Open Source Text Based Biovigilance

Madhav Erraguntla, Ph. D.[1], P.E., Larissa May, MD[2], MSPH, Belita Gopal[1], Richard J. Mayer, Ph. D.[1], Perakath C. Benjamin, Ph. D.[1]
[1]Knowledge Based Systems, Inc., 1408 University Drive East, College Station, TX 77840, USA
[2]George Washington University, Washington, D.C., USA

Abstract–*Timely detection of disease outbreak events is of paramount importance for the defense against infectious diseases and biological threats. Internet-based communications can provide good situational awareness for countries where public data collection is inadequate, unreliable or missing. The key challenge is to sift through this vast amount of unstructured text to identify relevant reports and to extract disease related information into a structured format suitable for analysis. In this work, Natural Language Processing (NLP) techniques are used on data from news feeds, websites, and medical publications to extract key biological event data. We developed the Threat Assessment Dashboard (BioTHAD™) in order to improve detection and monitoring of biological events. We demonstrate that disease outbreak incidence and timing can be effectively extracted from open news sources using NLP. The BioTHAD™ application could serve as a model for tracking not only infectious, but chronic diseases and other types of events worldwide.*

Keywords: Biovigilance, Open Sources Based Surveillance, Disease Outbreaks, Natural Language Processing, Text Mining

1 Background and Significance

Most emerging infectious diseases and agents of bioterrorism present as nonspecific "flu-like illness," thus early detection requires the use of alternative sources of data rather than waiting for reporting of laboratory-confirmed diagnoses [1]. Because syndromic surveillance systems access and analyze data streams not typically available to departments of health, syndromic surveillance has the potential to identify unusual patterns of illness prior to definitive diagnoses and thus potentially closer to "real-time" than traditional surveillance systems [1]. Detection of aberrant activity using statistical tools to identify unusual spatiotemporal distributions of symptoms for further public health investigations augments traditional surveillance [1, 2, 3].

A rich source of data for bio-surveillance is the wealth of news articles posted to or available through the web. While widespread availability is a major advantage for these reports, the sheer volume of information routinely available on the web makes its utilization difficult. The detection of disease outbreaks and other potential biological threats is extremely challenging due to nuances of natural language used in news articles. Automated identification of relevant events would improve surveillance and increase research and response effectiveness. One approach to making that information more tractable for biovigilance community members is to automate the processing of news reports and to flag reports of potential interest, highlighting or tagging relevant elements within reports, and extracting data for subsequent analysis.

Extracting disease and biological event related information from unstructured text is a challenging problem because of the nuances of natural language [4, 5]. According to Kawazoe, et al., "in many cases, disease-related events are mentioned with verbs (e.g., 'infect'), verbal nouns (e.g., 'infection') and verb phrases. There are many synonymous event expressions; for example, infecting events can be expressed by many verbs and verb phrases such as 'infect,' 'transmit,' 'contract,' 'communicate (pathogen/ disease),' 'catch (pathogen/disease),' 'get (pathogen/disease),' as well as verbal nouns such as 'infection,' 'transmission,' 'contraction.'" Event recognition (in this case, finding reports about outbreaks of diseases) also occurs at the noun phrase ("outbreak of ...") and clause ("people died from ...") levels [4]. Also, sometimes, an infecting event is not mentioned directly, but only implied, as in a sentence like "A man died of bird flu" [6].

2 Approach

The BioTHAD™ technology has software agents that can be configured to monitor information sources such as news feeds, and medical publications on a daily basis. Currently, the BioTHAD™ technology monitors and downloads data from 15 sources including ProMed Mail [7], WHO, BBC Health News, CDC Morbidity and Mortality Weekly Reports, The Lancet Infectious Diseases, and BMC Infectious Diseases. HTML files are downloaded from these sites, text is extracted from these HTML files, converted to an XML format, and then processed by an NLP pipeline to extract the relevant information.

Figure 1 shows the process used to extract information from text sources. The first part of the process, shown in the top box, uses our in-house NLP pipeline. The steps involved in this initial part of the process are sentence boundary

detection, tokenization, part-of-speech tagging, phrase chunking, Subject-Verb-Object (SVO) assignment, clause segmentation and finally, named entity recognition. The first four steps internally use OpenNLP (http://incubator.apache.org/opennlp/), while the SVO assignment, clause segmentation and named entity recognition modules have been developed in-house. The SVO and Clause assignment stage involve splitting the sentence into clauses and finding SVOs in the sentence.

