MiTAP for SARS Detection

Laurie E. Damianos, Samuel Bayer, Michael A. Chisholm, John Henderson, Lynette Hirschman, William Morgan, Marc Ubaldino, Guido Zarrella
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730
{laurie, sam, chisholm, jhndrsn, lynette, wmorgan, ubaldino,jzarrella}@mitre.org

James M. Wilson, V, MD and Marat G. Polyak
Division of Integrated Biodefense
ISIS Center, Georgetown University
2115 Wisconsin Avenue Suite 603
Washington, DC 20007
{wilson, mgp5}@isis.imac.georgetown.edu

Abstract

The MiTAP prototype for SARS detection uses human language technology for detecting, monitoring, and analyzing potential indicators of infectious disease outbreaks and reasoning for issuing warnings and alerts. MiTAP focuses on providing timely, multilingual information access to analysts, domain experts, and decision-makers worldwide. Data sources are captured, filtered, translated, summarized, and categorized by content. Critical information is automatically extracted and tagged to facilitate browsing, searching, and scanning, and to provide key terms at a glance. The processed articles are made available through an easy-to-use news server and cross-language information retrieval system for access and analysis anywhere, any time. Specialized newsgroups and customizable filters or searches on incoming stories allow users to create their own view into the data while a variety of tools summarize, indicate trends, and provide alerts to potentially relevant spikes of activity.

1 Background

Potentially catastrophic biological events that threaten US national security are steadily increasing in frequency. These events pose immediate danger to animals, plants, and humans. Current disease surveillance systems are inadequate for detecting indicators early enough to ensure the rapid response needed to combat these biological events and corresponding public reaction. Recent examples of outbreaks include both the HIV/AIDS and foot and mouth pandemics, the spread of West Nile virus to and across the US, the escape of Rift Valley Fever from Africa, SARS, and the translocation of both mad cow disease (BSE) and monkey pox to the United States.

Biological surveillance systems in the United States rely most heavily on human medical data for signs of epidemic activity. These systems span multiple organizations and agencies, are often not integrated, and have no alerting capability. As a result, responders have an insufficient amount of lead time to prepare for biological events or catastrophes.

Indications and Warnings (I&Ws) provide the potential for early alert of impending biological events, perhaps weeks to months in advance. Sources of I&Ws include transportation data, telecommunication traffic, economic indices, Internet news, RSS feeds (RSS) including weblogs, commerce, agricultural surveillance, weather, and other environmental data. Retrospective analyses of major infectious disease outbreaks (e.g., West Nile Virus and SARS) show that I&Ws were present weeks to months in advance, but these indicators were missed because data sources were difficult to obtain and hard to integrate. As a result, the available information was not utilized for appropriate national and international response. This illuminates a critical need in biodefense for an integrated system linking I&Ws for biological events from multiple and disparate sources with the response community.

2 Introduction

MiTAP (Damianos et al. 2002) was originally developed by the MITRE Corporation under the Defense Advanced Research Projects Agency (DARPA) Translingual Information Detection Extraction and

Summarization (TIDES) program. TIDES aims to revolutionize the way that information is obtained from human language by enabling people to find and interpret relevant information quickly and effectively, regardless of language or medium. MiTAP was initially created for tracking and monitoring infectious disease outbreaks and other biological threats as part of a DARPA Integrated Feasibility Experiment in biosecurity to explore the integration of synergistic TIDES language processing technologies applied to a real world domain. The system has since been expanded to other domains such as weapons of mass destruction, satellite monitoring, and suspect terrorist activity. In addition, researchers and analysts are examining hundreds of MiTAP data sources for differing perspectives on conflict and humanitarian relief efforts.

Our newest MiTAP prototype explores the integration of outputs from operational data mining (anomaly detection), human language technology (information extraction, temporal tagging, machine translation, cross-language information retrieval), and visualization tools to detect SARS-specific I&Ws in Asia, with relevance to pathogen translocation to the United States. Using feeds from English and Chinese language newswire, weblogs, and other Internet data, the system translates Chinese text data and tracks keyword combinations thought to represent I&Ws specific to SARS outbreaks in China. Analysts can use cross-language information retrieval for retrospective analysis and improving the I&W model, save searches to use as filters on incoming data, view trends, and visualize the data along a timeline. Figure 1 shows an overview of the prototype.

Warnings generated by this MiTAP prototype are intended to complement traditional biosurveillance and communications already in use by the international public health community. This system represents an expansion of current US surveillance capabilities to detect biological agents of catastrophic potential.

