Approved for Public Release; Distribution Unlimited
Case #04-0804

The MiTAP System for Monitoring Reports of Disease Outbreak

Laurie E. Damianos, Guido Zarrella, Lynette Hirschman
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730
{laurie, jzarrella, lynette}@mitre.org

The MiTAP system[1] [Damianos et al. 2002a, 2002b, 2003a, 2003b; MiTAP 2001] was developed as an experimental prototype using human language technologies for monitoring infectious disease outbreaks and other global disasters. MiTAP is designed to provide timely multi-lingual information access to analysts, medical experts, health services, and individuals involved in humanitarian assistance and relief work. Every day, thousands of articles from hundreds of global information sources are automatically captured, filtered, translated, tagged, summarized, categorized by content, and made available to users via a news server and web-based search engine. Information extraction technology plays a critical role in many of these processes, presenting information in a variety of time-saving mechanisms to facilitate browsing, searching, sorting, and scanning of articles. Machine translation provides analysts with access to foreign language information otherwise unavailable. We have created a novel prototype by integrating MiTAP with an expert system to help analysts and public health officials deal with overwhelming amounts of data and information in the biomedical domain, specifically relating to disease outbreaks. By providing the analyst with alerts to indications of disease-related activities, the prototype attempts to detect early signs of disease outbreak in non-traditional data sources, giving the analyst more time to focus on potentially interesting data while reducing the time spent investigating false alarms and insignificant events.

Background

Potentially catastrophic biological events are increasing in frequency. In recent years, the world has experienced SARS outbreaks, smallpox and anthrax scares, and the rapid spreading of West Nile Virus (see figure 1). Experts fear that other infectious diseases, like Rift Valley Fever, could arrive in the US in the very near future or that outbreaks of SARS could re-occur. Factors such as international trade and travel increase the potential economic and political impacts of major disease outbreaks. Yet it is difficult to detect a disease outbreak early enough to respond and contain it adequately. Modern disease surveillance systems rely on human medical

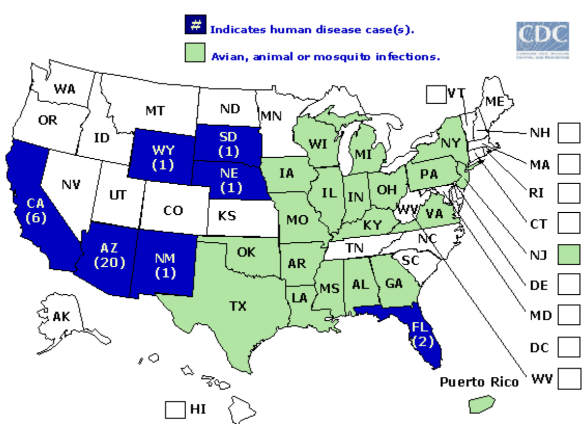

Figure 1 The 2004 spread of West Nile Virus across the US through June 2004. Courtesy of the CDC.

[1] The MiTAP system was funded, in part, by the Defense Advanced Research Projects Agency (DARPA) Translingual Information Detection Extraction and Summarization (TIDES) program under contract numbers DAAB07-01-C-C201 and W15P7T-04-C-D001. The extensions to support disease detection were funded by a MITRE Special Initiative for Rapid Integration of Novel Indications and Warnings.

data to track epidemic activity; an outbreak cannot be detected until patients start appearing in hospitals for treatment of symptoms and until those cases are reported to both local and global authorities.

The Problem

Even with accurate records of patient symptoms, the information is not always shared or integrated in a well-coordinated and timely manner. A few cases of flu-like symptoms at a single hospital or clinic will not raise an alarm. However, multiple hospitals in an area reporting cases of the same symptoms could indicate a major disease outbreak. This can only be understood if the hospitals communicate to an oversight organization that integrates data from multiple sources for a complete perspective.

After an outbreak has been detected, details may be reported in local news media and through other information channels. A disease surveillance system in other parts of the world may try to monitor global news reports to stay abreast of the latest outbreak information, but foreign language news is not always available. Even when it is, it may not be current by the time it has been translated and is in the hands of a medical analyst. With or without foreign language data, the overwhelming amount of data in English makes it nearly impossible for public health officials and medical analysts to acquire, manage, and digest critical information needed to detect and assess biological events in a timely manner. As a result, responders and health care workers have too little lead time to prepare for potentially catastrophic events.

