QUARTERLY REPORT #2

PROJECT: Research and Development for the Declassification Productivity Initiative

Grant #: DE-FG01-96IS50317


INTRODUCTION: Highlights of the second quarter include a trip by one of the OCR investigators to University of Nevada, Las Vegas to study the preparation of test suites and to explore a possible collaboration with the Information Science Research Institute. Some results and recommendations are presented for Knowledge Representation and Inferencing. Methodological strategies have been identified for the Logical Analysis research component. Preliminary findings on Tipster Technology will be reported in the third quarter.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED
a) Knowledge Representation and Inferencing: Cary deBessonet

During the second quarter efforts were focused on targeting areas in the domain that could be used as sources of information about the problems that classifiers frequently face. Because of delays in achieving reports on the progress made in TAP (Text Analysis Processor) in automated text analysis, which required extensive use of natural language understanding and lexicographical techniques, efforts in the quarter were spent assessing the methods employed in the classification process. The goal of the research has been to discover ways to improve the process within a reasonable time frame and to determine which technologies would be appropriate for the tasks. The results reveal that the classification process is far from what one might consider to be an analytical endeavor in which discretion in decision making is held to a minimum. The process appears to be more like rule application in an open-textured environment such as the legal domain.

Although most of the research has been conducted without access to actual classified information and to the actual, official guidelines used by derivative classifiers, the general methodologies seem appreciable in the guidelines and exercises that have been made public. The study conducted in the second quarter has yielded the results described below, stated as conclusions.

There are two areas in which the investigator is applying his expertise to enhance the classification process. The first area is the domain of classification guidelines. The guidelines themselves are not drafted in a way that supports optimally analytical application. A system by which to produce a more tightly drafted set of guidelines needs to be developed. The more closely drafted the guidelines, the more unpredictable the classification results will be that employ them. The investigator has written extensively in the field of scientific codification and automated systems and has gained particular expertise in this area by servicing as the Reporter for the Coordinating Committee of the Louisiana State Law Institute and as the director of the Law and Artificial
Intelligence Project of that Institute. In those roles he has been charged with the review of proposed
civil code materials and of developing automated systems to assist in the process. Theoretically, a
civil code is a comprehensive set of legal principles that govern a variety of types of behavior, status
and situation. A civil code is said to be a more tightly drafted body of principles (cf. guidelines) than
would be an ordinary statute covering the same matters. It appears that the process of producing
guidelines for classification could benefit significantly by employing a more structured and
analytical approach such as that used in scientific codification. Consequently, in the last quarter, the
investigator has worked on the specification of the design of an automated system that could be used
to help structure the process.

The second area of focus is closely related to the first one. To be able to produce a suitable
set of guidelines for impromptu use by classifiers, a system of classifying information types needs
to be developed as a technology that can be accessed by both those who draft classification
guidelines and those who use the guidelines. The technology would also be used extensively by
original classifiers to define the scope of the classifications they make. The development of such
a technology is crucially dependent on the creation of models that capture essential features of
classified information. This part of the work was described in the progress report for the first quarter
and thus will not be repeated here. Progress in this phase has been retarded by the lack of
availability of classified information and guidelines. Recently, some material has been made
available to the investigator, and the next quarter will be spent studying that material to determine
whether there is sufficient information contained therein to make significant progress specifying the
required models of classified information types.

The investigator has concluded that perhaps the best way to improve the classification
process in the time allowed under the grant is to focus on the development of technology suitable
for use in expert systems that have been carefully built to operate in somewhat narrow domains. Plans for the next quarter include a consideration of how best to implement automated assistance for the derivative classifier of guidelines drafted for a particular area. Hopefully, sufficient information on classification will be made available to the investigator to make the effort worthwhile.
b) Optical Character Recognition: S. Sitharama Iyengar and Nathan E. Brener

