DOE/IS/50317--T2

QUARTERLY REPORT # 3

RECEIVED

NOV 0 8 1996

OSTI

PROJECT: Research and Development for the Declassification Productivity Initiative

GRANT# : DE-FG01-961S50317

PERIOD : July 1996 - September 1996

INTRODUCTION : The third quarter began with a full-day workshop on declassification by Mr. Scott R. Lowery referred by the Technical Officer, Tom Curtis. this workshop, at Southern University, Baton Rouge, was attended by all the principal investigators. Various degrees of program have been observed for this period in all sub-areas of the DPI project. An initial evaluation of TIPSTER has revealed that its relevance to declassification would depend on a set of questions identified for further investigation. Progress is reported in developing a segment of a representation language that could be used to enable classifiers to classify and enter information and ask questions. A prototype test suite with approximately 145 pages is being finalized to be forwarded to UNLV for testing and analysis. The progress on the computer programs developed for the logical analysis approval is also reported along with a timeline of specific tasks for the fourth 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

MASTER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

a) **Optical Character Recognition/Document Recognition**
S. Sitharama Iyengar and Nathan E. Brener

The goal of the first year of this project is to develop a taxonomy of factors that cause problems (loss of accuracy) when OCR systems are used to scan documents and convert the scanned images to text files that can be processed by a word processor. This taxonomy, which will have several levels for each factor (e.g., a scale from 0 to 10), will be used to develop an automated software system that will screen a document in advance of OCR processing to determine the level of each factor present in the document and to calculate a composite level which takes all of the factors into account. The system will then use the levels for the individual factors and the composite level to determine whether the document is a suitable candidate for OCR processing (i.e., whether the predicted OCR accuracy is sufficient for DOE/DPI purposes). If the predicted OCR accuracy is too low, then the document would not be submitted for OCR processing but would be processed manually. This system will be useful in automating the DOE/DPI declassification process.

A taxonomy of factors that cause OCR errors has been developed and has been described in detail in previous progress reports. Of the various factors in the taxonomy, we have found that broken characters, touching characters, and noise are the most important in determining OCR accuracy. These three factors are currently being used to develop a prototype software system to screen documents in advance of OCR processing. In this prototype system, which is being developed in collaboration with Mike Cannon and Pat Kelly at Los Alamos National Laboratory, a page from a document is first processed to create a list of all of the connected black regions on the page. This list, which contains the size (number of pixels) in each connected region, is then used to compute the level of each of the three major factors on a scale of 0 to 10. Various methods are currently being explored for using the information in the list to compute the level of each factor. One method being investigated is to determine the average size (number of pixels) of the characters on the page (which will depend on the font) and then determine how many connected regions are significantly larger than the average size, how many are significantly smaller, and how many are very small (on the order of 5 to 10 pixels). The percentage of connected regions that are significantly larger than the average size will indicate the level of touching characters, and the percentage of regions that are very small will indicate the level of noise (since noise usually consists of dots or speckle).

In order to determine how the various levels of these factors affect OCR accuracy, we have developed a prototype test suite with approximately 145 pages that contain taxonomy factors at various levels and in various combinations. A collaboration has been arranged with Tom Nartker at UNLV in which Dr. Nartker's group will run the prototype test suite through several state of the art OCR systems and send a detailed error analysis to LSU. We will then use the results of the UNLV test to make a detailed determination of the effect that various levels of the taxonomy factors have on OCR accuracy. We have completed the preparation of the ground truth for the test suite and are currently running a zoning program to prepare the suite for OCR processing. Once the zoning has been completed, the suite will be sent to UNLV for testing.

b) **Classification/Declassification of Text Documents via a Logical Analysis Approach**
Evangelos Triantaphyllou

This report summarizes the progress of the research project "*Classification/Declassification of Text Documents via a Logical Analysis Approach*" for the period ending October 22, 1996. This research project is confined to the development of a computerized system to automatically classify/declassify DOE. The primary objective of this study is to extract the lexical compounds from a document (i.e., single words, word-pairs, phrases, etc.) that would determine its classification.

Currently, we are working on the *first problem*, of the two problems presented in our previous progress report. The first problem was defined as: *how to use a set of classifications guides (given or extracted) in classifying a given document*. Our current status is now targeted to identify lexicosemantic patterns (i.e., groups of words that appear adjacent to one another and have a meaning that is distinctly different from the individual words) in the documents to be processed.

