MiTAP, Text and Audio Processing for Bio-Security: A Case Study

Laurie Damianos[†], Jay Ponte[†], Steve Wohlever[†],
Florence Reeder[‡], David Day[†], George Wilson[‡], Lynette Hirschman[†]

The MITRE Corporation
[†]202 Burlington Road; Bedford, MA 01730
[‡]7798 Old Springhouse Road; McLean, VA 22101
{laurie, ponte, wohlever, freeder, day, gwilson, lynette}@mitre.org

Abstract

MiTAP (MITRE Text and Audio Processing) is a prototype system available for monitoring infectious disease outbreaks and other global events. MiTAP focuses on providing timely, multi-lingual, global information access to medical experts and individuals involved in humanitarian assistance and relief work. Multiple information sources in multiple languages are automatically captured, filtered, translated, summarized, and categorized by disease, region, information source, person, and organization. Critical information is automatically extracted and tagged to facilitate browsing, searching, and sorting. The system supports shared situational awareness through collaboration, allowing users to submit other articles for processing, annotate existing documents, post directly to the system, and flag messages for others to see. MiTAP currently stores eight hundred thousand articles and processes an additional 2000 to 10,000 daily, delivering up-to-date information to dozens of regular users.

Global Tracking of Infectious Disease Outbreaks and Emerging Biological Threats

Over the years, greatly expanded trade and travel have increased the potential economic and political impacts of major disease outbreaks, given their ability to move rapidly across national borders. These diseases can affect people (West Nile virus, HIV, Ebola, Bovine Spongiform Encephalitis), animals (foot-and-mouth disease) and plants (citrus canker in Florida). More recently, the potential of biological terrorism has become a very real threat. On September 11th, 2001, the Center for Disease Control alerted states and local public health agencies to monitor for any unusual disease patterns, including chemical and biological agents. In addition to possible disruption and loss of life, bioterrorism could foment political instability, given the panic that fast-moving plagues have historically engendered.

Appropriate response to disease outbreaks and emerging threats depends on obtaining reliable and up-to-date information, which often means monitoring many news sources, particularly local news sources, in many languages worldwide. Analysts cannot feasibly acquire, manage, and digest the vast amount of information available 24 hours a day, seven days a week. In addition, access to foreign language documents and the local news of other countries is generally limited. Even when foreign language news is available, it is usually no longer current by the time it is translated and reaches the hands of an analyst. This is a very real problem that raises a very urgent need to develop automated support for global tracking of infectious disease outbreaks and emerging biological threats.

The MiTAP (MITRE Text and Audio Processing) system was created to explore the integration of synergistic TIDES language processing technologies: Translation, Information Detection, Extraction, and Summarization. TIDES aims to revolutionize the way that information is obtained from human language by enabling people to find and interpret needed information quickly and effectively, regardless of language or medium. MiTAP is designed to provide the end user with timely, accurate, novel information and present it in a way that allows the analyst to spend more time on analysis and less time on finding, translating, distilling and presenting information.

On September 11th, 2001, the research prototype system became available to real users for real problems.

Text and Audio Processing for Bio-Security

MiTAP focuses on providing timely, multi-lingual, global information access to analysts, medical experts and individuals involved in humanitarian assistance and relief work. Multiple information sources (epidemiological reports, newswire feeds, email, online news) in multiple languages (English, Chinese, French, German, Italian, Portuguese, Russian, and Spanish) are automatically captured, filtered, translated, summarized, and categorized into searchable newsgroups based on disease, region, information source, person, organization, and language. Critical information is automatically extracted and tagged to facilitate browsing, searching, and sorting. The system supports shared situational awareness through collaboration, allowing users to submit other articles for

processing, annotate existing documents, and post directly to the system. A web-based search engine supports source-specific, full-text information retrieval. Figure 1 represents a graphical overview of the services provided by the MiTAP system.

