AI Magazine Volume 23 Number 4 (2002) (© AAAI)

MITAP for Biosecurity
A Case Study

Laurie Damianos, Jay Ponte, Steve Wohlever, Florence Reeder, David Day, George Wilson, and Lynette Hirschman

■ 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, multilingual, 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 over 1 million articles and processes an additional 2,000 to 10,000 daily, delivering up-to-date information to dozens of regular users.

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 11 September 2001, the Center for Disease Control alerted states and local public health agencies to monitor for any unusual disease patterns, including the effects of chemical and biological agents (figure 1). 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, 7 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 problem is very real and raises an 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:[1] 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 11 September 2001, the research prototype system became available to real users for real problems.

Figure 1. The Potential of Biological Terrorism Has Become a Very Real Threat.

Health officials need tools to monitor and track biological and chemical events.

Figure 2. MITAP Provides Timely, Multilingual, Global Information Access to Analysts, Medical Experts, and Individuals Involved in Humanitarian Assistance and Disaster Relief Work.

Text and Audio Processing for Biosecurity

MITAP focuses on providing timely, multilingual, global information access to analysts, medical experts, and individuals involved in humanitarian assistance and relief work (figure 2). Multiple information sources (epidemiological reports, newswire feeds, e-mail, 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. Additional "views" into the data facilitate analysis and can serve as alerts to events, such as disease outbreaks. Figure 3 represents a graphic overview of the services provided by the MITAP system, and figure 4 illustrates the three phases of the underlying architecture: (1) information capture, (2) information processing, and (3) user interface.

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 television news broadcasts, making the text transcriptions available to MITAP by a SOAP-based interface.[2] The data from all these sources are then sent to the processing phase, where the individual TIDES component technologies are used.

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 and article body. 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. 1996, 1995; Vilain 1999; Vilain and Day 1996), which is based on the

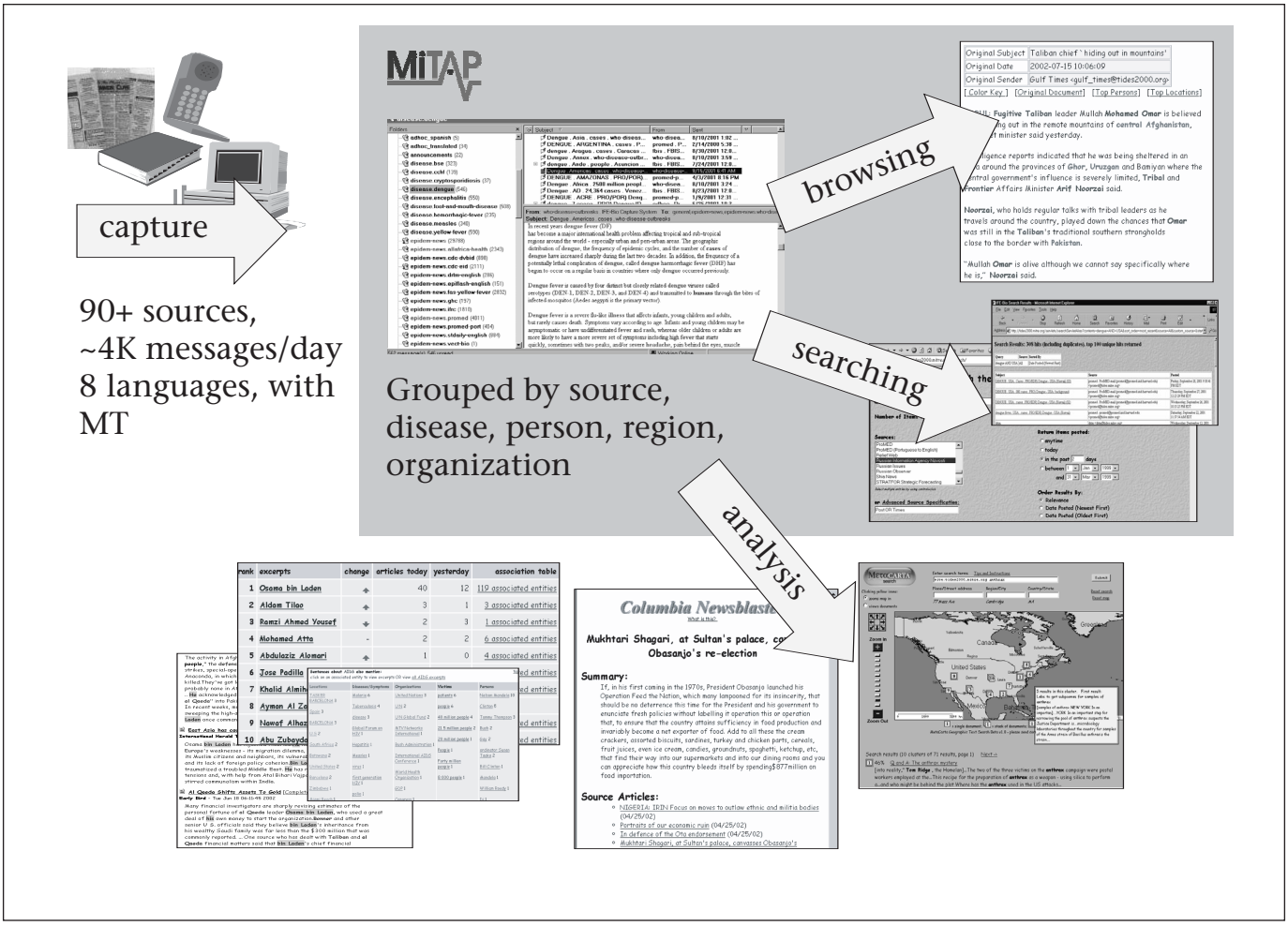

Figure 3. About 4000 Stories in Multiple Languages Are Captured Daily.

Articles are then made available for browsing with a news server, searching with a web-based search engine, and analyzing through various tools.

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,[3] 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 in the same way 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,