It is important to mention here that some preliminary experiments showed that word pairs were better indicators to address this first problem than were single words. Thus suggesting to us that the lexicosemantic patterns might be a much better choice to solve the first problem. To identify these patterns, our current approach consists of the following three phases: text preprocessing, text processing, and guide (rule) inferring. These three phases will be explained in detail in the Future Work section.

WORK COMPLETED

After the submission of the first progress report we have accomplished the following tasks:

- 1) We have extensively reviewed related literature. About thirty works on classification of news, e-mails, telex, etc. have been reviewed. A similar amount of works on information retrieval (IR) systems, text analysis, machine learning and data-mining have also been reviewed extensively. Our conclusion on this literature revision is that most of the reviewed works avoid semantic analysis for the classification of documents. Most of the reviewed works conclude that semantic analysis is a very hard exercise. More surprisingly was the fact that the reviewed literature always referenced Salton's works: The Vector Space Model (VSM), proposed about twenty years ago, and Automatic Indexing.

The VSM model selects the words that best describe one document (or keywords) based on the frequency of occurrence of these words. This model suggests that the words conveying more contextual meaning are those words with "reasonable" frequency. Further, the model suggests that the words with high frequency and those words with low frequency should be discarded for future analysis. Our initial approach to solve the first problem followed the VSM guidelines. We developed three preliminary programs (in Turbo Pascal 7.0) to test the model with three DOE sample documents. The first program extracted single words, the second program extracted word-pairs, and the third program extracted also word-pairs, but this time preselected stop words (or "words used in common English with little meaning") were eliminated.

- 2) The first computer program for extracting single words was completed on 9/07/96. The results on three DOE sample documents were disappointing. We realized that words loose

their contextual meaning whenever they were separated. Further, we found that the more meaningful words were the least frequent ones, which surprisingly were the majority of the words in the three DOE tested documents. These last observations contradicted the VSM claim because according to the model, these words conveyed little or not meaning at all and, therefore, should be discarded for further analysis.

Given these observations, we formulated a new hypothesis: *Consecutive words should retain the contextual meaning*. To test this hypothesis a second program was written.

- 3) The second preliminary computer program for extracting word pairs was completed on 9/21/96. We tested this program with the same three DOE sample documents; this time the results were very encouraging. These results showed that lexical compounds (in this case word-pairs) were better indicators to characterize the documents than were single words. For example, we found that the term "nuclear weapon" was more indicative than were the individual terms "nuclear" and "weapon." As another example, we believe that the term "Kim IL" (the name of a North Korean leader) also outperformed the individual terms "Kim" and "IL."
- 4) The results were more encouraging when the second computer program was modified (completed on 10/04/96) to eliminate 421 *stop words* (or words found in common English with little meaning). This program eliminated 15-20% of unwanted patterns (i.e., "would have"), patterns which for our purposes conveyed no meaning.

Although the results were very encouraging, they also revealed some deficiencies of our approach. For example, we found paired words like "L45T is." The first word "L45T" is the technical name of a bomb, while "is" is one form of the verb "*to be*." This suggested to us that technical (or generalizing for non English words) should be treated as single terms (i.e., single words, tuples, triplets, etc.) and that further lexical compounding of these terms should be avoided.

Regarding the computing power requirement, these results were also very enlightening in two aspects. They showed that by eliminating the stop words, the CPU time requirements increased significantly. They also showed us that the memory requirements for small documents would quickly exceed the memory limits of the programming language.

FUTURE WORK

These preliminary results also showed that the three DOE sample documents in these experiments had very few words in common. Thus, we could foresee that by creating clusters or families of documents with similar content we will improve our approach to tackle down our list problem. We suggest the following three phases to solve our first problem: Text preprocessing, text processing, and rule inferring.

In **Phase 0 or Text Preprocessing Phase (planned for 11/07/95)**, our goal is to create document categories (or families of documents) with similar content with 80% of the sample documents. Examples of these families, for a given period of time, might be (i) documents on nuclear materials, (ii) documents of weapon construction, etc. These families will be created by extracting the frequency of semantic patterns from (possibly selected sections of) the documents. We will test three hypotheses for extracting these patterns from a document:

Ho: The patterns in the first paragraph held enough content to create the categories.