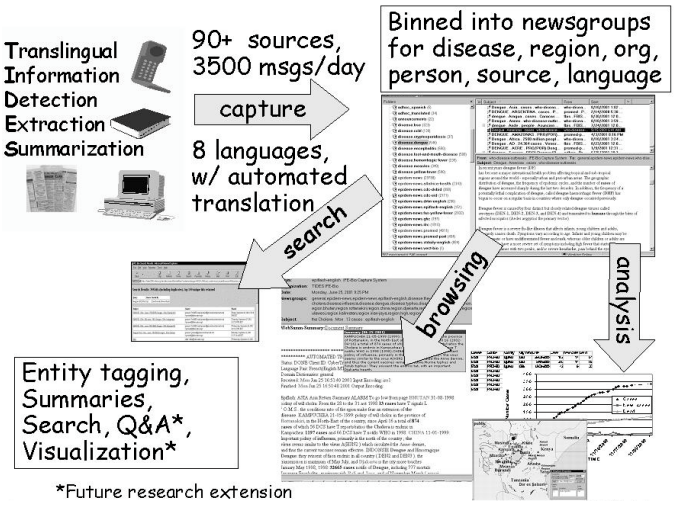

Figure 1 MiTAP Overview

Figure 2 illustrates the three phases of the underlying architecture: information capture, information processing, and user interface.

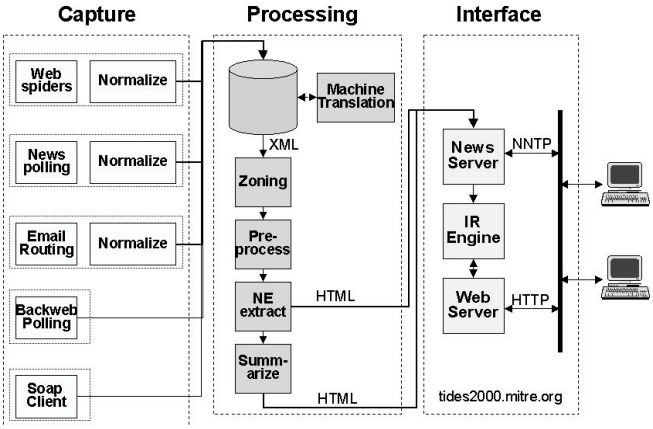

Figure 2 MiTAP Architecture

Information Capture

The capture process supports web sources, electronic mailing lists, newsgroups, news feeds, and audio/video data. The first four of these categories are automatically harvested and filtered, and the resulting information is normalized prior to processing. The ViTAP system (Merlino 2002) captures and transcribes TV news broadcasts, making the text transcriptions available to MiTAP via a SOAP-based interface (SOAP Version 1.2 Part 1 2001). The data from all of these sources are then sent on to the processing phase, where the individual TIDES component technologies are employed.

Information Processing

Each normalized message is passed through a zoner that uses human-generated rules to identify the source, date, and other fields such as headline or title, article body, etc. The zoned messages are preprocessed to identify paragraph, sentence, and word boundaries as well as part-of-speech tags. This preprocessing is carried out by the Alembic natural language analyzer (Aberdeen et al. 1995; Aberdeen et al. 1996; Vilain and Day 1996; Vilain 1999) which is based on the Brill (1995) tagger and uses machine-learned rules. The preprocessed messages are then fed into the Alembic named entity recognizer, which identifies person, organization and location names as well as dates, diseases, and victim descriptions using human-generated rules. This extended set of named entities is critical in routing the messages to the appropriate newsgroups and is also used to color-code the text so users can quickly scan the relevant information. Finally, the document is processed by WebSumm (Mani and Bloedorn 1999), which generates modified versions of extracted sentences as a summary. WebSumm depends on the TempEx normalizing time expression tagger (Mani and Wilson 2000) to identify the time expressions and normalize them according to the TIDES Temporal Annotation Guidelines, a standard for representing time expressions in normal form (Ferro 2001; Ferro et al. 2001). For non-English sources, the CyberTrans machine translation system (Miller et al. 2001), which "wraps" commercial and research translation engines and presents a common set of interfaces, is used to translate the messages automatically into English. The translated messages are then processed as the English sources are. Despite translation errors, the translated messages have been judged by users to be useful. There is generally enough information for users to determine the relevance of a given message, and the original, foreign language documents remain available for human translation, if desired. Without the machine translation, these articles would effectively be invisible to analysts and other users.

