email: collier@nii.ac.jp

ISSN 1744-1692 print/ISSN 1744-1706 online
© 2012 Taylor & Francis
http://dx.doi.org/10.1080/17441692.2012.699975
http://www.tandfonline.com

deliver results in customised forms so that users can rapidly synthesise situations of interest (Feldman and Sanger 2006).

Whilst dictionary-based search techniques certainly have their role to play, text mining usually goes far beyond keyword searching used by traditional search engines to find needles in the proverbial haystack. Rather the task can be characterised as a race to find a needle with a particular colour, weight and length. Uncovering documents on the topic of malaria for example, is no guarantee that the information contained in them is relevant to discovering a new epidemic. What is needed is to condense the facts contained in the document into a fixed format – an event frame – that embodies all aspects of interest to the expert. Is there a case reported, what are the symptoms and how severe are they? Where and when did the event happen? By incorporating sophisticated knowledge models, text mining aims to understand the meaning – the semantics – of texts, albeit in a limited area of human expertise.

While text mining has application in many real life scenarios as diverse as business intelligence, patent searching and market surveying, my focus here will be to highlight its contribution to the alerting of public health hazards in the online media and to briefly categorise the relevant methods and resources available. I conclude this article by discussing possible future trends and research issues.

Background

As shown by Hartley *et al.*'s survey paper (2010), event-driven surveillance systems are now widely used by national and trans-national public health organisations such as the WHO, the Centers for Disease Control and Prevention (CDC) and the European Centre for Disease Prevention and Control (ECDC), Public Health Agency of Canada (PHAC) and many other agencies. In November 2002, at the start of the SARS epidemic, the Global Public Health Intelligence Network (GPHIN) system (Mawudeku and Blench 2006) at PHAC was among the earliest, along with the ProMED network (Madoff and Woodall 2005), to provide early warning of the impending near-pandemic starting in Guandong Province in Southern China. During the A(H1N1) influenza pandemic in 2009, a number of systems are credited with the timely discovery of early events including MedISys (Steinberger *et al.* 2008), Veratect (Wikipedia 2009), HealthMap (Brownstein *et al.* 2008) and BioCaster (Collier *et al.* 2008). Tools such as Riff from InSTEDD (Fuller 2010) were used to enhance decision support by integrating signals from virtual teams of experts with multiple streams of data from EI systems such as EpiSpider (Tolentino *et al.* 2007), SMS and electronic medical records in OpenMRS. Additionally, the MEDCollector system aims to integrate multiple Web-based sources (Zamite *et al.* 2010). Of historical interest are two early systems: Proteus-Bio (Grishman *et al.* 2002) and MiTAP (Damianos *et al.* 2002).

Figure 1 illustrates the range of services available in the BioCaster EI system, produced by an international team based in Japan. As an example of the power of semantics driven text mining considers the following scenario. A public health expert is interested in finding out about a possible fatal case of person-to-person transmission of A(H5N1) in a family in Thailand. The expert who is in the field logs into a public Web portal on her smartphone and enters A(H5N1) as the search term along with *Thailand*, the date range of interest and requests only English language news articles. Internally the system recognises that the first term is an

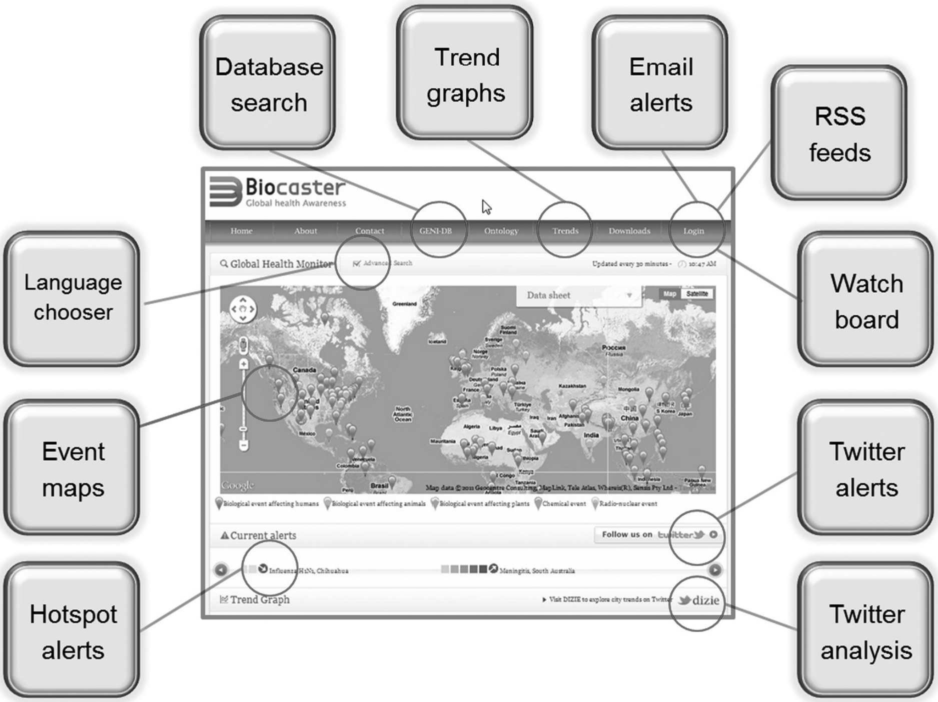

Figure 1. The BioCaster portal (http://born.nii.ac.jp) is a 24/7 system designed to deliver a variety of methods for enhanced access to epidemic events reported in news and social media.

English variant of an index term in its disease ontology (*highly pathogenic H5N1 avian influenz*a). The search is performed over thousands of possible events stored in the database but the results do not appear relevant to the expert's need. The system then offers the user the choice of searching using the disease symptoms. The user selects to search using symptoms such as *cough*, *high fever, pneumonia*, *acute respiratory distress* and all their synonyms. This time an article is found but the report is already two weeks out of date and missing some vital pieces of information about the name of the district and hospital. The user then chooses to search the Thai news and the search is automatically repeated using Thai term equivalents. A structured table is produced summarising each event in English with a flag indicating high priority items. The expert then finds the event that she is searching for and initiates a risk analysis procedure by transferring the event data to a secure watchboard for sharing with colleagues. In summary, the key component in this system is the analyst herself, but the technology has enabled her to increase her productivity by rapidly gaining insight into the context of a cluster outbreak so she can help her colleagues make a more informed decision. The EI system has enabled her to supplement whatever indicator-based information sources might have been available to her and to communicate better with her human network of contacts. Though I do not claim that mining the Web for reports is the only viable solution to EI, it is possible that without this service the expert might initially have had to rely on word-of-mouth, circulated news clippings or hit-and-miss ad hoc searches.