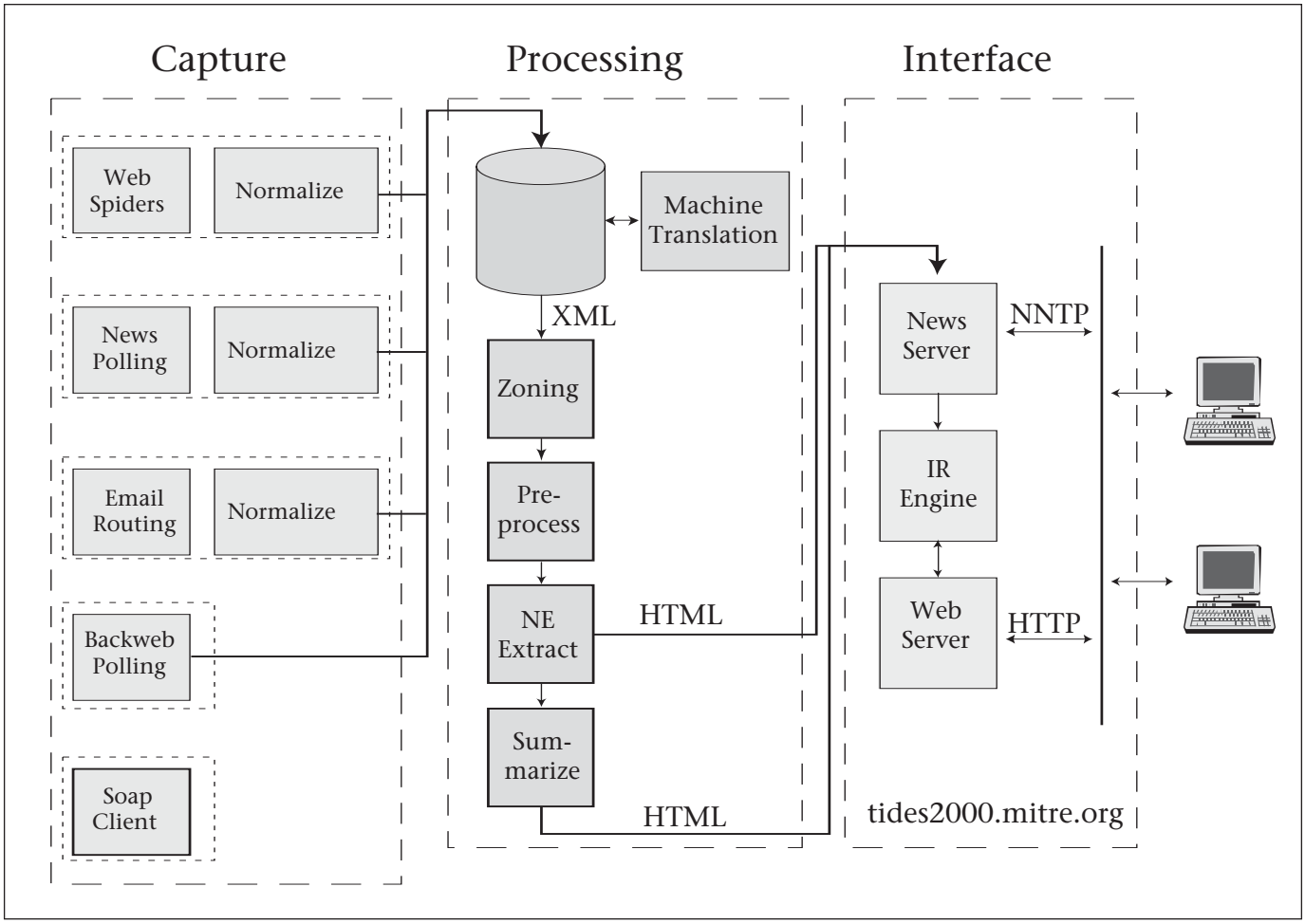

Figure 4. MITAP Architecture: Information Capture, Information Processing, and User Interface.

named entities, and routed to newsgroups hosted by a network news transport protocol (NNTP) server, INTERNETNEWS (figure 5).[4] The newsgroups are organized by category (that is, source, disease, region, language, person, and organization) to allow analysts, with specific information needs, to locate material quickly. The article summaries are included with 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, shows lists of named entities found in the document. Machine-translated documents contain a web link to the original foreign-language article. Figure 6 shows a sample message with color-coded, named entities and a pop-up summary of top locations mentioned in the 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 functions such as alerting to new messages on specified topics, flagging messages of significance, and accessing local directories that can be used as a private work space. 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,[5] allowing users to do full-text, source-specific Boolean queries over the entire set of messages. Because the relevance of messages tends to be time dependent, we have implemented an optimized query mechanism to do faster searches over time intervals.

MITAP Development and Deployment

The initial MITAP system was put together over a nine-month period. Our goal was to build a

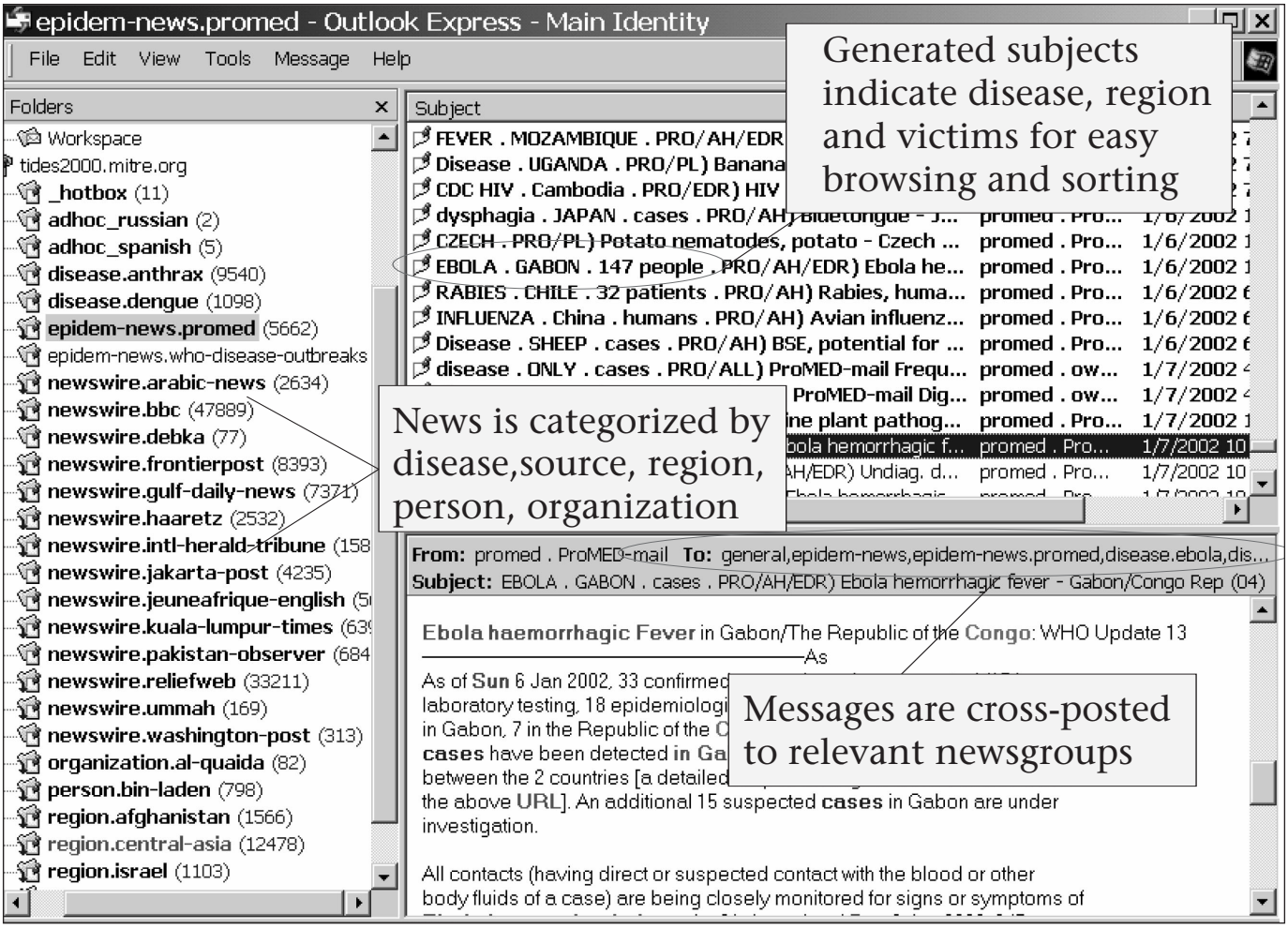

Figure 5. Users Can Access MiTAP through Any Standard News Reader and Customize Their View by Subscribing to Specific Newsgroup Categories Based on Their Needs.

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 toward our developing architecture. For the initial implementation, we chose components based on availability as well as ease of integration and modification. Thus, 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, for example, 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 (for example, Columbia's NEWSBLASTER [McKeown et al. 2002] and ALIAS I'S TOP 10 DISEASES[6]) to redesign the architecture and 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 and Burger 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, for example, 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.