User Interface

The final phase consists of the user interface and related processing. The processed messages are converted to HTML, with color-coded named entities, and routed to newsgroups hosted by a Network News Transport Protocol (NNTP) server, InterNetNews (INN 2001). (See figure 3.) The newsgroups are organized by category (i.e., source, disease, region, language, person, and organization) to allow analysts, with specific information needs, to locate material quickly. The article summaries are included via a web link and JavaScript code embedded in the HTML that displays a pop-up summary when the mouse is dragged over the link. Another type of summary, pop-up tables, show lists of named entities found in the document.

Machine-translated documents contain a web link to the original foreign language article. Figure 4 shows a sample message with color-coded named entities and pop-up summary.

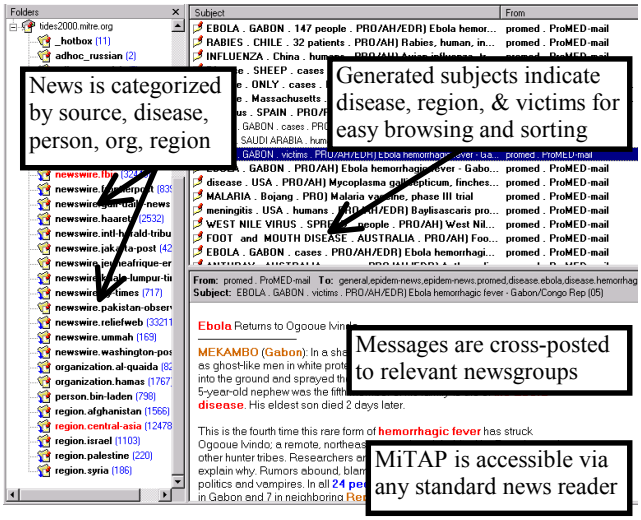

Figure 3 MiTAP viewed through standard newsreader

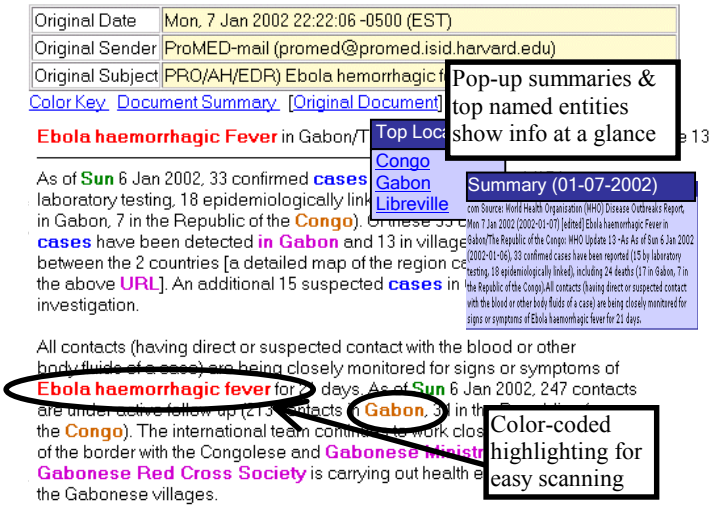

Figure 4 Sample MiTAP article

One major advantage to using the NNTP server is that users can access the information using a standard mail/news browser such as Netscape Messenger or Outlook Express. There is no need to install custom software, and the instant sense of familiarity with the interface is crucial in gaining user acceptance - little to no training is required. Mail readers also provide additional functionality such as alerting to new messages on specified topics, flagging messages of significance, and local directories that can be used as a private workspace. Other newsgroups can be created as collaborative repositories for users to share collective information.

