U.S. DEPARTMENT OF ENERGY

PNNL-18370

Prepared for the U.S. Department of Energy

under Contract DE-AC05-76RL01830

Annotated Bibliography for the DEWPOINT Project

Christopher Oehmen

April 2009

Pacific Northwest NATIONAL LABORATORY

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-ACO5-76RL01830

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@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.gov
online ordering: http://www.ntis.gov/help/ordermethods.aspx#online

PNNL-18370

Annotated Bibliography for the DEWPOINT Project

Christopher Oehmen

April 2009

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

Pacific Northwest National Laboratory
Richland, Washington 99352

Contents

Annotated Bibliography

This bibliography covers aspects of the Detection and Early Warning of Proliferation from Online INdicators of Threat (DEWPOINT) project including 1) data management and querying, 2) baseline and advanced methods for classifying free text, and 3) algorithms to achieve the ultimate goal of inferring intent from free text sources. Metrics for assessing the quality and correctness of classification are addressed in the second group. Data management and querying include methods for efficiently storing, indexing, searching, and organizing the data we expect to operate on within the DEWPOINT project. The overarching principles of classification center on information extraction, which is loosely defined as filling in a predefined template with entries derived from each record in a text corpus, and information retrieval, which is the task of locating relevant documents within the same text corpus. These are implemented using a variety of approaches, including keyword identification and searching, regular expression matching, and more sophisticated methods based on a variety of mathematical approaches including Bayesian statistics, clustering, Markov models, and machine learning (including Support Vector Machines). The references in this bibliography are grouped by expected relevance to DEWPOINT, categorically labeled as 1) CRITICAL - references which directly form the basis of expected DEWPOINT project work, 2) USEFUL - references that have algorithms or techniques that could be used in DEWPOINT should any of the primary sources prove insufficient or problematic, and 3) REFERENCE - material that is important prior work or concomitant work, but which is not expected to be directly incorporated into DEWPOINT.

Critical Literature

Culotta A and A McCallum. 2004. "Confidence Estimation for Information Extraction." In *Companion Volume: Short Papers, Student Research Workshop, Demonstrations, Tutorials Abstracts* of the *Human Language Technology Conference and North American Chapter of Association for Computational Linguistics*, pp.109-112. Boston, Massachusetts. Performance of standard confidence estimation techniques for field confidence (probability that a single field in a record is labeled correctly) and record confidence (probability that an entire record is labeled correctly) are benchmarked in this paper against one another and random and worst-case baselines. Four additional methods are tested that are specific to either field prediction or record prediction. All of these confidence estimation techniques were applied to predictions made using a single method of information extraction derived from linear-chain random field model (which uses a Markov model algorithm to assign labels to fields and aggregately to entire records). CFB and Maximum entropy consistently outperformed the other methods for both field confidence and record confidence using Pearson and precision as metrics. The primary interest of this paper for DEWPOINT is that it provides a set of algorithms for assessing confidence in information extraction predictions. Since the proposed method also can provide statistical confidence estimates in predictions, the algorithms described in this paper can serve as a baseline of comparison. The main drawback is that the confidence estimates were only applied to a single information extraction method so their performance on other methods is unclear.

Dean J and S Ghemewat. 2004. "MapReduce: Simplified Data Processing on Large Clusters." In *6th Symposium on Operating System Design and Implementation (OSDI)*, pp. 137-150. San Francisco, California. This paper describes a programming paradigm that allows operations on large datasets (terabytes or larger) to be efficiently scheduled using a very powerful library developed for Google searches. This approach has been used for many different applications including many statistical

operations on Web site content. This is the most straightforward way to implement a variety of approaches for this study.

Fisher D, S Soderland, J McCarthy, F Feng, and W Lehnert. 1995. "Description of the UMASS System as used for MUC-6." In *Conference on Message Understanding*, pp 127-140. Columbia, Maryland. This paper describes the UMASS system used at the Message Understanding 6 competition. It details algorithms for extracting meaning from free text sources that combines (in addition to other elements) 1) RESOLVE, a system to recognize multiple references to a single entity, 2) CRYSTAL, a concept node creation algorithm, and 3) WRAP-UP, a system for resolving how relational links apply to entities. All three of these system elements use machine learning to automate the tasks associated with extracting meaning from text. Subject matter expert input is needed, but only during construction of the underlying framework of these tools before their actual use. During the application of these tools to real text, machine learning steers decision making. This system is adaptable to specific domain areas. This sort of system will be a good test algorithm for DEWPOINT.

Freitag D. 1998. "Multistrategy Learning for Information Extraction." In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, ed. JW Shavlik, pp. 161-169. Madison, Wisconsin. This paper makes the argument that information extraction is optimized when a combination of approaches is used. Specifically, the authors demonstrate that combining brute-force word frequency methods with their SRV method, which adds features about tokens and their physical relations to other tokens. Interestingly, the authors also describe a linear regression-based method for assessing confidence in the predictions made using multiple methods. This is an interesting possibility for DEWPOINT as many more information extraction algorithms exist now than were available at the time of this publication, but they might be combined in a similar way, and assessed using a similar set of metrics.

Frigui H and O Nasraoui. 2002. "Simultaneous Categorization of Text Documents and Identification of Cluster-dependent Keywords." In *Proceedings of the 2002 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE '02)*, pp. 1108-1113. Honolulu, Hawaii. An algorithm to automatically cluster documents and derive keywords from a text set are described in this paper. The document clustering method is based on standard clustering algorithms (K-Means clustering). The significance of this approach for DEWPOINT is that keywords are automatically generated for each cluster. This is expected to be a straightforward and rigorously defined method for generating a keyword set from DEWPOINT data by focusing on keywords that appear in clusters that are also indicative of malicious intent. This approach may also be used to bootstrap annotated training and testing sets from a small hand-annotated seed set.

Henriksen JG, JL Jensen, ME Jørgensen, N Klarlund, R Paige, T Rauhe, and A Sandholm. 1995. "Mona: Monadic Second-Order Logic in Practice." In *Tools and Algorithms for the Construction and Analysis of Systems: First International Workshop TACAS '95*, pp. 89-110. Aarhus, Denmark. This paper describes an implementation of monadic second-order logic as an alternative to regular expression matching for text analysis. This algorithm might be useful where search patterns can be expressed by a human user, but not in the language of a regular expression. The basic implementation is built on finite state automata, and, surprisingly, exploits their computational complexity to generate efficient automata for finding the patterns. This could be used for

DEWPOINT as a more complex method, and hence, another baseline test, of finding patterns than regular expressions or simple keyword matching.