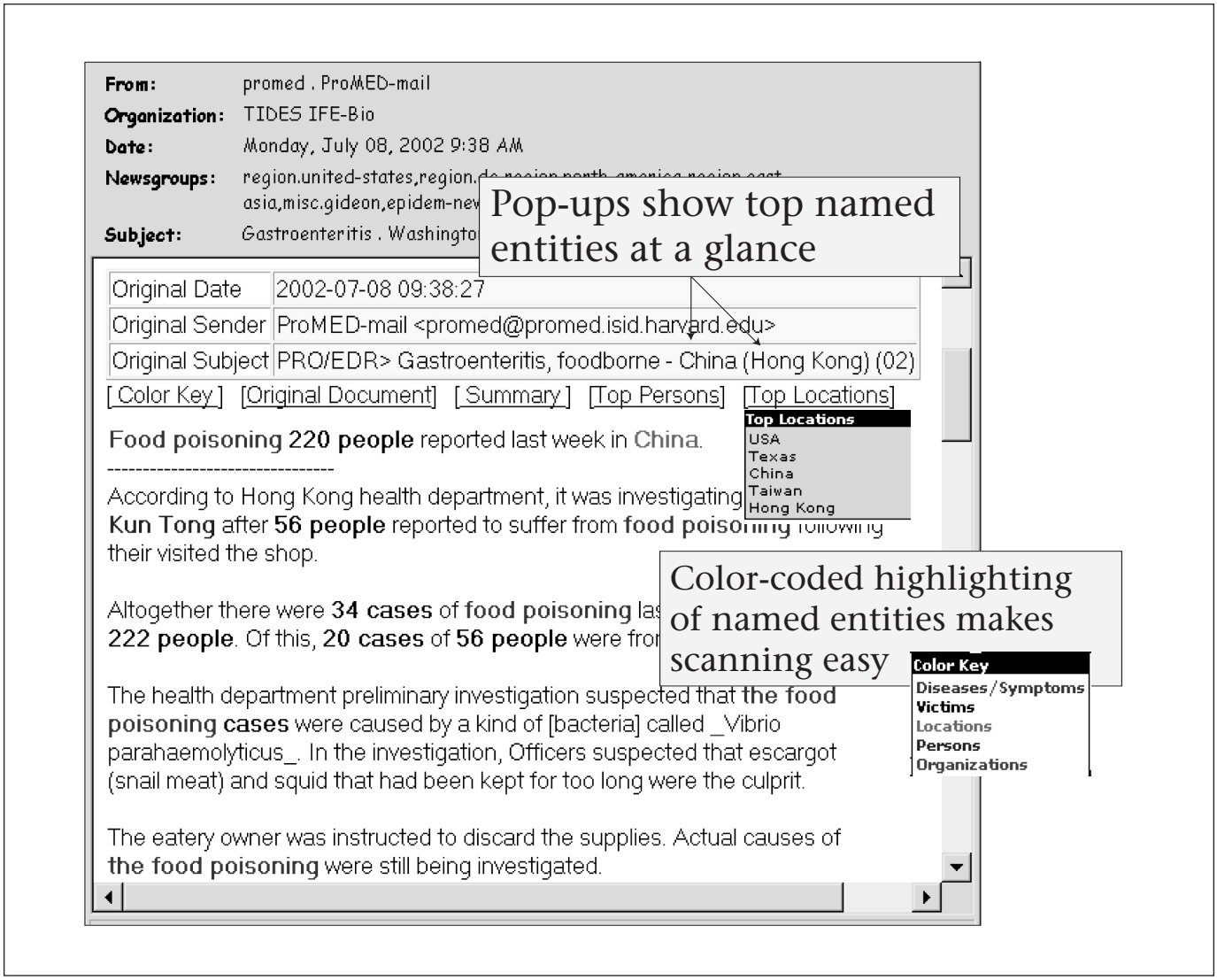

Figure 6. Sample MITAP Article.

Color coding of named entities makes it easy to scan long documents for specific types of information. Various summaries provide high-level views of individual articles.

For example, with inline XML, the tags must strictly be nested. If two language-processing modules do not agree on, say, sentence boundary detection, there is the potential for "crossing brackets" in the markup, which is a problem for inline XML markup but not for standoff annotation. Second, when annotations are kept separate from the signal, system components receive customized views of the signal. Thus, a component need not ever receive annotations that it does not explicitly require, making systems both more efficient and easier to test and debug. Efficiency is greater, sometimes considerably so, because 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

AI technology and techniques pervade MITAP to support its multifaceted, multilingual, and multifunctional requirements. From automated NLP techniques to information retrieval, the NLP modules use AI extensively. The techniques used fall predominantly into the data-driven camp of methods. Later, we describe the components, roughly in their order of processing flow.

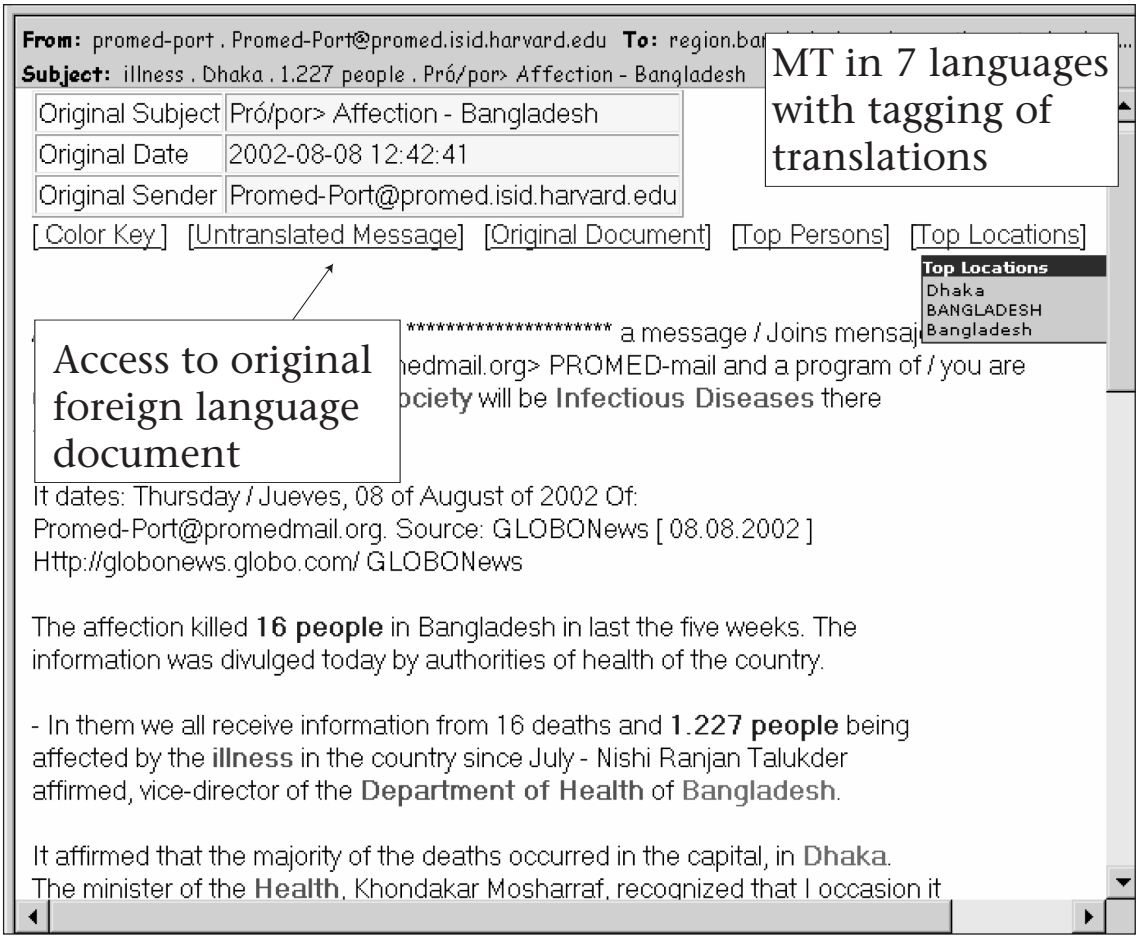

Figure 7. Although State-of-the-Art Machine Translation Is Not Perfect, It Provides Users with Access to Foreign-Language Data That Might Otherwise Be Unavailable to Them.