To supplement access to the data, messages are indexed using the Lucene information retrieval system (The Jakarta Project 2001), allowing users to do full text, source-specific queries over the entire set of messages. As the relevance of messages tends to be time dependent, we have implemented an optimized query mechanism to do faster time-constrained searches.

MiTAP Development and Deployment

The initial MiTAP system was put together over a 9-month period. Our goal was to build a prototype quickly to demonstrate the results of integrating multiple natural language processing (NLP) technologies. The longer-term strategy is to upgrade the components progressively as better performing modules become available and to migrate towards our developing architecture. For the initial implementation, we chose components based on availability as well as ease of integration and modification. This meant that we used components developed at MITRE (extraction, summarization) or developed with MITRE involvement (translation support), or commercial off-the-shelf (COTS) components (translation engines, information retrieval, news server, news browser interface). In cases where no component was readily available, we developed a minimal capability for MiTAP, e.g., scripts for capture of news sources, or use of named entity extraction for headline generation and binning of messages into appropriate newsgroups.

Since July 2000, we have been working to incorporate modules from other groups (e.g., Columbia's NewsBlaster, McKeown et al. 2002), to redesign the architecture, and to specify a protocol to support service-based access to other modules, such as information extraction, summarization, or topic clustering.

As part of the long-term efforts, we have been concurrently developing a framework known as Catalyst (Mardis et al. 2001). Catalyst provides a common data model based on standoff annotation, efficient compressed data formats, distributed processing, and annotation indexing. Standoff annotation (see, for example, Bird et al. 2000) means that the linguistic annotations are kept separate from the original text or audio as opposed to e.g., inline XML markup, where the annotations are added to the underlying signal. The advantages of standoff annotation are threefold. First, the limitations of the markup language do not limit the allowable annotations. For example, with inline XML, the tags must be strictly nested. If two language processing modules do not agree on, say sentence boundary detection, there is the potential for 'crossing brackets' in the markup. This is a problem for inline XML markup but not for standoff annotation. Second, when annotations are kept separate from signal, system components receive customized views of the signal. That means that a component need not ever receive annotations that it does not explicitly require. This makes systems both more

efficient and easier to test and debug. Efficiency is greater, sometimes considerably so, since annotations can account for a large proportion of the data in a complex language processing system. New modules added to the system do not affect existing modules unless explicit dependencies are also added, simplifying the testing process. Finally, standoff annotations are easy to compress and index, making further optimizations possible.

Uses of AI Technology

Artificial Intelligence (AI) technology and techniques pervade MiTAP to support its multi-faceted, multi-lingual and multi-functional requirements. From automated natural language processing techniques to information retrieval, the NLP modules utilize AI extensively. The techniques utilized fall predominantly into the data-driven camp of methods. Below we describe the components, roughly in their order of processing flow.

The CyberTrans machine translation server employs a combination of AI techniques to optimize the performance of COTS machine translation (MT) systems. Since system developers have only the most basic insight into the MT systems, we will not describe related AI techniques in depth here, and interested readers are referred to (Arnold et al. 1994; Hutchins & Somers 1992). MT systems in the last 30 or so years have been marvels of knowledge engineering, from the encoding of the lexical entries to the writing of grammatical rules. The simplest form of MT is word-for-word substitution, and all knowledge is encoded in the lexicon itself. While this type of system is easy and quick to construct given a translation dictionary, it also provides a hard-to-read translation, imposing a greater burden on the users of the system. To provide more well-formed output, systems perform increasingly sophisticated levels of analysis of the source language text using grammatical rules and lexicons. This analysis produces an intermediate structure which is then transformed by another set of rules to a format sufficient for generating the target language. The level of analysis increases in sophistication – from words to syntax to semantics to pragmatics with the "holy grail" of MT being a language-independent representation or interlingua. At this level, there is increasing overlap with traditional knowledge-base and ontology engineering, hence the increased reliance in computational linguistics on AI techniques (see Yakama and Knight 2001 for an example).