Ho: The patterns in the first and last paragraph held enough content to create the categories.

Ho: The patterns in the entire text are required to create the categories.

The performance measures for these three hypotheses will be the **Recall** and **Precision** measures. "Recall" is the ratio of correct system responses to the total number of categories. "Precision" is the ratio of correct systems responses to the total number of systems responses. These two measures will be our indicatives of the term overlapping between the different families, overlapping that must be minimized. The validation of these measures will be conducted first with the 80% of the documents and at a later stage with 20% of the remaining documents.

In **Phase 1 or Text Processing Phase (planned for 11/25)**, our goal is to analyze *lexicosemantically* one family of documents at a time. In this phase, however, the whole text will be scanned to extract the lexical compounds (i.e., single words, word pairs, phrases, etc.) and their frequency will be irrelevant; the presence or absence of those lexical compounds will be the decisive parameter. Those lexical compounds (or patterns) will be then converted into the attributes of classification rules. This process will be repeated for all the document categories created in *phase 0*. Similarly with *phase 0*, all the lexical compounds will be tested for recall and precision.

In **Phase 2 or Rule Inferring Phase (planned for 12/05)**, our main goal is to generate the rules (i.e., combination of patterns) that will predict if a given documents is classified or declassified. This extraction of rules will be conducted independently for each document category. The expectation of these rules is to allow our system to corroborate the documents' preassigned tags with the minimum error. In addition, using those rules, our system should be able to read through a new document and assign the (hopefully) appropriate tag.

CONCLUSION

The results obtained so far are very encouraging, specially the results obtained with the second program for extracting word-pairs. However, our new approach to extract the semantic compounds require us to modify our initial computer programs to allow for the detection of non-English words and for the compounding of three, four, and five English words. We expect to have the modified programs (along with additional ones) with more results by the end of November 1996. We also expect to have a preliminary demonstration of the system to tackle the first problem by the middle of December 1996.

c) **Tipster Technology Evaluation**
Alonzo Johnson

The declassification of a document starts with the classification of that document. Classification is the process of identifying information we need to protect in the interest of national security. Documents are currently classified using Executive orders or statutes. Atomic Energy Act of 1954, as amended, Department of Energy Organization Act of 1977, Department of Energy Delegation Order No. 0204-2, of 10-1-77 are a few statutes that specify responsibilities, authorities, policy, and procedures for the management of the Department of Energy (DOE) program to identify classified information, documents, or material. The statutes defines the categories for classifying information, such terms as restricted data (RD) and formerly restricted data (F.D.). The term "Restricted Data" means all data concerning the:

- (1) Design, Manufacture, or Utilization of Atomic Weapons;
- (2) The Production of Special Nuclear Material; or
- (3) The Use of Special Nuclear Material in the Production of Energy.

But shall not include data declassified or removed (transclassified) from the Restricted Data category pursuant to section 142. The category "Formerly Restricted Data" (section 142.d) mean classified information that has been removed from the Restricted Data category after DOE and DOD have jointly determined that it relates primarily to the military utilization of atomic weapons, and can be adequately safeguarded as defense information. National Security Information is information that pertains to the national defense or foreign relations (National Security) of the United States and has been classified in accordance with an Executive order. An Executive order is an order signed by the President. Executive Order 12958 was signed April 17, 1995 (President Clinton), effective October 16, 1995 signed to uniform a system for classifying National Security Information (NSI). To meet NSI requirements the item must meet two criteria. First Criteria, must be in at least 1 of 7 classifiable categories:

- (1) Military plans, weapons systems, or operations;
- (2) Foreign government information;
- (3) Intelligence activities, intelligence sources or methods, or cryptology;
- (4) Foreign relations or foreign activities of the United States, including confidential sources;
- (5) Scientific, technological, or economic matters relating to the national security;
- (6) United States Government programs for safeguarding nuclear materials or facilities; or
- (7) Vulnerabilities or capabilities of systems, installations, projects, or plans relating to the national security.

The Second Criteria must cause identifiable damage to National Security.

A duration of 10 years or less may be specified at the time of original classification for some NSI documents. There are 10 years and 25 years automatic declassification, while others may use one of 8 exemption categories to be exempted from the automatic declassification.