The main objective of the initial stages of text processing is to identify and classify the phrases which may be the constituents of event patterns. The steps shown in the green/top box are mostly domain-independent. The named entity recognition module recognizes generic named entities such as Persons, Locations and Organizations. On the other hand, the processes shown in the yellow/bottom box in Figure 1 are components that are domain-specific. The four components required are the domain concept tagger, event pattern matcher, inference of missing data, and data normalization. The following sections briefly outline each component.

2.1 Event Extraction

2.1.1 Domain concept tagging

The first step in domain event extraction (bottom box in Figure 1) is the Domain Concept Tagging. Unlike the named entity recognition module in the NLP pipeline (top box in Figure 1), the domain concept tagger tags concepts that are domain-specific. We identified the following items as key concepts associated with disease and biological events in the context of information extraction: Diseases, Symptoms, Pathogens, Antibiotics, Location, Date, Outbreak terms, Resistance terms, Victim Type and Severity. After identifying such domain-specific concepts, the extraction of domain-specific event frames becomes simplified. For example, the pattern 'disease killed victim' would match the string "cholera killed 7 inhabitants." In a similar fashion, noun phrases that could denote an unidentified disease will also be tagged. This helps to capture reports on outbreaks when the disease type is still unknown.

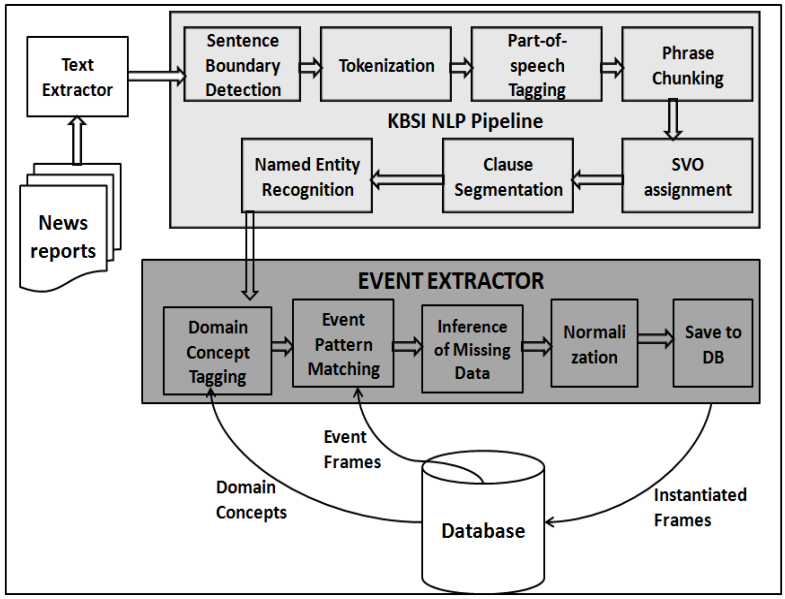

Figure 1. BioTHAD Information Extraction Pipeline

2.1.2 Event pattern matching

The next important step is the Event Pattern Matching which consists of feature extracting and the frame matching. The event pattern matching component will read all the pre-defined event patterns and compare them with the processed text to see if there are any matches. If a pattern is matched, an 'event' is generated

The Feature Extractor extracts relevant textual features that are generally very domain specific from the NLP output. Features such as noun phrases, disease, date, location, verb phrases, root form of verbs, etc., are extracted to produce a simplified representation of the processed text. This is illustrated in Figure 2.

Figure 2 shows an example of how text is processed as it proceeds through the various stages of the process. The input to the Feature Extractor is the tagged parse tree which comes from the NLP pipeline and the Domain Concept Tagger. By this stage, concepts such as disease and location have already been identified and appropriately tagged. The output of the Feature Extractor is:

NP SEGMENT(disease) + VP SEGMENT (surface) + PP SEGMENT (location)

The above feature vector means that this particular sentence contains the following components:

- A noun phrase which contains a disease, followed by,
- A verb phrase which contains the verb "surface," followed by,
- A Prepositional Phrase which contains a location.