COTS MT systems are designed primarily for interactive use in situations where users have control over the language, formatting and well-formedness of the input text. In adapting CyberTrans for real users and real-world data, the necessity for supporting technologies was quickly apparent. Three of these are of particular interest: automated language identification, automated code set conversion, and automated spelling correction, particularly for the incorporation of diacritics. The resulting tools can be used individually and eventually as standalone modules, but are currently integrated into the CyberTrans processing flow.

The first, most essential, part of automated processing of language data is to determine both the language and code set representation of the input text. While it would seem obvious that users know at least what the language of a given document is, this has proven not to be the case, particularly in non-Romanized languages such as Arabic or Chinese. In these situations, documents appear as unintelligible byte streams. In addition, some of the data sources contain documents in a mix of languages, so knowledge of the source does not necessarily determine the language. This is a classical categorization problem with a search a space of N*M where N is the number of languages to be recognized and M the number of code set representations. The categories are determined by a combination of n-graph measurements using the Acquaintance algorithm (Huffman 1996) with simple heuristics whittling down the search space.

Once the code set has been determined, it is converted into a standard representation. This process is not without information loss, so spelling corrections are applied. The most straight-forward spelling correction involves the reinsertion of diacritical markers where they are missing. This is treated as a word-sense disambiguation problem (Yarowsky 1994) and relies on both language spelling rules and trained probabilities of word occurrences. Here, the solution is a hybrid system where hand-coded rules are enforced using statistical measures of likely word occurrences.

"Tagging" refers to a range of natural language processing stages that associate information with a word or multi-word phrases. The tagging used in MiTAP relies on a combination of hand-crafted and machine discovered rules. Tagging operations begin with sentence and word boundary identification ("word segmentation"), most of which is manually created and relies on narrowly defined regular expression heuristics implemented as regular expression pattern transformations. This stage is followed by part-of-speech tagging, implemented as a "transformational rule sequence" (Brill 1995). A transformational rule sequence can be viewed as set of cascaded finite state transformations. This restrictive computational model allows a range of machine learning techniques to be applied iteratively to derive the rules during training. The rules for part-of-speech tagging are heavily influenced by pre-computed word lists (lexicons), in which words are associated with parts-of-speech derived from a large corpus of annotated textual data. In Alembic, part-of-speech tagging is followed by a separate set of rule sequences, developed through a mixture of manual and machine learning methods. These rule sequences perform "named entity tagging" that identifies such things as personal names, place names and times. These have been manually

extended to capture nominal expressions that refer to diseases and victims.

In addition, a specialized tagging operation occurs, that of temporal resolution. While dates such as *09 September 2000* are relatively unambiguous, many time references found in natural language are not, for instance *last Tuesday*. To get the time sequencing of events of multiple stories correct, it is necessary to resolve the possible wide range of time references accurately. In this case, the resolution algorithm also combines basic linguistic knowledge with rules learned from corpora (Mani and Wilson 2000).

Similarly, place names are often only partially specified. For example, there are a great many places in South America named La Esperanza. We are currently developing a module to apply a mix of hand-written rules and machine learning to metadata and contextual clues drawn from a large corpus to disambiguate place names.

This range of tagging procedures represents a strong shift in natural language processing research over the past fifteen years towards "corpus-based" methods. This work begins with the manual annotation of a corpus, a set of naturally occurring linguistic artifacts, by which some level of linguistic analysis (word segmentation, part-of-speech, semantic referent, syntactic phrase, etc.) is associated with the relevant portion of text. The resulting data provides a rich basis for empirically-based research and development, as well as formal evaluations of systems attempting to re-create this analysis automatically. The availability of such corpora have spurred a significant interest in machine learning and statistical methods in natural language processing research, of which those mentioned above are just a few. One of the benefits of the rule-sequence model adopted in MiTAP's Alembic component is its support for easily and effectively combining automatically derived heuristics with those developed manually. This was a key element in successfully modifying the Alembic NLP system for MiTAP in the absence of any significant annotated corpus.