Table 1. Summary of steps in text mining systems for epidemic intelligence.

Data ingestion is usually the first stage with a variety of textual sources such as emails, homepages, Really Simple Syndication (RSS) feeds, Microsoft office files and Portable Document Format (PDF) documents.

Data cleaning is vital in practice to remove unwanted noise from the text (such as advertisements or links to unrelated news stories) and to join together broken sentences. At this stage systems often try to breakdown large documents that talk about multiple topics into separate sections in a process called Zoning in order to remove noise or reclassify the document (Chanlekha *et al.* 2010).

Data triage assigns the document a topic category for either trashing – in the case of non-relevant documents – or subsequent processing using detailed fact extraction. At this stage redundant information – multiple reports of the same event – are detected through document clustering. This stage is also intended to remove the most obvious true negatives but systems may struggle to handle the more subtle cases on the borderline of their task definitions leading to high numbers of false positives.

Fact extraction obtains structured information about an event such as the name of the disease, the type of agent, the number of victims and time and location where the event happened. With this information the computer can then begin to answer questions such as what happened, to who, where and when.

Ranking is done by applying rules on the results of earlier stages of processing. High-end systems will use sophisticated statistical analysis to assign an alerting level based on a comparison of aggregated data in the present and past. In practice, this is often the most difficult stage for systems to perform automatically with high levels of accuracy.

Human judgement is a key stage in the process. It is almost always needed to understand what is abnormal, to discovery rare events that the system may have missed, to make the final decision about vague reports and to link together disparate events. The limitations of the system will be most visible to the user at this stage and they have to apply their own judgments to correct for nuances of meaning that are clear to people but opaque to the computer software. Human analytical skills will also be able to discovery regularities in the data that can lead them to investigate new paths not available to current automated approaches.

The aforementioned scenario represents the high-end of automated EI systems but is feasible by fully applying today's technology. The availability of Web 2.0 services such as mapping (e.g., Google Maps[1]/Bing Maps[2]), news aggregation (e.g., Google News[3]), photo sharing (e.g., Flickr[4]), video sharing (e.g., YouTube[5]), social media (e.g., Twitter[6]), text mining services (e.g., Open Calais[7]) and data converters (e.g., Google Translate[8]) along with traditional Linux–Apache–MySQL–Python (LAMP) architectures has made it possible to rapidly and cheaply deploy systems that can ingest, filter and visualise news data and individual reports posted on microblogging sites like Twitter. As I illustrated in the example, high-end systems combine such generic services into so-called Web 2.0 *mashups* together with specialised knowledge of the domain in order to reduce ambiguity and increase precision. Interfaces often employ web-mapping services such as Google Maps to organise data simply across time and space. Users can then explore domain-specific relations, drill down, aggregate across events and communicate their findings and interpretations to colleagues.

Text mining services running on the back-end of such systems incorporate a rich fusion of technologies from natural language processing, machine translation (MT), ontologies and reasoning. The challenges to these technologies are to make accurate interpretations of massive volumes of multilingual text in near real-time and then make judgements about whether the detected events violate domain norms. Seemingly innocuous contexts such as vaccination campaigns, bursts of media interest in politicians/pop idols such as *Obama Fever/Bieber Fever*, and vague reports of mystery illnesses are all challenge areas for automated text understanding. Trying to see through the fog of media interest to extrapolate case counts is also a challenge area complicated by the seeming lack of correlation with published news reports.

In the remainder of this article I will look in more detail at some of the issues surrounding text mining services which lie at the heart of semantic data extraction from free text at the same time as synthesising my group's research in this area over the last six years.

Core technologies

In this section I aim to give a broad impression of the automated technologies involved in text mining for EI. Events start with the biology in the real world and then through a process we still know too little about, media organisations report some of these events in digital form. From this point text mining systems have a chance to pick up the story in a trawl of the Web and convert the free text data into a structured event frame for sharing (see Table 1). The news story as a structured event frame is then analysed using both statistics and human analysts. This might lead to the event being flagged as an immediate alert for verification, put on a watch list or archived for future reference.

While my focus is on automated methods, human users naturally have a vital role to play at many levels: (1) skilled human analysts perform risk analysis and verification, (2) the general public can help suggest or rate reports in a process called crowdsourcing, e.g., in HealthMap and (3) users of social media sites can comment on their own health conditions on open access social media sites such as Twitter which can be aggregated for trend detection, e.g., in BioCaster's DIZIE project (Collier and Doan 2011).

Data sourcing

Whilst accurate statistics are hard to find, the World Wide Web (Web) is now one of the primary information sources for people seeking information (Janson and Spink 2006). Anyone with Web browsing software has almost instant low-cost access to an extensive range of electronic news reports, blogs, search, academic bulletins, etc. EI systems can tap into this data in a variety of ways.

The lowest cost option for computers to systematically work through this wealth of information is to harness a Web crawler. When pointed at a list of news sites this software will systematically trawl the links and download any pages that are new. Such an approach though incurs a hidden cost in the maintenance of software to decode the HTML template for each Web site so that informative content can be separated from non-relevant content such as metadata, adverts, images, headlines for other stories and hyperlinks. Given the huge variety of templates and their constant revision the manual effort in maintaining such software is considerable. Several groups have developed generic content discovery algorithms based on heuristic rules and statistical models, e.g., (Lin and Ho 2002), but ready to use software may be difficult to find in the public domain.

A more efficient approach to locating news is to use the power of really simple syndication (RSS) feeds – syndicated news provided in a structured XML format. This option allows EI systems to regularly poll news servers, pull-out links to new stories and download their content. The issue of content discovery on the news page is still a problem, though.

