Ames Automated Literature Support: BREM 500 Pilot Demonstration

Prepared for the Department of Energy

Synthesis Partners, LLC

Contract Number: DE-DT0002121

July 2011

11250 Roger Bacon Drive ~ Suite 2 ~ Reston, Virginia 20190 USA
Tel 703 318 6511 ~ Fax 703 318 9553 ~ Email info@synthesispartners.com
www.synthesispartners.com

Acknowledgements

Synthesis Partners (SP) wishes to thank Mr. Steven Boyd of the Vehicle Technology Program (VTP) at the Department of Energy (DOE) for his sponsorship of this task. SP also appreciates the efforts of Dr. Ichiro Takeuchi, Department of Materials Science and Engineering, University of Maryland for his careful guidance to maximize the effectiveness of this targeted demo for materials science research. SP acknowledges Dr. Iver Anderson and Dr. Bill McCallum of Ames Lab, and the Beyond Rare Earth Magnets (BREM) research team for providing the document set for this demonstration.

Background and Objectives

SP was tasked by the DOE to demonstrate the application of proprietary intelligent agent technology to data-mining, report the results, and assess the potential impact on DOE business processes of applying such an approach. SP worked with the DOE's BREM research team, which is led by the Ames Laboratory. This work was performed under contract number DE-DT0002121.

The impetus for this demonstration is the well-recognized need for more efficient approaches to access and learn from of ever-growing quantities of scientific and technical information. Scientific researchers often lack the time to identify, access, and read pertinent published materials and subsequently leverage relevant findings.

SP's work involved demonstrating, evaluating, and providing the outputs of a pre-existing proprietary intelligent agent technology applied against a set of scientific and technology journal articles in the materials science domain.

The potential implications of this process include increasing the efficiency and productivity of DOE research activities by orders of magnitude. If successful, expedited access and analysis of large sets of published scientific and technical materials could act as a research accelerator by enabling researchers to more quickly:

- Triage large sets of articles to focus on those most relevant to their work
- Identify promising scientific and technical approaches
- Avoid approaches which have been determined to be less productive

Scope

The corpus of inputs used for this task consisted of 505 scientific articles in the material science domain provided to SP by personnel on the BREM team at DOE's Ames Laboratory. All of the articles were in the PDF format.

The work involved the application of pre-existing and proprietary intelligent agent technology to automatically clean, parse, assess, and extract data of interest. Specific concepts, functions, and metrics of interest were identified by the BREM research team. These were chosen to afford maximum information gain. For example, all instances of

Curie temperature associated with a particular material or compound (e.g., Nd6Fe13Cu or Nd6Fe13-xal1+x), were to be extracted and reported.

The technology was applied to the all of the text-based information in these documents, including the captions for tables, diagrams, pictures, and graphs. This task did not address instances of non-text-based information contained in the journal articles, including mathematical formulae, pictures, charts, and diagrams.

Deliverables

Synthesis has provided:

1. Verbal status updates on the progress and results of this task at BREM workshops, conference calls, and team meetings.

2. Verbal and written status updates to Steven Boyd at DOE headquarters.

3. The results of machine-cleaning, parsing, assessing, and extraction of the specified functions and metrics from the 505 articles, as detailed in this report.

    a. These results include numerous Excel spreadsheets, per the request of BREM personnel, as well as other back-up data in an electronic data archive. Annotated analysis of electronic data is available upon request that contains a more detailed breakdown of the problems and issues with the various PDF files covered by this report.

    b. Recommendations regarding further application of the technology.

Timeframe

Synthesis began this task in October 2010 and completed it in July 2011.

October 2010 to February 2011:

- Preparatory Work
    - Conversations with Steven Boyd and Dr. Ichiro Takeuchi of the BREM team
    - Framing the technology test parameters
- Discussion of the Pilot Demo at BREM Workshop III Atlanta
- Preparatory BREM-5 (an initial 5 article corpus) Automated Search Pilot
- Dr. Ichiro Takeuchi review
- BREM-5 Excel Output, Review & Refinement

March 2011

- Receipt of BREM 500 corpus from Ames Lab

March to June 2011:

- Document Assessment

- PDF-to-Text Cleaning
- Initial PDF Processing & Extraction
- Dr. Ichiro Takeuchi review
- June 7 BREM Team WebEx Review

July 2011
- Final BREM 500 Processing & Extraction
- Validation
- Final Report

Technology

The pre-existing and proprietary technology that SP applied to the BREM 500 Pilot Demo is based on societies of intelligent agents. The technology is novel in the world of Natural Language Processing, as intelligent agents operate in concurrent, parallel, distributed societies to analyze the semantics in texts. The technology is based on ISO Standard Common Logic 24707, and employs the Conceptual Graphs knowledge representation pioneered by Dr. John F. Sowa.

The intelligent agent logical architecture has been shown to address quantitative and qualitative semantic data in a language- and domain-independent manner. This technology demonstrated the ability to both clean and recognize the correct stoichiometric chemical compound patterns, formulae, and equations in the course of this task.