Summarization is achieved through several machine learning techniques including Standard Canonical Discriminant Function (SCDF) analysis (SPSS 1997), C4.5 rules (Quinlan 1992) and AQ15c (Wnek et al. 1995). The feature set is an interesting twist on the summarization problem where the abstracts of documents are treated as queries that represent the user's information needs. In essence, the features being trained on are constructed from the criteria for successful summarization (Mani and Bloedorn 1999). Summarization features then use information retrieval metrics such as tf.idf, which measures the likelihood that a given phrase or word is relevant to the topic at hand, in combination with other more fine-grained metrics such as number of unique sentences with a synonym link to the given sentence.

Information retrieval services are provided by the Lucene information retrieval engine. Our search interface provides Boolean queries and relevance based queries. Since our users require timely access to information, we have developed an optimized search algorithm for relevance ranked searches within date ranges. The default behavior of Lucene was to produce the entire ranked list and then re-sort by date. An entire relevance ranked list can be quite large, and so the optimized algorithm for small date ranges does repeated searches by date for each date in the range and presents the results in relevance ranked order. For the small ranges of dates that our users prefer, we realize a significant savings in query latency through the optimized algorithm.

The utilization of classical AI techniques is a surface just being scratched in the computational linguistics community. Like many domains, the field has hit the wall of knowledge engineering familiar to most AI practitioners. We are therefore looking for corpus-based learning techniques akin to data mining and data modeling for gaining language knowledge quickly without pools of experts. It then follows that we are also learning some of the hard lessons from AI – that no one technique is a silver bullet for complex problems like translation or summarization. In addition, we eventually find ourselves up against the knowledge-engineering bottleneck as well as the fact that eventually all knowledge is encoded in a "language" and must be read and understood.

MiTAP Maintenance

One or two individuals are typically responsible for the daily maintenance of the MiTAP system. This includes a number of administrative tasks, such as adding new user accounts as they are requested, informing users (via an e-mail distribution list) of changes to the system (e.g., new data sources, outages for planned maintenance, etc.), and obtaining user feedback via online surveys. The other major tasks deal with adding new data sources to MiTAP and maintaining the processing elements that make up the system.

When a useful online news source (i.e., a web site) is identified, assuming there are no copyright issues, it can take as little as a half hour to build a custom capture script to start capturing data from the source. Feeding a new e-mail list into the system is even faster. Data sources that deliver content via a mechanism other than the web or e-mail may require more time to integrate (e.g., a subscription-based data feed). There is a wide range of methods by which such data may be delivered, and a general solution for feeding such data into the system is not always available. However, these types of sources are rare. Most of the sources that are currently connected to MiTAP are either web sites or e-mail lists. Of the various types of data sources, web-based sources require the most

maintenance. Each web capture script is designed for a specific web site. If the format of that site changes, the web capture may not perform as expected, and the capture script has to be updated.

Perl and Unix shell scripts make up most of the "glue" that connects the various NLP components into a processing pipeline. These scripts require little maintenance although we occasionally modify them to improve the formatting of the posted messages or to fix a minor bug when a new source is added. Only general knowledge of the underlying scripting languages is needed to maintain the non-NLP portions of the system.

Infrequent updates to the various NLP components (e.g., Alembic, CyberTrans, or WebSumm) usually require the assistance of an individual with more specialized knowledge of the relevant component. For example, in order to improve our named entity tagging (e.g., to better handle Arabic names), a programmer or linguist familiar with Alembic needs to develop new tagging rules and, working with one of the general MiTAP administrators, upgrade the running system.

MiTAP Usage, Evaluation and Utility

Usage

MiTAP has been accessible to users since June 2001. Data can be accessed in two ways: via newsgroups or through a web-based search engine. No special software is needed - just a standard news reader or web browser and an account on the system. The number of users that the system can support at one time is limited only by the loads that can be handled by the web and news servers. At the time of this writing, we have close to 100 user accounts. Up to 18 people have used the system on any particular day, with a daily average of seven regular users, including weekends. Our user base includes medical analysts, doctors, government and military officials, members of non-governmental organizations, and members of humanitarian assistance/disaster relief organizations. They access the system for updates on disease outbreaks as well as to read current news from around the world.