Users can identify critical information in documents and have them translated by a human, if desired.

The CyberTrans machine-translation server utilizes a combination of AI techniques to optimize the performance of COTS machine-translation systems. Because system developers have only the most basic insight into the machine-translation systems, we do not describe related AI techniques in depth here, and interested readers are referred to Hutchins and Somers (1992).[7] Machine-translation 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 machine translation is word-for-word substitution, and all knowledge is encoded in the lexicon itself. Although 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 that 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 machine translation 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 Yamada and Knight [2001] for an example).

COTS machine-translation 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 technologies are of particular interest: (1) automated language identification; (2) automated code set conversion; and (3) au-

tomated spelling correction, particularly for the incorporation of diacritics. The resulting tools can be used individually and eventually as stand-alone 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. Although 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 problem is one of classical categorization with a search space of *N***M*, where *N* is the number of languages to be recognized and 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 straightforward 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. Figure 7 shows an article translated from Portuguese. Once the documents are available in decoded form, the next stage is tagging.

Tagging refers to a range of natural language-processing stages that associate information with a word or multiword 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*), the rules of which are mostly manually created and rely 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 precomputed 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*, which identifies such things as personal names, place names, and times. These rules 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*. Although dates such as 09 September 2000 are relatively unambiguous, many time references found in natural language are not, for example, 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 NLP research over the past 15 years toward *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, and so on) is associated with the relevant portion of text. The resulting data provide a rich basis for empirically driven research and development as well as formal evaluations of systems attempting to recreate this analysis automatically. The availability of such corpora have spurred a significant interest in machine learning and statistical methods in NLP research, of which those methods mentioned earlier 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 element was key in successfully modifying the ALEMBIC NLP system for MITAP in the absence of any significant annotated corpus.

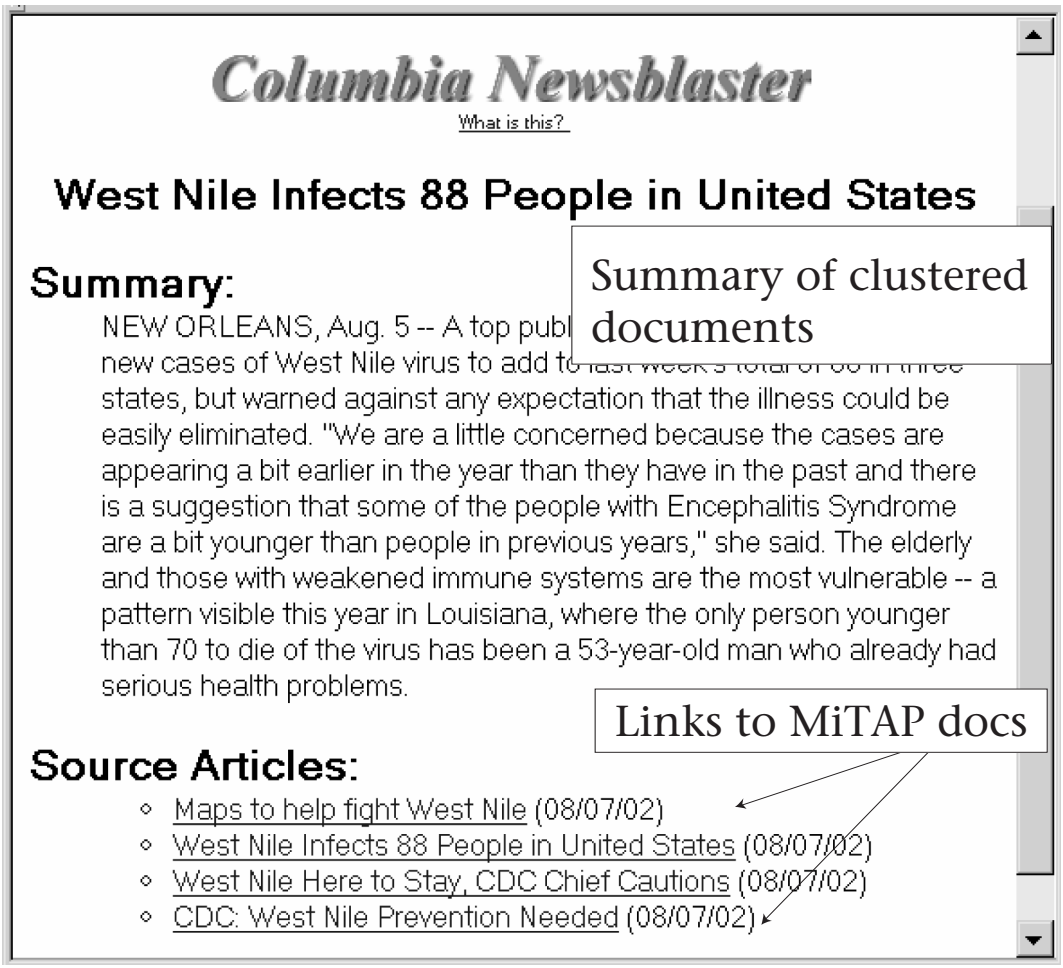

Figure 8. NEWSBLASTER *Automatically Summarizes Clusters of Documents.*
Users can access the complete MITAP articles for context or further reading.

Figure 6 depicts a message with the tagged entities color coded for easy scanning. Extracted entities are also used to create pop-up lists of people and locations mentioned in the article.

Summarization is achieved through several machine learning techniques, including standard canonical discriminant function (SCDF) analysis,[8] C4.5 rules (Quinlan 1992), and AQ15c (Wnek, Bloedorn, and Michalski 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 the number of unique sentences with a synonym link to the given sentence.

In addition to single-document summarization, we have incorporated two types of multi-document summarization into the MITAP system. NEWSBLASTER (McKeown et al. 2002) automatically clusters articles and generates summaries based on each cluster (figure 8).[9] These summaries are posted to the MITAP news

