Real-Time Event Extraction for Infectious Disease Outbreaks

Ralph Grishman Silja Huttunen Roman Yangarber

Computer Science Department
New York University
New York, NY 10003

grishman@cs.nyu.edu silja@cs.nyu.edu roman@cs.nyu.edu

ABSTRACT

We describe a system for creating and automatically updating a data base of information on infectious disease outbreaks. A web crawler is used to retrieve current news stories; potentially relevant stories are fed to an information extraction engine, whose output is used to update the data base. A web-based browser allows users to examine the data base and select records based on date, disease, location, and other fields. The browser can also display the source text from which the data base information is derived.

The data base provides an alternative to conventional search engines for retrieving documents in this domain; the advantages and disadvantages of this approach are discussed.

1. INTRODUCTION

New York University has developed a system, Proteus BIO, for maintaining, in real time, a data base of information on infectious disease outbreaks, linked back to the reports from which they are derived. This system has been developed under the TIDES program as an offshoot of the IFE-BIO Integrated Feasibility Experiment [3].

The goal of our system is to provide a simple but rich index structure for access to both current and historical reports on infectious disease outbreaks. The tool is intended to allow retrospective summaries to be quickly prepared by analysts in response to current events. Searches can be performed using any combination of date, location, and disease information. The data base (event index) is automatically updated on a daily basis from on-line news reports.

The system consists of

1. a Web crawler and mail reader
2. a text zoner
3. an information extraction system
4. an interface to a relational data base
5. a Web-based data base viewer, linked to the original texts

We describe each of these briefly in turn.

2. WEB CRAWLER / MAIL READER

The web crawler gathers daily news from multiple sources, including general news sources (such as newspapers) and medical sources such as the communicable disease reports from the World Health Organization and ProMed, a moderated electronic forum where medical professionals exchange reports on the spread of infectious diseases. For added coverage, the crawler also invokes a news search engine, using a small set of keywords that are relevant to epidemics of diseases. In addition, the system incorporates a mail reader interface, so that information from mailing lists can be added to the data base as it is received.

For general news sources, the system applies a keyword filter, requiring that one of a set of keywords (derived from the event patterns used by the extraction system, as described in section 5) be present before the document is sent to the extraction engine. This is done for efficiency purposes, to avoid further processing of totally irrelevant articles. The final discrimination between relevant and irrelevant is made by the extraction engine, using the full event patterns.

3. TEXT ZONER

The text zoner divides the incoming documents into zones for use by the information extraction engine. The zones include headline, source, date, and text body. For documents incorporating multiple stories, story boundaries are marked, for separate processing by the extraction engine (to block incorrect use of contextual information across stories.)

The zoner is rule-based. For HTML text, we parse the HTML and rely primarily on the mark-up for identifying zones. For plain text (such as the ProMed mailing list input), the rules make use of

layout and appearance (such as blank lines, underscores, all-cap lines) as well as keywords.

4. STRUCTURE OF EVENT DATA BASE

To record the information on disease outbreaks, we have developed a relatively simple event structure, whose primary components are the range of dates, the location, the name of the disease, the number and type of organisms affected, and whether the organisms died. A disease report may give rise to a substantial number of such elementary events [4]. For example, "7 people were infected and 2 died" generates two events in the data base; so does "15 people were infected last week, bringing the total for 1999 to 100". Events may contain other (sub-) events, either geographically or temporally (as in the last example). The data base includes linking fields which are intended to capture such containment relations.

Normalization of data is critical to support effective retrieval of information. Dates and date expressions (including dates relative to the report time, such as "last week") are normalized to a standard form, with explicit day, month, and year ranges. City and area names are tagged with the country, using both a gazetteer (for major cities) and information from elsewhere in the document. Disease names are normalized using a list of disease synonyms.

The events are stored in a relational data base, implemented using mysql.

5. EXTRACTION ENGINE

The extraction engine is a multi-phase finite-state transducer, previously used on several MUC evaluations and other extraction tasks. Processing begins with tokenization and lexical look-up. This is followed by finite-state pattern matching to recognize names (of people, locations, organizations, and diseases), noun and verb groups, select noun phrases, and finally select clause structures. Event recognition (in this case, finding reports about the incidence of infectious diseases) occurs at both the noun phrase ("outbreak of ...") and clause ("people died from ...") levels. Reference resolution is used to resolve anaphoric noun phrases, and event merging rules combine information about a single event from several clauses or sentences. Links back to the original text are maintained throughout the processing. The events extracted from the document, along with the links to specific phrases, are used to update the data base.

To monitor the combined performance of the text zoner and extraction engine, we maintain test corpora derived from the ProMed and WHO sources. Our current performance on these corpora is F=0.54, which is comparable to the best performance on MUC tasks, but this performance is continuing to improve as we enlarge our knowledge sources and training corpora

Customization of this system to the domain of infectious-disease outbreaks — primarily, the development of the domain-specific concept hierarchy and patterns — was done using our example-based customization tool [1]. Additional patterns have been found using the ExDisco pattern-discovery procedure [2].