Results

The results of this task are discussed below, and contained in two appendices to this report:

- Appendix A (electronic file): Results of processing the text of the 505 articles. Instances of Curie temperature are provided with article source information.

- Appendix B (electronic file): Results of processing the captions and in-line text references to tables and figures (including, charts and diagrams) contained in the 505 articles, referenced with article source information.

From the start of this task, SP appreciated that PDF files were corrupted at the text layer, making effective machine-extraction impossible. Issues identified included corrupted text, columnar text intermixed, words concatenated into long unintelligible sequences, and unrecoverable formula and stoichiometric data. Our assessment of the machine readability of the 505 documents is shown in Chart 1.

Chart 1: Corpus Assessment

Directory Name	Description	Percent of Total
AMESPDF	Complete set (505)	100.0 %
AMES_FCTD	Corrupted but may be correctable	~45%
AMESTXT_BAD	Very corrupt, need extensive cleaning	~26%
AMESTXT_UNCLEAN	Unclean, but easily correctable	~20%
AMESTXT_USABLE	Perfectly clean and usable files	~7 %
AMES_OCR	Only correctable with OCR	~2%
AMESPDF_CORRUPT	Digitally corrupt, unusable	~<1 %

Note: Data does not add to 100% due to rounding.
Source: Synthesis Partners (2011)

Based on these findings, we began with the approximately 7% "perfectly clean and usable files" to run initial tests. Due to the small number of clean files, we were able to hand-validate the machine knowledge extraction results relative to the content in the articles. We confirmed through this process that there were in fact very few instances of the required Curie temperatures present in the text.

To complete this task, we processed the full document set in two stages, as described in the tables below. Table 1 shows the results of our follow-on cleaning efforts. Table 2 shows results of the analytical processing and knowledge extraction from the resulting machine readable PDFs.

Table 1: Machine Cleaning Summary

TEST RUN #	TOOLS USED (No Materials Science Ontologies Used)	NUMBER OF MACHINE READABLE FILES OBTAINED (Files with sufficient machine readable text extracted from the PDFs)	MACHINE READABILITY SCORE (Manual Validation)	Main Issues
1	Adobe PDF-X[1]; SP technology	35 documents out of 505	7%	Few PDFs are machine readable.
2	PDFBOX[2] + SP technology	171 documents out of 505	34%	Text recovered using glyph-list (hardcoded mapping) and error analysis based on PDF source creator behind the journal, regression tests with custom-compiled version of open-source PDFBOX tool.
3	PDF-AGENTS™3 +PDFBOX+SP technology	496 out of 505	98%	Added proprietary SP software agents to handle the "logic" of the layout and the logic of the errors from Run#2.

Source: Synthesis Partners (2011).

Achievement of the 98% document machine readability benchmark for the 500 PDFs appears to be a significant accomplishment.[4] This was necessary to begin parsing and extracting knowledge relevant to BREM, including very-difficult-to-obtain stoichiometry-based information.

The results of the automated knowledge extraction application follow. The total runtime for the knowledge extraction effort on a quad-core Mac PRO laptop was 20 hours for the 505 articles; of which 496 were successfully processed. This equates to a rate of approximately 25 articles per hour.

[1] Adobe-PDF-X is Adobe® Acrobat® X software that permits saving digital document as PDF files and conversion back to text, in conformity with ISO and a variety of industry-specific standards.
[2] The Apache PDFBox™ library is an open source Java tool for working with PDF documents. It allows creation of new PDF documents, manipulation of existing documents and the ability to extract content from documents.
[3] PDF-Agents™ are custom software agents designed by the Synthesis team to convert PDF documents to machine readable text.
[4] Research has not revealed commercial tools which clean PDFs to an equivalent degree, and thus enable the automated extraction of stoichiometry-based information.

Table 2: Summary Results on Knowledge Extraction from Machine Readable PDFs

TEST RUN #	TOOLS USED (No Materials Science Ontologies Used)	MACHINE EXTRACTION OF KNOWLEDGE (From machine readable files, extracted Curie temperature and Captions of Figures, Tables, Charts and Diagrams)	KNOWLEDGE EXTRACTION SCORE (Non-SME Manual Validation)	Main Issues
1	Adobe PDF-X; SP technology	Zero knowledge extracted	0%	Text corrupted throughout.
2	PDFBOX + SP technology	Limited useful knowledge extracted	~15%	Text recovered using glyph-list (hardcoded mapping) and error analysis based on PDF source creator behind the journal, regression tests with custom-compiled version of open-source PDFBOX tool.
3	PDF-AGENTS™ +PDFBOX + SP technology	Significant Curie temperature and captions discovered and extracted (included in the Appendices to this report)	~70%	Added custom agents to handle the "logic" of the layout and the logic of the errors from Run#2, and agents that performed extensive semantic error detection and correction (i.e. spelling recovery and de-multiplexing of mixed or concatenated words or lines).

Source: Synthesis Partners (2011)

The machine identified 363 references to Curie temperatures in the document set.

