THE HOOKAH INFORMATION EXTRACTION SYSTEM

Chris Barclay, Sean Boisen, Clinton Hyde, and Ralph Weischedel
BBN Systems and Technologies
70 Fawcett Street
Cambridge, MA 02138
sboisen@bbn.com
617-873-4309

0. ABSTRACT

This paper describes Project HOOKAH, a TIPSTER Implementation Project with the Drug Enforcement Administration to extract information from the DEA-6 field report. The paper overviews Project HOOKAH, describes the system architecture and modules, and discusses several lessons that have been learned from this application of TIPSTER technology.

1. PROJECT HOOKAH

1.1 Overview

Project HOOKAH is a TIPSTER Implementation Project with the Drug Enforcement Administration to extract information from DEA field reports in support of populating a database. Its goal is the partial automation of DEA operations by moving information extraction technology into the DEA fileroom, where these reports are currently manually processed.

HOOKAH has been supported by Congressional "Dual Use" funding for transferring TIPSTER technology to civilian agencies. The prototype development effort has been managed by Mary Ellen Okurowski and Boyan Onyshkevych of the Department of Defense. The deployment effort is being jointly managed by DoD and DEA, with DEA responsible for life cycle maintenance.

1.2 Domain: DEA-6s

The focus of Project HOOKAH is to improve the processing of the DEA-6 report, a semi-formatted report generated primarily by field agents, as well as legal staff, analysts, and others. DEA-6s are organized into case files, and are composed of multiple sections with varying amounts of formatting. Header fields are normally highly formatted, and indicate the subject, case, date, time, etc. There is a semi-formatted index, which contains references to most subjects to be added to the database and some information about them. There is also unformatted text, where much of the useful information is found.

1.3 Current DEA Operation

Project HOOKAH will augment an existing work flow that depends on substantial manual data extraction. Currently, DEA personnel read hardcopy DEA-6s and other forms to manually identify extractable data. Using this information, they then attempt to uniquely identify the subject in the NADDIS database, which contains millions of subjects and is widely used throughout the law enforcement community. Once the subject is identified, the data extracted from the document is retyped into NADDIS. Quality control is provided by an independent group of analysts who review the extraction and database update.

The manual extraction of information from DEA-6 documents represents a major expense, which is contracted out to more than 100 analysts working in two shifts. Thousands of documents are processed per week by these personnel, a substantial volume of data. DEA is in the initial stages of converting to softcopy report dissemination.

Several factors make this an ideal task for the application of TIPSTER information extraction technology:

- a constrained domain makes information extraction feasible
- high traffic volume increases the payoff for reducing manual processing
- an established base of analysts are available to support system development and perform testing
- the NADDIS database already exists and has high value to the customer
- the need for timely database update is an appropriate match for state-of-the-art TIPSTER technologies

1.4 Concept of Operations

The envisioned configuration for HOOKAH would give each analyst a HOOKAH "workstation", for most simply an X-terminal or comparable X-based display. DEA-6s in softcopy form will arrive daily over the network, and will be automatically grouped by case or file number.

The HOOKAH system will analyze each DEA-6 off-line, doing an initial match of DEA-6 subjects against NADDIS records, extracting information for each subject, and performing a provisional merge against any existing NADDIS information to provide an initial set of suggested database updates.

The analyst will then verify the system's NADDIS match, and review the proposed database updates, correcting them as required and entering any additional information. An audit trail mechanism will track the status of each DEA-6 in the system. Completed DEA-6s and their associated NADDIS updates will go to Quality Control for review.

2. ARCHITECTURE

The HOOKAH architecture is presented in Figure 1.

2.1 HOOKAH Modules

Preprocessor The HOOKAH preprocessor converts incoming electronic DEA-6s in a specially-coded format to SGML markup. It also adds markup for certain sub-fields in the text which are not encoded in the original format.

Extraction Processor The bulk of HOOKAH processing is done off-line by the extraction processor, which is responsible for processing of the indexing section of the document to determine which subjects should be processed, automatic (non-interactive) matching of subjects against NADDIS, and extracting information from the body of the text. This is performed during off-hours to increase user productivity and maximize use of machine resources.