During the week of April 15, Dr. Brener attended part of the Fifth Annual Symposium on Document Analysis and Information Retrieval, sponsored by the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas (UNLV), and met with Dr. Tom Nartker, the director of ISRI. After the symposium, Richard Brooks, a Ph.D. graduate student at LSU, joined Dr. Brener for a visit to Dr. Nartker's research group at ISRI. During the visit, Brener and Brooks observed in detail the procedures used by Nartker's group to select document pages for OCR test suites and to prepare the "ground truth" for comparison with the OCR output. The preparation of the ground truth is a labor intensive process, requiring that every page in the test suite be entered manually by each of several different typists working independently. They then observed the software developed by ISRI to run the test suites through the various OCR systems that have been provided to ISRI by commercial vendors for ISRI's annual test of OCR accuracy. For each OCR system tested, the ISRI software compares the OCR output with the ground truth and delivers a detailed analysis of the OCR errors.

The knowledge gained from the visit to ISRI will be useful in the preparation of the DOE declassification test suite at LSU. Once this test suite has been prepared at LSU, it will be used to test current OCR systems. At this point in the project, a collaboration with Dr. Nartker's group would be advantageous, as the OCR testing and evaluation can be done using the existing facilities at ISRI. Doing this testing without the facilities at ISRI would require that LSU acquire OCR software and develop tools equivalent to those at ISRI. While this is definitely possible, it would be time consuming and the cost may be prohibitive. It is doubtful that LSU would be able to acquire the range of systems available at ISRI without it being prohibitively expensive. Thus the
use of the existing OCR test facilities at ISRI has some clear advantages. The possibility of such a collaboration was explored during the discussions with Dr. Nartker.

We also have compiled a preliminary list of factors that can cause errors when OCR software is used to convert DOE declassification documents to text. The most important factors on the list include:

- Speckle or noise
- Broken characters
- Touching characters
- Text thinning
- Blur
- Clumping
- Skew
- Baseline variation
- Resolution of scanner
- Threshold of scanner

The last two factors will be kept constant during the initial phases of this study. The list also contains a number of less important factors including:

- Vertical streaking
- Horizontal streaking
- Bleed through
- Width variations of characters
- Height variations of characters
- Jitter (variation of the arrangement of the output pixel photo-receptors from a square grid)

DOE has provided us with 530 pages which are representative of the DOE documents that are waiting to be declassified. We have analyzed these pages to determine which of the above factors they contain. For the purposes of selecting the initial test suite, each page has been marked as either None, Low, or High for each of the major factors and some of the minor factors on the list. The results of this analysis are currently being used to assemble the initial test suite, which will contain between 100 and 200 pages.
c) **Declassification through a Logical Analysis Approach: Evangelos Triantaphyllou**

The first phase was slow because we spent a lot of time in understanding and defining the problem. The recent seminar given by Mr. Scott R. Lowery was decisive in giving us a clear idea of the type of classification/declassification problem at DOE. Now we expect to proceed much faster. Next week we will receive the first computers which will be exclusively used in this research and more hardware is on its way.

We defined two closely related problems. A close examination reveals that one is the reverse of the other. The **first problem** is how to use a set of classification guides in classifying a given document. From the limited illustrative examples we were given during the seminar and the DOE cite on the World Wide Web (WWW) it seems that the antecedent part (conditions to be satisfied) of each guide can be translated into a Boolean expression. The parameters of these Boolean expressions could be conditions such as: "a specific date is given", "types and amounts of nuclear materials used", "location of a military exercise," etc. The system should be able to read through a document and decide which of these conditions are satisfied or not. We propose that this be accomplished through a second lawyer of rules. In this way, we can build an expert system with natural language processing capabilities. Currently, we are in the process of defining some key conditions (like the above ones) and soon we will be able to have a test system capable of analyzing text (ASCII) documents.

The **second problem** is how to use source documents (for which a classification status is given) and extract the underlying classification guides. Ultimately, these guides will be used to enhance the existing DOE guides.
We are involved in addressing the first problem. We expect to have a prototype system (in
a simple form) ready before the end of the summer. We will need some declassified guides and
documents, to make it more realistic.