Motivation and Hypothesis

The initial SARS outbreak occurred in southern China in late November of 2002. The global World Health Organization (WHO) alert was not issued until almost four months later on March 12, 2003. Within a month after the announcement, the epidemic had spread to over 3000 cases in 20 countries on all continents with over 100 deaths, and a worldwide panic had begun [WHO 2003]. Public health officials had not received timely information about the crisis from hospitals or governments and therefore were powerless to contain the disease or control public reaction.

Yet there were other, less direct indicators that may have provided advance clues of a disease outbreak. These included abrupt changes in purchasing patterns of staples or over-the-counter pharmaceuticals, or events such as hospital closings [Asia Times 2003a & 2003b, FBIS 2003b, Manila Bulletin 2003, Miami Herald 2003]. Had there been a comprehensive monitoring system for these types of activities, analysts may have been alerted to the gravity of the situation earlier, and public health officials may have been able to contain the rapidly spreading disease.

Our hypothesis is that, by identifying and automating the monitoring of indirect indicators of disease outbreak, we will be able to provide faster advance warning of a potential epidemic before it spreads. These non-traditional types of indicators usually appear in local media before global awareness of an event is ordinarily possible. It remains to be seen which of these indicators are most relevant and whether or not they can be discovered in the massive collections of real-time, noisy data available from international news organizations in electronic form.

Approach

To demonstrate the feasibility of using novel data sources to detect pre-event indicators related to disease outbreak, we are developing a semi-automated environment which captures data and allows analysts to investigate alerts and build a model iteratively.

The prototype automatically collects electronically accessible global data over the web in multiple languages. This includes newswire, Internet data, stock indices, environmental information, and transportation data. Using human language technology, the data is then filtered and translated, and relevant information is automatically extracted and annotated. News articles are categorized by topic and binned accordingly.

Analysts investigate the captured data retrospectively to detect indirect indicators for signs of disease-related activity as well as explicit mentions of disease. The indicators are then fed into the prototype as filters on incoming data. As data come in, the prototype monitors matches against the indicators. When thresholds are reached (as pre-determined by an analyst), the system triggers alerts which are fed into an expert reasoning component. The expert system consists of human-tailored, weighted rules which compare and correlate events from the data capture component and from other external sources to alert analysts to suspect activity. The analysts are presented with the related context and tools with which to investigate the alerts. Analysts can view the data in aggregate form and use their time investigating only the most promising and important events.

Indicators of Disease-Related Activities

Through retrospective analysis, we have produced a draft list of potential indicators of events surrounding disease outbreak. These pre-defined terms and keywords are used as filters on large quantities of data (roughly 5,000-10,000 articles a day in the current prototype). Combinations of these keywords are used as input to human-generated rules in the reasoning engine.

Exploratory Prototype

Our first semi-automated prototype for disease detection consists of a data collection component, an expert system (based on [JESS]), and a graphical user interface to analytical tools and the underlying data. Figure 2 depicts the interaction of these components and a sample of data sources ingested by the prototype.

Open source, electronic data is collected in multiple languages from local, regional, and national news sources. The articles are ingested by the MiTAP system, a suite of integrated human language technologies combined for the purpose of enabling humans to find and interpret relevant information quickly and effectively, independent of language or medium. The current system is limited to written material; however, the technology is being developed that would make it possible to collect broadcast news (TV and radio) in future versions [Palmer et al. 2004]. MiTAP was initially created for general tracking and monitoring of infectious disease as part of an experiment in biosecurity and later expanded to other domains for tracking of global events. For this work, MiTAP has been further configured to the disease domain [Damianos et al. 2004].