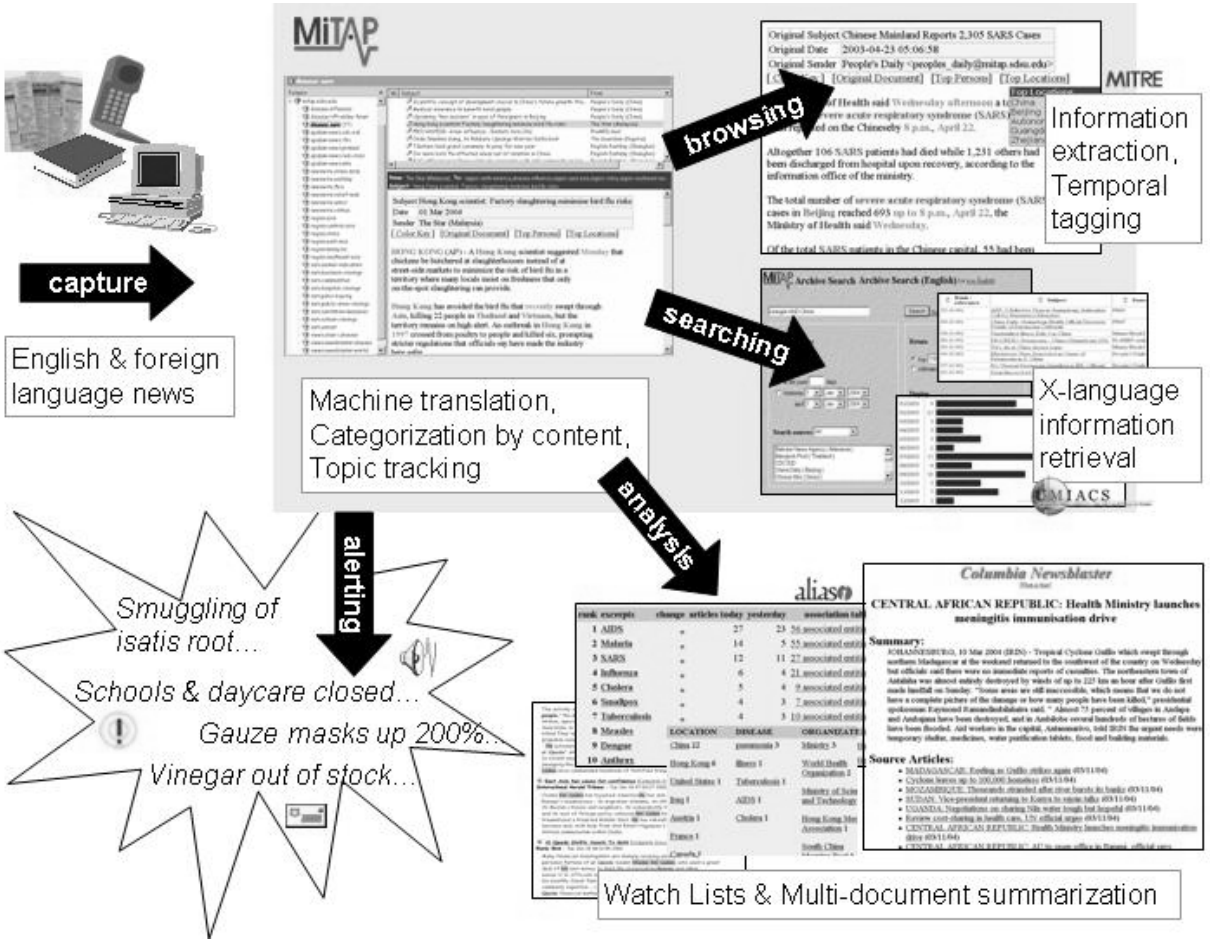

Figure 1 Overview of the MiTAP prototype for SARS detection.

3 Component Technologies

The MiTAP prototype relies extensively on human language technology and expert system reasoning. Below, MiTAP capabilities are described briefly along with their contributing component technologies.

3.1 Information Processing

After Internet news sources are captured and normalized, they are passed through a zoner using human-generated rules to identify source, date, and other information such as headline, or title, and content. The Alembic natural language analyzer (Aberdeen et al. 1995; Vilain and Day 1996) processes the zoned messages to identify paragraph, sentence, and word boundaries as well as part-of-speech tags. The messages then pass through the Alembic named entity recognizer for identification and tagging of person, organization, location, and disease names. Finally, the article is processed by the TempEx normalizing time expression tagger (Mani and Wilson 2000).

For Chinese and other non-English sources, the CyberTrans machine translation system (Miller et al. 2001) is used to translate articles automatically into English. CyberTrans wraps commercial and research translation engines to produce a common set of interfaces; the current prototype makes use of the SYSTRAN Chinese-English system.

