Optimizing LLE Information Operations through Natural Language Processing

Brandon E. Avila
Allendale Columbia School
Advisor: Rick Kidder

Summer 2011

Abstract:

Research in natural language processing (NLP) was conducted to determine the feasibility of its use in optimizing information processing for large scale laser facility operations at the Laboratory for Laser Energetics (LLE). NLP was used to extract and link the information from several sources used to house operational knowledge. A vocabulary of common terms found in LLE's documentation, including diagnostic system documents and LLE staff records, was created and used for developing increasingly complex data relationships among the data in the decentralized knowledge base. Extensible Markup Language (XML) data serialization was chosen to store portable information. Using XML, expansion by NLP provided a method for faster knowledge-base construction than any previously implemented technique at LLE.

Introduction:

Benefits of NLP

In recent years, an effort has been made to centralize the information within the LLE systems, creating large repositories of relationally linked information gathered from the databases and tables within the inherently disorganized server system. Methods of creation of these knowledge bases have required manual work in adding object data and explicitly defined descriptions of relationships rather than automatic inference through table and document reading. Adding object data to these repositories has relied upon time-consuming manual information look-up, searching through internal and external documentation and phonebooks to complete entries on associated scientists and engineers as well as laser and lab components.

Furthermore, a manual centralization can never be entirely exhaustive. Data exist in documentation in hundreds of directories throughout many servers that would be impractical to manually read, discern, and enter into a central location, such as the ontology created in a separate project [1]. In addition, some information is hard to gather, as it is not included in the easily accessible databases. Multiple levels of permissions are required to find any kind of information when reading through newly added documentation, and it is often difficult to keep track of which directories have been read through and entered into the centralized system. Moreover, directories are dynamic, and their contents may be modified or removed, making a record of completed directories impossible.

To assist in the process of recording and analyzing this information, NLP is designed to read, interpret, and understand these data on a more abstract level than a simple database parse, simulating a human reader in discerning the meaning of English statements within the LLE documentation. NLP algorithms are designed to be a system of cumulative learning as demonstrated in Figure 1, taking information from previously scanned documents (a), searching for more in-depth information (b), storing the data in a dynamic vocabulary of terms (c), and using newly learned information to repeat the cycle with a greater knowledge of the meaning of words.