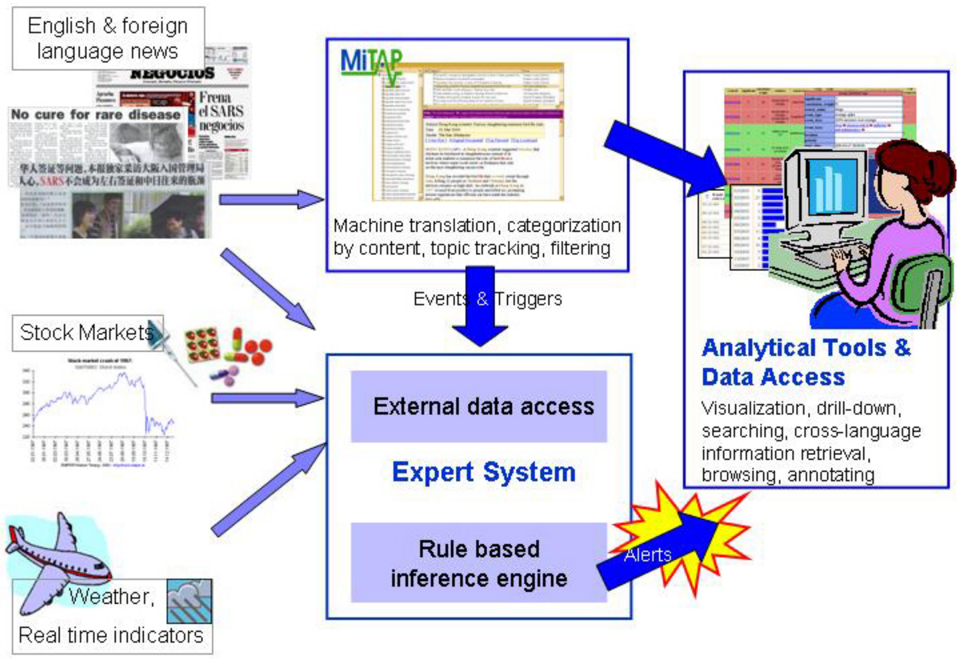

Figure 2 Early prototype for disease detection. Open source data is processed by the MiTAP system. Detected events are fed into the expert system which fuses, analyzes, and correlates data from multiple sources. Analysts have access to data and tools for further exploration. Based on retrospective analysis and iterative processing, analysts can modify the filters, rules, and models that system components use.

The foreign language articles are machine translated into English [Miller et al. 2001], and then all English articles (original language, human translated, and machine translated) are normalized into a common format, tagged, summarized, categorized by content, and made available to analysts via a news server [INN 2001] and web-based search engine [Jakarta Project 2001]. Information extraction technology [Aberdeen et al. 1996, 1995; Ferro 2001; Ferro et al. 2001; Mani & Wilson 2000; Vilain 1999; Vilain & Day 1996] plays a critical role in many of these processes, presenting information in a variety of time-saving mechanisms to facilitate browsing, searching, sorting, and scanning of articles. Information extraction is also used in novel ways to provide high level views of multiple documents, for example, by presenting the most frequently mentioned diseases each day [Baldwin et al. 2002. Figure 3 shows how information extraction is used to highlight critical keywords and entities in articles, providing information at a glance.

Subject	Dengue Fever Claims Another Life in Jeddah
Date	27 Jul 2004
Sender	Arab News (Jedda)

[Color Key] [Original Document] [Top Persons] [Top Locations]

Top Locations
Jeddah
Asia
Caribbean
Saudi Arabia

Author: Staff Writer

JEDDAH, 27 July 2004 — A second Saudi has d
result of dengue fever in a private hospital, Okaz daily reported yesterday.

The cases of suspected dengue fever reported in Jeddah hospitals has now reached 251. Of these 145 cases were confirmed by laboratory tests. Most of these cases were treated and only 11 patients remain in hospital.

A health care official in Jeddah said the health department was monitoring the situation, which was brought under control thanks to preventive measures adopted by the disease-control team. The official said a number of precautionary measures have been taken to prevent the disease from spreading. These include spraying mosquitoes and their breeding areas and increasing public awareness about the disease.

Dengue fever is a mosquito-borne disease caused by a virus. Dengue fever may occur in people of all ages who are exposed to infected Aedes mosquitoes.

The disease occurs mainly in tropical Asia and th
usually during the rainy seasons in areas with high
infected mosquitoes.

Color Key
Date/Time
Person
Location
Organization
Disease/Symptom
Victim

Figure 3 Information Extraction technology plays a critical role in the system. Here, keywords are highlighted in the text of an article, making scanning for critical information quick. Pop-ups list people and locations mentioned in the article.

After standard processing, each article is filtered on the indicator keywords, entities, or simple Boolean expressions of the indicators. These matching article counts are recorded and tracked daily. A count exceeding a given threshold triggers an event which is then fed into the expert system. For example, if a day's count of articles mentioning drug sales is 200% higher than the daily average, this might indicate an activity worth investigating and will trigger further analysis. Figure 4 graphs daily counts of articles that match the *drugs* filter which monitors news on pharmaceuticals, antibiotics, anti-inflammatories, etc.