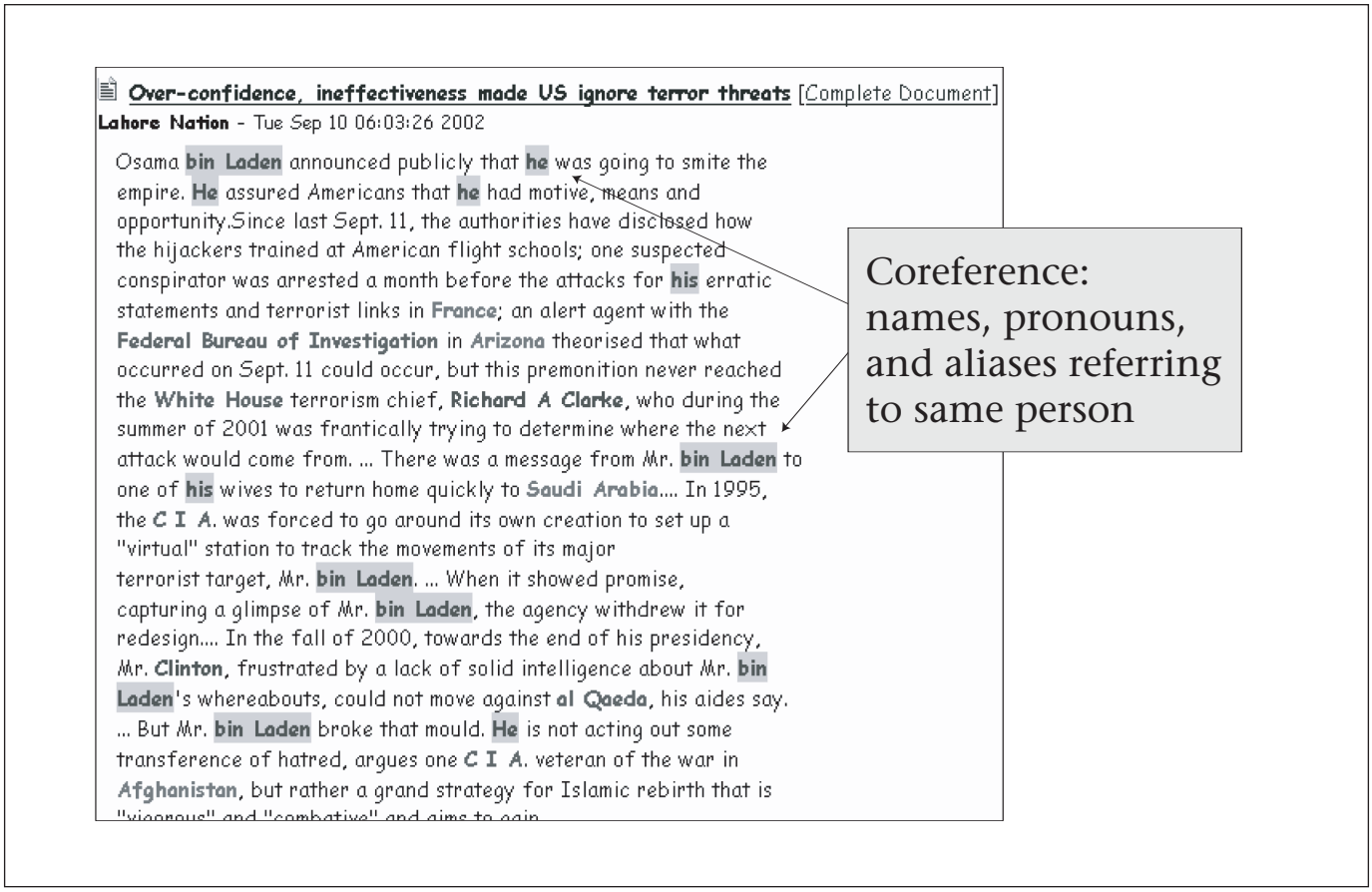

Figure 9. Daily Summaries of Named Entities, such as Osama bin Laden, Are Posted to MITAP.

The summaries consist of all sentences in the news that mention a specific name as well as any aliases and pronouns that refer to the same person.

server on a daily basis and act as both high-level views of the day's news and as points of entry into the system.

ALIAS I produces summaries on particular entities (for example, a specific person in the news) and also generates daily top 10 lists of diseases in the news and watch lists of people (for example, terrorist suspects).[10] These summaries are in the form of extracted sentences that reference (by name, alias, or pronoun) the particular entity. See the examples in figures 9 and 10. Association tables also provide relevant information in the form of lists of related entities.

Information-retrieval services are provided by the LUCENE information-retrieval engine. Our search interface provides Boolean queries and relevance-based queries (figures 11 and 12). Because 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 resort by date. An entire relevance-ranked list can be quite large, 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 use 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; for example, that no one technique is a silver bullet for complex problems such as 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.

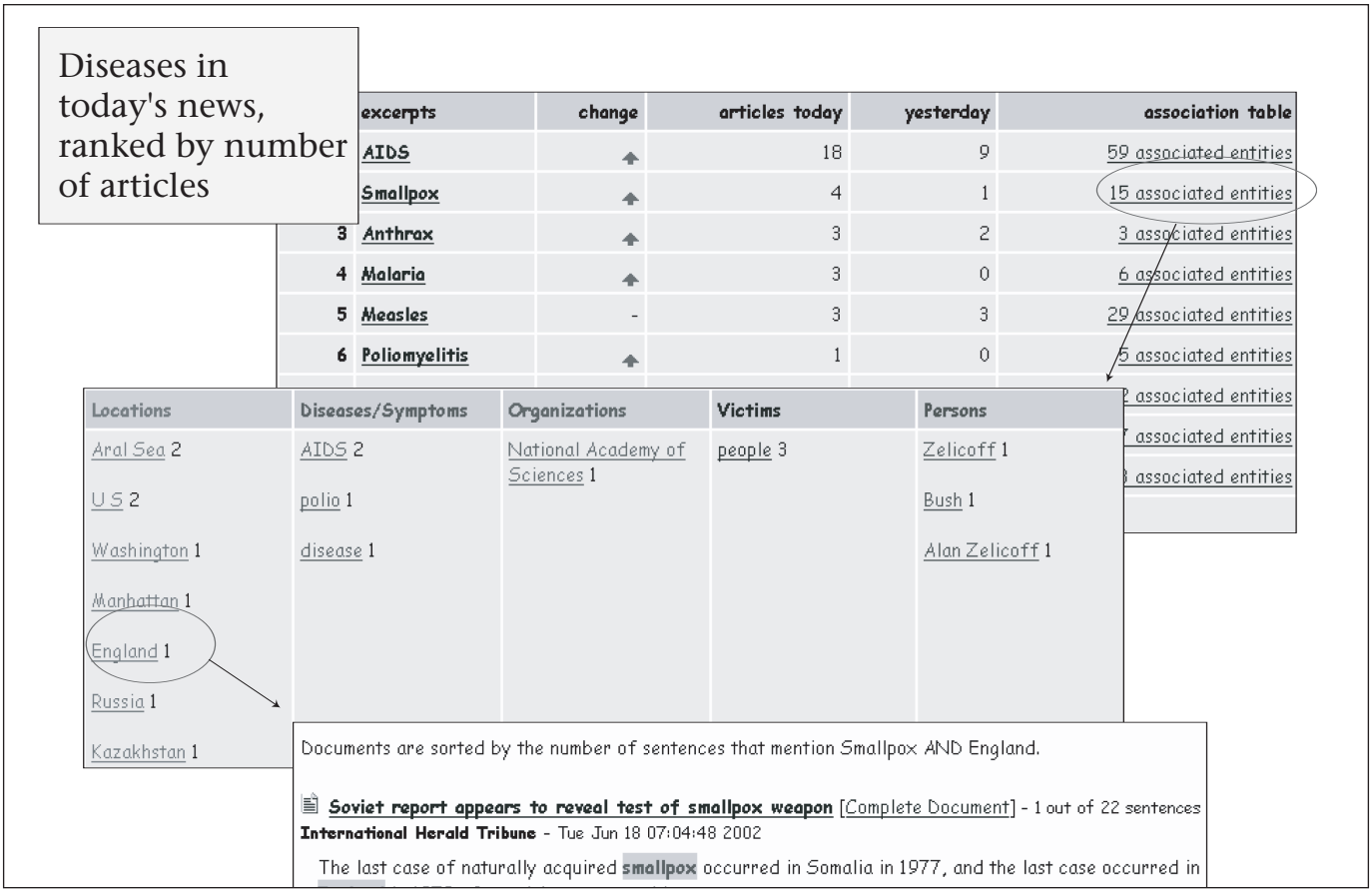

Figure 10. A Daily Ranked List of Top 10 Diseases Can Alert to Disease Outbreaks or New Developments in Cures or Vaccines.
Links provide access to summaries on specific diseases. Association tables show related entities.

