DOE/IS/50317--T4

QUARTERLY REPORT

PROJECT: Research and Development for the Declassification Productivity Initiative

GRANT #: DE-FG01-961S50317

PERIOD: JANUARY 1998-JUNE 30, 1998

INTRODUCTION: Since new funding was unavailable for third year project activities, this report reflects the intended deliverables on the SU/LSU DPI project at the end of the second year "carryover"project period, December 21, 1998.

Except for the Tipster component which has been concluded (see report #3), the other three components are proceeding with significant developments and the lack of continuous funding would severely marginalize the results.

We are proposing to use the balance funds for the TIPSTER component to conclude the Knowledge Representation research. If this is acceptable, then an additional deliverable on the latter, will be forthcoming with the final report. Otherwise, this report will be the final report for the Knowledge Representation component.

RECEIVED
JUL 09 1998
OSTI

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER

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.

a) **Knowledge Representation and Inferencing**
Cary deBessonet

The Knowledge Representation Component progressed in accordance with submitted work plans to develop models of various types of information with a view towards employing them in classification technology. Work over the 1996-1997 term focused on the design of technology that could be used to assist both derivative and original classifiers in performing their daily tasks. The bulk of the results of the research are included in the Report for the 1997 DPI Conference that was submitted to DOE in March of 1997. A summary and an assessment of the results are given below.

The primary product of the research was the design of an interactive support system that would combine and coordinate classification and representation technologies. The system, called ACSS (Automated Classification Support System), would be used to represent structural and semantic content of various textual domains while simultaneously building conveniently accessible knowledge bases for them.

The technology was designed based on the following findings produced during the course of the research:

- that the system could take advantage of recent advances in object-oriented programming and windows to produce a suitable interface for the technology;

- that the system could employ independent modules for various tasks and could manipulate the results by means of coordinating modules or systems;

- that a Lisp environment would accommodate both the operating modules and the coordination modules;

- that by employing a uniform representation scheme, the inference engine of the system could be made to operate over multiple modules;

- that a version of the representation language SL (Symbolic Language) developed by the principal investigator would be suitable for representing document content for ACSS;

- that useful results could be produced using a limited set of quantifiers and operators of SL;

- that an interpreter similar to the one designed for SMS (Symbolic Manipulation System) by the principal investigator could be used in an interactive mode to build knowledge bases that could be queried about their contents; and

- that ACSS could be built incrementally to perform useful operations at each level of development.

The research has yielded very promising results in a very short time and should be continued. It is quite unfortunate that the research was interrupted for lack of funding. Some of the theory developed has already drawn the attention of NASA, which is now funding part of the development of the inclusive theory of classification as used in SMS. In that research, SL and SMS-like technology are being employed by the principal investigator to advance the theory of android (robot) epistemology and cognitive (epistemic) inference. The research planned for the DPI Project of DOE bears a complementary relationship with the work for NASA. If continued, the work for DOE can draw upon the theory of classification being developed within SMS for NASA, and this should enable ACSS to acquire advanced reasoning capabilities of the sort that could ground a sophisticated declassification system.

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

This project will culminate with the following itemized deliverables at the end of the project period.

1. 144-page test suite used to test four current OCR devices.

   The deliverables for this test suite will include the following:

   A. Original copies of the pages in the suite

   B. TIF file for each page in the suite

   C. Zone file for each page in the suite

   D. Ground truth file for each page in the suite

2. Written report giving results of OCR test on the 144-page test suite.

   For each page in the suite, this report gives character accuracies and word accuracies for each of the four OCR devices tested. The report also gives character and word accuracies for the entire 144-page suite.

3. Publication which analyzes the results of the OCR test and draws conclusions as to which of the four OCR devices is best suited for processing OD typewriter-era documents.

4. Computer program which processes the TIF file for a page and determines the cell width (character width) for the text on the page. This program assumes that fixed pitch, rather than proportional spacing, was used to produce the page of text. Once the cell width has been determined, it could potentially be used to improve OCR accuracy.

c) Classification/Declassification via Logical Analysis

Evangelos Triantaphyllou

1. ELEMENTARY PROBLEM DESCRIPTION

At first, I would like to briefly describe the central problem we studied in this project. {Important notice: In the following paragraphs by "document classification" we mean the placing of a given document into the appropriate class. Such classes may be documents in the "secret", "top secret", "declassified", etc. categories}

Suppose that given are collections of documents which belong to different categories. That is, we assume that these documents have already been classified by human experts and have been placed into the appropriate categories (such as "secret", "top secret", "declassified", etc.).