Lehnert W and B Sundheim. 1991. "A Performance Evaluation of Text Analysis Technologies." *AI Magazine* 12 (3): 81-94. The Message Understanding Conferences (MUC) were a series of open challenges in text analysis scored by a committee of experts using a single dataset and a single scoring system for each year the MUC was held. For each competition, a variety of industry and academic participants provided fully automated systems to analyze the given dataset. This paper is a description of the third MUC competition, the data set used, and the scoring methods used. The scoring methods used in this competition provide a good starting point for defining metrics for assessing the methods benchmarked under DEWPOINT study. These metrics can also be used as part of validation and verification of DEWPOINT algorithms developed.

Peshkin L and A Pfeffer. 2003. "Bayesian Information Extraction Network." In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pp. 421-426. Acapulco, Mexico. The algorithm presented in this paper uses Dynamic Bayesian Networks to extract information from free or semi-structured text. The method is shown as an improvement to previous models, which employ primarily Hidden Markov Modeling (HMM) approaches. This method does not appear to rely too much on local contextual information, so it may perform well on small and large information transactions. This would be a good baseline algorithm for DEWPOINT.

Webb-Robertson BJM, W Cannon, and C Oehmen. 2007. "Support Vector Machine Classification of Probability Models and Peptide Features for Improved Peptide Identification from Shotgun Proteomics." In *Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA '07)*, pp. 500-505. Cincinnati, Ohio. This paper describes how the performance of correctly classifying homology in biological sequences is significantly enhanced using support vector machines (SVM). The SVM-based HOmology Tool (SHOT) is demonstrated to greatly improve sensitivity of finding similar biosequences at a low false positive rate; a task that will be at the core of the proposed DEWPOINT system. This work was done in the area of bioinformatics, but the approach can be straightforwardly applied to other domains, such as that of interest to DEWPOINT where similarity in text strings is indicative of conserved behaviors in general.

Useful Literature

Arimura H, J Abe, R Fujino, H Sakamoto, S Shimozono, and S Arikawa. 2001. "Text Data Mining: Discovery of Important Keywords in the Cyberspace." In *Proceedings of the IEEE International Conference on Development and Learning (ICDL)*. Kyoto, Japan. This paper describes an algorithm that is shown to automatically derive keywords to describe Internet text. The algorithm makes use of a collection of methods, including string-based, statistical (maximum enropy, minimum entropy, frequency of word occurrence), and computational geometry. The authors present an algorithm for deriving keywords from ordered patterns that runs in near linear time with respect to the total size of text that must be searched. A more real-world applicable algorithm for unordered patterns is also presented that runs in quadratic time with respect to the size of random text that must be searched to find frequently occurring word patterns. This method does not use clues from markup languages to enhance the contextual information used in classification. This may be a method for deriving keywords from DEWPOINT datasets to use as a baseline for comparison to more advanced methods.

Baeza-Yates R, EF Barbosa, and N Zivian. 1996. "Hierarchies of Indices for Text Searching." *Information Systems* 21(6):497-514. The method described in this paper is for generating a hierarchical set of indices that can be used to dramatically enhance the rate at which textual data can be retrieved from a free text database using a query. The approach is to create an inverted file (or word frequency table) for each of several equivalent-sized file segments, and a condensed representation of the entire structure to minimize memory usage. The method is well described in this paper from an algorithmic and complexity standpoint. There is much attention paid to the balance between memory reuse and file access, a serious consideration for searching against large-scale data such as is expected for DEWPOINT. Results are presented on a variety of file sizes (all less than 1GB) and partition parameters. In general, this method shows a many fold improvement in retrieval time compared to other methods.

Bell B, JE Santos, and SM Brown. 2002. "Making Adversary Decision Modeling Tractable with Intent Inference and Information Fusion." In *Proceedings of the 11th Conference on Computer Generated Forces and Behavioral Representation*. Orlando, Florida. The methods described in this paper are focused on modeling adversary's intent for military applications. However, what is relevant to DEWPOINT is how the authors applied the principles of modeling user intent in the context of operating systems (achieved by combining behavior models and context information) to the area of inferring adversary's intent. A similar process of extending user intent inference through online behavior might be applied within DEWPOINT. However, behavioral models would need to be developed outside the scope of this project.

Chang CC and CJ Lin. 2001. "LIBSVM: a library for support vector machines." Access at http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (updated April 1, 2009). This Web site describes LIBSVM, which is an open source library containing an implementation of the Sequential Minimization Optimization (SMO) approach to support vector machine training. LIBSVM is one of many possible SVM tools that could be used to train a classifier to recognize patterns of interest to DEWPOINT. LIBSVM employs some simple enhancements to the basic SMO algorithm to improve convergence behavior, but unfortunately has too many parameters which must be set by the user (including the fraction of vectors that will be returned as support vectors).

Chen CH and V Honavar. 1999. "A Neural Network Architecture for Syntax Analysis." *IEEE Transsactions on Neural Networks* 10(1): 94-114. The majority of this paper describes applications of neural nets to language processing relevant to analyzing software. However, there are two sections on lexical analysis that pertain to this project. Lexical analysis is shown to be possible using neural nets where the system could be taught to parse large texts looking for particular patterns it has learned to associate with particular terms in a database. This may be a more sophisticated option than keyword searching that can still handle "fuzzy" grammatical constructs of interest.

Chen Z, F Lin, H Liu, Y Liu, WY Ma, and L Wenyin. 2002. "User Intention Modeling in Web Applications Using Data Mining." *World Wide Web* 5(3):181-191. Identifying user intention is a critical aspect of optimizing Internet-based searching. This paper describes methods for modeling user intention by accumulating data on responses to query results. Naive Bayes classifiers were used to demonstrate that user intent could be inferred automatically nearly as well as it could be predicted manually by a human. Keywords alone were consistently the poorest performing indicators of true intent. Though DEWPOINT is not geared to delivering optimal results to users, it may be able to take

advantage of information about user responses to queries in much the same way as this paper to augment the assessment of intent beyond what can be learned from content alone.