MITAP Maintenance

One or two individuals are typically responsible for the daily maintenance of the MITAP system, including a number of administrative tasks, such as adding new user accounts as they are requested, informing users (with an e-mail distribution list) of changes to the system (for example, new data sources, outages for planned maintenance) and obtaining user feedback by 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 (that is, 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 by a mechanism other than the web or e-mail can require more time to integrate (for example, a subscription-based data feed). There is a wide range of methods by which such data can 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 a site changes, the web capture might 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 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 (for example, ALEMBIC, CYBERTRANS, or WEBSUMM) usually require the assistance of an individual with more specialized knowledge of

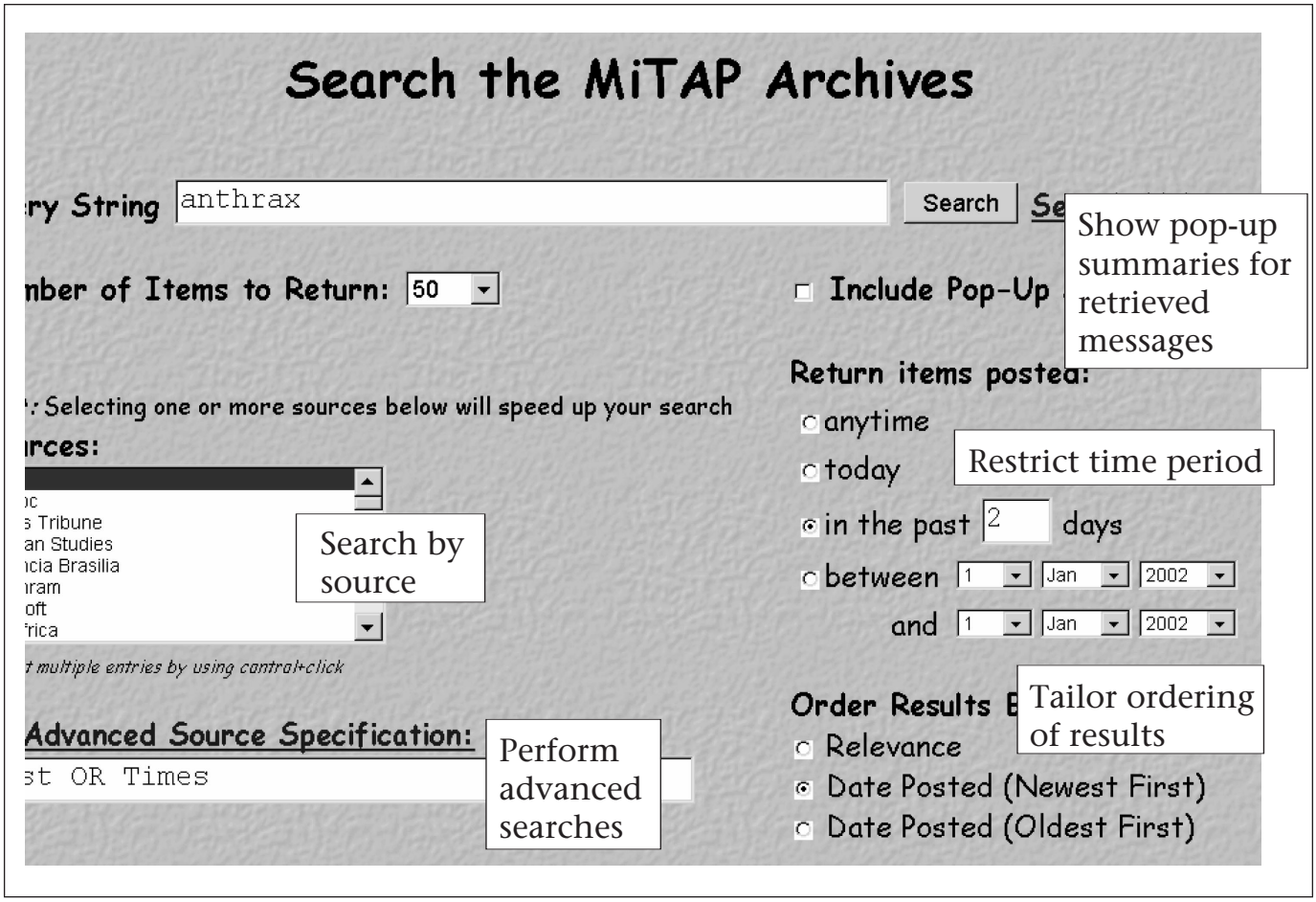

Figure 11. A Web-Based Search Engine Supports Source-Specific, Full-Text Information Retrieval across All Documents in the News Server and in the Archives.

the relevant component. For example, to improve our named entity tagging (for example, 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

MITAP has been accessible to users since June 2001. Data can be accessed in two ways: (1) by newsgroups or (2) 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 150 user accounts. Dozens of people use the system on any particular day, with a daily average of 10 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 (figure 13). Users access the system for updates on disease outbreaks as well as to read current news from around the world.

Figure 14 illustrates averaged daily MITAP activity from July 2001 through the end of June 2002. The dark line, on the left axis, shows the number of messages processed and posted to the system, and the lighter line, on the right axis, shows user access by 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 our subscribers read the contents of hotbox every time they log on to the system.

One regular user, a consultant to an organization of medical professionals, spends 1 to 2

