

Innovation

Air Force SBIR/STTR Innovation Story

SBIR Topic Number:
AF05-090

SBIR Title:
Enabling Visualization of Event Information from Unstructured Text

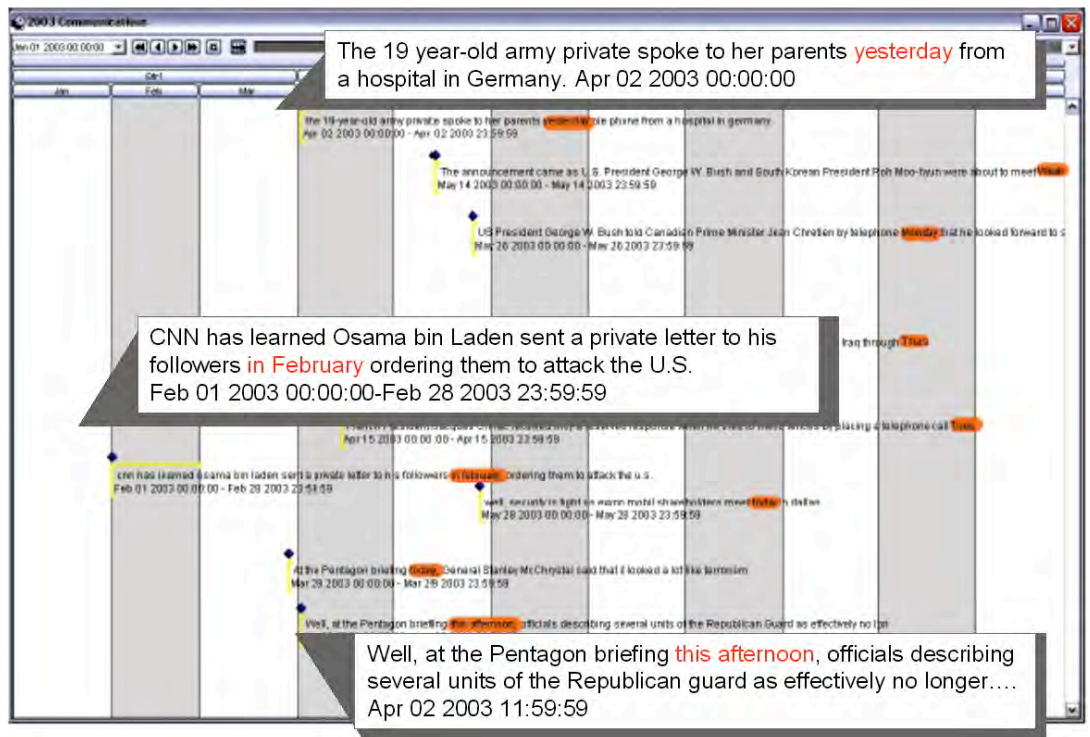
Contract Number:
FA8750-06-C-0056

SBIR Company Name:
Janya Inc.
Amherst, NY

Technical Project Office:
AFRL Information
Directorate, Rome, NY

This Air Force SBIR/STTR Innovation Story is an example of Air Force supported SBIR/STTR technology that met topic requirements and has outstanding potential for Air Force and DoD.

Timeline Visualization of Events



Enabling Visualization of Event Information from Unstructured Text

- Existing visualization tools require input to be in a structured database; however, most valuable information is in free text
- Current technology for event extraction from free text is based on verbs; many events are not expressed in verbs, but in nominal phrases (e.g., the attack took place) leading to low accuracy
- SBIR developed technology supports the automatic or semi-automatic analysis of large volumes of electronic information
- Using machine learning, Semantex™ has been extended to detect nominal events, thereby greatly increasing the accuracy of event detection and enabling visualization of information from free text

AFRL/IF 081607

A

DISTRIBUTION A:
Approved for public
release; distribution
unlimited.