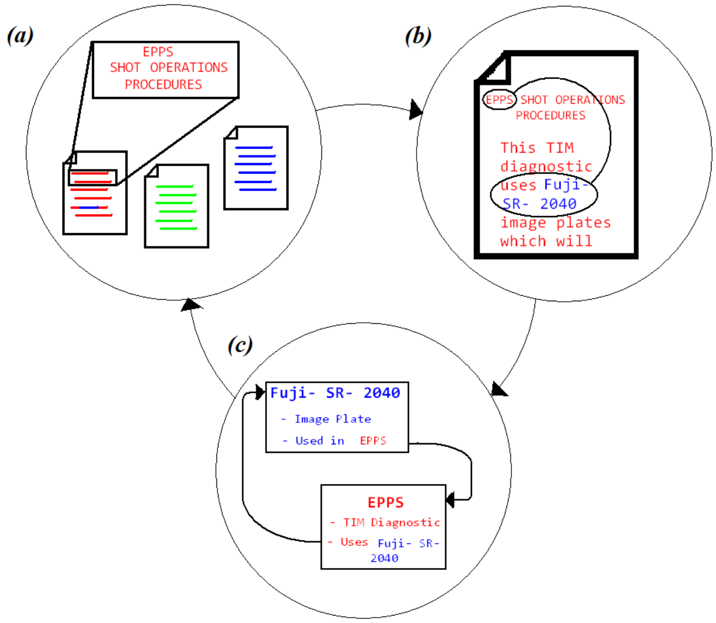


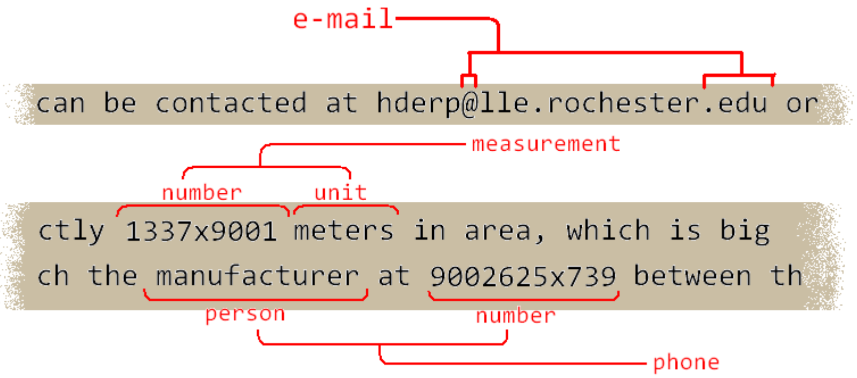
Figure 1: The NLP model. Relevant information is (a) extracted based on search parameters, (b) classified, and (c) inserted into the vocabulary's web of relationships. This information can be (a) reused to extend future searches.

NLP eliminates several major problems involved in the process of manually scanning documentation. First, scanning a directory for documents containing relevant information only takes a few seconds, while a human reading a single operations manual or procedural checklist may take several minutes or longer to find no information at all. Automatic parsing decreases time spent searching by orders of magnitude. Second, the efficiency of the program depends on how much it knows. No time scanning documents is wasted, because even in scanning the same document twice, new information can be inferred based on what was learned during the last parse. It takes a significant number of scans to cease finding new data to gather. Finally, a well-constructed program does not have to know what specific data are being sought, but can register any and all new information in its vocabulary. While a human may search for the phone number of an employee on a page containing his information, a NLP-based algorithm may be able to infer his room number, job title, and team without any additional effort. A perfectly constructed algorithm is able to search any document, and learn all knowledge directly connected to any information it already knows.

Foundations of NLP Algorithms

An algorithm for parsing and interpreting documents through NLP is based on the recognition, classification, and definition of certain words, as well as the inference of their relationships with other terms within a vocabulary [2]. At the base of an algorithm using NLP is word recognition, shown in Figure 2. To recognize a word by type or shape, the particle is compared to a set of regular expressions that provides a format for any meaningful categorization (e.g. name, number, email). Based on character matches between the examined word and the templates, a word may be categorized under one of these types and saved for further context processing. Once a phrase or group of words is complete, a higher-level matching algorithm may search the context, recognizing the structure of the phrase in terms of its previously defined categories, and infer more perceptive information about the meaning of the word.

When words are recognized in context, and can be related to information already known, they are added to a web of terms, or relational vocabulary, that is capable of explaining these terms in relation to the others. It is the analog of an English dictionary. Terms can be accessed by the NLP algorithms or third party programs, but they mean nothing alone, and the relationships by which they are defined must be understood in order to find any intelligible meaning within them. For this reason, a search through a relational vocabulary is able to provide a more comprehensive explanation of an object by providing partial definitions of related terms that can be used to recursively search. Figure 3 shows how upon searching for a term, one may find its


can be contacted at hderp@lle.rochester.edu or

ctly 1337x9001 meters in area, which is big
ch the manufacturer at 9002625x739 between th

Figure 2: Demonstration of NLP parsing techniques. Words are tagged by examining their character composition and comparing them to regular expressions. These words are given meaning based on the context in which they are found.