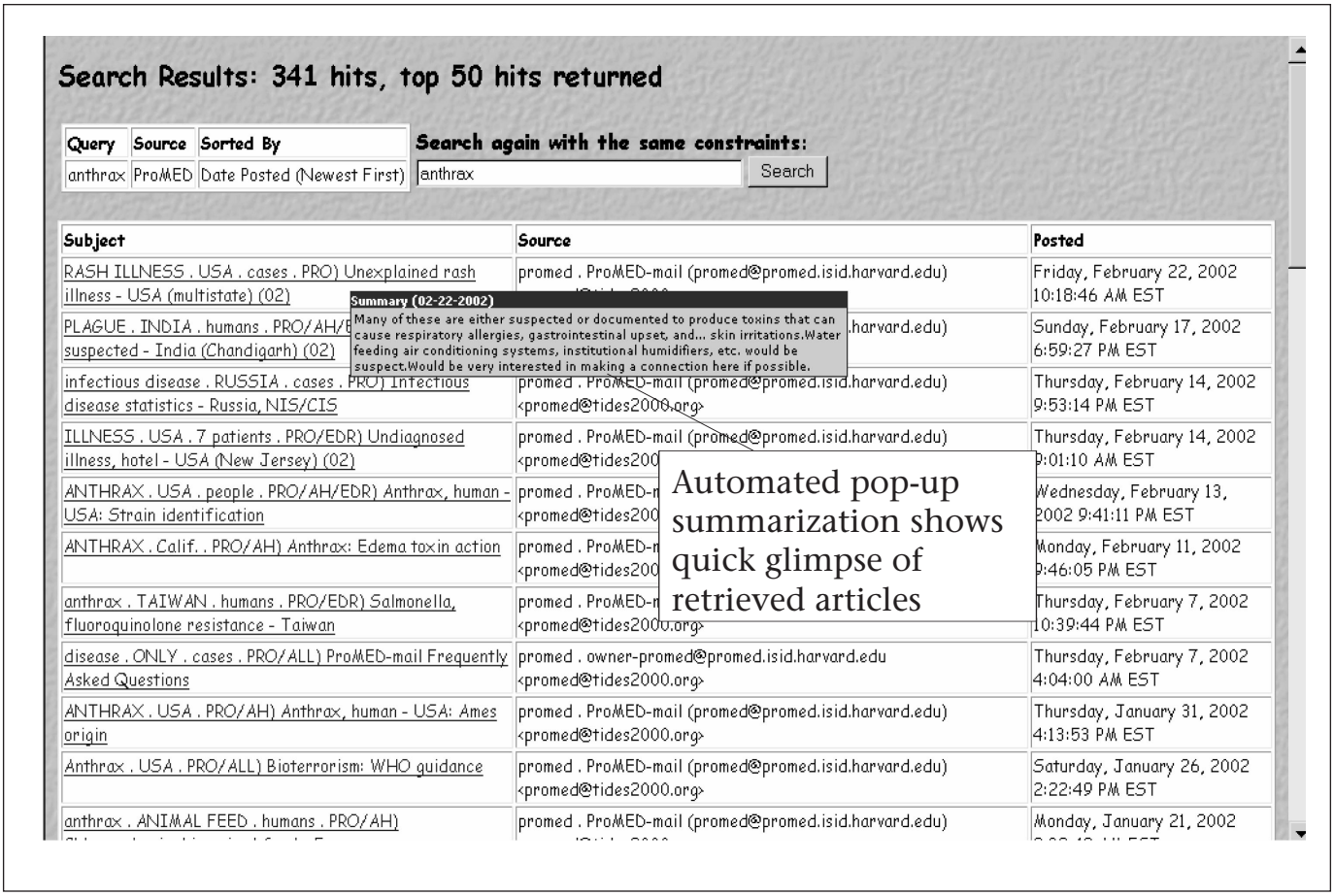

Figure 12. Search Results Can Be Sorted by Relevance or Date.
Pop-up summaries provide a glimpse into the retrieved articles.

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 5 to 15 articles of high importance. These selected articles are used to create the TIDES World Press Update,[11] 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.

MITAP Evaluation

The Disease of the Month Experiment, a series of user-centered, task-based minievaluations, was designed to assess utility, evaluate usability, measure progress, and provide iterative feedback to MITAP developers. We chose a scenario familiar to analysts (that is, 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 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: (1) efficiency, (2) task success, (3) data quality, (4) user satisfaction, and (5) usability. These categories were adopted and modified from those established by Walker et al. (2001) for the Defense Advanced Research Projects Agency (DARPA) COMMUNICATOR Project.

In our experiments, MITAP users provided

Figure 13. MiTAP's User Base Includes Members of the American Red Cross, the United Nations, and the European Disaster Center.

Our results … showed that the test groups were able to find a larger number of relevant articles in fewer searches. In fact, … test groups … 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 using MiTAP articles.

more detail and more up-to-date information on disease outbreaks than the web alone; however, they did not necessarily spend less time doing so. Our results also showed 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 using MITAP articles. Over the course of this experiment series, the feedback has enabled the MITAP development team to improve the overall system performance (for example, throughput increased by a factor of 2.5, yet 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.

MITAP 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 percent) are repeat users, with 63 percent 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 redistribute 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.

Future Directions

Our core research on MITAP will shift focus to biomedical translingual data and associated features. In partnership with other researchers and developers, we plan to integrate new capabilities, including cross-language information retrieval and enhanced machine translation. Longer-term plans include incorporation of question-and-answer technology, additional summarization, clustering, temporal tagging and normalization, fact extraction, and alerting capabilities. Continued emphasis will be put on user requirements and user-centered evaluations to provide real functions, utility, and usability to the end user. Evaluations will focus on real analysts using multilingual biomedical data.

There are numerous, critical issues remaining to be addressed as we move forward and incorporate new technology. State-of-the-art human-language technology is errorful. A challenge is to design around the current limitations and allow users to benefit as the technology improves. For example, we would like

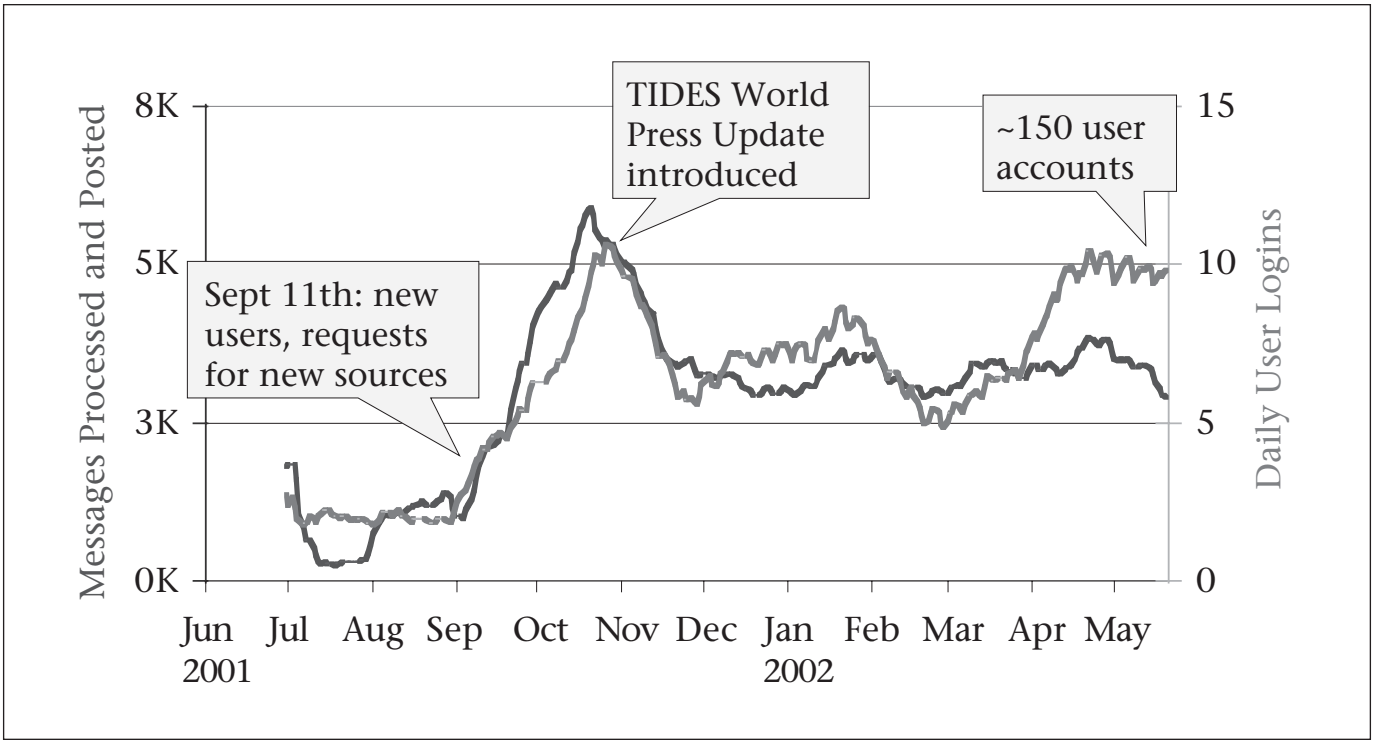

Figure 14. Daily MiTAP Activity over the Span of One Year.

September and October brought new users and new data sources to the system. The volume of news slowed in November, which also coincided with the introduction of a distributed MiTAP newsletter. MiTAP currently has about 150 user accounts, with an average of 10 people logging onto the system on any given day.

to design a mechanism to enable users to "correct" misinformation, such as mistagged entities. The corrected information could then be fed back into the system as training data. User feedback also suggests that temporal and geospatial normalization would provide improved function, not just for summarizing by visualization but also for searching over spatial and temporal constructs.

Another challenge is transforming technology into value-added utilities for end users. For example, how can relation extraction benefit the user? Perhaps we should provide browsing capability, link analysis, or something entirely different.

Although the list of open issues does not end here, one important requirement emerges from our experience. How can we enable users, with specific needs, to tailor systems to unknown or changing domains, given the current limitations and capabilities?[12]

Acknowledgments

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

Notes

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

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

3. K. Miller, F. Reeder, L. Hirschman, and D. Palmer. 2001. Multilingual Processing for Operational Users. In NATO Workshop on Multilingual Processing at EUROSPEECH. www.mitre.org/support/papers/tech-papers-01/miller_multilingual.