The data spike, indicated by the arrow, is what triggered an event. Note that the thresholds were arbitrarily set by the developers in this early work. One of the interactive roles of the analyst is to refine these thresholds and tune the system iteratively.

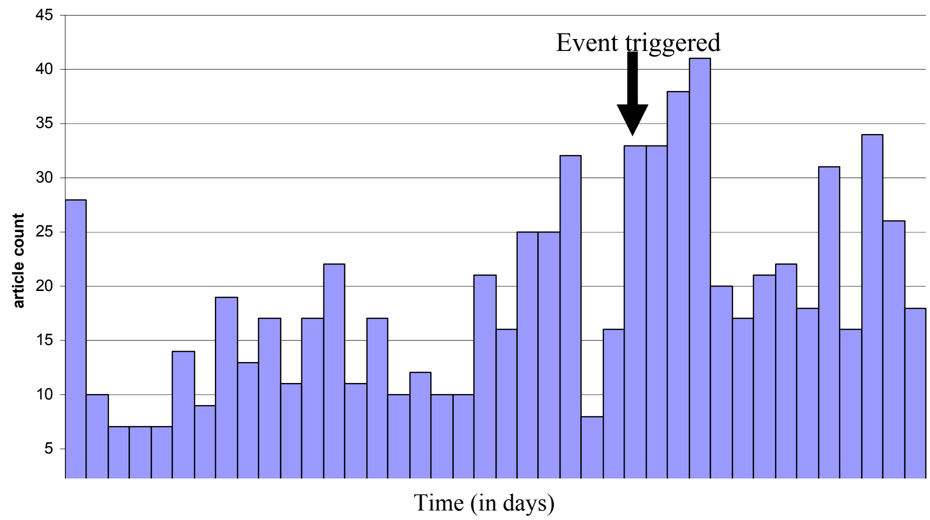

Figure 4 The prototype monitors the number of articles that match a specific filter pattern on a daily basis. If the number exceeds a specified threshold, an event is triggered and sent to the expert system for correlation with other pieces of information and then further analysis by a human. Note that data are not currently normalized to volume, weekends and holidays, or seasonal changes; this will be part of future work to validate the indicators and alerts.

The expert system performs data fusion on the events from MiTAP and other sources, compressing a large volume of information into a smaller, but more significant, set of events or alerts. In addition to the output from MiTAP, the expert system can monitor information events in [RSS] feeds and other dynamically updated information on the web or other electronic sources. The rule-based reasoning engine (with hand-authored rules) analyzes these events, identifies related events, correlates or invalidates other events, and generates an estimate of significance, complete with an audit trail of supporting and negating evidence. Figure 5 provides an example rule that was fabricated for purposes of demonstration. The reasoning engine can also supplement its knowledge base by performing a directed search via the query management component, allowing retrieval of information from a wide variety of sources including databases and web pages (e.g., stock market indices, transportation data, weather information, etc.). For example, the rule in figure 5 retrieves information about the state of the stock market in real-time from sources on the World Wide Web. The alerting component of the engine disseminates the resulting conclusions and associated references to the analyst. Figure 6 illustrates one possible interface to this information.

```
if: event is of type pharmaceuticals
and: the event indicates a spike in the count of articles
and: the pharmaceutical stock index has recently increased
then: increase the confidence in the event's significance by 85%
and: add a stock quote to the audit trail
```

Figure 5 Example rule, written in plain English, demonstrating how the prototype collects and combines context for significance of each event. This rule was fabricated for demonstration purposes.