The feature vector is then fed to the next module, the Frame Matcher. The Frame Matcher is the key module of the whole processing pipeline. Its main function is to match incoming feature vectors against a set of predefined event frames or templates. Figure 2 also shows the event frame that matches the feature vector from this particular sentence. One important input needed for this step is the repository of event frames that capture disease incidents in news report. The generation of these frames is described in more detail below.

The next step in the BioTHAD™ Information Extraction Pipeline is inferring missing data. Natural language is efficient and hence often key information is not explicitly mentioned because the human can infer it from the context. This is especially true for dates and locations.

2.1.3 Handling incomplete date specification

Heuristics are used to fill out missing or incomplete dates.

- Infer the year. Consider the example: "On 22 August the Ministry of Health (MoH) of the Central African Republic reported a laboratory confirmed case of yellow fever." To complete the date "22 August," the year from the published date is added to it.
- Infer the date from the context. Consider the example "Southern Missouri man dies of rabies." Since the date is completely missing from this sentence, it is assumed that the date is the same as the published date.
- Dereference the relative date. Consider the example "Southern Missouri man died of rabies yesterday." The actual date can be computed relative to the published date. Therefore, in this case, "yesterday" = Published Date - 1

2.1.4 Missing or ambiguous location information

Very often, an instantiated frame does not contain a location. In addition, many frames that do contain an explicit reference to a location, the source text does not contain the context of that location; whether the location is a city, a state or a country. For this reason, it is necessary to tag these three categories based on lookup lists. We created a lookup list of countries and in the case of the United States, the list also includes cities and states. The location lookup list is used to resolve a location and identify its country or state.

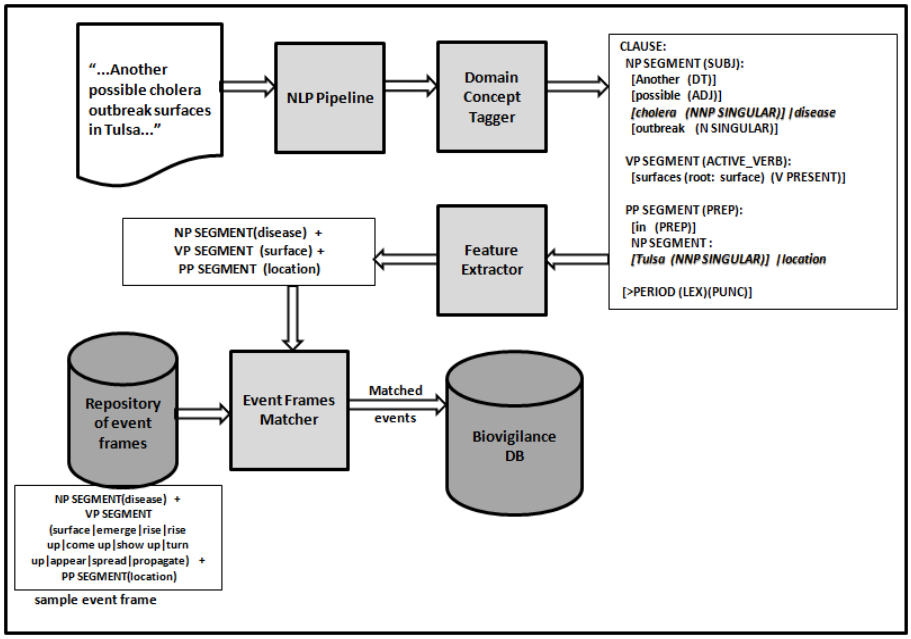

Figure 2. Text Processing and Event Extraction Process

2.1.5 Data Cleaning and Normalization

Once an event is generated, some of the instantiated event slots will need to be further refined or cleaned to generate consistent information. The following examples illustrate the need to normalize the data before saving the instantiated event to the database.

- Dates and date expressions (including dates relative to the report time, such as "last week") should be normalized to a standard form, with explicit day, month, and year.
- One disease could have many names. We designate one canonical name for each disease and normalize all

variants to that standard e.g., "Ebola," "Ebola hemorrhagic fever," "EHF."