The main problem then is to use the above groups of documents as training instances and extract any pertinent information, in the form of patterns of keywords, which in turn can be used to accurately classify new documents (without the assistance of the human experts).

This problem is a critical one for a number of reasons. First, it is rather easy to have collections of documents which have been already classified by human experts. Therefore, it makes sense to try to classify new documents, based on the results of the past classifications. Of course, this assumes that enough documents have already been classified to have captured the most representative cases.

2. SOLUTION APPROACH USED

An intuitive approach for the previous problem might be to try to design a system which can analyze a given document the way humans do and then try to examine which classification guide(s) is(are) applicable. Such an approach implies the development of a parser- based syntactic recognition system. That is, the system should be able to parse a document into sentences, identify the key parts (verb, subject, object, etc.) of each sentence, and eventually infer the precise meaning of a given document.

Although such systems have already been developed and used with some success, they are also too rigid to be used in complex real- world applications. The reason is that such systems assume that the text has been developed in perfect or near perfect accordance with the syntactic and grammatical rules of the (English) language. If the author of a text, made mistakes or used too complicated grammatical/syntactic schemes, then chances are that the parser-based system will get confused and make errors. The best well known example is the failure of developing computerized text translation systems which can translate text written in one language to text written in another language.

The text analysis approach which we proposed is both flexible and robust. We had already developed an approach which can extract patterns from the characteristics of the training examples (documents grouped into similar classes in our case). That approach is very accurate and it can also explain its decision making process because it is based on mathematical logic.

In order to apply our approach to the document classification / declassification problem, we followed the following steps.
First, we processed all available training documents in order to extract all keywords. For this task, we used procedures suggested in the literature (i.e., we dropped the common words and kept the rest). Next, each document was represented by a binary vector of size N (where N is the total number of all keywords). If for a given document the i-th keyword was present, then the binary vector representation for that document had the i-th digit equal to 1, 0 otherwise. These binary vectors are known in the literature of text analysis as "document surrogates" or just "surrogates".

Therefore, by using the above document surrogates, a set of documents in a given group can be represented by a set of binary vectors.

The next step was to apply to methodologies for classifying new documents. The first method was the Vector Space Model (VSM). The second method was a method developed by us at LSU and it is called the One Clause At a Time (OCAT) approach. The VSM is supposed to be the most widely used method for text analysis (Salton, 1989). We tested these two methodologies on a total of 3,000 text documents retrieved from the TIPSTER collection of documents. The results of these experiments are briefly described in the next section.

It is important to state here that we have two versions of the OCAT approach. One is based on a branch-and-bound (B&B) algorithm while the other is based on a fast heuristic with quadratic time complexity. The B&B approach is more time consuming but it returns patterns of near minimal size. On the other hand, the fast heuristic returns small size patterns but in a fraction of the time. Thus, it is more practical for large scale inference problems as is the case for this project. Therefore, this was the approach we used in this project.

3. SOME EXPERIMENTAL RESULTS

We used documents from the TIPSTER collection. The TIPSTER collection is comprised by text documents which belong to different categories. Such categories are: DOE documents, Associated Press (AP) documents (i.e., news stories from AP), Wall Street Journal (WSJ) articles, and ZIPFF documents (technical papers from various journals). Since we do not have access to genuine DOE documents distributed into different classes, we used the previous four classes to test our procedures.

We performed three types of experiments. The first type is what is known as the Round-Robin test. For these tests we first randomly selected 60 documents grouped into two classes. We formed pairs of classes from the previous four categories of documents in the TIPSTER collection. We allocated 30 documents in each class (i.e., a total of 60 documents).

Then a document was taken out of the collection. The remaining documents were used to derive the patterns of keywords (for the OCAT case) or the centroids (for the VSM case). The excluded document was used as a testing case against the centroids or the patterns derived above. Its inferred classification was examined in regard to its actual class membership. The above procedure was repeated until all the documents had been used, one at a time, for testing the accuracy of the centroid and patterns of keywords derived from the remaining documents. The numbers of successes and failures under each method were recorded as well. The above Round-Robin test was repeated 10 times with different data each time.

