Multilingual NameTag^{TM}
Multilingual Internet Surveillance System
Multimedia Fusion System

SRA International
4300 Fair Lakes Court
Fairfax, VA 22033
703-502-1180 (tel)
703-803-1793 (fax)
http://www.sra.com

1 Multilingual NameTag^{TM}

SRA's Multilingual NameTag is name tagging-software that can handle multiple languages, including English, Japanese, Spanish, Chinese and Thai. It finds and disambiguates in texts the names of people, organizations, and places, as well as time and numeric expressions with very high accuracy. The design of the system makes possible the dynamic recognition of names: NameTag does *not* rely on long lists of known names. Instead, NameTag makes use of a flexible pattern specification language to identify novel names that have not been encountered previously. This makes maintenance and porting to new domains very easy. In addition, NameTag can recognize and link variants of names in the same document automatically. For instance, it can link "IBM" to "International Business Machines" and "President Clinton" to "Bill Clinton."

NameTag incorporates a language-independent C++ pattern-matching engine along with the language-specific lexical, pattern, and other resources necessary for each language. The Japanese, Chinese, and Thai versions integrate word segmenters to deal with the challenges of these languages.

NameTag is an extremely fast and robust system that can be easily integrated with other applications through its API. For example, it can be integrated with an information retrieval system to improve retrieval accuracy and with a machine translation system to prevent name translation mistakes. Versions are available for UNIX and Windows 95 and NT platforms.

2 Multilingual Internet Surveillance System

The Multilingual Internet Surveillance System uses SRA's NameTag, the powerful SQL capability of an RDBMS, and a Java-enhanced Web-based GUI to provide an intelligent surveillance capability. The special features include:

- **Built-in Java-based Web crawler:** By using this built-in Web crawler, the user can choose key WWW sites for surveillance. It automatically retrieves Web documents for intelligent indexing. The crawler has a built-in scheduler and make uses of multiple threads for the quickest possible acquisition of documents.
- **Concept-based intelligent indexing by NameTag:** SRA's NameTag indexes retrieved Web documents and extracts the most important information, i.e. the proper names. In addition, NameTag can be customized to identify collections of other domain specific terms which are of interest to a particular Internet surveillance system (e.g., financial, legal, medical or military terms).
- **Pro-active monitoring and alert capabilities:** Using a variety of data mining techniques, the system can monitor daily activities on the Internet (what's new and hot today?) and alert the user to unusual activity as it is happening.
- **Powerful SQL queries through an easy-to-use Web-based GUI:** Once alerts go off, the user can perform more in-depth analysis by retrieving relevant information through the user-friendly GUI. Powerful SQL capability along with concept-based indexing ensures high precision and time saving.
- **Automated hyperlinking for intelligent browsing:** Another way to analyze the information effectively is to browse texts by following hyperlinks automatically created by the system. Hyperlinks are added for each proper name and custom term found by NameTag.
- **Multilingual capability for monolingual speakers:** By incorporating multilingual versions of NameTag and machine translation modules, monolingual speakers can also retrieve, browse, and analyze the content of foreign language documents.

The multilingual capability allows the user to gather and assimilate information in foreign lan-

guages without further effort. For example, by simply clicking on one of the hyperlinks, the user can view a list of other articles in any language that contain the same term (either original and translated). By entering queries in English, the user can obtain all documents in any language that contain the English terms or their translations.

The Multilingual Internet Surveillance System provides a truly unique way to analyze and discover necessary information effectively and efficiently from a vast information repositories on the Internet. For example, it can answer types of questions which cannot be asked of traditional search engines, such as "Which companies are mentioned along with Internet and Netscape?" or "Which people are related to the Shinshintou Party?"

In addition, the concept-based indexing allows high-precision search; the user can ask for documents that contain "Dole," the former senator, instead of "Dole," the pineapple company. In short, the system can eliminate most of the noise associated with traditional search engines and focus attention on precisely the information of interest.

The Web-based client runs on multiple platforms. The server currently runs on a SUN Solaris platform (other server ports are underway).

3 Multimedia Fusion System

The Multimedia Fusion System (MMF) combines an automated clustering algorithm with a summarization module to automatically group multimedia information by content and simultaneously determine concise keyword summaries of each cluster. MMF assists the user who must assimilate a vast amount of information from different sources quickly and effectively. As MMF generates clusters in an unsupervised fashion, i.e., no pre-defined user profile need be used, the system can adapt to new and changing world events with no extra effort.

Specifically, the system takes newspaper articles and CNN Headline News, and creates a hierarchical cluster tree in which related stories are clustered together at tree nodes regardless of their sources. MMF consists of four main components: keyword selection, document clustering, cluster summarization, and cluster display. The resulting cluster tree is visualized in a Java-based interactive GUI. The user can follow a cluster tree hierarchy and expand clusters all the way down to individual documents. For newspaper articles, the text is shown while for CNN Headline News, both the closed-captioned text and the captured video are displayed in-line with a browser plug-in. Each displayed cluster also has its concise keyword summary next to the corresponding tree node.

In addition to its clustering capabilities, the MMF server is also responsible for capturing video, audio, and closed-captions from a live satellite feed in real time. The data is segmented as it is received and can be simultaneously stored and forwarded to viewers on the network. The server also handles data input through textual newswire feeds.

The Web-based client runs on multiple platforms. The server currently runs on a SUN Solaris platform.

Contact:

Chinatsu Aone	Dave Conetsco
(technical)	(administrative)
aonec@sra.com	dave_conetsco@sra.com