- Similarly, some countries have many names; therefore, there is a need to normalize country names, e.g., "U.S.A.," "U.S.," "United States."

Further analysis of the stored data using techniques such as data mining will be performed on this atomic event extracted from the text.

After the frames are extracted, the information extracted undergoes further processing. This involves data cleaning and data filtering.

Open sources like news feeds and medical publications often have information related to statistical updates and summaries, prevention, negative outbreaks (i.e., documents reporting that an outbreak is abating) and reports about general information related to drugs and research, in addition to reports about disease incidents. It is necessary to filter out reports that have information content other than disease incidence reporting. We detail some of the types of filtering and aggregation below.

a) Filtering reports on statistical updates: Reports that provide information about statistical updates, such as total count of cases for a specific disease for a whole year and the worldwide situation of specific diseases are filtered out. This elimination consists of two steps. The first step removes reports that contain the pattern "update yyyy" and "world-wide" in the title. The second step filters out reports that contain the word "update" in the title and more than three locations in the text. Another example of statistical updates is temporal generalizations. For example, "Every year, cholera causes 100 deaths in Zimbabwe." Such instantiated events containing repeating time intervals are filtered out.

b) Filtering reports on prevention: The second type of report that has nothing to do with the outbreak of diseases is prevention related reports. After going through some sample reports on prevention, a list of keywords was created to denote that a report concerns prevention.

c) Filtering reports on multiple diseases: It was also decided that reports that discuss multiple diseases could potentially be noisy reports for the purpose of disease incidence detection. Therefore, these are also filtered out based on the number of diseases extracted from the text. To date we have not processed any report that contained more than three instances of a disease in the text.

d) Filtering reports on negative outbreaks: Another category of documents that must be filtered from the repository is negative outbreaks; i.e., documents reporting that an outbreak is either abating, or being controlled. A report is considered to fall in this category if the title contains one or more of the following words or phrases: under control, is over, waning, officially over, eradicate, recover, etc.

e) Filtering reports on drugs and medical research: Finally, another category of documents to be filtered out from the repository was reports focused exclusively on pharmaceutical and medical research. Reports whose titles contained any of the following concepts were eliminated: research, experiment, drug testing, pharmaceutical research, etc. Some of the reports eliminated might contain important information like approval of a drug or vaccine for a pandemic, but it is not within the current scope of the BioTHAD™ technology and those reports are filtered out.

2.1.6 Spatio-temporal aggregation

Since the disease outbreak information is being extracted from multiple sources, it is likely that multiple records are generated for the same incident by different sources. It is also possible that the same source reports information about an incident over multiple days. While displaying situation awareness information to analysts, the multiple incidents are aggregated and presented as single, major events. The aggregation is based on spatio-temporal proximity. Link list and nearest neighborhood-based clustering techniques are used to form the aggregate events [8, 9, 10].

3 Results

We validated the effectiveness of our approach against well known disease outbreaks. We present results from two of the validation studies in this section. The first analysis was to verify the data against the Haiti cholera outbreak that began in late October 2010 following the earthquake that shook the country in January 2010 [11]. The date range of the cholera outbreak events in Haiti extracted from news reports is shown in Figure 3 along with the actual date range of the outbreak as reported by CDC. The timeline analysis reveals that the date range of the extracted events coincides very well with the recorded date range for the 2009-2010 Haiti cholera outbreak.