Cho J and S Rajagopalan. 2002. "A Fast Regular Expression Indexing Engine." In *Proceedings of the 18th International Conference on Data Engineering*, p. 419-430. San Jose, California. This paper describes a method for accelerating regular expression matching by using a pre-indexing scheme. This is directly relevant to efforts such as DEWPOINT baseline performance where regular expression matching may be a rate-limiting step in natural language analysis. One complication in this method is that regular expression queries against the dataset must be processed so that the proper indices are used for lookups, and only the highly relevant segments of the data are searched deeply. The algorithm is implemented in a package called FREE. The authors present results that FREE most significantly outperforms conventional regular expression matching when the target of the query is a very small subset of the overall data. This is an expected feature of DEWPOINT datasets, as real data will be overwhelmingly benign in nature and even malicious events will fall into highly specific categories.

Dennis S. 2003. "A Comparison of Statistical Models for the Extraction of Lexical Information from Text Corpora." In *Proceedings of the 25th Annual Meeting of Cognitive Science Society*, eds. Alterman R and D Kirsh. Boston, Massachusetts. Syntagmatic Paradigmic Model (SP) and Pooled Adjacent Context (PAC) Model are statistical models for extracting information about similarity of word context and word substitution within natural language text. SP and PAC are described and compared in this paper. The two methods are shown to have similar performance for extracting syntactic and semantic information, and SP is shown to have superior performance for associating words. Since DEWPOINT is focused on specific terms and their meanings rather than their usage patterns, SP would be more appropriate to use. It is not clear how either method would perform on semi-structured text where contextual information may be lacking or obscured intentionally.

Deutsch A, M Fernandez, and D Suciu. 1999. "Storing Semi-structured Data in Relations." In *Proceedings of the Workshop on Query Processing for Semistructured Data and Nonstandard Data Formats*. Jerusalem, Israel. This is a companion paper to the STORED approach by Deutsch et al. 1999. Where STORED is an algorithm for mapping semi-structured views onto existing relational data (and database management systems), this paper describes how one would cast existing semi-structured data into a relational system using a procedure that effectively reverse-maps the STORED algorithm. Since much of our data will be unstructured or semi-structured, this may be a useful way to store and manage the data. One useful feature of this approach is that it can automatically generate the relational schema from the semi-structured data representation. Relational storage has been reported to preserve 90% of the original data. The primary concern with this method would be its ability (or inability) to handle data at a very large scale.

Feigenbaum J, S Kannan, M Strauss, and M Viswanathan. 1999. "Streaming Algorithms for Distributed Massive Data Sets." In *40th Symposium on Foundations of Computer Science (FOCS)*. New York, New York. This paper describes algorithms for rapidly calculating difference function on large-scale, high-throughput data streams. If we successfully demonstrate that the DEWPOINT approach can be applied to our target dataset, eventual deployment of a system may rely on algorithms such as this for rapidly computing on that datastream. This paper does not include benchmark results of the method, so performance measures are not yet available. This paper describes the mathematics of the approach only, and proves the assertion that for small differences in the data streams (i.e., when looking for

small deviations between objects or between a stream and a template object as in our case) this one-pass approach reliably detects those differences regardless of the ordering in the streams, or delocalized nature of them.

Giles J, L Wo, and M Berry. 2004. "GTP (General Text Parser) Software for Text Mining." In *Statistical Data Mining and Knowledge Discovery*, ed. H Bozdogan, pp. 457-473. CRC Press LLC, Boca Raton, Florida. General Text Parser (GTP) software and algorithm is described in this paper. GTP provides a method for automating information retrieval from free text data documents. This paper is a good reference for users of GTP as it describes syntax of the interface as well as a cursory description of the underlying algorithms. The relevant aspect of this tool for DEWPOINT is that it vectorizes documents by the number of instances terms appear in a corpus. This could be the basis for vectorizing network transactions by payload content for automating content discovery. Using this with simplistic keyword queries could also be a baseline method against which machine learning approaches could be benchmarked.

Henzinger M, BW Chang, B Milch, and S Brin. 2003. "Query-Free News Search." In *Proceedings of the 12th International Conference on World Wide Web*, pp. 1-10. Budapest, Hungary. This paper describes several methods for retrieving relevant text articles relating to given queries. It falls primarily in the area of information retrieval (as opposed to information extraction), and as such could be used as a method for classifying text. The main concern with this method is reliability as the paper demonstrates less than 70% of the matches are exact, even for the best algorithm tested. This could be used as a cursory pass over DEWPOINT data to roughly categorize text as a baseline for performance gain metrics for more sophisticated techniques.

Kang K, C Domeniconi, and D Barbara. 2005. "Categorization and Keyword Identification of Unlabeled Docments." In *Proceedings of the 5th IEEE International Conference on Data Mining*, pp. 677-680. Houston, Texas This paper describes a local clustering algorithm devised for inferring relevant keywords associated with categorizing free text. The method is demonstrated on news services and email spam detection and is shown to infer keywords that are highly predictive for classification tasks. This method could be used by DEWPOINT for keyword identification or as a means simply to group data by contextual relevance. It is not clear how well this would work on data for which little context is provided.

Lewis DD and WA Gale. 1994. "A Sequential Algorithm for Training Text Classifiers." In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3-12. Dublin, Ireland. This paper describes methods for enhancing the performance of text classifiers through an iterative process with subject matter experts. Sampled training data is used to create a simplistic classifier. Rather than annotating points throughout the vector space, an analyst is asked to annotate or refine the classification only near the classifier boundary. This refinement leads to additional round of training on the improved training set. The consequence of this approach is to improve the performance of the sampled training without the need for training on the full dataset. The classification methods in the paper are heavily influenced by Bayesian statistics. For DEWPOINT, it would be more likely to use such a sampling approach with support vector machines as the training system. We would still have to independently devise and implement a vectorization scheme.

Lewis DD and M Ringuette. 1994. "A Comparison of Two Learning Algorithms for Text Categorization." In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pp. 81-93. Las Vegas, Nevada. Two common implementations of text categorization are described in this paper. A Bayesian method and a decision-tree-based method are compared using two sets of free-text newsfeed data. Both methods perform with nearly the same precision vs. recall for a given dataset suggesting that neither has a clear advantage for text categorization. The paper outlines some of the potentially incorrect assumptions used when applying these categorization methods to natural language, including the assumption of independence of word occurrences and the implicit assumption that natural language trends are static. The most relevant aspect of the paper to DEWPOINT is the discovery of time-dependence in categorization of text. In this case, training sets did not overlap with the end of a quarter, resulting in high degree of incorrect categorization for quarterly financial reports. It is expected a similar time sensitivity will occur in the data of interest to this project, so training data must be found that specifically represents the temporal relationship between an entity's information gathering and their current technological intent.