NADDIS (Database) Interface All communication with the NADDIS database proceeds through the NADDIS interface. Since a program interface to NADDIS has not been provided, all interactions take place through a screen-oriented forms interface. Queries from HOOKAH are translated into commands to this interface, and the resulting display screens are parsed by the NADDIS interface module into normalized data structures. This module also transmits updates to NADDIS, and handles certain database error

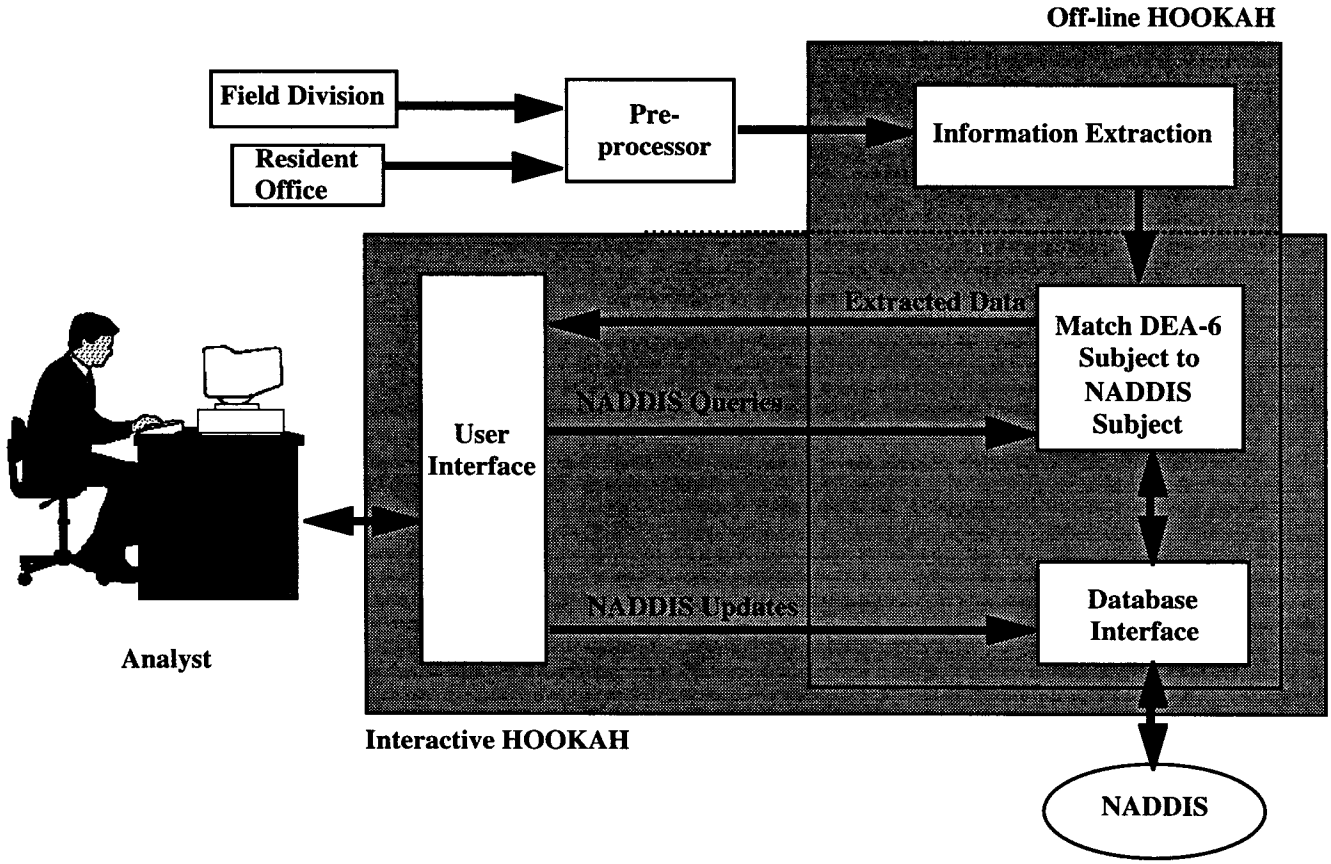

Figure 1: HOOKAH architecture.

conditions.