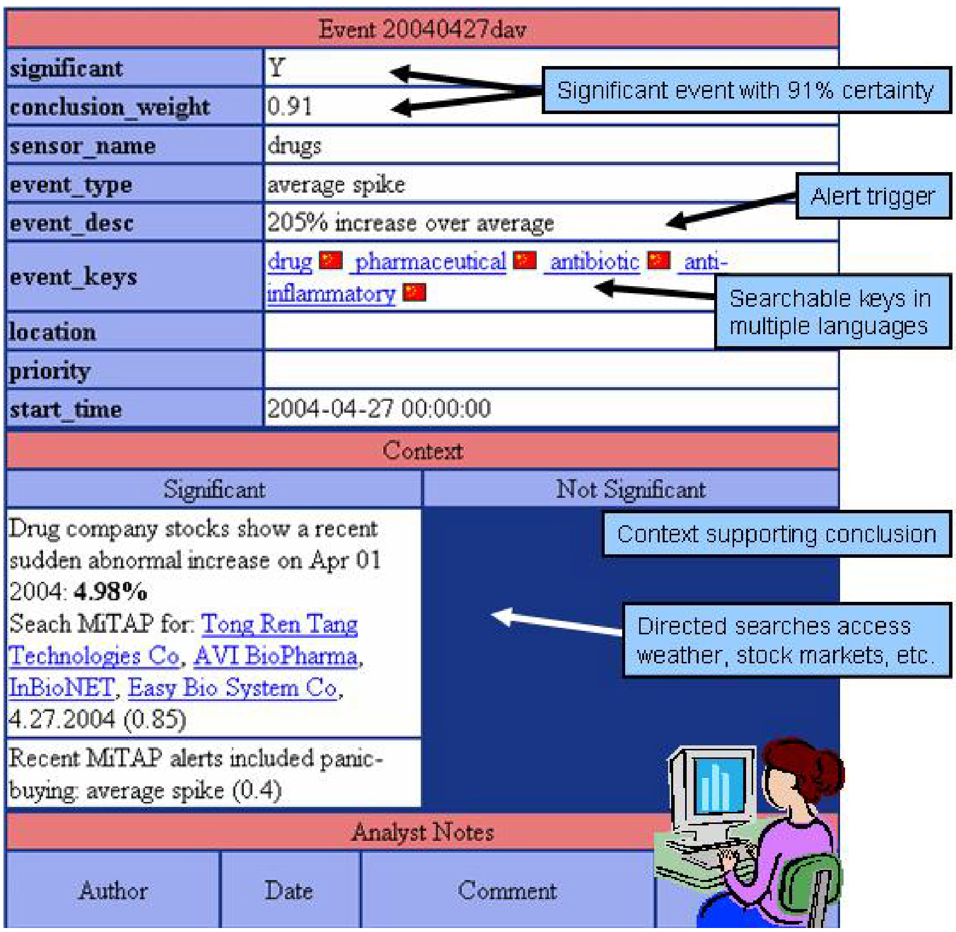

Figure 6 Example of a significant event presented to an analyst by the expert reasoning system. The event was triggered by a 205% increase over average in the number of daily articles referring to drugs, pharmaceuticals, antibiotics, and anti-inflammatories in a specific region. In support of this conclusion of 91% certainty, the expert system has directed searches of drug company stock markets which indicate a correlated sudden increase. Related events include a data spike in the *panic buying* filter. The prototype provides the analyst with access to the underlying data; keywords are searchable across both English and foreign language databases. Graphs (similar to that in figure 4) show the sensor data over time, and other links provide quick access to relevant data sources.

The analyst can drill down into the data, directing specific searches to follow leads and draw his own conclusions. An analyst can search on any of the filter keywords, in multiple languages, simply by clicking on the term. Results from these or any other directed search can be retrieved in list form or plotted along a timeline. Visualizing the data in a graph sometimes reveals patterns, trends, or peaks indicative of interesting activity. The graph of retrieved search results in figure 7 shows a data spike corresponding to a surge in mentions of disinfectant sales during the SARS outbreak in February 2003 – a month before the WHO announced the epidemic.

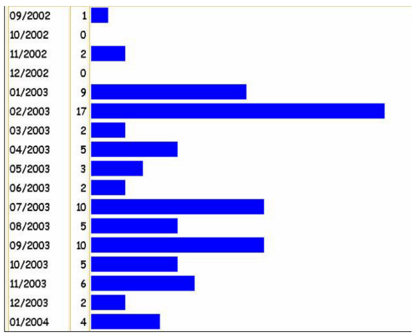

Figure 7 Search results can be visualized by plotting the aggregated number of articles along a timeline. Viewed in such a way, peaks and trends are more easily identified. This particular graph depicts articles about disinfectant sales and usage. The data spike during February 2003 corresponds to panic buying of disinfectant during the SARS outbreak – before WHO announced the epidemic publicly.

Relevant or interesting searches can be saved and used as new filters on incoming data. The analyst can also specify a threshold, or trigger, for alerting purposes and input into the reasoning system. In this way, the analyst can tweak the system and build a more accurate model for detection.