Although freely available public news aggregators such as Google News and Yahoo News have access to a very wide range of sources, for mission critical systems as well as to ensure coverage, several EI systems have contracts with private news aggregation companies such as Factiva and LexisNexis. These companies offer the widest possible range of sources across a variety of languages with clean content. A practical question for system builders is to ensure quality of geographic coverage. This is not always so simple to achieve given the inherent biases in each media source.

Text analysis

Once news articles have been captured, the first stage of semantic analysis is to filter them for topical relevancy. The techniques used here that have enjoyed the most success are usually data driven based either on supervised (Conway *et al*. 2009), semi-supervised (Torii *et al*. 2011) or unsupervised machine learning. These techniques are distinguished by how much use they make of pre-classified example data.

Text mining systems are designed around a clearly defined task specification such as a case definition. For example, ‘Identify all infectious disease outbreak reports that contain evidence for human to human transmission’, or ‘Identify all events consist with the International Health Regulation Annex 2 Decision Instrument’.

To convert the unstructured data from a Web document into a structured event frame the computer requires knowledge about the syntactic and semantic structure of the language as well as the target output structure. This requirement tends to make text mining a language and domain-specific technology requiring interdisciplinary collaboration to develop system rulebooks. Building expert knowledge into a computer system for a specific task is economical only if the text collection is very large – such as the Web – and the nature of the information being found makes it very valuable to users. In addition to custom-built EI systems such as BioCaster, HealthMap, Epispider and MediSys, several private companies market generic text mining solutions including SAS, SPSS, Nstein and LexisNexis. Widely used open source toolkits include NLTK[9], the R project’s text mining package[10] and Sheffield University’s GATE project[11].

For computers to extract high quality information from text requires some degree of linguistic understanding. Systems typically require two sets of knowledge – domain knowledge that show the classes of objects of interest and their relationships and the patterns that show how these relationships are realised in the language of an actual text.

Most text mining systems start with a specialised module for recognising the names of important entities in the text – a process called named entity recognition (NER) (Nadeau and Sekine 2007), which can be done using either data driven techniques such as support vector machines (SVMs) or rule-based techniques. We illustrate this with an example from the BioCaster system's rule book which has the following pattern:

D21:- name(disease) {list(%virus) 'outbreak'}

In the language of SRL (Collier *et al*. 2010) this rule indexed as D21 identifies objects of type DISEASE. It states that a sequence of words should be labelled as a DISEASE type if it matches to an entry in the virus list and is followed by the string 'outbreak'. The output of this rule is to insert information into the text in the form of inline XML annotation for use in later processing steps. For example, the text 'The AH1N1 outbreak occurred in communities across the region' would be recognised internally as 'The < DISEASE > AH1N1 outbreak < /DISEASE > occurred across the region'. Following from NER is usually a stage of normalisation so that surface forms of names get linked to a unique identifier in a dictionary or ontology (i.e., a structured conceptual representation of the terms and relationships in the domain).

In SRL more sophisticated rules can be made to identify relations consisting of one or more objects like DISEASE, VIRUS, PERSON, SYMPTOM, ORGANIZATION, LOCATION and so on. For example:

FW99: farm_worker('true'):- 'death' 'of' name*(person,P) {list(@farming_occupation)}

Rule FW99 is another string matching rule that looks for sequences of words showing the death of farm workers. If the rule matches then it outputs 'farm_worker("true")', i.e., the left hand side of the ':-'. The rule states that the string must match with a PERSON type containing a farming occupation listed in the dictionary such as abattoir workers, breeders, livestock handlers, veterinarians, ranchers etc. So, for example, the text 'The ministry announced the death of <PERSON>2 slaughterhouse workers</PERSON> from the virus' would successfully match this rule.

While regular expression patterns like SRL can be quite effective, they are vulnerable to sensitivity constraints due to the large variety of surface patterns that need to be explicitly modelled. As in biomedical applications, more robust solutions are expected to come from full sentence parsing to uncover grammatical relations between words and phrases. Full parsing will also help to capture subtle aspects of the event such as polarity, certainty and temporality that can be hard to capture using regular expressions. However, full parsing may come at a cost to computational efficiency, potentially creating a bottleneck when timeliness is one key criterion for usability. This is particularly important during bursts of information that can occur during major epidemics.

Understanding time and location are key foundations for high quality EI (Chanlekha *et al*. 2010). In practice, though, there are many pitfalls. Document time stamps for example, are not necessarily the best guides to deciding on the time when a reported event took place. For example a document dated 2 October 2008 might report 'Last Tuesday avian influenza virus A was identified as the cause of an outbreak in two southern provinces of Viet Nam'. We would expect the text mining system to record the date of the case as the 30 September 2008.

In practice location names are also often highly ambiguous. For example, an equine influenza outbreak in Camden during the summer of 2007 would have to be identified as Camden near Sydney, Australia and not as Camden in London, UK. Equally confusing for automated systems is the fact that an outbreak of Venezuelan haemorrhagic fever might not be taking place in Venezuela and an outbreak of a food-borne disease from eating Satsuma's would probably have no relation to Japan. Much research has taken place on identifying geo-political named entities such as countries and cities in general news texts, e.g., (McCallum and Li 2003), with performance for English place names generally in the 1980s to low 1990s F-score on unseen texts, where F-score is the harmonic mean of recall and precision. Keller *et al.* (2009) provide a review of the issues for epidemic surveillance and present a new method for tackling the identification of a disease outbreak location based on neural networks trained on surface feature patterns in a window around geo-entity expressions. The resulting 64% F-score appears at first sight to be lower than we might have expected. The performance gap may be due to the variety of contexts in which geographic expressions for disease outbreaks occur and the lack of training data available. Contextual information for deciding on whether one of many mentioned locations mentioned in a report is the actual disease outbreak location is often dependent on contextual clues outside the scope of a single sentence. For example, a local hospital may be mentioned as the place of treatment and the attributable source may be mentioned as a health ministry spokesperson from the country's government. Since local names tend to be highly ambiguous both within and across countries, an EI system has a high chance of making a mistake in geo-coding the event based only on this first piece of information. It requires a combination of clues from the health ministry name and the local name to fix the actual specific location.