The levels of classifications are as follows:

Top Secret - Exceptionally Grave Damage to the Security of the United States
Secret - Serious Damage to the Security of the United States

Confidential - Damage to the Security of the United States

Information when classified is placed in one of three levels (Top Secret (TS), Secret (S), or Confidential (C)) and one of three categories (Restricted Data (RD), Formerly Restricted Data (F.D.) or National Security Information (NSI)). It has been stated that if all the classified documents were stacked, it would be about three times the height of the Washington Monument. If divided into categories, about 70% of the documents would be Restricted Data, 20% Formerly Restricted Data and 10% National Security Information.

Classification and declassification Authorities are divided into three areas, original classification, derivative classification and derivative declassification. Original classification is the initial determination that information is classified. It applies only to information and only to NSI. An original classifier must be a Federal employee and makes initial determination that information is classified. Derivative classification determines that documents (not information) are classified. It is based on written guidance and applies to RD, F.D., and NSI. Derivative declassification is the determination that previously classified documents no longer contain classified information. This operation must use written guidance — not source documents.

Classification guides are used to promulgate approved classification guidance. The guides are published by subject matter and list what is classified. The guides are used by authorized classifiers for derivative decisions. See Appendix A for DOE Classification Guide Structure. Appendix B contain a flowchart for derivative classification and original classification. Before any document may be declassified a mandatory review must be done. Appendix C contain a flowchart for NSI Systematic Review processing flow, Freedom of Information Act (FOIA) and mandatory review processing flow, FOIA/Mandatory Review flow for other agency document, and for DOE document. The presence of such a document for flowcharting the step for declassification is not known to exist.

The procedures for classifying an information is well documented. It is clearly defined who may classify what, how long some of it may be classified, an index on the subject matter and a list of the classification guidance, a DOE Classification Guide Structure, flowcharts for Original Classification and Derivative Classification. There are guides for declassification but no step by step flowchart.

It is the policy of DOE to release as much information as possible within the bounds of applicable statutes and Executive orders. For a document to be declassified it must first go through declassification reviews. A determination that a document is unclassified does not mean that it can be released to the public. Many statutes and Executive orders govern the release of the information (i.e., Freedom of Information Act (FOIA), Privacy Act, etc.). Some documents require the physical removal of both classified and unclassified controlled information from a document, usually for the purpose of making the document publicly releasable.

As indicated in SA-232, Document Review Procedure, September 18, 1992, item 2 states.

The reviewer clearly marks with brackets ([]) all classified information and UCNI in the document and provides the exemption category for each deletion identified, as appropriate.

This is only one of seven items listed in this communication "Document Review Procedures for Classified Information and Unclassified Controlled Nuclear Information" it makes a statement for minimum requirements for any declassification procedure. The statement is that any successful declassification procedure must first contain a successful classification procedure.

The question is how well could the Tipster Technology operate in a Classification environment. To perform in this environment Tipster must have the ability to understand text. To understand text or to extract information Jerry R. Hobbs of the Artificial Intelligence Center says the following.

The input is the sequence of words or lexical items that constitute the sentence. The output is a parse tree of the sentence. Any system will be characterized by its own set of modules, but generally they will come from the following set, and most systems will perform the functions of these modules somewhere.

1. Text Zoner, which turns a text into a set of text segments.
2. Preprocessor, which turns a text or text segment into a sequence of sentences, each of which is a sequence of lexical items where a lexical item is a word together with its lexical attributes.
3. Filter, which turns a set of sentences into a smaller set of sentences by filtering out the irrelevant ones.
4. Preparser, which takes a sequence of lexical items and tries to identify various reliably determinable, small-scale structures.
5. Parser, whose input is a sequence of lexical items and perhaps small-scale structures (phases) and whose output is a set of parse tree fragments, possibly complete.
6. Fragment Combiner, which tries to turn a set of parse tree or logical form fragments into a parse tree or logical form for the whole sentence.
7. Semantic Interpreter, which generates a semantic structure or logical form from a parse tree or from parse tree fragments.
8. Lexical Disambiguation, which turns a semantic structure with general or ambiguous predicates into a semantic structure with specific, unambiguous predicates.
9. Coreference Resolution, or Discourse Processing, which turns a tree-like structure into a network-like structure by identifying different descriptions of the same entity in different parts of the text.
10. Template Generator, which derives the templates from the semantic structures.