Figure 5 illustrates averaged daily MiTAP activity from July 2001 through February 2002. The bold line, on the left axis, shows messages processed and posted to the system while the broken line, on the right axis, shows user access via newsgroups or search engine.

To support collaborative work, there is a newsgroup, called *hotbox*, to which users can submit news, messages of current interest, personal opinions, and annotations. Almost all of our subscribers read the contents of *hotbox* every time they log on to the system.

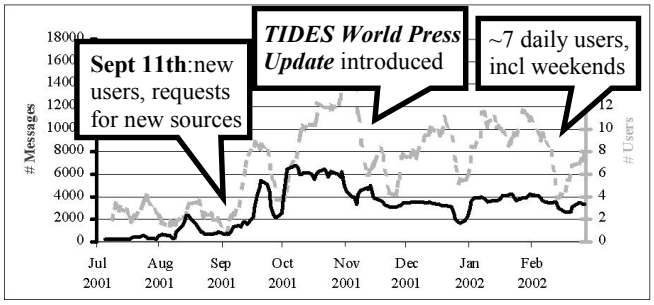

Figure 5 Daily MiTAP activity

One regular user, a consultant to an organization of medical professionals, spends one to two hours a day reading 800 to 1000 MiTAP articles from over 50 mostly foreign sources. He is able to isolate between 20 and 50 articles of significance and five to 15 articles of high importance. These selected articles are used to create the *TIDES World Press Update* (TIDES World Press Update 2001), a daily newsletter available to users of MiTAP (through *hotbox*) and distributed to an internationally-wide list of readers. The consultant considers MiTAP a "*labor-saving and intelligence gathering tool*" and credits the accurate headline extraction and color-coded highlighting of named entities for his ability to extract critical information quickly.

Evaluation

The *Disease of the Month Experiment*, a series of user-centered, task-based mini-evaluations was designed to assess utility, evaluate usability, measure progress, and provide iterative feedback to MiTAP developers. We chose a scenario familiar to analysts (i.e., research a current disease outbreak and prepare a report) to help minimize dependent variables and reduce training. Test groups were compared monthly to control groups in order to measure the utility of the system. Comparing MiTAP to the web and its vast amount of information, we hypothesized that 1) MiTAP users can produce better analytic reports in a shorter amount of time, where "better" means more up-to-date and more complete, and 2) MiTAP users spend less time reading documents and can read more in a given period of time. Test groups were also compared across iterations to measure the progress of development. Simultaneously, we performed independent usability studies.

For purposes of contrasting and comparing test versus control and test versus test across months, we defined five categories of metrics: efficiency, task success, data quality, user satisfaction, and usability. These categories were adopted and modified from those established by Walker et al. (2001) for the DARPA Communicator project.

In our experiments, MiTAP users provided more detail and more up-to-date information on disease outbreaks than just

the web alone; however, they did not necessarily spend less time doing so. Our results also show that the test groups were able to find a larger number of relevant articles in fewer searches. In fact, the test groups, who were also permitted to use the web to find information, cited MiTAP articles in their reports an average of three times more than articles found on the web, and often the links to the relevant web information were found via MiTAP articles. Over the course of this experiment series, the feedback has enabled the MiTAP development team to improve the overall system performance (e.g., throughput increased by a factor of 2.5 while source integration time decreased by a factor of 4). As a result, we have been able to add a multitude of new sources, producing a significantly richer, broader, and larger data collection.

This ongoing evaluation series has proven to be an invaluable method of measuring utility, usability, and progress of MiTAP. The results of the experiments have guided development, improved the system on many levels, inspired creative thinking, and given us a more comprehensive understanding of what our real users do and how we can better help them. User surveys as well as unprovoked feedback from our users have supplemented our evaluation efforts.