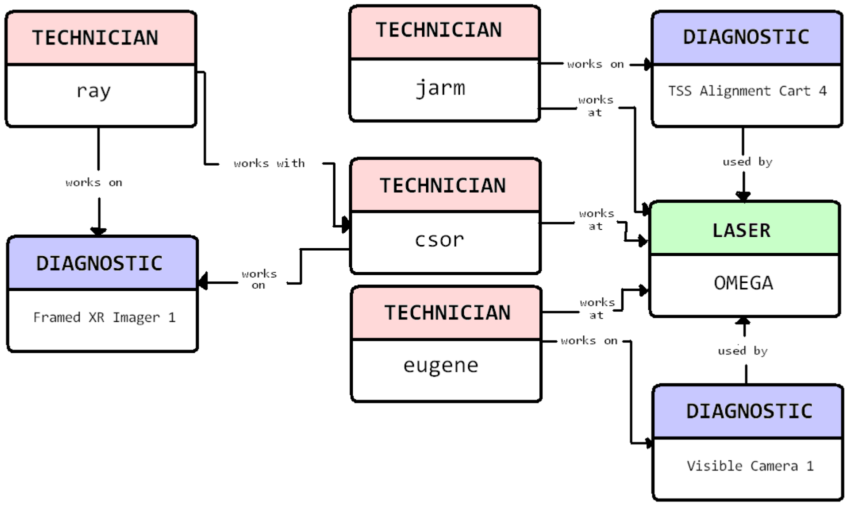


Figure 3: An example relational vocabulary. Known terms share relationships that link the vocabulary into a network. When a program queries the knowledge base, these relationships can be used to provide a large amount of relevant information.

type, as well as the objects to which it is related. A query whose search term includes any specific technician will provide information on his project team as well as the diagnostics and lasers with which they work.

Research and Development:

Creating and Storing the Vocabulary

To begin seeding the vocabulary to be used at the LLE, many databases and tables were first parsed manually to retrieve some information to which newly added terms could be related. Through a combination of manual definition of data-types such as people, diagnostics, and laboratories, and a simplistic automatic scan of these tables, approximately 200 terms were linked with about 900 relationships among them. This initial seed was used to further accumulate terms by parsing through a large directory of presentations and reports containing information about certain diagnostics, projects, and the scientists and engineers maintaining them. The vocabulary was continually being manually seeded from other sources while information was extracted automatically to accelerate the data acquisition. A total of five major parsing algorithms were completed, each consisting of several phases of extraction of different sets of data from a major directory. Data and relationships extracted included, but were not limited to, Project Assistants, Principal Investigators, System Engineers, experts, and specialists, and were extracted from procedural checklists, Operational Readiness Reviews, and project reports, among other sources of information.