Our non-subject matter expert (SME) manual review of the Curie temperatures extracted in the third run provided the following results for an 11 file sample:

- Five files showed accurate extraction results
- Four files showed examples of where the lack of an ontology prevented the machine from interpreting subtle clues to relevant Curie temperatures, including concepts like "large increase of," "decrease in," and ferromagnetic or other types Curie temperatures that may or may not be relevant to BREM.
    - These errors are fully expected. It is a straightforward matter to integrate a concise materials science ontology with the machine's knowledge base.
- Two files showed examples of where an identified software bug caused the machine to miss Curie references.
    - These software bugs are expected, and once identified, can be easily addressed in future iterations.

Our manual review of a selection of articles regarding the machine extraction of captions and related references to tables, figures, charts and diagrams discovered in the full-text provided the following results:

- Caption recognition is greater than 95%.

- There is a need for basic lexical resources (e.g., dictionaries) and light ontologies to condense the results and eliminate unnecessary references.

- The system identified many references:
  - 827 lines/rows of information in tables, and
  - 5,094 lines/rows of information in figures (including charts and diagrams) were identified.

- An improved interface with Excel is required to ensure the stoichiometric and mathematical information extracted is properly transcribed.

- Automated extraction in the materials science domain will need to process information in tables, figures, charts, and diagrams (addressed in the Recommendation section):
  - Articles concerning structural changes in materials science properties (e.g., phase changes) present as much as 80% of the key information in figures or tables, with very little found in the text.
  - Articles concerning experimental processes, methods, and summary properties present approximately 80% of the relevant information in the text and tables, and only 20% in figures and diagrams.

Observations

- Effective text cleaning is required for an information extraction system to successfully produce structured knowledge from unstructured data residing in PDFs.

- This technology approach successfully cleaned 98% of the PDFs in the BREM 500, and extracted the key figures of merit provided to within ~70% accuracy.
  - With ten-known bug fixes, we estimate results will improve from 70% to ~85%.
  - Adding a light materials science, physics and/or chemistry ontology will also increase the quality of the results to a significant degree. Such ontologies, which are not difficult to craft with SME support, would define, for example, why Curie temperature concepts are not all the same, why under certain circumstances linguistic variables (e.g. “x increases”, “y decreases”, “greater than”, “less than” etc…) are significant, and under what circumstances a particular variables in a table, figure, chart, and diagram should be highlighted.

- Other approaches to the issue of machine cleaning and extraction of knowledge from PDFs do not appear to be effective. Extensive research revealed no commercial technologies which successfully extract scientific knowledge from PDF documents. Other tools that claim to perform chemical compound pattern recognition were found to be very brittle, in that these systems progressively failed as text inputs become degraded. For example, chemical patterns such as (Mn1-xPd1x) were not recognized by any of the tools we tested, including:
    - University of Cambridge
        - OSCAR
    - ChemAxon
        - Chemicalize.org
    - InfoChem
        - Annotator and ICN2S
    - ChemMantis
        - SureChem (NER) with ACD/Name
        - ACD/NTS / Batch (N2S)
    - MPirics
        - Chemical Content Recognition

  Other tools were also found lacking. These were not directly tested, but were assessed either directly or indirectly (i.e., through their search portals) included:
    - CambridgeSoft
        - Name=Struct
    - OpenEye
        - Lexichem TK
    - Accelrys
        - ChemMining
    - TEMIS/MDL
        - Chemical Entity Relationships Skill Cartridge

- One key issue identified is that extraction from PDFs in the materials science and chemistry domains is several orders of magnitude more difficult than other disciplines because of the density of chemical and mathematical symbols, as well as the methods of representing critical information in the text along with information in figures, charts, diagrams, and tables.

- In most of the 505 documents, information in tables was highly relevant but was not extracted, as special intelligent algorithms for recognizing a tabular structure are needed, which were beyond the scope of this demonstration.

- Finally, we find that very little work, if any, addresses the potentially high value domain of semantic processing or information extraction in materials science. Basic tasks required for ontology construction in support of information extraction do not appear to be addressed.

Recommendations

There are a number of potential paths forward beyond this highly targeted effort. Some of these have been discussed with Dr. Ichiro Takeuchi and others. Based on the potential identified in this task, SP provides the following recommendations for consideration:

1. DOE initiate a large scale demonstration of this and other semantic text extraction technologies to assess:
    a. Scalability
    b. Speed, accuracy, and precision
    c. Domain and language independence
    d. Capability to extract information from text, tables, figures, charts and diagrams

2. DOE institute a transparent, objective process (potentially at a DOE Lab facility) to perform objective quantified assessments of semantic text-extraction technologies. This could involve the:
    a. Development of a technical data "sand-box" in which semantic full-text extraction technologies may be neutrally tested
    b. Establishment of criteria to ensure technologies are objectively validated
    c. Periodic recommendations generated for DOE leadership

3. Assess this technology for potential integration with pre-existing data-mining activities at Ames Lab and other DOE facilities.

4. Forecast the effects of implementing this capability could have on R&D efficiency and efficacy.
    a. SP believes that if this technology were to increase R&D efficiency and decrease time-to-discovery, the resulting accelerated development and fielding of novel and innovative technologies would produce large benefits to the US scientific and research community, and the country as a whole.