U.S. DEPARTMENT OF ENERGY

PNNL-21871

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Text Analysis Capablities

EB Bell
SJ Rose
S Choudhury

October 2012

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

This document was printed on recycled paper.
(9/2003)

PNNL-21871

Text Analysis Capability

EB Bell
SJ Rose
S Choudhury

October 2012

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Text Analysis Capability

Information Extraction

Broadly, this section explores functionality critical to the task of generating structured information from unstructured text. This functionality centers around the capability of processing human language in a machine readable form and producing a representation that allows for computation and reasoning against the text.

Named Entity Extraction

The task of named entity extraction consists of combing through text to find atomic mentions of typed elements in text. Common types of named entity extraction involve the extraction of people, organizations, geopolitical entities, time expressions, monetary values and percentages. In general, most of the research in this area involves taking a sentence like "Abdel Wadoud did not shoot Kamel Bourgass on Wednesday" and recognizing that both Abdel Wadoud and Kamel Bourgass are names and that the expression "on Wednesday" refers to a specific time. It is well understood that hand-generated grammar-based solutions for named entity extraction produce higher precision at the tradeoff of lower recall and a significant time investment. On the other hand, the typical statistical based approaches in this area require a significant volume of annotated training data.

PNNL has active ongoing research in this area. We have looked at defining new classification strategies for recognizing and extracting people, locations, and organizations. We've also explored the combination of COTS, GOTS, and open source solutions to produce a wider span of types with greater accuracy. This capability is largely an early step in a much larger pipeline for knowledge representation and information extraction. We have also conducted modest annotation efforts in new domains for providing statistical extraction systems with the necessary training and test datasets.

Concept/Domain Extraction

PNNL has the capability to take raw, unstructured text, and use the content of that text to extract ontological concepts and academic domains. This capability is based on GOTS software, with some adaptations we've made for computation efficiency and for combinations of concepts and domains across large document collections. PNNL has developed the capability to easily extend the base ontology to new domains and into new concept spaces.

Semantic Role Labeling

Semantic Role Labeling is the capability to take a series of arguments associated with a verb in a sentence and assign those arguments into fulfillment of specific roles (sometimes referred to as slots). This capability often requires a syntactic parse of text and then recognition of semantic arguments. Again, this capability is often one step in a larger pipeline employed by PNNL to solve information extraction and knowledge representation problems.

Continuing with the example above, semantic role labeling would form a representation:
Event: Shoot

Agent: Abdel Wadoud
Patient Kamel Bourgass

And map that representation to a more meaningful relationship as:
Attack: Shoot
Attacker: Abdel Wadoud
Attacked: Kamel Bourgass

Note: The final representation would also represent the negation of the verb shoot and the time of the event as Wednesday.

Entity Disambiguation

Entity disambiguation is a critical step for building a knowledge base of named entities. It consists of resolving the mention of entities in a free-form document (blog, newswire or random web text) to an entity in the knowledge base and updating the profile of an entity in the knowledge base by accumulating newly arriving information over time. Clustering named entities such as people, organization, location that are extracted from documents using natural language processing techniques is a major step in the entity-disambiguation process. This clustering is a challenging task because of the large number of entities (spanning into millions or billions), extremely large number of features for these entities and high computational complexity of the algorithm itself.

The research at PNNL is distinguished by its graph based approach and focus on multithreading-based parallel computing. Graphs are attractive for representing very large sparse feature vectors and capturing the relationships between documents, named entities, features and associated contexts. However, computation on large graphs introduces unique computational challenges. We are leveraging on large-scale graph analytics capabilities developed through the CASS-MT project (http://cass-mt.pnnl.gov) to address these challenges. Specifically, we are exploring massively multi-threaded architectures such as the Cray XMT and multi-core CPU based systems to develop scalable solutions for the entity-disambiguation problem and the comparison of solutions on these platforms vs. standard commodity hardware clusters.

Reasoning

This section reflects a number of capabilities that require a significant degree of statistical reasoning. The techniques described herein are algorithmic in nature and produce results with a degree of mathematical certainty.

Keyword Extraction (RAKE) / Theme Extraction and Exploration (CAST)

PNNL has developed Rapid Automatic Keyword Extraction (RAKE), a feature-extraction method that automatically extracts a document's essential single- and multi-word keywords. By improving the identification and selection of features from text collections, the utility of existing visual analytics systems is significantly improved.

RAKE automatically extracts single- and multi-word keywords from individual documents and applies several metrics to select a set of high-value features that characterize the content of documents within

the collection. RAKE can be applied to individual documents and large text collections and is easily configurable through its input parameters to tune its performance for specific domains, document types, and languages other than English. In practice, RAKE typically processes over 500 documents per second or 30,000 documents in one minute.

RAKE's methods are based on several observed conventions of manually assigned keywords. Those keywords assigned by professional indexers and librarians or provided by authors often comprise multiple words, occur frequently within their respective document, and typically do not contain stop words or punctuation. RAKE applies these conventions to parse out and rank candidate keywords as sequences of contiguous words.