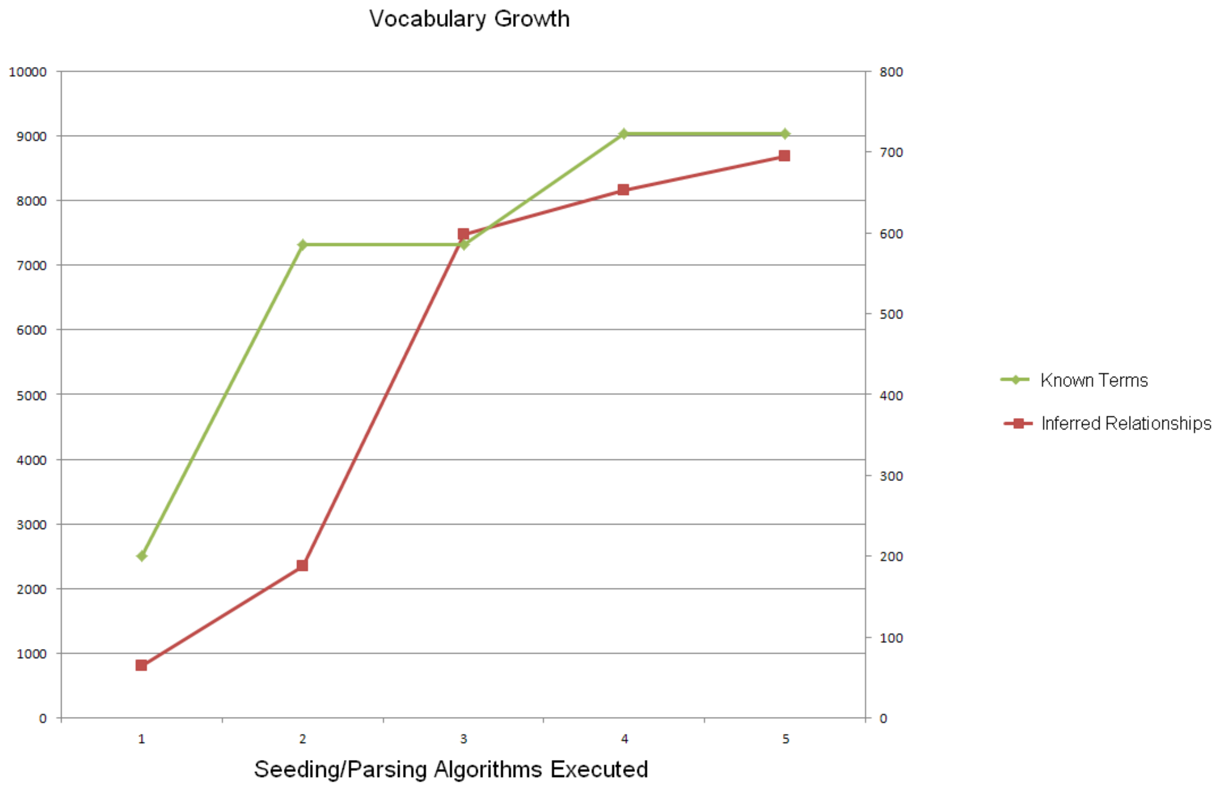


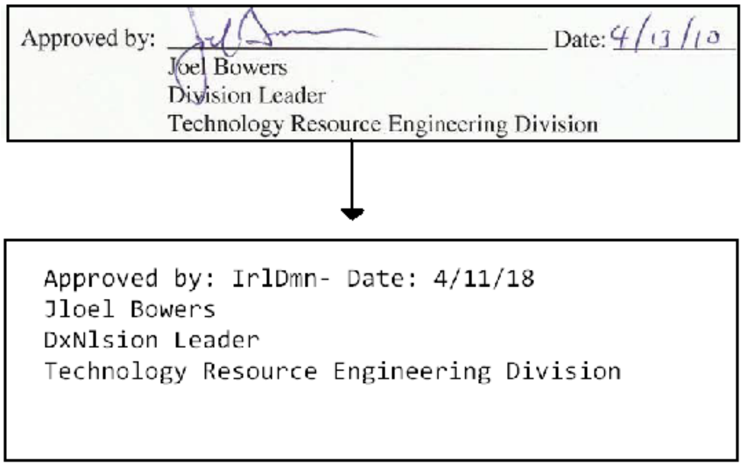
Figure 4: Growth of the established vocabulary. After initial seeding, the vocabulary contained approximately 200 terms and is seen to have grown to over 700 by the end of the project. This was achieved through both continuous seeding and relationship inference. The right and left axes correspond to the number of Known Terms and Inferred Relationships respectively.

As seen in Figure 4, the algorithms developed eventually reached a point where each algorithm run on a directory brought fewer inferred relationships. While the graph demonstrates only a portion of all documentation available for parsing, the concept applies to larger sets of documentation. Without more sophisticated algorithms or a larger pool of information available to organize, the amount of knowledge not yet gathered continues to decrease until every logical relationship has been inferred.

To make the vocabulary available for reuse by the NLP algorithms and other programs dependent on LLE terms and relationships, it was stored in portable, accessible files. XML was chosen to serialize the information in an easy-to-read format that could quickly be searched and altered through simple programs and scripts. While XML was efficiently updated and read by the NLP algorithms, it was also effective in serving as the knowledge base of a newly designed ontology, fabricated in a separate project.

<u>Problems Faced in Reading Documentation</u>

Many LLE documents include signatures, forms, or other hand-written information. When these documents are signed or filled out, they are scanned and uploaded to the servers in Portable Document Format (PDF) files. These can be easier to use than simple image files, and most PDF readers contain a built-in feature that attempts to convert these hand-written lines into machine-readable text. These programs, however, are certainly not comprehensive, and often fail to accurately read any sort of information: errors like the one in Figure 5 occur frequently. It is thus quite difficult to extract anything valuable from the handwritten forms.