Martinez-Fernandez JL, A Garcia-Serrno, P Martinez, and J Villena. 2004. "Automatic Keyword Extraction for News Finder." In *First International Workshop on Adaptive Multimedia Retrieval*, eds. Nurnberger A and M Detyniecki, pp. 99-119. Hamburg, Germany. This paper focuses on information retrieval from a collection of text documents containing a variety of media types and languages. Automatic Keyword Extraction (AKE) is the central algorithm and is based on the frequency of word occurrences or, for phrases, co-occurrences. The effectiveness of word stemming vs. non-stemming (i.e., finding a word fraction that is highly conserved rather than considering all variants of a single word) is discussed. The mathematics of the method are very simplistic and are not adequately compared to other methods in terms of performance. This method could be used by DEWPOINT as a simple weighting scheme for deriving keyword stems.

McCallum A and K Nigam. 1999. "Text Classification by Bootstrapping with Keywords, EM and Shrinkage." In *ACL99- Workshop for Unsupervised Learning in Natural Language Processing*, pp. 52-58. University of Maryland, College Park, Maryland. This paper describes a method for producing an annotated training class by bootstrapping using keywords. This is highly relevant to DEWPOINT because we expect to have some datasets that are marginally annotated because subject matter expert curation is expensive and time consuming, plus the datasets are expected to be large. The method presented uses keywords to create annotation classes on a subset of the data. Keyword searching creates a first pass annotation that is used as the input to a naive Bayes classifier, which, combined with Expectation-Minimization and hierarchical shrinkage (to smooth the probability of estimates of the same term on different hierarchical levels of the classifier), results in a more accurately annotated set. For DEWPOINT, this might provide a good baseline method for comparison as well as a means for annotating the training dataset. This method could be simply converted by substituting Bayesian classification with more sophisticated machine learning methods.

Miikkulainen R. 1997. "Natural Language Processing with Subsymbolic Neural Networks." In *Neural Network Perspectives on Cognition and Adaptive Robotics*, ed. A Browne, pp. 120-139. IOP Publishing Ltd, Bristol, UK. The methods described in this paper are theoretical in nature, having only cursory implementations for proof-of-concept. However, the concepts behind sub-symbolic neural networks for natural language processing have potential to add a useful dimension to automated understanding of text because it uses neural nets to essentially model the human process of language cognition. Rather than assume each word has a fixed meaning and usage, it develops

satistical models of how words are used and what they mean. This makes it easier to understand text in the presence of errors or complex semantic constructs, which might cause grief for other systems. The paper's stated primary target is a more human-like interface to knowledge extraction from text. However, for DEWPOINT, the main interest is in that content obfuscated by intentionally misusing terms or grammatical constructs may be possible to detect and understand using this automated method.

Miklau G and D Suciu. 2007. "A Formal Analysis of Information Disclosure in Data Exchange." *J. Computer and System Sciences* 73(3): 507-534. The central problem addressed in this paper is the need to verify that public information cannot not be used to infer information which is not meant to be public. Verifying the security of single points of information disclosure is not difficult, however, situations easily arise when combining multiple data transfers, each to legitimate sources, that together could unintentionally (or otherwise) reveal sensitive information. The mathematics of this approach may lead us to identify combinations of information sources in our own datasets that would be suspected targets of entities of interest, or at least to some patterns of access that may indicate intent.

Navarro G. 2003. "Regular Expression Searching on Compressed Text." *J. Discrete Algorithms* 1(5-6): 423-443. This paper describes an algorithm for regular expression matching in the Ziv-Lempel class of text compression techniques (including LZ78 and LZW compression). Though this algorithm's worst case performance is slower than O(2^m), where m is the size of the regular expression, it is still shown to be more efficient than decompression followed by conventional regular expression matching. A derivative method for approximate matching is also presented. The significance of this algorithm to the DEWPOINT project is that it is possible some information will be exchanged in compressed form. Finding the patterns of interest in such transactions may require algorithms such as this one to avoid having to decompress (or ignore altogether) compressed text. This algorithm only applies to regular expression matching, so alternative methods would have to be identified or developed for finding patterns using more advanced extraction approaches in DEWPOINT. The algorithm was also only benchmarked on very small text segments. Scalability would be a major concern, but this method is predicted to scale in the worst case linearly with respect to size of text, and linearly with respect to number of matches to the pattern.

Riloff E. 1993. "Automatically Constructing a Dictionary for Information Extraction Tasks." In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pp 811-816. Washington, D.C. The AAAI Press, Menlo Park, California. To deal with some of the challenges of creating domain-specific dictionaries (a prerequisite for many natural language processing systems), the author has developed an automated system for constructing such a dictionary. This automated dictionary creation can be done for new domains in less than a single day, by a single person, and has been reported to enable 98% of the performance for downstream natural language systems when compared to those enabled by an expert-created dictionary. Results from the MUC-4 conference are reported in support of this. The main limitation of this method is that it was created with news-feed textual data sources in mind. In fact, the central design relies on stylistic features of news articles. This may limit the usability of this system for DEWPOINT, however, it may have value for some the intended information sources.

Soderland S. 1999. "Learning Information Extraction Rules for Semi-structured and Free Text." *Machine Learning* 34(1-3): 233-272. WHISK, a system for extracting information from semi-

structured or free text, is described in this paper. The significance of handling semi-structured text is that semantic parsing will fail when natural language constructs on which it relies are absent, such as in “want ads”' or in network transactions where natural language content is obscured or highly abbreviated. The limitation of WHISK is that it is based on regular expression pattern matching. This allows WHISK to operate on many domains (where the patterns are learned based on training), regardless of content. But in the case of DEWPOINT, the domain is well defined and it is likely that domain specific information is crucial to properly understanding the content of transactions. The priority for DEWPOINT will be in developing a highly reliable and specialized classification system as opposed to a general purpose one. Nevertheless, some of the aspects of how WHISK deals with semi-structured text may be relevant.

Spyrou T and J Darzentas. 1996. “Intention Modeling: Approximating Computer User Intentions for Detection and Prediction of Intrusions.” In *Information systems security: facing the information society of the 21st century*, eds. Katsikas S and D Gritzalis, pp. 319-335. Chapman and Hall, Ltd, London, UK. This paper describes a modeling approach for inferring the intent of users in the context of malicious intrusions. The significant aspect of this work for DEWPOINT is the notion that any single transaction (or behavior) may be “legal” or “not indicative of an intrusion,” but an aggregate set of “legal” actions may itself be an intrusion. The algorithm described in the paper is based in Intention Models for inferring user intention by constructing a coarse model of user behaviors (Cognitive Task Modelling, CTM) and combining this model with a priori information about the knowledge a user must have to perform these actions (represented in Task Knowledge Structures, TKS). This paper describes only the theoretical foundation of the approach, and does not discuss a particular implementation, nor its performance on real tasks. Such an approach could be adapted to DEWPOINT, but TKS and CTM would have to be specially developed to be relevant in our application domain.