RSS feeds can provide a high volume textual gestalt. Weblogs, in particular, are a good source of timely text, some of which is topical and all of which is based on personal observations and experiences. Aggregate measurements on these feeds can provide indications of public health-related phenomena. Consider the relative rates of words and phrases such as "stay home from" or "pneumonia." Geotemporal location of non-seasonal spikes in relative rank of these strings can establish suspicion for further investigation by I&W experts.

3.2 Browsing

English language data and pairs of foreign language documents and their translated versions are made available on a news server (INN 2001) for browsing. The system categorizes and bins articles into newsgroups based on their content. To do this, the system relies on a combination of the information extraction results as well as human-generated rules for pattern matching. Newsgroups are created to provide multiple perspectives on the data; analysts can subscribe to specific disease tracking newsgroups, regional newsgroups, specific data source newsgroups, or to customized topic tracking newsgroups that may be based on several related subjects.

Tagged entities in each article are color-coded to enable rapid scanning of information and easy identification of key names. The five most frequently mentioned locations in each article as well as the top five people are presented as a list for quick reference.

3.3 Information Retrieval

To supplement access to the articles on the news server and to allow for retrospective analysis, articles are indexed using the Lucene information retrieval system (The Jakarta Project 2001) for English language documents and using PSE (Darwish 2002) for foreign language documents. Web links are maintained between foreign language documents and their translated versions to allow for more accurate human translations of selected documents.

Analysts can perform full text, source-specific queries over the entire set of archived documents and view the retrieved results as a relevance-ranked list or as a plot across a timeline. A cross-language information retrieval interface allows users to search in English across the Chinese language sources.

Users can also save specific search constraints to be used as filters on incoming data. These saved searches provide a simple analytic capability as well as an alerting feature. (See below.)

3.4 Analysis

To assist analysts in identifying relevant and related articles, we have integrated multi-document summarization and watch lists. Columbia University's Newsblaster (McKeown et al. 2002) automatically detects daily topics, clusters MiTAP articles around those topics, and generates multi-document summarizations which are made available on the news server. Multiple technologies (e.g., coreference, information extraction) from Alias I, Inc. (Baldwin et al. 2002) produces comprehensive views on specific named entities (i.e., people or disease) across MiTAP documents. These views are summarized through ranked lists, highlighting important topics of the day and activities which might indicate disease outbreak.

Finely-tuned searches can be saved and applied as filters or topic tracking mechanisms. These saved searches are automatically updated at specific intervals and can be aggregated and displayed visually as bar graphs to reveal spikes of activity that otherwise might go undetected.

3.5 Alerting

The MiTAP prototype has two separate alerting capabilities: saved searches and an integrated expert

system. The saved search functionality allows analysts to set thresholds for alerting purposes. For example, MiTAP can send email when any new article arrives, when a specified maximum number of articles arrives, or when the daily number of new articles increases by some percentage of the total or moving average.

The Human Language Indication Detector (HLID) performs data fusion on a number of disparate sources, compressing a large volume of information into a smaller but more significant set of alerts. HLID monitors a variety of sources including MiTAP articles, information events in RSS feeds, and other dynamically updated information on the World Wide Web. HLID analyzes events from these sources in real time and generates an estimate of significance for each, complete with an audit trail of supporting and negating evidence. This allows an analyst to direct a search for indicators towards interesting data while reducing the time spent investigating false alarms and insignificant events.

HLID is composed of four major components. The first is an event collector, which monitors a data source and triggers action when an event is observed. These events are sent to the rule based reasoning engine, an expert system shell (JESS 2004) with hand authored rules. The engine performs vetting and initial investigation of each event by identifying correlated events, corroborating or invalidating evidence, and references to supporting information. The engine can also supplement its knowledge base by performing a directed search via the query management system, which allows retrieval of information from a wide variety of sources including databases and web pages. Lastly, the alerting mechanism disseminates the conclusions reached by the system and provides an interface that allows an analyst to launch a deeper search for indicators and warnings.

4 Acknowledgments

This work has been funded, in part, by the Defense Advanced Research Projects Agency Translingual Information Detection Extraction and Summarization program under contract numbers DAAB07-01-C-C201 and W15P7T-04-C-D001, the Office of the Secretary of Defense in support of the Coalition Provisional Authority in Baghdad, and a MITRE Special Initiative for Rapid Integration of Novel Indications and Warnings for SARS.

5 References

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)*.

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

Damianos, L., Ponte, J., Wohlever, S., Reeder, F., Day, D., Wilson, G., Hirschman, L. 2002. *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.

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

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 2004 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), 69-76.

McKeown, K., Barzilay, R., Evan, D., Hatzivassiloglou, V., Klavans, J., Sable, C., Schiffman, B., Sigelman, S. 2002. *Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster*. In Proceedings of HLT 2002: Human Language Technology Conference.

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

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

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.