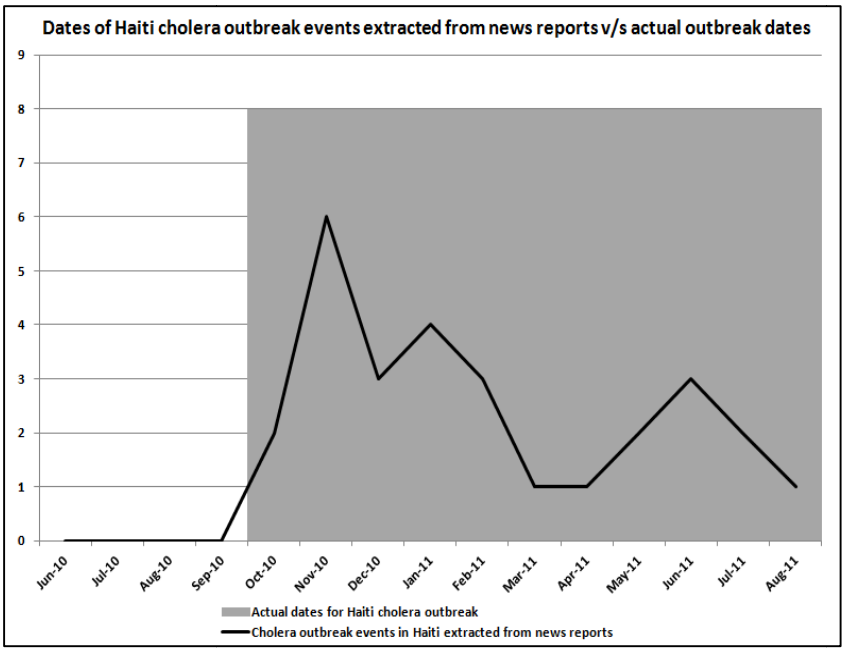

Figure 3. Trend of Cholera Outbreak in Haiti

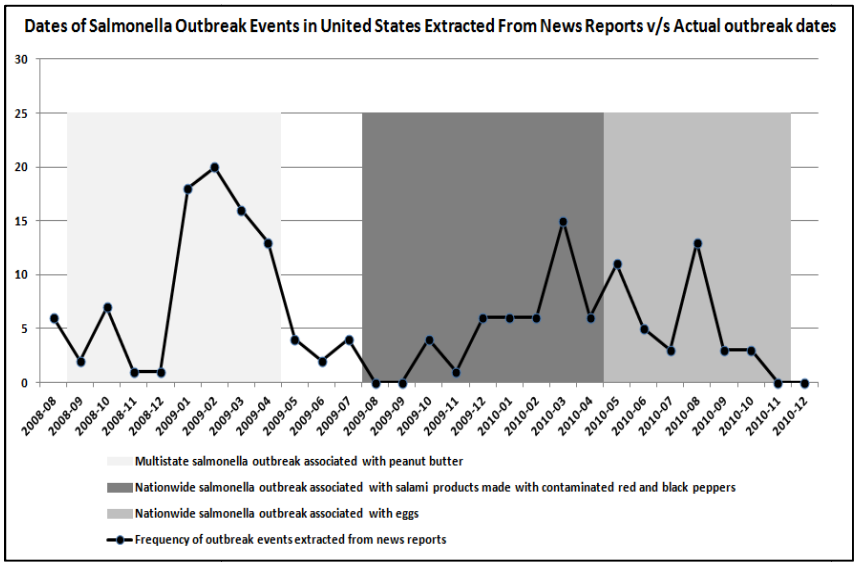

Figure 4. Salmonella Outbreak Events in the United States Extracted From News Reports

The next objective was to verify the trend for salmonella outbreaks in the United States. For this analysis, the date ranges of the extracted disease outbreak events were compared to three different known outbreaks of salmonella in the United States. The results of this comparison are illustrated in Figure 4, which also shows the three different actual salmonella outbreaks, namely:

- Multistate salmonella outbreak associated with peanut butter [12],
- Nationwide salmonella outbreak associated with eggs [13], and
- Nationwide salmonella outbreak associated with salami products made with contaminated red and black peppers [14].

Figure 4 shows a fairly close correlation between the three known outbreaks with the date ranges of the extracted disease outbreak events for salmonella in the United States.

4 Conclusion

Timely detection of disease outbreak events is of paramount importance for the defense against infectious diseases and biological threat events. News reports and medical publications of diseases and biological events are a

valuable source for collecting and organizing information regarding potential disease outbreak around the world. We present the methodology for extracting disease and biological events from open sources. In this paper, we have presented the design and implementation details of Biovigilance Threat Assessment Dashboard (BioTHAD™) that implements this capability. The BioTHAD™ technology provides good situational awareness into the status of disease and biological events across the world and is capable of detecting diverse biological events from open source reports. Analysis supported by the BioTHAD™ technology includes spatio-temporal aggregation and time-line visualization of disease outbreaks. Our approach is validated using a number of known disease outbreaks in recent years.