Wu T, LE Holzman, WM Pottenger, and DJ Phelps. 2003. “A Supervised Learning Algorithm for Information Extraction from Textual Data.” In *Proceedings of the Workshop on Text Mining, 3rd SIAM International Conference on Data Mining*, pp. 60-71. San Francisco, California. The algorithm described in this paper is designed to automatically learn regular expressions from natural language text for the task of information extraction. Specifically, this method uses supervised learning to derive finite automata that describe regular expressions. This paper is a supervised learning companion to another Wu paper (also included in this bibliography) that focuses on semisupervised methods.

Wu T and WM Pottenger. 2003. “A Semi-supervised Algorithm for Pattern Discovery in Information Extraction from Textual Data.” In *Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 117-123. Seoul, Korea. This paper describes a semi-supervised algorithm for discovering regular expressions specific to information extraction tasks from natural language text datasets. If one defines the information extraction and provides a subset of annotated text, this process will derive the regular expression rules that can be used to populate information extraction templates from the remainder of the text. This paper presents benchmark results on very small training and testing sets and for a small number of extraction template fields. It is not clear how this algorithm would scale to large datasets expected within DEWPOINT, but the method for deriving the regular expression patterns may still prove useful for training. This paper contains an incomplete description of the algorithm, which can be found in a complete form only in a Lehigh University technical report.

Reference Literature

Balkan L, D Arnold, and S Meijer. 1994. "Test Suites for Natural Language Processing." In *Language Engineering Convention*, pp. 17-22. Puteaux, France. This paper describes an evaluation process for test suites and test corpora used to develop, benchmark, and assess natural language processing approaches for English, German, and French. The work is centered in Europe and is driven in part by the need for machine translation. Advantages of test suites (collections of annotated real-world text of interest to a user) are described, including the fact that they arise from naturally occurring text and so are a more realistic test of the performance of natural language processing systems. However, when they reveal shortcomings within a system, the exact nature of the shortcoming is often not transparent. By contrast, test corpora (collections of usually contrived text snippets most often used by developers to test their systems for a particular response) can reveal much more specific information about the shortcomings of various natural language processing systems, but are not representative of real-world text.

Bell B, J Franke, and H Mendenhall. 2000. "Leveraging Task Models for Team Intent Inference." In *Proceedings of the 2000 International Conference on Artificial Intelligence.* Las Vegas, Nevada. This paper presents a model and model platform for inferring intent of an operator based on historical actions and a model of tasks the operator engages in. The goal is to make it possible to create interfaces that proactively retrieve and filter relevant information based on what the system believes an operator will want to do next, and on the information that will be required to inform decisions relevant to that action. Though DEWPOINT is not geared to inferring intent for proactive queries, it may be useful to employ some modeling techniques to infer intent of operators in our context. It is possible a system like the one described in this paper could be adapted for that use.

Brown SM, E Santos Jr, SB Banks, and MR Stytz. 1998. "Intelligent Interface Agents for Intelligent Environments." In *Proceedings of the 1998 AAAI Spring Symposium on Intelligent Environments*, pp. 145-147. Stanford University, Palo Alto, California. This paper describes an interactive knowledge environment for real-time knowledge acquisition from streaming data. It is implemented using agent-based programming model. This approach is interesting in that it partitions the computing tasks needed for processing data partially by using input from the user. The goal is to allow the human to direct the system to do the most useful computing possible to present the data in the most useful way. This paper is too short to describe the task partitioning or the visual aspects of the algorithm in any detail.

Bunescu R and RJ Mooney. 2004. "Relational Markov Networks for Collective Information Extraction." In *Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields (SRL-2004)*, Banff, Alberta, Canada. This paper describes a process for enhancing the construction of information extraction systems by including correlations between extracted entities. The approach makes use of Relational Markov Networks to represent relationships between extractions as the information extraction strategy is learned. This method is compared in the paper to the conditional random fields method, the state of the art in information extraction when the research in this paper was performed. The authors demonstrate their method outperforms conditional random fields on two different datasets, but with F-measures that indicate these methods may still not be good enough to detect the small signals we will be searching for in DEWPOINT without overwhelming analysts with false positives.

Bunescu R and R Mooney. 2007. "Statistical Relational Learning for Natural Language Information Extraction." In *Introduction to Statistical Relational Learning*, eds. Getoor L and B Taskar, ch. 19. MIT Press, Cambridge, Massachusetts. This paper describes follow-on research to Bunescu and Mooney (2004), referenced above, that extends the utility of Relational Markov Networks to the challenge of collective information extraction. Where information extraction is satisfied by correctly filling the slots in an information template, collective information extraction is concerned with finding all instances of the template within a document. The method is demonstrated in a biological document context where the challenge is to find all references to particular proteins, even as they have multiple names within the same document. This is applicable to DEWPOINT in that it may be necessary to find many references to the same concept using a widely varying set of terminology. Relational Markov Networks were shown to improve collective information extraction where co-references to the same identity are highly delocalized. This is done using a hierarchy of templates that capture both global and local relationships in the text.

Clarke C and G Cormack. 1997. "On the Use of Regular Expressions for Searching Text." *ACM Transactions on Programming Languages and Systems* 19(3): 413-426. This paper is a good introduction to how regular expressions can be used to search for patterns in text. It contains a brief introduction to regular expressions and the problem of text searching, including algorithmic time and space complexity. This paper also outlines some of the more derived aspects of regular expressions, such as containment, and substring matching.

Cristianini N and J Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK. This book describes the theory and some implementation details for Support Vector Machines. It is a good reference, including pseudocode for an implementation of the sequential minimization optimization SVM approach. This method has been shown to dramatically enhance the convergence properties of SVM training over other methods because it decomposes the basic problem into subproblems of very small size each having analytical solutions. This is the likely algorithm that will be employed by DEWPOINT to learn what patterns in information exchange are indicative of malicious activity.