Because geo-temporal disambiguation is so difficult and because of the variety of ways in which cases are described across different news reports, it is challenging to completely de-duplicate news reports about events and obtain accurate tracking of case counts. An approach that might begin to tackle this was the spatio-temporal event calculus proposed by Chaudet (2006). Although the knowledge representation seems stable and repeatable, it is not clear yet how easily this can be operationalised.

Ontologies

It is clear that some a priori knowledge over and above that supplied in the media report is necessary for the text mining system to make sense of the report, e.g., to resolve sense ambiguities such as knowing that A(H1N1) influenza, swine flu and swine flu A all refer to the same disease, understand idiomatic expressions such as Venezuelan Hemorrhagic Fever and to exclude implausible contexts such as vaccination campaigns. Where does domain knowledge come from? Working systems often incorporate a fusion of knowledge both statistical and symbolic. For example, Keller *et al.*'s (2009) use of a neural network to detect the focus location of the outbreak is a statistical approach, and BioCaster's SRL rules for resolving the focus disease agent is a symbolic approach. Here I focus on the role of ontologies in EI, which is to help automate human understanding of key concepts and relations so that the desired level of filtering accuracy can be achieved.

One of the most important functions of ontologies is to decide how alike two concepts are to each other. Biomedical ontologies minimally contain lists of terms and their human definitions, which are then given unique identifiers and arranged into classes with common properties. These classes are then structured according to principles of classification such as the subsumption (is a) relation. For example, the Medical Subject Headings (MeSH) ontology (Lowe and Barnett 1994) says that the term 'influenza, human' is a type of *respiratory tract infection*. Other widely known examples of ontologies for human understanding include SNOMED Clinical Terms (Price and Spackman 2000), the Foundation Model of Anatomy (Rosse and Mejino 2008) the Unified Medical Language System (UMLS) (Humphreys and Lindberg 1993) and AGROVOC (Soergel *et al.* 2004). Community efforts such as the Open Biomedical Ontologies (OBO 2011) have come a long way in recent years towards forming standards for ontology construction, highlighting common pitfalls in their construction and promoting inter-operability.

In the domain of EI it is necessary to identify and link term classes such as DISEASE, SYMPTOM and SPECIES in order to separate reports about human, animal or crop diseases. We might also include a CHEMICAL class if knowledge of chemical or nucleotide agents were important. In order to capture geospatial reference we also need to define types for COUNTRY, PROVINCE and CITY. This would help to integrate information from the system with geospatial browsers such as Google Maps or NASA's World Wind.

Currently there are few dedicated publicly available ontologies that contain all the terms necessary for EI systems. In addition to the general purpose biomedical ontologies mentioned earlier, the commercial knowledge management tools Gideon,[12] has extensive coverage, contains a sophisticated reasoning engine and is widely used to support expert diagnosis but is closed source and not designed to interoperate with automated text analytics. Within open source resources, we have provided the BioCaster ontology (BCO) version 3 (Collier *et al.* 2010) in the OWL Semantic Web language to support automated reasoning across technical and laymen's terms in 12 languages for 336 conditions. The BCO supports a variety of relation types including term equivalence across languages, preferred term, causality between agents and conditions and between agents and symptoms. For example, if we find that a news document contains the disease 'chicken pox' then the ontology informs the system that the causal agent is the 'varicella-zoster virus', or if the news article mentions a disease outbreak of 'swine flu' and another of 'swine influenza A' then the ontology can provide a unifying root term of 'A(H1N1) influenza'. Another application for the ontology is in helping to choose appropriate levels of generality for disease names. For example, if the document mentions both 'Highly pathogenic H5N1 avian influenza' and 'avian influenza' then the event will be designated as the more specific of the two. In addition to human diseases it also covers animal diseases where the disease is a potential zoonotic threat to humans or can have severe economic consequences for society.

As a final note it is important to consider how to keep the ontology up to date. Although disease vocabulary is relatively stable, when new types of diseases strike such as 'swine flu' during 2009 the nomenclature can evolve surprisingly rapidly. In the future we would like to explore community efforts to harness expertise for solving this issue.

Machine translation

Given the very large volumes of media reports and the variety of human languages in which they are written, high throughput MT (Wilks 2009) is usually required in order to make sense of news events in the timeliest manner. MT systems have been in widespread use for many years, e.g., the Systran system used by the European Commission, or Yahoo!'s Babelfish used for Web page translation. The fidelity of MT output generally varies from high for cognate language pairs such as English–French to mediocre for non-cognate pairs such as English–Japanese or English–Arabic. One issue complicating the choice of MT system is that it is not clear yet how quality of output impacts on the final performance of the EI system although we have seen in our own evaluations that MT output has proven useful for improving the timeliness and sensitivity of alerting (Eysenbach 2002).

A variety of general purpose MT systems exist from commercial companies such as Google Translate or Microsoft's Bing Translate each allowing a wide range of language pairs at a cost that is typically based on the volume of text translated per month. Systems that can be installed and run on a local server such as the commercial Systran or the freely available MOSES have at least one advantage over general purpose MT systems which is that they can usually be customised to the domain vocabulary if sufficient quantities of example texts exist in both the source and target languages.

Machine translation is often employed before text analysis – translating all languages to a common target language such as English so that rule books do not need to be developed and maintained for each language. MT is also useful to help analysts make a first pass at understanding the topicality and significance of news reports. However, in the absence of fully automated high quality MT, end users will need access to bilingual analysts who can interpret the content and context of the source language directly.

Aberration detection

Being able to detect a news report about a public health event is not enough to make an EI system useful. In order to have value EI systems must be able to differentiate between mundane and unusual reports in a timely manner and supply this information to people who can initiate the appropriate actions. Such systems must be flexible to adapt themselves to changing patterns of diseases without any bias for a particular country or language. In practice, human experts with familiarity of the country concerned will almost always be necessary to analyse and interpret warning signals. The question for text mining researchers and users is how far can the technology be trusted to detect aberrations and what kind of aberrations are capable of automated analysis? Given that the state of the physical world with regard to disease incidence is always changing and that new pathogens are constantly evolving this is not a problem that can be tackled solely using the static ontologies I discussed earlier.

