SRA PROJECT FOR ARPA / USACOM

David Oppenheimer
SRA International, Inc.
4300 Fair Lakes Court
Fairfax, VA 22033
david_oppenheimer@sra.com
(703) 803-2412

I. INTRODUCTION

The USACOM project is part of the larger SAIC/BBN Advanced Joint Planning (AJP) Advanced Concepts Technology Demonstration (ACTD). The primary objective of the project is to demonstrate the effectiveness of Tipster-sponsored data extraction technologies. Although this project is just now getting underway, we plan to investigate the applicability of data extraction technologies in a number of scenarios, including:

- Improved accuracy of text filtering based on integrating natural-language based extraction with detection (text retrieval)
- Using data extraction technology to automatically categorize incoming texts into predefined folders
- Integrated structured information extracted from texts as an additional source of input to the planning process
- Enabling the collection of aggregate information derived from texts for information presentation
- Methods of displaying text to focus on area of interest explicitly

II. NAMETAG OVERVIEW

NameTag consists of a software engine that applies name recognition rules to text, supported by lexical resources and limited lists of proper names. Users implement an application driver that invokes the engine and configures the processing. NameTag can either generate a document that has the names annotated with SGML (Standardized Generalized Markup Language), or provide a table of the names with indices (offsets) to the text -- "annotation" mode. A name recognition rule consists of a pattern and an action. The pattern is similar to a regular expression and consists of special operators and operands that match portions of text. Typically, patters recognize structural or contextual indicators of names and thus perform dynamic recognition.

The action performs operations on the text, such as tagging a name with a classification. The rules are partitioned to form processing phases that primarily recognize one class of name. For example, NameTag has separate phases for recognizing personal names and organizational names. This feature allows for the recognition of certain classes of names to impact the recognition of other names. The lexical resources contain information about words, such as their part-of-speech and their meaning.

III. PROJECT PLANS

This project will integrate with two community-developed reference architectures. These are the Joint Task Force (JTF) Reference Architecture and the Tipster Architecture. With respect to the JTF Reference Architecture, SRA's core data extraction product (NameTag) will be integrated into the Generic Server layer, with a customized data extraction prototype integrated into the Application Layer. With respect to the Tipster Architecture, a Tipster-compliant version of NameTag will be used. Additional portions of the developing architecture will be evaluated for the appropriateness of insertion into the demonstration (for example, the document manager or text annotator). The entire project will be evaluated by the Tipster Engineering Review Board (ERB) for documentation and certification as appropriate.