User Interface Once data is prepared off-line, it is available for user review. In addition to supporting displaying and editing information, the user interface also supports manual browsing of NADDIS in case a different subject match is required. All system processing results, including user corrections, are stored as annotations once the document is complete.

3. LESSONS LEARNED

The project has completed an operational prototype, and the next phase will deploy the system operationally on a small-scale. Several lessons for transitioning TIPSTER technology have emerged from the project so far.

3.1 Database Integration

A crucial component of the HOOKAH problem is comparing extracted information to a legacy database (NADDIS). This aspect has not been addressed in the TIPSTER program or the Message Understanding Conference (MUC) evaluations, so there is only limited previous experience to draw on for this aspect of the project.

In particular, interfacing with a database has required several new technology components:

- matching information extracted from text against existing database records, and determining when no suitable match exists and a new database record must be created
- once a match is made, extracted information must be fused with database records to isolate new information. Since the database information is normalized according to different standards than are used in preparing the reports, this aspect is challenging.
- updating the database with the new information

In the HOOKAH application, interfacing with the NADDIS database has been a difficult systems engineering problem. In addition, the fact that much information in the text already exists in the database, and that the text and database can disagree, provides additional challenge.

3.2 User Interface

Project HOOKAH has provided considerable experience in the importance of the user interface for extraction systems. There are both generic requirements for the interface (for example, viewing extraction results and correcting them) and application-specific ones (providing a database browser, and a tool for manual subject matching): we therefore spent significant effort in designing the interface to match the needs of the DEA analyst. In particular, the analyst's focus is updating the database, and extraction output is merely a tool to support that activity. For this reason, the user interface displays proposed NADDIS updates, not templates or other direct representations of extraction results.

Our experience so far indicates that the user interface can affect analyst throughput and performance even more than the quality of the extraction itself. No matter how good the extraction system performs, a poor interface can make the entire system unusable. This suggests a implementation strategy of spending early effort on interface design, and incrementally improving the quality of the extraction engine over time, a strategy we have pursued for HOOKAH.

Another lesson learned from the user interface pertains to the tradeoff between recall and precision. We have found that users are readily able to correct erroneous proposals from the system (improving precision), but they are less likely to skim the document for themselves to ensure no important information has been omitted (improving recall). Mechanisms to help users in supervising system recall are still an open area of investigation in the HOOKAH system.

3.3 User Involvement

For Project HOOKAH, user involvement has been essential for getting the information extraction application done correctly. As development has proceeded, it has become clear that HOOKAH will change the job description for the analyst. Previously, analysts shared both an analytical role and a data entry role: HOOKAH substantially reduces the data entry task and shifts the balance toward "supervising" and correcting the extraction system. Retraining analysts for this new job function may prove to be costly.

We have also learned that unforeseen technology mismatches can arise, complicating user involvement in testing and development. For example, the HOOKAH user interface relies on the now-standard features of window-based systems: scroll bars, buttons, and mouse operations. However, the existing analyst interface relies on character-based "dumb" terminals: for those analysts who are not familiar with window-based systems, the transition to using a mouse rather than cursor movement commands may be a more significant change than using an information extraction system.

3.4 Other Issues

We are still investigating ways to evaluate the performance impact of HOOKAH on the DEA analyst. One obvious metric is to measure the change in

productivity, though it is harder to then determine whether that change results from the use of extraction, or simply from a different user interface design. We would also like to be able to measure the effort required to correct extraction results, which relates more directly to the performance of extraction technology *per se.*

In general, our experience suggests there is still a significant gap between laboratory-grade extraction software and operational applications. There are still numerous issues to address in integrating extraction technology into useful operational systems: in general, extraction has not been the hardest problem for HOOKAH.

4. PROJECT STATUS

An operational prototype currently exists and is undergoing testing at DEA. In addition, DEA is preparing for initial operational implementation on a small-scale to test the feasibility of the system.