Detecting aberrations relies on identifying metrics that strongly correlate to the target objectives of the system designers – the discipline of infodemiology that was coined by Eysenbach (2002). News reports push the limits of what can be achieved using early warning data because of their biases, inaccuracies and vagueness. For example, the data can be strongly driven by fear and socio-economic biases which need to be compensated for. In addition to natural language processing, making sense of underlying trends draws on several established empirical disciplines: (1) knowledge discovery in databases (Fayyad *et al*. 1996) and, (2) time series analysis

(Wagner *et al.* 2001, Buckeridge *et al.* 2005) for change point detection. Many algorithms exist in both areas that can be adapted to the task at hand and compared.

The first stage in modelling begins by deciding on the objectives of the system such as coverage, alerting speed or low false alarm rates. A set of features are then identified, for example, the name of the disease and the country or province where it occurred, before establishing strong temporal and spatial baselines based on aggregated counts of these features over a history period. Deviation from such baselines by a significant margin constitutes an alert. Deciding on how to calculate the baseline and deviation, e.g., using statistical process control methods, is an on-going research topic (Buckeridge *et al.* 2005).

My previous work in BioCaster has looked at flagging aberrations for a broad range of diseases using features from the structured event frame, specifically the disease and country where the event took place. By using aggregated counts of news events I was able to obtain high levels of alerting performance on a range of diseases and outbreak sizes against ProMED as the silver standard baseline. I could also compare a range of models and feature types. Since the actual state of the physical world is not usually known, I considered ProMED's human moderated network to be a reasonable standard for event alerting. My comparisons of English and multilingual news (Collier 2010, 2011) showed high levels of performance for the CDC's Early Aberration and Reporting System's (EARS) C2 and C3 models (Hutwagner *et al.* 2003) with a 7 day baseline and 2 day buffer period. Both algorithms showed a good balance of F-score, timeliness and false alarm rates.

A different approach is adopted by (von Etter *et al.* 2010) who uses supervised classification on textual features using naive Bayes and SVMs to categorise outbreak events on a 0–5 scale of relevance (F-score 79.24% on SVM with an RBF-kernel).

Dissemination

Notifying alerts to users and other systems is the final key stage. At present no interoperable standard for message structure, semantics or vocabulary appears to have been agreed internationally among Web-based EI systems. Although standards such as the Common Alerting Protocol have been proposed, the most popular format currently in use may be GeoRSS, a lightweight XML format for syndicating links to Web content that encodes geographic information. Minimal necessary elements might include for example, a unique message identifier, the time of the message, the time of the event, a uniformly agreed name for the disease, the outbreak location, the species affected, a description of the reporting source, the degree of certainty, the level of confidentiality of the report, the status of the report (e.g., a trial exercise), message type (e.g., an update or an error notification) and a unique identifier for the event by the reporting system.

Case study: BioCaster

Background

BioCaster is a fully automated experimental system for near real-time 24/7 global health intelligence based at the National Institute of Informatics in Tokyo. Major goals of the research are (1) to explore advanced algorithms for the semantic annotation of documents, (2) to acquire knowledge which can empower human language technologies and (3) to investigate early alerting methods from news and open access social media signals. Analysis and validation of signals is assumed to take place downstream of the system by the community of users.

The concept of BioCaster (Collier *et al.* 2008) began in 2006 when grant-in-aid funding from the Japan Society for the Promotion of Science enabled the construction of a core high performance system (Collier *et al.* 2007) for semantic indexing of news related to disease outbreaks. At the start BioCaster's focus was on Asia-Pacific languages due to the perceived risk of newly emerging and re-emerging health threats in the region (Jones *et al.* 2008) such as highly pathogenic A(H5N1) influenza. Work therefore began in 2006 on the construction of a multilingual ontology (Collier *et al.* 2006) that would form the conceptual framework for the system – a freely available community resource containing a structured public health vocabulary.

The core team involved in BioCaster's development at the National Institute of Informatics is usually three or four members with expertise in computational linguistics and software engineering. In 2006, collaboration with a network of academic partners was quickly established including groups at the National Institute of Infectious Diseases (Japan), Okayama University (Japan), the National Institute of Genetics (NIG, Japan), Kasetsart University (Thailand) and the Vietnam National University (VNU, Vietnam). These groups provide expertise in software engineering, public health, genetics and computational linguistics across several languages. Since 2007, BioCaster has partnered with the Early Alerting and Reporting Project of the Global Health Security Action Group, a G7 + Mexico + EC + WHO initiative bringing together stakeholders, EI experts, and system owners to share expertise and develop a common Web-based platform.

Funding

BioCaster is a non-governmental system developed with grant-in-aid support from national funding organisations. In 2009 BioCaster was awarded a 3-year grant-in-aid by the Japan Science and Technology (JST) agency under the Sakigake programme to investigate enhanced health threat understanding by computers.

Output

BioCaster's implicitly intended users are analysts working at national and international public health agencies but there has also been considerable interest from physicians, veterinarians, researchers and the general public. Unique user numbers tend to be in the thousands per month but can rise substantially during major epidemics such as pandemic A(H1N1) and cholera in Haiti. As shown in Figure 1 BioCaster makes its output available in several formats such as Google maps, graphs, GeoRSS feeds and email alerts. The Web portal operates in two modes: (1) a freely accessible mapping and graphing interface called the Global Health Monitor (see Figure 1) and (2) a password restricted alerting interface which is currently used by a small test community of public and animal health experts. Additionally the open access multilingual ontology provides structured term sets in 12 languages and has been downloaded by over 250 academic, industrial and public health groups worldwide including the WHO.

Coverage

On a typical day BioCaster processes 30,000 reports. Of these approximately 55% will be in English, 11% in Chinese, 7% in German, 7% in Russian, 6% in Korean, 5% in French, 3% in Vietnamese, 2% in Portuguese, 2% in Chinese and the remainder in Thai, Italian and Arabic. Approximately 200 reports will be considered relevant after full analysis has taken place. About 80% of these reports will pertain to human cases and the remainder to animals with a very small number of plant diseases.