If an analyst chooses to perform a cross-lingual search [Darwish 2002] on the foreign language data, she can specify the search terms in English. The query is automatically translated and expanded, and the results are returned in English with links to the original foreign language documents. State of the art machine translation does not rival human translation but usually provides enough information for an analyst to determine whether a particular article is of significance, thus giving the analyst access to data otherwise unavailable. If a human translator is available, the original foreign language versions of the selected subset of articles can be retrieved and translated, reducing the need for a human to translate hundreds or thousands of articles.

The event interface also allows the analyst to visualize graphs of the filters, browse related data (e.g., weather, stock market indices, transportation data), and make annotations and judgments on the significance reported by the system.

Early Results and Discussion

As we integrated the components of the prototype, we arbitrarily set thresholds and created rules based on the exploratory keyword combinations. We selected just a subset of indicators to use as filters on the data and created some very simple rules in the expert system. Our goal was to demonstrate the proof of concept. Coincidentally, during the month we ran our exploratory prototype on open source data, there was another reported disease outbreak. We had unintentionally stumbled across a very real world situation in which to test our hypothesis. We ran the system in the real world environment and monitored the data for indicators of events that might be related to the disease we were investigating.

Surprisingly, even our poorly tuned demo system ingesting noisy, non-normalized data was able to detect potential outbreak activity before this information was made public. Figure 8 shows a high-level view of multiple events after analysis via the expert system. Each row represents a single event; red indicates significance as deemed by the expert system, green indicates non-significance. The entire set of events was detected during a single month. Note that several events occurred on a single day, and not every day had an event. What is most interesting about this screen capture is that it shows significant events occurring more frequently over time – as we approach the reported incident. The arrow indicates when the first new disease case appeared in the press.

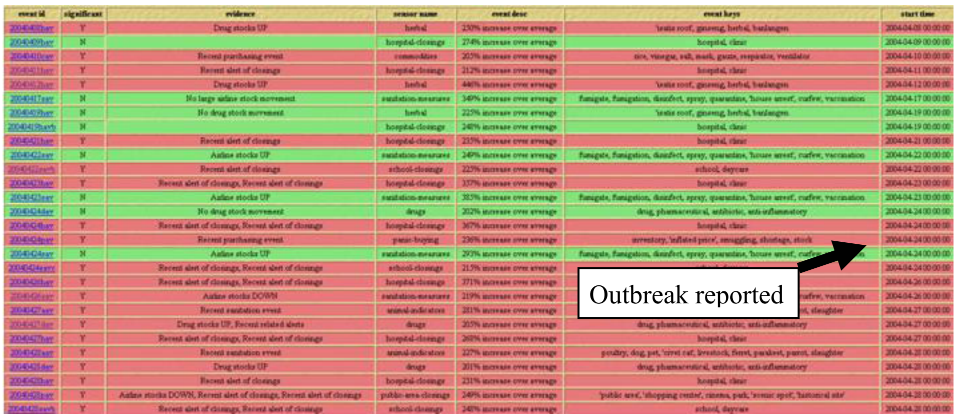

Figure 8 Multiple events during a single month. The increase in frequency and number of significant events (red rows) coincides with the reported disease outbreak.

To visualize the change in significant events over time, we have also plotted the alerts (both significant and non-significant) on a graph in figure 9. Again, it is clear that the alerts became more frequent and more significant approaching the date of the case reports.

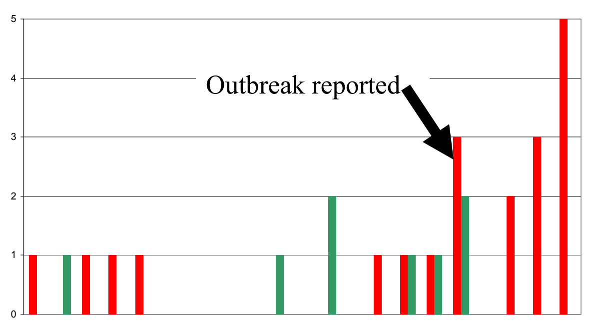

Figure 9 Number of events occurring during a single month, plotted along a timeline to visualize the increasing frequency and number of significant events coinciding with the announcement of the case reported.