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

5. The Jakarta Project. 2001. jakarta.apache.org/-lucene/docs/index.html.

6. Alias I, Inc. 2002. New Technology for Information Access. www.alias-i.com.

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

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

9. Columbia Newsblaster. 2002. www.cs.columbia.-edu/nlp/newsblaster.

10. Alias I, Inc. 2002. New Technology for Information Access. www.alias-i.com.

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

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

References

Aberdeen, J; Burger, J.; Day, D.; Hirschman, L.; Palmer, D.; Robinson, P.; and Vilain, M. 1996. MITRE: Description of the ALEMBIC System as Used in MET. Paper presented at the TIPSTER 24-Month Workshop, 5–8 May, Tysons Corner, Virginia.

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. Paper presented at the Sixth Message Understanding Conference (MUC-6), 6–8 November, Columbia, Maryland.

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, 1699–1706. 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* 21(4): 543–565.

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

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

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., and Somers, H. 1992. *An Introduction to Machine Translation*. San Diego, Calif.: Academic.

McKeown, K.; Barzilay, R.; Evan, D.; Hatzivassiloglou, V.; Klavans, J.; Sable, C.; Schiffman, B.; and Sigelman, S. 2002. Tracking and Summarizing News on a Daily Basis with Columbia's NEWSBLASTER. Paper presented at HLT2002: Human-Language Technology Conference, 24–27 March, San Diego, California.

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 Thirty-Eighth Annual Meeting of the Association for Computational Linguistics (ACL'2000), 69–76. Menlo Park, Calif.: AAAI Press.

Mardis, S., and Burger, J. 2002. Qanda and the Catalyst Architecture. Paper presented at the AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, 25–27 March, Stanford, California.

Merlino, A. 2002. ViTAP Demonstration. Paper presented at HLT2002: Human-Language Technology Conference, 24–27 March, San Diego, California.

Quinlan, J. 1992. *C4.5: Programs for Machine Learning*. San Francisco, Calif.: Morgan Kaufmann.

Vilain, M. 1999. Inferential Information Extraction in Information Extraction. In *Lecture Notes of the 1999 SCIE Summer School on Information Extraction,* ed. M. Pazienza, 95–119. New York: Springer Verlag.

Vilain, M., and Day, D. 1996. Finite-State Phrase Parsing by Rule Sequences. Paper presented at the 1996 International Conference on Computational Linguistics (COLING-96), 5–9 August, 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.; and Whittaker, S. 2001. DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection. Paper presented at EUROSPEECH 2001, 3–7 September, Aalborg, Denmark.

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.

Yamada, K., and Knight, K. 2001. A Syntax-Based Statistical Translation Model. In Proceedings of the Thirty-Ninth Annual Meeting of the Association of Computational Linguistics, 523–530. Menlo Park, Calif.: AAAI Press.

Yarowsky, D. 1994. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In Proceedings of the Thirty-Second Annual Meeting of the Association of Computational Linguistics, 88–95. Menlo Park, Calif.: AAAI Press.

Laurie E. Damianos is a lead artificial intelligence Engineer in MITRE's Intelligent Information Access Department at the Center for Integrated Intelligence Systems in Bedford, Massachusetts. Damianos has been with MITRE for more than five years. Her interests focus on the field of human-computer interaction, including design and development of usable and useful systems for real problems and real users, applications of user-centered evaluation, and research into evaluation methodologies. Damianos is currently project lead of MiTAP, a system for monitoring infectious disease outbreaks and other global threats. She is involved in several user-centered evaluation efforts for this project as well as a question-and-answer system. Damianos received a B.S. in math and computer science and a B.S. in the biological sciences from Carnegie Mellon University. Her e-mail address is laurie@mitre.org.

Jay M. Ponte received a B.S. with honors in computer science from Northeastern University in 1993. He received an M.S. and a Ph.D. in computer science from the University of Massachusetts at Amherst in 1996 and 1998, respectively. While in graduate school, he worked on classification of text-based medical records, Chinese natural language processing, topic segmentation, and information retrieval. His dissertation work in probabilistic language modeling for information retrieval has been influential in the information-retrieval field. After graduate school,

he joined GTE Laboratories, where he managed the SuperPages Advanced Development Group and was awarded two patents in the areas of web search and text classification. In 2000, he joined the MITRE Corporation where he continues to work on probabilistic approaches to natural language processing and information retrieval and also on software architectures in support of these technologies. His e-mail address is ponte@ mitre.org.

Steven Wohlever has worked at the MITRE Corporation for the past five years. His work has focused on the areas of distributed system design and systems integration. He received his M.S. in computer science in 1997 from the University of Rhode Island, where he was a member of the Real-Time, Distributed Computing Research Group. His e-mail address is wohlever@mitre.org.

Florence Reeder is currently a Ph.D. student in information technology at George Mason University, where she is pursuing research in the intersection of language learning, computer-assisted language testing, and machine translation evaluation. At the MITRE Corporation, she is coleader of the Human Language Technologies TAT with the MITRE Technology Program and is working on the Defense Advanced Research Projects Agency Translingual Information Detection, Extraction, and Summarization (TIDES) Project in translingual information access. She has been the project lead on a lexicon development research project and has led the CyberTrans and Quick-MT Machine Translation projects. Previously she worked in signal processing for E-Systems/Raytheon Corporation developing a system for mobile cellular communications exploitation. Her e-mail address is freeder@mitre.org.

David Day is associate department head for the Intelligent Information Access Department and deputy program manager of the Northeast Regional Research Center. He is currently involved in a variety of projects involving natural language processing, information extraction, and linguistic annotation technologies. Day received a B.A. in philosophy from Hampshire College in 1978. He earned an M.S. (1985) and a Ph.D. (1991) in computer science from the University of Massachusetts at Amherst. Since 1995, Day's research at MITRE has focused on natural language processing, including applying neural network algorithms to anaphora resolution; using n-grams for adaptive e-mail filtering; and developing the machine learning component for acquiring phrase rule sequences in Alembic, an information-extraction system. His current research focus is on multilingual information-extraction (English, Chinese, Spanish) and cross-document entity tracking. His e-mail address is day@mitre.org.

George Wilson is a lead artificial intelligence engineer and a group leader at the MITRE Corporation. He is also a professorial lecturer in the Department of Linguistics at Georgetown University. He received an A.B. and M.S in mathematics from the University of Chicago and a Ph.D. in mathematics from Brandeis University. His current research interests include information extraction and the use of large untagged corpora in natural language processing. His e-mail address is gwilson@mitre.org.

Lynette Hirschman is chief scientist for the Information Technology Center at the MITRE Corporation in Bedford, Massachusetts. She received a B.A. in chemistry from Oberlin College in 1966, an M.A. in german literature from the University of California at Santa Barbara in 1968; and her Ph.D. in formal linguistics from the University of Pennsylvania in 1972 under Aravind Joshi. Her research interests span text processing and spoken language understanding, with a strong emphasis on evaluation. She is principal investigator of two large Defense Advanced Research Projects Agency–funded programs at MITRE: (1) Translingual Information Detection, Extraction, and Summarization (TIDES) and (2) communicator, an architecture for spoken-language interfaces. Her current research interests include natural language processing for biomedical data, looking at issues of information access for biology and creating common evaluations and standards; her other main interest is on reading comprehension—developing systems that can take and pass standardized reading comprehension tests, such as those given in elementary school or high school. In this project, we have developed automated grading for short-answer tests, and we have explored the use of such systems in interactive teaching environments, for both English and other languages. Her e-mail address is lynette@mitre.org.