The range of health threats in BioCaster were prioritised according to notifiable diseases at health ministries in major countries in the Asia-Pacific region, Europe and North America as well as discussions with veterinarian and CBRN experts. In October 2011 the BioCaster database (GENI-DB) (Collier 2011) contained news event records (without personal identifiers) for over 176 infectious diseases and chemicals while the rulebook has the potential to find 182 human diseases, 143 zoonotic disease, 46 animal diseases and 21 plant diseases. Additionally 40 chemicals and 9 radio-nucleotides are also under surveillance.

Signals

In addition to direct signals on 18 concept types such as DISEASE, VIRUS, BACTERIUM, SYMPTOM and LOCATION names, BioCaster also looks for various event features such as international travel, drug resistance as well as a number of STEEP (Social Technological Economic Environmental Political) indicators. These include school closures, shortages of vaccines and panic buying of commodities.

Data sources

Data are ingested on a 1 hour cycle with approximately 27,000 news items analysed per day from news sources at a commercial news aggregation company, Google News, as well as various NPO and official sources such as WHO, OIE and European Media Monitor alerts. Additionally BioCaster's sister project in social media analysis (DIZIE) is analysing syndromic signals from the Twitter microblogging service. After testing is completed we expect to integrate DIZIE alerts within BioCaster.

User feedback

BioCaster has been used by a variety of public health organisations including the ECDC, the US CDC, the WHO and the Ministry of Health in Japan. User feedback has been encouraging both about the quality of information the system provided and its scope. Public health analysts have asked for us to customise the system to monitor mass gathering events such as the Shanghai Expo in 2010 or the London Olympics in 2012 as well as possible outcomes of environmental disasters such as the Gulf of Mexico oil spill in 2010. Animal health analysts have begun to see the potential for systems like BioCaster and have asked us to expand the range of diseases we monitor to include notifiable conditions for animals.

The area where we receive the most requests is in user interface. In 2006 we focused information on a global bio-geographic map. As BioCaster's coverage has increased we have found that the map can easily overwhelm users and an adaptable alerting system was needed. In 2010 we therefore introduced hotspot alerts to draw the user's attention to specific reports. However, there is still much to be done, for example in removing duplication, clustering related events and integrating reports across languages and media types.

The information we provide is inevitably biased by BioCaster's input sources, which rely heavily on Google News. In recent years we have expanded BioCaster's language coverage to include news in several other languages such as Spanish, Vietnamese and Chinese but the source engine still appears to have a US-centric focus with significant gaps for sub-saharan Africa and parts of middle-Asia. We are currently trying to supplement the system with other sources such as news aggregators in China. In a seminal study of EI systems, Lyon *et al*. (2011) compared BioCaster, HealthMap and Epispider over the period from 2 to 30 August 2010 and found similar timeliness between the system alerts as well as complementarities in geographical and language focus between all three systems. The report highlighted the issue of automated location detection, e.g., BioCaster's missing of Pakistan during the study period. We have since corrected this anomaly but in the process discovered a number of issues stemming from the transliteration into English of place names in certain locations.

Future developments

Our current work on aberration detection has touched upon only the explicitly stated facts in news media reports. More sophisticated text mining techniques hold out the potential for greater accuracy. For example, using multi-variate features such as STEEP indicators, or symptom severity features might help to piece together seemingly disparate facts in order to better understand the significance of rare events. An improved model for spatial dispersion of events would also help. For example, a report of a mystery illness in two villages in north-eastern Italy might not in itself be significant enough to trigger an alert. However, the report could take on more significance if it were combined with the facts that (1) there were an unusually high number of cases, (2) several victims complained of mild to severe joint pain and severe headache, (3) the first cases included a traveller from Kerala, India, (4) there had been a recent severe outbreak of Chikungunya in Kerala and (5) the health authorities were recommending precautions to prevent contact with mosquitoes and suspended all blood donations.

As a first measure, coarse grained granularity of time and location needs to be improved so that events can be pinpointed down to at least a city and a day of occurrence, reducing the 'late warning' issue that I noted in (Collier 2010) where the tail of news reports about past events gets confused with newer events that share the same geographic feature.

On the issue of evaluation, other domains of text mining such as literature mining for bioinformatics (Hirschman *et al*. 2002) have made enormous progress in assessing quality, expanding participation and improving performance by organising shared evaluation challenges. In evaluations such as the DARPA sponsored TREC, TIPSTER and MUC, systems are compared against a common task-based benchmark, allowing for both technical comparisons as well as user-based evaluation However, adequate care needs to be taken to avoid ‘inbreeding’ of participating systems through over-sharing of methods and resources. In contrast, in Web-based EI there has been relatively little community organisation around evaluation or the sharing of tools and data. One recent study by Vaillant *et al.* (2011a) shows progress in this area by comparing seven EI systems for CBRN threats with a focus on sensitivity evaluation from a French public health perspective. Vaillant *et al.* show that by combining data from at least four systems over 94% sensitivity can be achieved. This result corroborates an earlier extrinsic evaluation highlighting high sensitivity and high timeliness perceived by users including international EI experts (Vaillant *et al*. 2011b).

So far I have implicitly assumed that digital news reports should be the main source of information for EI systems. In reality, the landscape of digital sources is much richer: search queries, micro-blogs, digital radio, discussion boards, images, livecasts etc. Several works have already appeared looking at the potential to make use of individual health reports in Twitter (Corley *et al*. 2010, Culotta 2010, Lampos and Cristianini 2010, Signorini *et al*. 2011) for tracking influenza-like illness. Pearson correlations with CDC surveillance reports from sentinel providers and UK GP reports have been very encouraging. Although microblogs have no editorial control, they contain a direct real-time view into the health conditions of individuals. Another source that has received attention are search engine query trends from Google and Yahoo! (Ginsberg *et al*. 2008, Polgreen *et al*. 2008). As with all short message sources the challenge here is to interpret the search query's context – a user may query about a particular drug or health condition for a variety of reasons, e.g., general interest, a school report or concern about a health condition. Ginsberg's study clearly showed the potential to closely correlate query counts with CDC influenza data but research questions remain, particularly about geographic coverage as well as coverage across particular age groups, e.g., the young or old who may not be familiar or have access to the Internet. Other sources such as digital radio, potentially useful for countries in parts of Africa, SMS and livecast reports have yet to be explored.