The rules which will drive such a system must be developed manually.

The preprocessor module would minimally determine the possible parts of speech for each word, and may choose a single part of speech. It handles unknown words, minimally by ignoring them, or more generally by trying to guess from their morphology or their immediate context as much information about them as possible. Spelling correction is done in this module as well. The filter uses superficial techniques to filter out the sentences that are likely to be irrelevant, thus turning the

text into a shorter text that can be processed more quickly. The set of keywords may be developed manually, or more rarely if ever, generated automatically from the patterns. In some cases a statistical profile may be generated automatically of the words or N-grams that characterize relevant sentences. The current sentence is evaluated by this measure and processed only if it exceeds some threshold.

Certain small-scale structures are very common and can be recognized with high reliability. The prepararsing module recognizes these structures, thereby simplifying the task of the Sentence Parser. Noun groups, verb groups and other grouping may be recognized by the preparser. The most frequent parsing algorithm is chart parsing. Sentence are parsed bottom-up, with top-down constraints being applied. For complex, real world sentences of the sort that are found in newspapers, no parser in existence can find full parses for more than 75% or so of the sentences. One solution for this is to parse the sentences in fragments, each fragment is translated into a logical form fragment and then combined to make a logical form of the whole sentence. The methods that have been employed so far for this operation are ad hoc. There is no real theory of it. The methods are developed manually.

The success of this approach can be found in results of the latest Message Understanding Conferences (MUC-6), November, 1995. Testing was conducted using Wall Street Journal text provided by the Linguistic Data Consortium. The test set for the two information extraction tasks consisted of 100 articles. A subset of 30 articles was selected for use as the test set for the two SGML annotation task. The evaluation began with the distribution of the scenario definition and training data. Sixteen sites participated in the evaluation; 15 systems were evaluated for the Named Entity task, 7 for Coreference, 11 for Template Element, and 9 for Scenario Template.

The variety of tasks were designed to reflect the interests of the community for near term and long term challenging text understanding problems. The tasks designed for MUC-6 were as follows:

1. Identification of names, which constitute a large portion of the Named Entity task and a critical portion of the Template Element task, has proven to be largely a solved problem. The majority of the system evaluated on Named Entity had recall and precision over 90%;
2. Recognition of alternative ways of identifying an entity constitutes a large portion of the Coreference task and another critical portion of the Template Element task; it has been shown to represent only a modest challenge when the referents are names or pronouns. All but two of the Template Element systems posted combined recall-precision scores in the 70-80% range. The top-scoring system had 75% recall, 86% precision.
3. The Scenario Template task concerned changes in corporate executive management personnel; the extracted information includes answers to the basic questions of "Who is creating or filling what vacancy at what organization?". The mix of challenges that the task represents — extraction of domain-specific events and relations along with the pertinent entities — yielded levels of performance that are similar to those achieved in pervious MUCs (40%-50% recall, 60%-70% precision), but with a much shorter time required for porting. The highest Scenario Template performance overall was 47% recall and 70% precision.

The Tipster technology has progressed to a level where it can extract information text with a predefined subject. The level of precision is high but the question is still what level is to be

accepted. The technology is getting better each year as indicated in the reports that are published. The missing questions in all of the reports is the time factor to solve the given problem and the processing power used. The complexity of classification is also a problem. The document could be published on one or more subjects and contain information about any number of other subjects and all subject may or may not be classified. The question of time performance enters the formula even more as the number of possible subjects increase.

The Tipster technology has already been used to design and implement digital library systems at a number of locations. The technology is working good in those environments. The declassification task go a step beyond search engines for library systems. The library systems is where a subject text is given and all related documents are retrieved. In the classification system, a document would be given and a search is performed on the guides for a match of subjects. Once a subject has been identified and determined to be classified then all information in the document which may direct or indirectly identify that subject must be recognized. This is similar to the Scenario Template described above. In this area the technology precision is only about 70%, but is improving. The ability to recognize external relationship will prove to be an advantage. External meaning text which may reference information external to this document which may lead to the identification of information within the document.

The answer to the question of, "If Tipster could be used in the declassification effort?", can not be completely answered at this time. More information is needed.

- 1) The question of how many guides?
- 2) What percentage of precision is required for acceptance by DOE?
- 3) What effects will an increase in the number of subjects have on the performance of the routines?
- 4) What time factor is acceptable (document size vs processing time)?