6. INTERFACE

To provide access to the data base, we use a Web-based interface which provides a spreadsheet-like presentation of data base records. Users can select subsets based on any combination of attributes, including date, location, and disease type. Each event is linked back to the corresponding text record; selecting an event on the spreadsheet brings up the text and highlights the event, with the components (date, location, disease) suitably color-coded. This allows for rapid validation of the extracted information and review of associated information.

A snapshot of the interface is shown below. The main window shows the data base itself, with one record highlighted, along with the record selection criteria (in this case, either 'anthrax' or 'smallpox' in the disease name field). The window in lower left shows the selected record in more detail. The window at lower right shows the text from which the record is derived, with the text used for specific fields (date, location, disease name, case descriptor) color-coded.

Our original interface was implemented as a client-side Java applet. Because this interface required a browser supporting Java Swing, we have recently completed an alternative interface implemented on the server side using JavaServer Pages (JSP), which can operate with all standard browsers.[1]

7. STATUS AND PLANS

The current data base incorporates

- all of ProMed (11,471 reports, 17,899 disease outbreak events)
- all of the WHO Infectious Disease reports

In addition, we are currently accumulating events reported by the Chicago Sun-Times, and from various sources retrieved through the Northern Light news search engine, and plan to add several other news sources.

Access to documents through the event data base created by information extraction has some significant advantages over keyword-based browsing. Requiring disease names to be in the context of outbreaks excludes incidental mentions, descriptions of treatments, etc. Having normalized dates allows searches over date ranges which would not be possible with the original documents; associating countries with city and area names similarly allows search in geographical regions. For example, a query such as "Which countries have reported anthrax outbreaks in the past year?", which could be answered quickly from the event data base, would be difficult to answer efficiently with a keyword-based search. In addition, the tabular structures provide a compact summary of outbreaks in a particular region or involving a particular disease. An informal evaluation we conducted in September 2001 (using a task designed by MITRE, which required us to prepare a report in a 2-hour period) indicated to us the benefit of such tabular structures for the rapid creation of summaries.

On the other hand, the set of extraction patterns will be incomplete and so will miss some instances of disease outbreaks. In circumstances where one can exhaustively review the documents returned by a keyword-based search for a disease, the results can be more comprehensive. Since a user may not always have time for such an exhaustive review, the combination of

[1] The client-side interface was based on an earlier version written by Troy Straszheim; the server-side interface was written by Joshua Rosenblatt.

keyword-based and extraction-based search can offer the user the most flexibility. This will be one avenue of development as we continue to work with MITRE, the lead IFE-BIO contractor, to integrate our event extraction technology with their keyword-based routing and retrieval system.

As we add more sources to our system, we are seeing multiple reports of the same outbreaks. These present both a problem and a potential strength. If the event entries from individual documents are kept separate, the user will be presented with redundant entries reporting the same event, a potential distraction. However, combining compatible entries can yield multiple benefits. Extraction patterns are likely to miss some events, but the chance of missing an event if stated differently in two sources is considerably reduced, so the data base will be more complete. Data which is successfully extracted from two sources can be reported with improved confidence. This will be one area we will focus on as we continue our development.

One other important direction for extension involves linking infectious disease reports to other news events for which we have developed an information extraction capability, such as "natural disasters" (hurricanes, floods, earthquakes). Natural disasters often have repercussions involving disease outbreaks. Linked data bases with common formatted fields (location, date) will allow for event search and summary along multiple dimensions.

8. ACKNOWLEDGMENTS

This research was supported by the Defense Advanced Research Projects Agency as part of the Translingual Information Detection, Extraction and Summarization (TIDES) program, under Grant N66001-001-1-8917 from the Space and Naval Warfare Systems Center San Diego, and by the National Science Foundation under Grant IIS-0081962.

This paper does not necessarily reflect the position or the policy of the U.S. Government.

9. REFERENCES

[1] Roman Yangarber and Ralph Grishman. Customization of Information Extraction Systems. *Proceedings of International Workshop on Lexically Driven Information Extraction*, Frascati, Italy, July 16, 1997.

[2] Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Automatic Acquisition of Domain Knowledge for Information Extraction. *Proc. 18th Int'l Conf. on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany, July-August 2000, 940-946.

[3] L. Hirschman, K. Concepcion, L. Damianos, D. Day, J. Delmore, L. Ferro, J. Griffith, J. Henderson, J. Kurtz, I. Mani, S. Mardis, T. McEntee, K. Miller, B. Nunam, J. Ponte, F. Reeder, B. Wellner, G. Wilson, A. Yeh. Integrated Feasibility Experiment for Bio-Security: IFE-Bio, A TIDES Demonstration. *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*, Morgan Kaufmann, San Francisco, 2001.

[4] Silja Huttunen, Roman Yangarber, and Ralph Grishman. Diversity of Scenarios in Information Extraction. To appear in *Proceedings of the Third International Conference On Language Resources And Evaluation*, Las Palmas, May 2002.

Snapshot of the Proteus-BIO Data Base Interface