The need for high performance computing to process data in real-time and adjust to surges during pandemics is a practical barrier to entry. Future systems may develop based around cloud computing services that are becoming available from companies such as Amazon, Google and Microsoft.

Conclusion

In this article I have just begun to uncover the surface of the complex technical aspects that Web-based EI system developers have grappled with over the last decade. Future developments in text mining will undoubtedly be necessary to harness the increasingly massive volumes of media and social network data and to combine this with non-media sources. Readers who wish to delve further into the issues raised here may find more detailed sources in several survey papers. Hartley *et al.* (2010) outline several active EI systems, Kosala and Blockeel's paper on mining the Web (Kosala and Blockeel 2000) raises many issues that are still relevant today and Howard

Burkom's tutorial slides[13] from ISDS 2008 provide an excellent foundation for getting to grips with aberration detection along with R project software packages[10]. Among text mining books two accessible sources include Berry and Kogan (2010) and Feldman and Sanger (2006). Data counts from the BioCaster system are available for study at GENI-DB database[14] (Collier and Doan 2012).

Notes

1. Google Maps: http://maps.google.com
2. Bing Maps: http://www.bing.com/maps
3. Google News: http://news.google.com
4. Flickr: http://www.flickr.com
5. YouTube: http://www.youtube.com
6. Twitter: http://twitter.com
7. Open Calais: http://www.opencalais.com
8. Google Translate: http://translate.google.com
9. The Natural Language Toolkit: http://www.nltk.org/
10. The R project: http://cran.r-project.org/
11. Sheffield University's GATE project: http://gate.ac.uk
12. Gideon: http://gideononline.com
13. Howard Burkom's 2008 ISID tutorial slides: http://isds.wikispaces.com/ISDS + Conference + Workshop + Materials
14. The GENI-DB database: http://born.nii.ac.jp/

References

Berry, M.W. and Kogan, M., 2010. *Text mining: applications and theory*. Edison, NJ: Wiley.

Brownstein, J., Freifeld, C., Reis, B., and Mandl, K., 2008. Surveillance san frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *Public Library of Science Medicine*, 5 (7), 1019–1024.

Buckeridge, D., Burkom, H., Campbell, M., Hogan, W.R., and Moore, A.W., 2005. Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, 38 (2), 99–113.

Chanlekha, H., Kawazoe, A., and Collier, N., 2010. A framework for enhancing spatial and temporal granularity in report-based health surveillance systems. *BMC Medical Informatics and Decision Making*, 10(1), e43.

Chaudet, H., 2006. Extending the event calculus for tracking epidemic spread. *Artificial Intelligence in Medicine*, 38 (2), 137–156.

Collier, N., 2010. What's unusual in online disease outbreak news? *Journal of Biomedical Semantics*, 1 (1), 2.

Collier, N., 2011. Towards cross-lingual alerting for bursty epidemic events. *Journal of Biomedical Semantics*, 2 (Suppl. 5), S10.

Collier, N. and Doan, S., 2011. Syndromic classification of Twitter messages. *Proceedings of eHealth*, 21–23 November, Malaga, Spain, arXiv:1110.3094.

Collier, N. and Doan, S., 2012. GENI-DB: A database of global events for epidemic intelligence. *Bioinformatics*, 28 (8), 1186–1188.

Collier, N., Doan, S., Kawazoe, A., Matsuda Goodwin, R., Conway, M., Tateno, Y., Ngo, Q., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., and Taniguchi, K., 2008. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24 (24), 2940–2941.

Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R., Takeuchi, K., and Kawtrakul, A., 2006. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language Resources and Evaluation*, 40 (3–4), 405–413.

Collier, N., Kawazoe, A., Shigematsu, M., Taniguchi, K., Jin, L., McCrae, J., Dien, D., Hung, Q., Takeuchi, K., and Kawtrakul, A., 2007. Ontology-driven influenza surveillance from

Web rumours. *Proceedings on Options for the Control of Influenza VI (Options 2007)*, 17–23 June, Toronto, Ontario, Canada.

Collier, N., Goodwin, R.M., McCrae, J., Doan, S., and Kawazoe, A., 2010. An ontology-driven system for detecting global health events. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 23–27 August, Beijing, China, 215–222.

Conway, M., Doan, S., Kawazoe, A., and Collier, N., 2009. Classifying disease outbreak reports using n-grams and semantic features. *International Journal of Medical Informatics*, 78 (12), e47–e58.

Corley, C.D., Cook, D.J., Mikler, A.R., and Singh, K.P, 2010. Text and structure data mining of influenza mentions in Web and social media. *International Journal of Environmental Research and Public Health*, 7, 596–615.

Culotta, A., 2010. Detecting influenza outbreaks by analyzing Twitter messages. *Southeastern Louisiana University Technical Report*. Available from: arXiv:1007.4748v1 [cs.IR] [Accessed 25 July 2012].

Damianos, L., Ponte, J., Wohlever, S., Reeder, F., Day, D., Wilson, G., and Hirschman, L., 2002. MiTAP for bio-security: a case study. *AI Magazine*, 23 (4), 13–29.

Eysenbach, G., 2002. Infodemiology: the epidemiology of (mis)information. *American Journal of Medicine*, 113 (9), 763–765.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., 1996. From data mining to knowledge discovery in databases. *AI Magazine*, 17 (3), 37–54.

Feldman, R. and Sanger, J., 2006. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Fuller, S., 2010. Tracking the global express: new tools addressing disease threats across the world. *Epidemiology*, 21 (6), 769–771.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L., 2008. Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012–1014.

Grishman, R., Huttunen, S., and Yangarber, R., 2002. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35 (4), 236–246.

Hartley, D., Nelson, N., Walters, R., Arthury, R., Yangarber, R., Madoff, L., Linge, Y., Mawudeku, A., Collier, N., Brownstein, J., Thinus, G., and Lightfoot, N., 2010. The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, 3, e3.

Hearst, M., 1999. Untangling text data mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 20–26 June 1999, Maryland, USA, 3–10.

Hirschman, L., Park, J.C., Tsujii, J., Wong, L., and Wu, C.H., 2002. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18 (12), 1553–1561.