As an additional test of our progress, we plotted the number of significant events related to indirect indicators on a timeline and compared their occurrence with actual news reports that specifically mentioned the disease. The results are shown in figure 10. The figure shows that alerts to potential indicators occurred at roughly the same time the report was made public. This lends some validity to the indicators we chose to monitor, demonstrating that our indirect cues do indeed appear in the media during fear of an outbreak. One confounding factor is that societies can become sensitized to particular diseases, and reactions to new incidences of that disease may be altered.

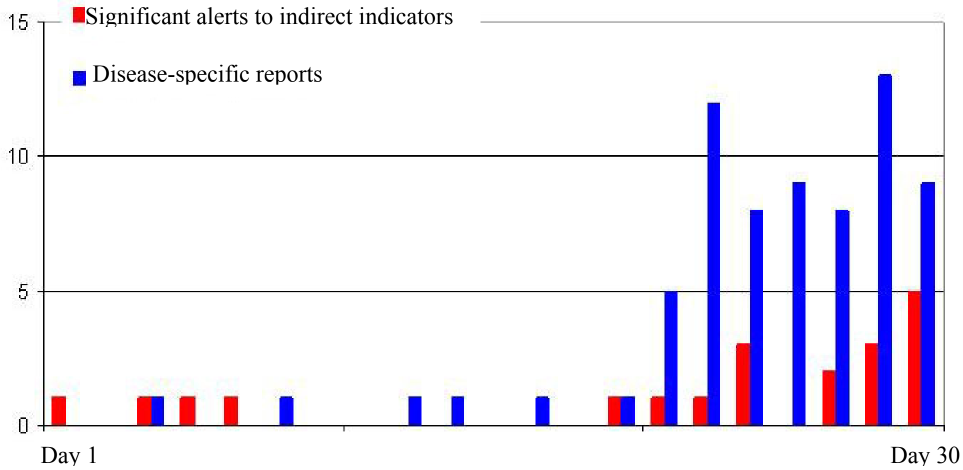

Figure 10 Number of indirect significant events (as determined by prototype) graphed with number of news reports specifically mentioning the disease. The prototype issued alerts before any specific mention of the disease.

Conclusions and Future Directions

By providing the analyst with alerts to indications of possible events surrounding an outbreak, the prototype attempts to detect early signs of the spread of disease, giving the analyst more time to focus on potentially interesting data while reducing the time spent investigating false alarms and insignificant events.

Although not a rigorous validation of our hypothesis that pre-event indicators of infectious disease outbreak are present weeks to months in advance of transnational spread of epidemics, the prototype shows promising potential for detecting and analyzing the indicators of disease-related activities, in non-traditional data sources. Despite using noisy and non-normalized data, non-validated indicators, and untested rules, the system demonstrated promising results in a real world situation. A medical analyst using the prototype before and during the disease reports would have had access to a collection of information and articles that would be useful in investigating the threat. This proof of concept will serve as a springboard for future iterative development, baseline establishment, and validation of the indicators and rules.

We plan to incorporate additional tools and visualization techniques to allow analysts a variety of ways to analyze the presented data. In addition, we plan on exploring alternate data sources such as Weblogs, weather and environmental data, and perhaps even transportation or telecommunication information, if accessible.

Once we have the opportunity to explore more fully and work closely with a team of analysts, we will devote our efforts to a formal evaluation of both the model and the system components. Eventually, a validated system would be an ideal complement to traditional biosurveillance and communications.

Acknowledgments

We would like to thank our MITRE colleagues for their contributions to this work: Samuel Bayer, Jason Bruzdzinski, John Henderson, Ann Huang, Tom McEntee, William Morgan, and Marion Warwick. We would also like to thank James Wilson and Marat Polyak of the Division of Integrated Biodefense ISIS Center at Georgetown University.

References

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Palmer, D., Robinson, P., Vilain, M. 1996. MITRE: Description of the Alembic System as Used in MET. In *Proceedings of the TIPSTER 24-Month Workshop*, 461-462, Tysons Corner, VA, 5-8 May 1996.

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., and Vilain, M. 1995. MITRE: Description of the Alembic System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 141-155, Columbia, MD, 6-8 November 1995.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., & Sadler, L. 1994. Machine Translation: An Introductory Guide. http://clwww.essex.ac.uk/~doug/book/book.html.

Asia Times: China vs SARS: A Good Dog is a Dead Dog. May 31, 2005.

Asia Times: HK Plays Down Pneumonia Fears. March 18, 2003.