RAKE has been integrated within IN-SPIRE to provide users a richer set of keyword features that provide information cues to aid users in exploration while providing an understanding of macro patterns and key details. These keyword features have the characteristics of accurately capturing relevant knowledge patterns for analytic work and have a strong cognitive value as they differentiate between parts of a text corpus and describe the salient characteristics of each part for the user. RAKE's simplicity and efficiency make it potentially useful in many applications where keywords can be leveraged.

The keywords that RAKE automatically extracts are keywords that a person would also identify as representing the document's essential content. Evaluation of RAKE in comparison to several comparable keyword-extraction methods on a benchmark dataset of short technical abstracts shows that RAKE achieves higher precision and recall in extracting keywords that match those assigned by professional curators for the same documents.

Rapid Automatic Keyword Extraction (RAKE) and Computation and Analysis of Significant Themes (CAST) are computational methods to support users' discovery and exploration within unstructured text collections. RAKE extracts single- and multi-word keywords from individual documents enabling fine-grained insight from documents and improved feature selection for large repositories. CAST computes themes as clusters of related keywords, providing a higher level grouping of documents to aid users in understanding the key components within collections. RAKE and CAST are compatible with the Apache Lucene/Solr search APIs and can be integrated into many search engines and text analysis pipelines. The RAKE and CAST APIs are currently integrated within IN-SPIRE, SRS.

Conceptual Clustering / Topic Modeling (LSI, LDA)

Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) are analytic techniques used to explore relationships among pieces of data. LSI is based on the mathematical technique of Singular Value Decomposition and rests upon the assumption that words with the same or similar meanings are frequently used in the same or similar contexts. The technique is used to establish associations between previously unseen or unknown terms based on their contextual occurrences.

Applied

This section describes a small number of techniques that reflect an application of natural language processing capabilities, but in and of themselves are normally not the solution to the presented analytic problem. These techniques are exploited in use in the final section, application areas.

Relationship detection and characterization

This capability is focused on statistical and rule-based techniques for looking at communications between two individuals. We explore techniques for characterizing communicants relationships based on unidirectional and bi-directional communications in static datasets and streaming conversations.

We focus on 3 core areas of relationship characterization:

1. Familiarity- to what degree do the two communicants know one another? Are they familiar with one another or complete strangers?
2. Dominance- what is the hierarchy between the two communicants? Is one a superior to the other? Are they peers?
3. Business/personal- what is the degree of personal friendliness between the two communicants? Are their discussions characterized by business matters or personal affairs?

We also explore affect (sentiment) in the context of communicants. To this end, we can describe the overall sentiment from each of the communicants toward the other. Sentiment analysis is covered in greater depth directly below.

Sentiment Analysis

PNNL has developed a number of tools and algorithms that operate on subjective information in text. These capabilities come from the domain of opinion mining (sentiment analysis) and focus on understanding the opinion of an individual or a document with respect to a given topic in the overall context of a conversation or paragraph. We explore sentiments based on affect or emotion, judgment, and evaluation. This is an area of active research, with significant application to the area of social media. From a marketing perspective, companies want to understand public sentiment toward their products.

Application Areas

This final section provides a wide-span summary of current application areas of the text analysis capability described thus far. The listed application areas are non-exhaustive, but are meant to be representative of a number of fairly common real world challenges.

Social Media (Cultural analysis, emergency management, marketing)

At PNNL we study online communities or groups who are linked through abstract interests. We have developed a number of algorithms for using structural and content information from social media to build an expectation footprint for groups, to place members into groups, and to understand how culturally defined behaviors influence the footprints of those groups.

Social networks can provide early indicators of security and health threats, public sentiment, and response – but the massive volume of streaming data complicates analysis. Algorithms for detecting emerging themes in social media are in use at emergency management organizations.

Lastly, we study social media data for understanding mentions of national laboratories in social media and traditional news. We're interested in understanding public perception and reaction to national laboratories and their accomplishments.

Knowledge Base Population

At PNNL we've developed a number of systems designed to automatically populate a knowledge base from both structured and unstructured text given an ontology. The system is designed as a modular end-to-end system that detects events, extracts information, maps it to an ontology, and disambiguates entities in the knowledge base. Our approach is created in such a way as to be easily adapted to new ontologies and domains. This allows for reasoning about events and entities involved in the events in a number of vastly different domain areas. This enables data mining- the exploration of previously unknown patterns over large groups of the data.

Dataset Visualization and Reasoning

In the research in visualization at PNNL, we make big, complex data useful through great visual design, compelling interaction, support for sound analytic methods, and solid engineering. We invent visual metaphors, create analysis algorithms, and deliver software products that put new analytic capabilities into our users' hands. As far as text is concerned, the world communicates in natural language, but manual analysis of large text collections to discover patterns and relationships is impossible. We have a software product called IN-SPIRE that is deployed across government and industry to organize text content by topic automatically.

Semantic Web

Cognitive, social, and behavioral science applied in the design of information technology systems enables efficient collaboration while supporting the natural social behaviors that must exist among groups of individuals and intelligent systems. Our Social and Collaborative Computing research leverages social computing concepts, such as virtual presence, shared annotation mechanisms, and trust/reputation protocols, to customize effective solutions.