Humphreys, B. and Lindberg, D., 1993. The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81 (2), 170.

Hutwagner, L., Thompson, W., Seeman, M.G., and Treadwell, T., 2003. The bioterrorism preparedness and response early aberration and reporting system (EARS). *Journal of Urban Health*, 80 (2), i89–i96.

Janson, B. and Spink, A., 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42 (1), 248–263.

Jones, E., Patel, N., Levy, M., Storeygard, A., Balk, D., Gittleman, J., and Daszak, P., 2008. Global trends in emerging infectious diseases. *Nature*, 451, 990–993.

Keller, M., Freifeld, C.C., and Brownstein, J.S., 2009. Automated vocabulary discovery for geo-parsing online epidemic intelligence. *Bio Medical Central Bioinformatics*, 10, 385.

Kosala, R. and Blockeel, H., 2000. Web mining research: a survey. *Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining Explorations*, 2 (1), 1–15.

Lampos, V. and Cristianini, N., 2010. Tracking the flu pandemic by monitoring the social web. *2nd IAPR Workshop on Cognitive Information Processing (CIP 2010)*, 14–16 June 2010, Tuscany, Italy, 411–416.

Lin, S. and Ho, J., 2002. Discovering informative content blocks from Web documents. *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 23–26 July 2002, Alberta, Canada.

Lowe, H. and Barnett, G., 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association*, 271, 1103–1108.

Lyon, A., Nunn, M., Grossel, G., and Burgman, M., 2011. Comparison of Web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap. *Transboundary and Emerging Diseases* [E-publication ahead of print]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1865-1682.2011.01258.x/abstract [Accessed 25 July 2012].

Madoff, L.C. and Woodall, J.P., 2005. The Internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED. *Archives of Medical Research*, 36, 724–730.

Mawudeku, A. and Blench, M., 2006. Global Public Health Intelligence Network (GPHIN). *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, 8–12 August, Cambridge, MA.

McCallum, A. and Li, W., 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the Seventh Conference on Natural Language Learning*, 31 May–1 June 2003, Edmonton, Canada, 188–191.

Nadeau, D. and Sekine, S., 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30 (1), 3–26.

Paquet, C., Coulombier, D., Kaiser, R., and Ciotti, M., 2006. Epidemic intelligence: a new framework for strengthening disease intelligence in Europe. *EuroSurveillance*, 11 (12), pii = 665.

Polgreen, P.M., Chen, Y., Pennock, D.M., and Nelson, F.D., 2008. Using Internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47 (11), 1443–1448.

Price, C. and Spackman, K., 2000. SNOMED clinical terms. *British Journal of Healthcare Computing & Information Management*, 17 (3), 27–31.

Rosse, C. and Mejino, J.L.V., 2008. The foundational model of anatomy ontology. In: A. Burger, D. Davidson and R. Baldock, eds. *Anatomy ontologies for bioinformatics: principles and practice*. London: Springer, vol. 6, 59–117.

Signorini, A., Segre, A.M., and Polgreen, P.M., 2011. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *Public Library of Science One*, 6 (5), 19467.

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., and Katz, S., 2004. *Reengineering thesauri for new applications: the AGROVOC example. Journal of Digital Information*, 4(4). Available from: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel

Steinberger, R., Flavio, F., van der Goot, E., Best, C., von Etter, P., and Yangarber, R., 2008. Text mining from the web for medical intelligence. *In*: F. Fogelman-Soulié, D. Perrotta, J. Piskorski, and R. Steinberger, eds. *Mining massive data sets for security*. Amsterdam, The Netherlands: IOS Press, 295–310.

Swanson, D.R., 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30 (1), 7–18.

The Open Biomedical Ontologies (OBO), 2011. *The open biomedical ontologies* [online]. Available from: http://www.obofoundry.org/ [Accessed 25 September 2011].

Tolentino, H., Kamadjeu, R., Fontelo, P., Liu, F., Matters, M., Pollack, M., and Madoff, L., 2007. Scanning the emerging infectious disease horizon – visualizing ProMED emails using EpiSpider. *Advances in Disease Surveillance*, 2, 169.

Torii, M., Yin, L., Nguyen, T., Mazumdar, C.T., Liu, H., Hartlet, D.M., and Nelson, N.P., 2011. An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. *International Journal of Medical Informatics*, 80 (1), 56–66.

Vaillant, L., Nys, J., Gastellu-Etchegorry, M., and Barboza, P., 2011a. Enhancement of sensitivity with gathering Internet-based systems for early threat detection within the global health security initiative (GHSI): the EAR project. *Proceedings of eHealth*, 21–23 November, Malaga, Spain, (in press). Available from: http://electronic-health.org/poster_abstracts/ehealth2011_poster_GHSAG.pdf [Accessed 3 July 2012].

Vaillant, L., Barboza, P., and Arthur, R.R., 2011b. Epidemic intelligence: assessing event-based tools and user's perception in the GHSAG community. *Proceedings of IMED 2011*, 4–7 February, Vienna, Austria.

von Etter, P., Huttunen, S., Vihavainen, A. Vourinen, M., and Yangarber, R., 2010. Assessment of utility in Web mining for the domain of public health. *Proceedings of NAACL HLT 2010 Workshop on Text and Data Mining of Health Documents*, 5 June 2010, California, USA, 29–37.

Wagner, M.M., Tsui, F.C., Espino, J.U., Dato, V.M., Sittig, D.F., Caruana, R.A., McGinnis, L.F., Deerfield, D.W., Druzdzel, M.J., and Fridsma, D.B., 2001. The emerging science of very early detection of disease outbreak. *Journal of Public Health Management Practices*, 7 (6), 51–59.

Wikipedia, 2009. *2009 flu pandemic timeline* [online]. Available from: http://en.wikipedia.org/wiki/2009_flu_pandemic_timeline [Accessed 25 September 2011].

Wilks, Y., 2009. *Machine translation – its scope and limits.* London: Springer.

Zamite, J., Silva, F.A.B., Couto, F., and Silva, M.J., 2010. MEDCollector: multisource epidemic data collector. *Lecture Notes in Computer Science*, 6266, 16–30.