These and other point of interest will be addressed and discussed in the coming months. Site visits to one or more of the MUC-6 participants are planned and the TREC conference next month will also be attended.

Information to prepare this document was taken from:

- 1) General Course on Classification/Declassification, Student Syllabus, Handouts, and Practical Exercises by Classification Training Institute.
- 2) A Generic Information Extraction System by Jerry R. Hobbs, Artificial Intelligence Center SRI International.
- 3) Message Understanding Conference (MUC-6) Proceedings.

APPENDIX A

DOE Classification Guide Structure

APPENDIX B

Original Classification

Derivative Classification

Appendix C

CHART 1: NSI SYSTEMATIC REVIEW PROCESSING FLOW

• See procedures Section C.2.c, page 16
 • See procedures Section D.6.a, page 22
 See Sections D.2 - D.5 for other non-standard cases

**CHART 1:
FOIA AND MANDATORY REVIEW
PROCESSING FLOW**

Note 1:
DOE review and Other Agency review are often done in parallel.
The final RD/FRD stamp consolidates results of both reviews.

**CHART 4:
FOIA/MANDATORY REVIEW FLOW
DOE DOCUMENT**

* See Table 1 for Classifiable Topics allowed.
 ** See Table 2 for 10-year exemption categories.
 *** See Table 3 for 25-year exemption categories.

Table 1: NSI Classifiable Topics from EO 12958

Reference	Classifiable Topic from EO 12958
1.5(a)	Military plans, weapons, or operations
1.5(b)	Foreign government information
1.5(c)	Intelligence activities, sources or methods, cryptology
1.5(d)	Foreign relations or foreign activities of the United States including confidential sources
1.5(e)	Scientific, technological, or economic matters relating to national security
1.5(f)	U.S. Government programs for safeguarding nuclear materials or facilities
1.5(g)	Vulnerabilities or capabilities of systems or installations relating to national security

Table 2: NSI 10-Year Exemptions from EO 12958

Exemption Reference	10-Year Exemption from EO 12958
X1	Reveal an intelligence source, method, or activity, or a cryptologic system or activity
X2	Reveal information that would assist in the development or use of weapons of mass destruction
X3	Reveal information that would impair the development of use of technology within a United States weapon system
X4	Reveal United States military plans, or national security emergency preparedness plans
X5	Reveal foreign government information
X6	Damage relations between the United States and a foreign government, reveal a confidential source, or seriously undermine diplomatic activities that are reasonably expected to be ongoing for a period greater than that provided in paragraph (b) above
X7	Impair the ability of responsible United States Government officials to protect the President, the Vice President, and other individuals for whom protection services, in the interest of the national security, are authorized
X8	Violate a statute, treaty, or international agreement

Table 3: NSI 25-Year Exemptions from EO 12958

Exemption Reference	25-Year Exemption from EO 12958
25X1	Identify the identity of a confidential human source, or reveal information about the application of an intelligence source or method, or reveal the identity of a human intelligence source when the unauthorized disclosure of that source would clearly and demonstrably damage the national security interests of the United States
25X2	Reveal information that would assist in the development or use of weapons of mass destruction
25X3	Reveal information that would impair U.S. cryptologic systems or activities
25X4	Reveal information that would impair the application of state of the art technology within a U.S. weapons system
25X5	Reveal actual U.S. Military war plans that remain in effect
25X6	Reveal information that would seriously and demonstrably impair relations between the United States and a foreign government, or seriously and demonstrably undermine ongoing diplomatic activities of the United States
25X7	Reveal information that would clearly and demonstrably impair the current ability of United States Government officials to protect the President, Vice President, and other officials for whom protection services, in the interest of national security, are authorized
25X8	Reveal information that would seriously and demonstrably impair current national security emergency preparedness plans
25X9	Violate a statute, treaty, or international agreement

d) Knowledge Representation and Inferencing
Cary deBessonnet, Principal Investigator

In this quarter research has focused on building upon prior results on the belief that to be most successful in the declassification area, several coordinated modules are needed, namely: an inferencing module, a tool collection, and an interface. The inferencing module(s) is to drive the overall system and thus must be capable of operating not only in a pure rule-based declassification environment, but also within a text processing environment generated by text processing programs, including any natural language understanding (NLU) systems that can be made available. Specific inferencing capabilities are to be distributed in automated tools that original and derivative human classifiers will use to carry out classification and declassification tasks. Each module is to be specially tailored for the domain in which it will be employed. A friendly and flexible interface is being developed to enable the classifiers to employ the tools effectively and efficiently.