The second type is a typical training and testing experiment done in such cases. That is, documents grouped into two classes were randomly selected from the TIPSTER collection. Next, 70% of these documents were assigned as the training examples, while the remaining 30% as the testing data. That is, the centroids and patterns of keywords were derived as above by using the training examples only. These centroids and patterns of keywords were next used to infer the actual class membership of the remaining (30%) testing examples. The accuracies of the two methodologies under different pairs of document classes were recorded.

In both types of tests, the OCAT method clearly outperformed the VSM method. At this point we would like to stress out that the OCAT method possesses a rather interesting characteristic. The way it makes classification decisions, allows for 3 possible outcomes. An unclassified object (i.e., a text document with hidden class membership in our case) is assigned either to the first class, or to the second class, or is set to be an UNDECIDED case. The last outcome is issued if the method is in doubt of the correct class membership of the new example (document).

It turned out that in our experiments with the TIPSTER documents, the VSM method was 55-60% accurate and 45-40% inaccurate. On the other hand, the OCAT method was 70-80% accurate, and only 2-6% inaccurate. The rest of the cases were assigned as undecided. The above accuracies are expected to get higher (for both methods) if larger collections of documents are used.

Finally, we performed a third type of experiments titled the "Guided Learning" tests as follows. Suppose that two collections of documents, which belong to two different classes, are somehow available. These documents may have been categorized by human experts. Thus, one can use a classification approach (such the OCAT method) to extract the pertinent classification knowledge (e.g., in the form of patterns of keywords).

Next suppose that a large number of uncategorized documents is also available. However, their actual class membership is not known yet. The system (which is based on the documents classified so far) can apply its procedures and propose a class membership for a given new document. A human expert can also derive the ACTUAL class membership of that new document. The problem now is which document to consider next for classification by the system.

From the above problem description it follows that if the system derived from the training examples is incorrect, then we would like the next example to reveal that fact. If the system is incorrect and new examples are classified identically by the human expert and the system, then no progress in our derivation of more accurate classification rules (i.e., patterns of keywords) is possible. That is, it is highly desirable for the case of having an incorrect system, the next example to reveal a case of contradiction between the human expert and the derived system.

The OCAT approach also offers an intuitive avenue for dealing with the above problem. As next example it uses documents which can be classified as UNDECIDED by the current pattern of keywords (i.e., the derived classification patterns of keywords). We used as the next document a randomly selected one and also one classified as UNDECIDED. The result is that under the OCAT approach and when as the next document to include for training is an UNDECIDED one, the extracted pattern of keywords can classify the remaining documents much more accurately than when the next example is considered randomly.

Please note that a full description of all of the above will be presented in two papers currently being prepared for submission and publication in referred journals.

4. SIGNIFICANCE OF THE EXPERIMENTAL RESULTS

The results from the first and second type of experiments demonstrated the the OCAT approach is superior to the VSM method for the following three reasons:

a. It is more accurate than the VSM method. Although we used rather small training sets (sets of size 60-170 documents), the accuracies achieved were between 2-6%. We expect for larger sizes to have even higher accuracy rates.

b. Under the OCAT approach, when there is not enough evidence to place a new document in a given class, it is classified as an UNDECIDED case. This in turn could be classified by an experienced human expert. That is, the OCAT approach provides for a flexible way for classifying documents in the appropriate category and when in doubt call for a human expert to assist. When a document is classified by a human expert, then the rule base (patterns of keywords) of the computerized system can be updated by including the new document in the training instances.

c. The patterns of keywords provide for an easy interpretation. Such patterns reflect the reasons why a document is placed into one class over another. That is, the OCAT approach can explain its decision making process. This can help the validation and verification process which should be an integral part of a computerized approach to the document classification / declassification process.

5. REMAINING TASKS TO BE PERFORMED

Next we are working on developing an incremental learning version of the OCAT algorithm.

Right now if a new document is presented to the OCAT algorithm and the current rule base needs to be modified, all the data have to be processed from the beginning. We are working on an "incremental learning" version of it. That is, we would like to do only a limited processing which will focus on the new training case (new document). This can make the proposed approach more time efficient. We have some algorithmic ideas that we want to implement. We are also going to test them empirically as we did with the previous algorithms.

6. DELIVERABLES

As it was mentioned above, we are preparing two papers to be submitted for publication in referred journals. These papers provide an in-depth description of all the findings of our research so far.

Furthermore, we will furnish DOE with our computer programs, (more than 2,000 lines of computer code) of all the programs developed for this project. These were written in Turbo Pascal version 7.0 for rapid development. However, for faster processing they need to be transferred (in the future) into the C++ programming language.