The BioTHAD™ technology is very effective in extracting disease incidents and their timing and has the potential for broad applications to the early detection and monitoring of not only infectious diseases, but chronic diseases and other events of national and global importance to biosecurity. We are currently validating and refining extracting victim counts and plan to extend our approach to extracting the etiology and nature of disease outbreaks (endemic, nosocomial, community acquired, etc). Another future direction is to design and develop a user interface for epidemiologic analysis and validation of intermediate results from BioTHAD™.

5 Acknowledgements

This work was performed under funding from two Office of Secretary of Defense (OSD) Small Business Innovative Research (SBIR) programs – Biosurveillance-based Integrated Outbreak Warning And Recognition System (BIOWARS) and Environment, Epidemiology, and Etiology Surveillance and Analysis Toolkit (E3SAT) [15, 16, 17]. We would like to acknowledge the support of our sponsor Dr. Kevin Montgomery.

6 References

[1] Stoto MA, Schonlau ML. Syndromic surveillance: is it worth the effort? *Chance* .2004; 17(1), 19-24.

[2] May L, Chretien JP. Pavlin JA. Beyond traditional surveillance: Applying syndromic surveillance to developing settings—opportunities and challenges. BMC Public Health 2009; 9:242.

[3] May L, Griffin BA, Maier Bauers N, Jain A, Mitchum M, Sikka N. Carim M, Stoto MA. Evaluation of emergency department chief complaint and diagnosis data for detection of influenza-like illness using an electronic medical record. The Western Journal of Emergency Medicine 2010;11 (1).

[4] Grishman R, Huttunen S, Yangarber, R. Real-time event extraction for infectious disease outbreaks. Proceedings of Human Language Technology Conference (HLT) 2002.

[5] Lu H, Zeng D, Chen H. Prospective infectious disease outbreak detection using Markov switching models. IEEE Transactions on Knowledge and Data Engineering 2010; 22 (4): 565-577.

[6] Kawazoe A, Chanlekha H, Shigematsu M, Collier, N. Structuring an event ontology for disease outbreak detection. BMC Bioinformatics 2008; 9 (3).

[7] Promed Mail. http://www.promedmail.org/

[8] Erraguntla M, Belita G, Ramachandran S, Mayer R. J. Inference of Missing ICD9 codes using text mining and nearest neighbor techniques. Submitted for Hawaii International Conference on System Sciences, 2012.

[9] Hoebe CJ, de Melker H, Spanjaard L, Dankert J, Nagelkerke N. Space-time cluster analysis of invasive meningococcal disease. Emerg Infect Dis. 2004; 10: 1621-1626.

[10] Si Y, Debba P, Skidmore AK, Toxopeus AG, Li L. Spatial and temporal patterns of global H5N1 outbreaks. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2008; 37: 69-74.

[11] CDC. Cholera Confirmed in Haiti, October 21, 2010. http://www.cdc.gov/haiticholera/situation/.

[12] CDC. Timeline of Infections: Multistate Outbreak of Salmonella Infections Associated with Peanut Butter and Peanut Butter-Containing Products — United States, 2008–2009. http://www.cdc.gov/salmonella/typhimurium/salmonellaOutbreak_timeline.pdf.

[13] CDC. Investigation Update: Multistate Outbreak of Human Salmonella Enteritidis Infections Associated with http://www.cdc.gov/salmonella/enteritidis/index.html.

[14] CDC. Timeline of Infections: Nationwide Outbreak of Salmonella Montevideo Infections Associated with Salami Products Made with Contaminated Black and Red Pepper United States, 2009 -2010. http://www.cdc.gov/salmonella/montevideo/montevideo_timeline2.pdf.

[15] BIOWARS. Biosurveillance-based integrated outbreak warning and recognition system (BIOWARS). OSD SBIR Phase II 2010; Contract No. W81XWH-08-C-0093.

[16] E3SAT. Environment, epidemiology, and etiology surveillance and analysis toolkit (E3SAT). OSD SBIR Phase II 2009; Contract No. W81XWH-08-C-0756.

[17] Erraguntla M, Ramachandran S, Wu C., Mayer, R J. Avian influenza data mining using environment, epidemiology, and etiology surveillance and analysis toolkit (E3SAT). Hawaii International Conference on System Sciences 2010.