Deutsch A, M Fernandez, and D Suciu. 1999. "Storing Semistructured Data with STORED." In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pp. 431-442. Philadelphia, Pennsylvania. This paper describes a method for mapping semi-structured data views into conventional relational database applications. Normally, this is either not possible (because of structure imposed by relational databases) or not efficient (where spatial and time penalties are incurred for handling semi-structured data). This paper describes a mapping technique (STORED) to bridge the gap between semi-structured data (or views of data) and relational database schema. Since the DEWPOINT project will focus primarily on unstructured or semi-structured text, it is unlikely we will produce relational databases at first. However, the companion paper to this one by Deutsch et al. describes a converse method that may be highly useful. This is different than data mining because patterns are found over nearly the entire space of data (rather than a small sample). This approach is provably lossless in that exact data can be reconstructed from any semi-structured mapping.

Espinoza M, J Gracio, R Trillo, and E Mena. 2006. "Discovering the Semantics of Keywords: An Ontology-based Approach." In *2006 International Conference on Semantic Web and Web Services (SWWS'06)*, pp. 193-201. Las Vegas, Nevada. This paper describes an algorithm for utilizing existing

ontologies to automatically discover the different meanings or uses of keywords in a body of text. It was developed for Web-based applications where large-scale data can contain many different uses of a single term, and multiple terms to describe the same content. Since this method relies on preexisting ontological orgnization, which does not already exist in the DEWPOINT domain, it could not be used without first creating an ontology. While this would potentially be very useful in the long term, it is beyond the scope of short-term activities for DEWPOINT.

Gaines BR and MLG Shaw. 1995. "Knowledge Acquisition and Representation Techniques in Scholarly Communication." *ACM SIGDOC Askerisk Journal of Computer Documentation* 19(2):23-26. This paper describes the use of concept maps and formal knowledge structures to facilitate scientific communication. The goal of these methods is to provide a mechanism for illustrating information and the complex relationship between concepts. Perhaps the key result of the paper is that concept maps and formal knowledge structures were shown to help automate text analysis in the context of scholarly communication. The methods section is severely lacking from this paper, so it is in effect a conceptual paper stating how these principles can be used to create a new paradigm for scientific discourse, as compared to the conventional media of static journals and conference proceedings.

Gaines BR and MLG Shaw. 1996. "WebMap: Concept Mapping on the Web." In *Proceedings of the 4th International World Wide Web Conference*. Boston, Massachusetts. This paper describes a concept mapping system for internet content. Concept mapping is a visual representation of the relationship between textual elements based on a tree structure. Several applications of this are described, mainly focused on building or presenting concept maps to users who then use them to browse internet content based on relationships displayed in the map. This implementation is specifically for Apple Macintosh computers. This is a very narrow implementation platform, but the paper is older so it is possible a newer implementation of this exists. The visual (GUI) nature of the representation is not amenable to automation or for handling data at the volumes we expect for the DEWPOINT project. However, the underlying idea of concept mapping may be useful for expressing the relationship between network transactions of interest in DEWPOINT after primary analysis is done.

Gamon M, A Aue, S Corston-Oliver, and E Ringger. 2005. "Pulse: Mining Customer Opinions from Free Text." In *Lecture Notes in Computer Science* Vol. 3646, *Advances in Intelligent Data Analysis VI*, pp. 121-132. Springer, Berlin/Heidelberg, Germany. Pulse is a tool for extracting the opinion of people from free text customer feedback. This method is presented as it applies to car reviews. Pulse mines free text data to derive a minimally hierarchical classification scheme of the data, in this case, make and model. It employs clustering on sentences to infer group and label (with derived keywords) responses relevant to this hierarchical view. The approach taken by Pulse could be applied to other forms of free text, as indicated in the paper. The relevant aspect of this paper for DEWPOINT is that clustering on sentences has been shown to be useful for characterizing the prevailing opinion, as opposed to intent, within a given context. However, since our goal is to identify and infer the intent of entities, prevailing opinion is not important. It is unlikely we will be able to infer opinion, nor necessarily do we want to, from our data since specific questions are not being answered.

Grishman R. 1997. "Information Extraction: Techniques and Challenges." In *Information Extraction: International Summer School : A Multidisciplinary Approach to an Emerging Information Technology table of contents*, pp. 10-27. Springer-Verlag, London, UK. This paper describes the concept of information extraction. One key aspect of information extraction highly relevant to this

project is that it focuses on understanding the meaning of text in a user-defined context. It is not meant as a method for complete automated understanding. This is advantageous for DEWPOINT in the sense that we wish to focus on analyzing text for a specific purpose, not for general understanding. This paper points out that some methods devised for the MUC conference series can be successful for understanding content of text when a small number of representations of the concepts are present and highly localized in text. For some aspects of DEWPOINT, these restrictions may hold. There are many facets of information extraction discussed, including lexical analysis, regular expression matching, and the use of templates for specialized patterns.

Hulgeri A, G Bhalotia, C Nakhe, S Chakrabarti, and S Sudarshan. 2001. "Keyword Search in Databases." *Bulletin of the Technical Committee on Data Engineering* 24(4): 22-32. This paper describes a method for keyword searching over multiple databases. The queries and the responses are represented as graphs. Ranking of the response graphs is done using metrics relating to interconnectedness. This paper contains little in the way of results, other than a brief mention of query times for a particular database collection requiring on the order of seconds. There were no metrics used to demonstrate this query system is more reliable or produces more relevant responses to a user query. The most useful concept that might be utilized by DEWPOINT is the notion of using query responses over multiple datasets as a tree structure to facilitate relevance ranking.

Karttunen L, JP Chanod, G Grefenstette, and A Schiller. 1996. "Regular Expressions for Language Engineering." *Natural Language Engineering* 2: 305-328. This paper introduces the calculus of finite state transducers and demonstrates how they can be used to construct fully finite-state-based tools to tokenize text, recognize parts of speech, mark-up entire phrases for grammatical analysis, and extract verbal syntax. The methods described in the paper are highly specific to finite state transducers, and are meant to illustrate the utility and breadth of their uses for natural language processing. The theoretical results in the paper could be adapted for DEWPOINT, but the lack of actual implementation of the ideas relegates these approaches to more of a fallback formalism if new methods have to be devised for any of these aspects of information extraction. The number of fully implemented solutions (for instance through the Message Understanding Conference participants) for the same problem space make this an unlikely need.

Kushmerick N, E Johnston, and S McGuinness. 2001. "Information Extraction by Text Classification." In *IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*. This paper describes a method for using text classification for the task of information extraction. This is implemented using hidden Markov models to exploit underlying structure, such as the order of information presented, in the text source. This is a novel application of text classification for information extraction, which had been historically unsuccessful. The drawback of this method for DEWPOINT is that there may not be enough structural information available in DEWPOINT dataset to create hidden Markov models.