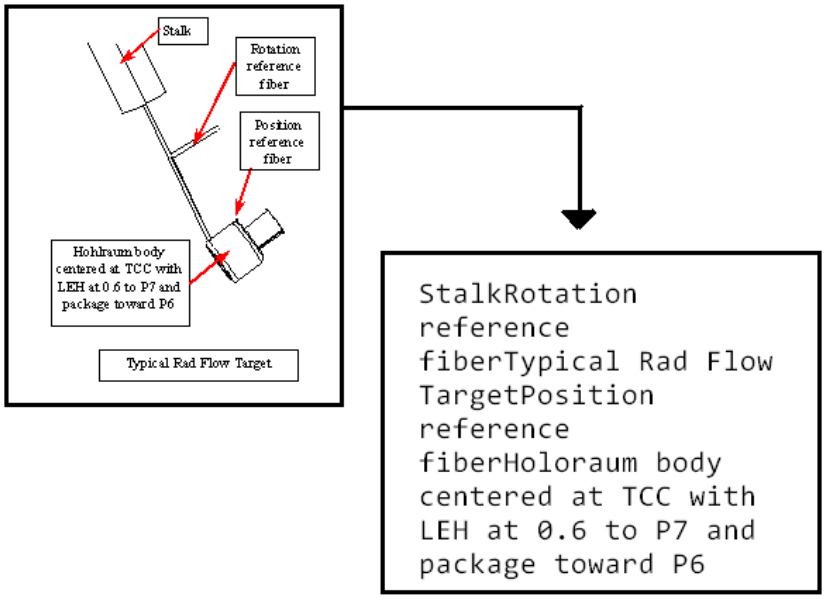


Approved by: Date: 4/13/10
Joel Bowers
Division Leader
Technology Resource Engineering Division

```
Approved by: IrlDmn- Date: 4/11/18
Jloel Bowers
DxNlsion Leader
Technology Resource Engineering Division
```

Figure 5: Inaccurate document reading. Current technologies are incapable of accurately reading hand-written or scanned documents.

In addition to the scanned documents, the majority of other documentation at LLE is also stored in PDF files. These files are effective in providing users with a comprehensive view of diagnostics and procedures at the lab, as they elegantly integrate text, graphics, and checklists. Often it is simple to read the text within these files through the use of external programs, but sometimes this information is jumbled or concatenated, making it difficult to extract coherent sentences, or even words, from the text. The most common cause of the problem proved to be



```
StalkRotation
reference
fiberTypical Rad Flow
TargetPosition
reference
fiberHoloraum body
centered at TCC with
LEH at 0.6 to P7 and
package toward P6
```

Figure 6: PDF reading errors. When trying to parse a figure with discernable text, the result is often jumbled and concatenated captions.

graphics labeled with spatially formatted blocks of text. As seen in Figure 6, words were present, but incoherent in the formatting retrieved from reading the documents.

Future Work

Work on this project was primarily in determining the feasibility of NLP algorithms as the primary method of knowledge base construction and information compilation. For this reason, many of the algorithms written were simplistic, and meant to test only for efficiency in scanning, reading, and extracting simple information from a relatively small collection of documents. A large-scale version of the project would involve creating programs capable of performing more complex operations on larger collections of information. Extending the project should be aided by the vocabulary now installed within the LLE systems.

Further work involves the creation or use of more effective document-reading programs. Much information was lost in inaccurate readings of text or hand-written information. The project would benefit most from a program granting the ability to accurately read these obfuscated documents.

Acknowledgments:

I thank Mr. Rick Kidder and Dr. Stephen Craxton for granting me the opportunity to spend my summer at the Laboratory for Laser Energetics. I further extend my gratitude to Mr. Kidder for his advice and support in designing and carrying out this project. Additionally, I thank Rob Cooper and Dustin Axman for their previous work on related projects, as well as their assistance and support in integrating this project into the established system at the LLE.

References:

1. R. Cooper, “Development of an Ontology for the OMEGA EP Laser System,” 2010 Summer High School Research Program at the University of Rochester's Laboratory for Laser Energetics
2. S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python* (O’Reilly Media, 2009).