Utility

The popularity of MiTAP and the *TIDES World Press Update* is growing monthly by word of mouth. Users request additional news sources, coverage of other areas, and more languages. The dynamic nature of the system has allowed it to become broader in scope and richer in detail over time. Most of our users (89%) are repeat users, with 63% logging in to the system at least once a week. We measure the success of the MiTAP system by the ever-increasing number of accounts requested, the high repeat user rate, the popularity of the *TIDES World Press Update* (read by MiTAP account-holders as well as 120+ subscribers, many of whom re-distribute the newsletter), and the overwhelmingly positive user feedback. An additional measure of success is the number of immediate complaints we receive the few times we have had temporary access or network problems.

Although MiTAP contains only open-source information, much of it is sensitive, as are its usage and the identity of our users. Below are several quotes from government users, decision makers on whom MiTAP and the *TIDES World Press Update*, a product of MiTAP, have made a major impact.

> I look it over each day and consider it a tool in the war on terrorism. - *Advisor on bioterrorism, in a White House office since September 13th, 2001*

> I use it to improve my understanding of opinions that affect our Nation's safety. - *Director of Homeland Security for the National Guard*

> Superb work and an asset to the new war in which information sometimes trumps fire and steel. My staff uses it daily. - *Vice-Admiral, Director of Warfighting Requirements (N7), the Pentagon*

> MiTAP has clearly become a national asset. It is enhancing our national security and, perhaps, altering the course of the war by informing those making decisions on topics critical to our understanding of the forces arrayed against us. This is a war, like all others, decided by information, but the opposition is obscure, affecting, and affected by, voices we cannot hear unless something like MiTAP amplifies and illuminates them. - *Former Fleet Surgeon, Third Fleet, US Navy*

For more information or to apply for an account on the system, go to http://tides2000.mitre.org.

Acknowledgments

This work is supported, in part, under DARPA contract number DAAB07-01-C-C201.

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*.

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)*.

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.

Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., and Liberman, M. 2000. ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation. In *Proceedings of the Second International Language Resources and Evaluation Conference*. Paris: European Language Resources Association.

Brill, E. 1995. Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging, Computational Linguistics.

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.

Huffman, S. 1996. Acquaintance: Language-Independent Document Categorization by N-Grams. In The Fourth Text Retrieval Conference (TREC-4), 359 – 371. Gaithersburg, MD: National Institute of Standards and Technology.

Hutchins, H., & Somers, H. 1992. *An Introduction to Machine Translation.* Academic Press.

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

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

Mani, I. and Bloedorn, E. 1999. Summarizing Similarities and Differences Among Related Documents. Information Retrieval 1(1), 35-67.

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.

Mardis, S., Burger, J., Anand, P., Anderson, D., Griffith, J., Light, M., McHenry, C., Morgan, A., and Ponte, J. 2001. Qanda and the Catalyst Architecture. In *Proceedings of the Tenth Text REtrieval Conference (TREC-10)*, Gaithersburg, MD.

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.*

Merlino, A. 2002. ViTAP Demonstration. In *Proceedings of HLT2002: 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.*

MiTAP (MITRE Text and Audio Processing) System 2001, http://tides2000.mitre.org/.

Quinlan, J. 1992. C4.5: Programs for Machine Learning. Morgan-Kaufmann. San Mateo, CA.

SOAP Version 1.2 Part 1: Messaging Framework. Eds. Gudgin, M., Hadley, M., Moreau, J., Nielsen, H. 2001. http://www.w3.org/TR/soap12-part1/ (work in progress).

SPSS Base 7.5 Applications Guide 1997. SPSS Inc. Chicago.

TIDES World Press Update 2001. http://www.carebridge.org/~tides/.

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.

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.

Walker, M., Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., Whittaker, S. 2001. DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection. In *EUROSPEECH 2001*.

Wnek, K., Bloedorn, E., and Michalski, R. 1995. Selective Inductive Learning Method AQ15C: The Method and User's Guide. Machine Learning and Inference Laboratory Report ML95-4. George Mason University, Fairfax, VA.

Yakama, K and Knight, K. 2001. A Syntax-Based Statistical Translation Model. In *Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics*.

Yarowsky, D. 1994. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics*.