Levy AY, A Rajaraman, and JJ Ordille. 1996. "Querying Heterogeneous Information Sources Using Source Descriptions." In *22nd International Conference on Very Large Data Bases (VLDB)*, pp. 251-262. Mumbai (Bombay), India. This manuscript describes techniques for querying across multiple, heterogeneous, text-base information sources. Key challenges in this area include non-uniformity of data representation, inconsistent nomenclature or object references, and missing data. The central methods of the paper are relatively straightforward set-theory-based approaches to understanding the structure and type of data within each source and building inheritance-based objects to help resolve ambiguities and omissions with these datasets. A second thrust of the paper is a development of

transformed queries from user-defined queries and the mathematical formulation for guaranteeing that the transformed query in fact answers the user's query. Query construction is based again on set theory, object attribute inheritance, and order in which query subtasks are dispatched. An implementation of this formulation is included, and is called Information Manifold System. Some of these concepts might be relevant to DEWPOINT if future directions allow us to operate on multiple network transaction types.

Lin TC, M Hsu, FY Kuo, and PC Sun. 1999. "An Intention Model-based Study of Software Piracy." In *Proceedings of the 32nd Hawaii International Conference on System Sciences*, p. 5030. The methods described in this paper focus on modeling the intent of individuals engaging in software piracy by first understanding situations that are commonly associated with piracy. The method includes psychological and behavioral determinants used in predictions of decision-making. There are similar dimensions of interest to the larger challenges surrounding DEWPOINT, but they are beyond the scope of our research. Nevertheless, implicit and explicit links between inferring intent (as proposed) and socio-technical aspects of nonproliferation research may lead to future interactions with other projects.

Liu P and W Zhang. 2003. "Incentive-based Modeling and Inference of Attacker Intent, Objectives, and Strategies." In *Proceedings of the 10th ACM Conference on Computer and Communications Security*, pp. 179-189. Washington, D.C. This paper describes a game-theoretic approach to inferring the intent of an attacker. The novel concept introduced is that of "utilities" that can serve as a link between an attacker's cost and incentive functions. The goal of this method is to provide a partial model of an attacker's true intent. This paper is primarily theoretical in nature, describing a few example instantiations of the methods. The primary value to DEPOINT is an explicit link between observing attacker behavior and inferring true intent. However, this game theory approach may not be highly applicable because it presumes real-time interactions between attacker and defender. We may not have access to an intruder's reaction strategy because there might not be one. Also, there may be no defensive action taken. Rather our goal is to characterize intent for now.

Lu T, S Sinha, and A Sudan. 2003. "Panaché: A Scalable Distributed Index for Keyword Search." Massachusetts Institute of Technology, Cambridge, Massachusetts. Panaché is a competitor of Gnutella for enabling peer-to-peer (P2P) file sharing. Though P2P file sharing is largely not allowed when cyber security is a priority, the peripheral methods in this paper geared toward efficient searches on distributed keyword sets may be relevant to DEWPOINT. One possible way of searching large-scale datasets we expect for this study will be to distribute the data, with the potential of having different keywords at each data location. However, the first approach will be to use MapReduce, an alternative algorithm that eliminates the need to distribute the keywords.

Mauw S and MA Reniers. 1994. "An Algebraic Semantics of Message Sequence Charts." *The Computer Journal* 37(4):269-277. Message Sequence Charts, defined and described in this paper, are a mathematical description of the order of operations in a sequence of communications. The algebra describing such messages takes into account concurrence of some operations, actions resulting from operations, timeouts (and timeout handling) and asynchrony. This could be a useful notation for devising high performance implementations of the classification strategies envisioned for DEWPOINT, regardless of whether they are for shared memory or distributed memory systems or components.

Neumann G, R Backofen, J Baur, M Becker, and C Braun. 1997. "An Information Extraction Core System for Real World German Text Processing." In *Proceedings of the 5th International Conference of Applied Natural Language*, pp. 208-215. Methods described in this paper focus on special algorithms for automated extraction of information from German text. The methods are highly specific for German language constructs and sentence structure. It does describe an interesting normalization strategy for international representations of date and time that might be relevant for DEWPOINT. However, the primary interest of this paper is as an outline of some of the considerations that must be made when analyzing non-English language text.

Nguyen XT. 2002. "Simulating Automated Intent Assessment." In *SimTecT 2002: Simulation Conference and Exhibition*. Melbourne, Australia. This paper describes a Bayesian network-based implementation of automated intent assessment and a test bed system for comparing different intent assessment approaches for a given problem. This test bed was developed in the context of assessing the intent of unfriendly airborne targets. Though this application area is very different than the target of DEWPOINT, the approach taken by the authors to classify general intent types, identify signal behaviors, and use the Bayesian network to link the probabilities of these could be used by DEWPOINT.

Noord Gv and D Gerdemann. 1999. "An Extendible Regular Expression Compiler for Finite-State Approaches in Natural Language Processing." In *Lecture Notes in Computer Science* Vol. 2214, *Revised Papers from the 4th International Workshop on Automata Implementation*, pp. 122-139. Potsdam, Germany. Springer-Verlag, London, UK. This paper describes an extended compiler tool for creating regular expressions embedded in finite-state automata. The tool, called FSA5, is shown to have extensions above previous versions that allow for the construction of regular expressions that are directly relevant for natural language processing tasks. This paper is presented much as the background of a user's guide. It describes implementations of extensions, and how these relate to natural language processing. No accuracy or timing benchmark results are reported and the methods herein are not directly compared to competing methods. From the perspective of DEWPOINT, such a tool might be used in the future if other regular expression methods outlined in this bibliography prove insufficient.

Palmer D and M Ostendorf. 2001. "Improving Information Extraction by Modeling Errors in Speech Recognizer Output." In *Proceedings of the First International Conference on Human Language Technology Research*, pp. 1-5. San Diego, California. This paper focuses on methods for improving information retrieval from auditory spoken language data sources. This is not directly relevant to DEWPOINT, but should be kept in mind for potential future enhancements as voice-based data transfer is another potential modality of interest to the larger community. This paper introduces the use of manually introduced and annotated errors with confidence estimates to enhance the correct extraction of information from auditory natural language data sets. This was shown to significantly improve information extraction, but the authors acknowledge that many more advances in this area are needed before automated information extraction from auditory signals is highly reliable.