Ideally, a representation language with expressive power comparable to the predicate calculus and a multipurpose interpreter for that language should be developed and used within the inferencing module. Research thus far indicates that for best results, both the language and the interpreter should be capable of being employed at multiple levels of the system. That, in turn, would make for efficiency in operations since uniformity of representation and processing could be preserved throughout the system. Progress has been made this quarter in developing a segment of a representation language that could be used in the system. At present, the plan is to avoid requiring human classifiers to use the representation language directly. Instead, a system of menus and prompts will be used to enable classifiers to classify and enter information and to ask questions. The reason for this is that since NLU capabilities are not available, the system must rely on formal representations and thus a classifier would have to participate in an extensive training program just to be able to use the language. Because of the limited funding available at this time, it does not seem appropriate to proceed on the assumption that such a training program would be implemented. Therefore, the inferencing and interpretation modules are being designed to operate behind the interface.

The inferencing module is to operate over both sentential objects and nonsentential input such as weighted keyword configurations produced by text processing programs. The goal in this respect is to have the inferencing module detect key semantic relations between the objects over which it operates. The system is being designed to be able to detect and construct nonstandard associations between objects. Nonstandard associations are those that are not understood in terms of deductive or probabilistic relations. Higher-level representations are being defined in terms of lower-level representations, and strategies are being developed to relate the various levels of representation. The goal in this respect is to determine the specifications of a knowledge base that would possess sufficient semantic scope, including adequate top-level and anchoring representations, to enable it to ground the interpretation module of the system. A scheme of multiple levels of abstraction is being developed to fix meanings of concepts at the basic level. Figure 1 schematically diagrams the relation between a higher level representation and a basic representation.

-
1. < a <gave<at t-1>> b g> <delighted<at t-2>> b>. [surface level expression]

 2. {[a (caused (bmo give)) (b (has (at t-1)) g)] [intermediate representation]
 caused
 (b (is (at t-2)) delighted)}

 3. (b (has-not (at t-0)) g) [lower-level representation]
 and
 {[a causes (b (has (at t-1)) g)]
 causes
 [(b (has (at t-2)) EMOTIONAL-STATE)
 and
 ((EMOTIONAL-STATE (the)) is POSITIVE)
 and
 ((EMOTIONAL-STATE (the)) is= delighted)]}
-

Figure 1. Levels of Expression

[Note: In Figure 1, a, b and g are variables (a and b ranging over persons and g ranging over things). The arrows indicate that the components of representations 1 and 3 map into representation 2, which functions as an intermediary causal formalism. The term "bmo" flags the term that describes the type of act at the root of the causal relation. The terms "t-0", "t-1", and "t-2" are sequential temporal indicators. The link "has-not" represents the negation of the link "has", and the term "the" attached to the term "EMOTIONAL-STATE" indicates that all the references are back to the first emotional state mentioned. The uppercase terms "EMOTIONAL-STATE" and "POSITIVE" are primitive terms that function to structure lower-level representations.]

The idea is to produce anchoring representations automatically from surface-level representations by employing syntactic functions. A syntactic function W can be defined to operate on the top-level representation Q in Figure 1 to produce an equivalent basic level representation that would be optimized for storage and retrieval. In other words, the following would hold:

$$W(Q) = D_i \langle a, b, g, tx, \uparrow ty, tz \rangle$$

where D is a name of a basic formalism indexed by i and where $\langle a, b, g, tx, \uparrow ty, tz \rangle$ is an ordered sequence associated with formalism D_i and where members of the sequence have mappings into the levels of representation in Figure 1. In this example tx, $\uparrow ty$, tz map to the temporal

indicators in Figure 1. The anchoring representations are to function as the semantic base of the system. The system is to use these anchoring formalisms and representations to assist it in detecting relations between key terms in classified documents. For this approach to produce the best results, techniques by which to parse document text into the basic formalisms will have to be developed. In the next quarter, an attempt will be made to determine how best to integrate the described capabilities and coordinate them with components of the interface.