Air Force Requirement

In order to make the information about events extracted from unstructured text truly useful to information analysts, the extraction process must be accurate (high precision) and thorough (high recall). It must present sufficient detail about particular events (times, locations, participants). It must allow for sufficient and relevant organization and consolidation of extracted event mentions either by type (e.g., all assassinations), identity (e.g., all mentions of the assassination of Rafik Hariri), or temporal or causal relations (e.g., all events leading up to the assassination of Rafik Hariri). If these conditions are achieved, extracted events can be presented visually in a way that will maximize their utility to analysts. High precision and recall will allow all and only actual event mentions to be passed to a visualization application. Detailed time and location information provides a basis for organizing event mentions on timelines and maps. Participant information allows analysts to see events relevant to particular entities (people, organizations, and locations) as they browse through entity profiles. Organization and consolidation of events allow simultaneous presentation of groups of event mentions of similar type, of mentions of a single event, and of mentions of related events.

SBIR Technology

Technology has been developed that enables the extraction of nominal events, event mentions anchored in noun phrases, such as "the attack" in the sentence "The attack took place last night in Baghdad." This technology extends the current capabilities of the Semantex™ information extraction system that Janya has previously developed. Semantex is a complete information extraction system that supports the automatic or semi-automatic analysis of large volumes of electronic information in order to detect entities, attributes, relationships and events. Semantex represents a hybrid model for information extraction, merging machine-learning and grammatical approaches, and leveraging the strengths of each to achieve a high degree of accuracy. Machine learning techniques for weakly supervised learning, specifically bootstrapping, have been developed to distinguish event mentions from non-events. This technique leverages a repository of data representing Semantex processing results on a large text corpus. The learning module is provided with some samples of true nominal events (seeds) based on a lexicon of words that reliably denote nominal events. The rich set of features generated by Semantex provides the context for learning new extraction patterns, hence enabling semi-supervised learning of nominal events.

A subsequent stage uses supervised learning to classify the resulting nominal events as being either generic or specific, or real or unreal. For example, in the sentence "Until his arrest in December 2002, the suspect smuggled about 200 compatriots into

the United States," the word "arrest" indicates a specific event that took place. Contrast this with the sentence "Police said further arrests were possible," where the event denoted by "arrests" is neither specific nor real. The ability to extract nominal events has significantly increased the accuracy of event detection, as well as providing key attributes (time and location) that enable visualization of information.

Potential Air Force Application

The ability to visualize event information from unstructured text in a more conceptual form, such as on timelines and maps, will make it easier and faster for warfighters to gain situation awareness and to detect potentially significant changes. It supports multi-intelligence fusion by providing event information from textual data sources, like human intelligence (HUMINT).

The primary impact of this work is on timeline visualization, a critical component in situation awareness systems. The results of this effort will lead to a significant increase in the number of time-stamped events produced by an information extraction (text analytics) system. The ability to time stamp critical event information is also vital to a business text analytics system. For example, it would enable a timeline plot of the significant events associated with a company; this can be useful in understanding the events leading to a lawsuit. Other benefits arise from the increased number of events detected; this can be used to constrain search results as well as perform data-driven clustering of events across documents.

Company Impact

Janya specializes in providing products and services to government markets to support information discovery from unstructured data. The ability to increase the accuracy of event detection has increased Janya's opportunities to offer complete solutions by partnering with companies that provide visualization and link analysis tools. Janya recently announced a partnership with Intelligent Software Solutions (ISS) paving the way for solutions combining the power of Semantex text analytics with Webtas visualization. This has already led to a new AFRL-sponsored prototype for visualizing information from U.S. message traffic data (USMTF), including HUMINT. Such solutions require high accuracy detection of event information which is enabled by this effort. This is also a major step towards transitioning Semantex into an operational environment.

In March 2007, Janya received the BETA Star award at the annual BETAS award dinner given by InfoTech Niagara, the leading trade association for the western New York technology industry. This award is given to an entrepreneurial firm bringing unique, innovative products or services to the market that show promise of becoming major economic forces.

Rohini K. Srihari, CEO
Janya Inc.
Tel: (716) 565-0401 x 303
E-mail: rohini@janyainc.com



SBIR/STTR

Air Force SBIR Program
AFRL/XR
1864 4th Street
Wright-Patterson AFB OH 45433

AF SBIR/STTR Program Manager: Steve Guilfoos
Website: www.sbirsttrmall.com

Comm: (800) 222-0336
Fax: (937) 255-2219
e-mail: afrl.xrs.dl.sbir.hq@wpafb.af.mil