Baldwin, B., Moore, M., Ross, A., Shah, D. 2002. Trinity Information Access System. *Proceedings of Human Language Technology Conference (HLT 2002)*, San Diego, CA.

Damianos, L., Ponte, J., Wohlever, S., Reeder, F., Day, D., Wilson, G., Hirschman, L. MiTAP for Bio-Security: A Case Study In *AI Magazine*, Volume 23, No. 4, pages 13-29, 2002(a).

Damianos, L., Ponte, J., Wohlever, S., Reeder, F., Day, D., Wilson, G., Hirschman, L. MiTAP, Text and Audio Processing for Bio-Security: A Case Study In *Proceedings of IAAI-2002: The Fourteenth Innovative Applications of Artificial Intelligence Conference*, Edmonton, Alberta, Canada, July 28 - August 1, 2002 (b).

Damianos, L., Wohlever, S., Kozierok, R., Ponte, J. MiTAP: A Case Study of Integrated Knowledge Discovery Tools in *Proceedings of the Thirty-Sixth Annual Hawaii International Conference on Systems Sciences (HICSS-36)*, Big Island, Hawaii, January 6-9, 2003 (a).

Damianos, L., Wohlever, S., Kozierok, R., Ponte, J. MiTAP for Real Users, Real Data, Real Problems in *Proceedings of the conference on Human Factors of Computing Systems (CHI 2003)*, Fort Lauderdale, Florida, April 5-10, 2003 (b).

Damianos, L. E., Bayer, S., Chisholm, M. A., Henderson, J., Hirschman, L., Morgan, M., Ubaldino, M., Zarrella, J., Wilson, J., Polyak, M. MiTAP for SARS Detection In *Companion Volume to Proceedings of HLT-NAACL 2004*, Boston, MA, May 2-7, 2004, pages 241-244.

Darwish, K. *PSE: A Small Search Engine written in Perl* 2002
http://tides.umiacs.umd.edu/software.html

FBIS. 5 Killed by Virus in Guangdong; Authorities Call for Emergency Measures. February 11, 2003 (a).

FBIS. SCMP on Official Efforts To Curtail 'Panic Buying' in Southern China. February 14, 2003 (b).

Ferro, L. 2001. Instruction Manual for the Annotation of Temporal Expressions. MITRE Technical Report MTR 01W0000046. McLean, Virginia: The MITRE Corporation.

Ferro, L, Mani, I, Sundheim, B, and Wilson, G. 2001. TIDES Temporal Annotation Guidelines: Version 1.0.2, MITRE Technical Report MTR 01W0000041. McLean, Virginia: The MITRE Corporation.

INN: InterNetNews, Internet Software Consortium 2001, http://www.isc.org/products/INN.

The Jakarta Project, 2001 http://jakarta.apache.org/lucene/docs/index.html.

JESS: the Rule Engine for the Java™ Platform http://herzberg.ca.sandia.gov/jess/

Mani, I. and Wilson, G. 2000. Robust Temporal Processing of News. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)*, AAAI Press, Menlo Park, CA, 69-76.

Manila Bulletin: SARS Downs 60 More People in HK; 13th Fatality Reported. March 30, 2003.

Miami Herald. Unidentified Illness Kills 5 in China. February 11, 2003.

Miller, K., Reeder, F., Hirschman, L., Palmer, D. 2001. *Multilingual Processing for Operational Users*, NATO Workshop on Multilingual Processing at EUROSPEECH.

MiTAP System 2001, http://mitap.sdsu.edu/.

Palmer, D., Bray, P., Reichman, M., Rhodes, K., White, N., Merlino, A., Kubala, A. Multilingual Video and Audio News Alerting In *Companion Volume to Proceedings of HLT-NAACL 2004*, Boston, MA, May 2-7, 2004, pages 245-246.

RSS RDF Site Summary http://purl.org/rss/1.0/spec

Vilain, M. 1999. Inferential Information Extraction in Pazienza, M., ed., *Information Extraction*, Lecture notes of the 1999 SCIE Summer School on Information Extraction. Springer Verlag, 95-119.

Vilain, M. and Day, D. 1996. Finite-state phrase parsing by rule sequences. In *Proceedings of the 1996 International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark, 5-9 August 1996.

World Health Organization. Communicable Disease Surveillance and Response: Update 83 – 100 Days into the Outbreak. http://www.who.int/csr/don/2003_06_18/en/ , 2003.