Riloff E. 1999. "Information Extraction as a Stepping Stone Toward Story Understanding." in *Understanding Language Understanding: Computational Models of Reading*, Eds. Ram A and K Moorman, pp. 435-460. MIT Press, Cambridge, Massachusetts. This book chapter excerpt describes, at a very high level, basic concepts of information extraction and how they are related to the much larger task of automated story understanding. The difference is that information extraction is highly

specific to a particular use case (e.g., wanting to extract weather information from a news feed). Whereas in story understanding, one must also infer what the use itself is (e.g., realizing that the primary reason for a story is to describe the weather). The excerpt describes how participants on the MUC conference series had advanced the state of the art in information extraction and how that is relevant, but not sufficient for story understanding.

Rindflesch T. 1996. "Natural Language Processing." *Annual Review of Applied Linguistics* 16:70-85. This paper is a very high-level review of natural language processing terms and research. It briefly introduces the concepts of corpus linguistics, lexicon, automatic tagging, parsing, word-sense disambiguation, semantics, discourse analysis, and gives references for research in each of these areas. This paper also lists several applications in natural language processing with references. This paper includes a brief annotated bibliography and a reasonably extensive unannotated bibliography. This review is somewhat dated, so newer work must be found by other means.

Schulzrinne H, A Rao, and R Lanphier. 1998. *Real Time Streaming Protocol (RTSP)*. The Internet Society, Reston, Virginia. This document is a draft of the specification for online streaming media, real-time streaming protocol (RTSP). It contains details of how devices communicate streaming information and how they utilize various protocols for this purpose. This specification does not include how the payload itself is handled, but rather all of the control aspects of RTSP communications.

Soderland S. 1997. "Learning to extract text-based information from the World Wide Web." In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pp. 251-254. Newport Beach, California. The methods described in this paper include CRYSTAL, a natural language processing system, and WebFoot, an application that uses structure of web pages to infer context rather than relying on grammatically correct text. This application tandem is shown to perform well together for basic information extraction from web pages. Since it relies on the structural tags in web pages, it is unlikely this process would perform well on any non-web text source. However, it is suggestive of an alternative method for natural language processing applied to semi-structured data sources. CRYSTAL is described in more detail in multiple MUC proceedings, and stands alone as a natural language processing system for normal text. CRYSTAL learns its information extraction rules via training on labeled data.

Strzalkowski T, G Stein, G Wise, J Perez-Carballo, P Tapanainen, T Jarvinen, A Voutilainen, and J Karlgren. 1998. "Natural Language Information Retrieval: TREC-7 Report." In *7th Text Retrieval Conference*, pp. 217-226. Gaithersburg, Maryland. This paper describes topic expansion, a method for enhancing natural language processing systems that query text sources for particular content. The central idea behind topic expansion is similar to boot-strapping, in that a user query is augmented using related but more detailed examples that seem to fit the initial query. The method was shown to more than double performance of retrieval approaches in some cases where the augmenting was done manually. Automation of this method is described as well. Automated topic expansion is not as effective as manual topic expansion, but it still appears to add value for some datasets. The concept of topic expansion could be used in DEWPOINT to create a more effective template for finding information of interest.

Sutcliffe RFE, P Boersma, A Bon, T Donker, MC FerrisZ, P HellwigX, P Hyl, HD Koch, P Masereeuw, A McElligott, D O'Sullivan, L Relihan, I Serail, I Schmidt, L Sheahan, B Slater, H Visser, and PJTM Vossen. 1995. "Beyond Keywords: Accurate Retrieval from Full Text Documents." In *Proceedings of the 2nd Language Engineering Convention*. London, UK. This paper reports an early research effort aimed at providing an interface for finding information in text beyond simple keywords. One of the main weaknesses identified with keywords is that natural language text can have many terms that mean the same thing (or many spellings of the same term) that are easily missed using keywords alone. This application is highly specific to extracting information from technical manuals, as section headings are utilized to provide context. The algorithm presented attempts to create an ad hoc mapping between terms based on semantic distance measures, which performs well for nouns, but does not capture the meaning of verbs well. The paper can serve as a reverence for DEWPOINT in that it describes the reasoning process behind how extensions to keyword searching were developed to meet a special purpose.

Tomokiyo T and M Hurst. 2003. "A Language Model Approach to Keyphrase Extraction." In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 33-40. Sapporo, Japan. This paper introduces the concepts of "phraseness" and "informativeness" as they apply to deriving keywords from natural language text. "Phraseness" is a measure of the extent to which assumptions of word independence degrade the ability to extract meaning from a group of words, leading to the notion that the words should indeed be regarded as a phrase or single unit. "Informativeness" is a measure of the extent to which a phrase captures the essence of a segment of text. These concepts are shown to have nearly no correlation, so the authors make an argument that they must be combined to get a global measure of the quality of keywords or phrases. While the concepts and the results presented are intuitive, the lack of statistical verification (or a goodness metric, for that matter) limit the utility of these concepts to real-world applications where reliability and scientific verification of correctness are paramount.

Vapnik V. 1998. *Statistical Learning Theory*. John Wiley, Indianapolis, Indiana. This book describes the theory and some implementations of support vector machines as an instance of statistical machine learning. The central idea is to use categorical training data to calculate the high-dimensional boundary between classes. This is achieved even for data that is not linearly separable by projecting the observation vectors into a kernel space, which is a user-defined transformed dot product space (usually linear, quadratic, or radial).

Whittington D and H Hunt. 1999. "Approaches to the Computerized Assessment of Free Text Responses." In *Proceedings of the 6th International Computer Assisted Assessment Conference*, ed. Danson M, pp. 207-219. This paper is a survey of several methods developed for automated assessment of free text essays. Though the domain is very different than that of DEWPOINT, automated assessment has led to the development of an interlingual representation of free text that can be used to compare texts of different sources. The reliability of this interlingual representation needs further development to be useful in a context where organizational responses will rely in part on the assessment of text.

Distribution

No. of

Copies

Name
Organization
Address
City, State and ZIP Code

Organization
Address
City, State and ZIP Code
Name
Name
Name
Name
Name (#)

Name
Organization
Address
City, State and ZIP Code

No. of

Copies

 # Foreign Distribution

Name
Organization
Address
Address line 2
COUNTRY

 # Local Distribution
Pacific Northwest National Laboratory

Name	Mailstop
Name	Mailstop
Name	Mailstop
Name	Mailstop
Name	(PDF)

Pacific Northwest
NATIONAL LABORATORY

U.S. DEPARTMENT OF
ENERGY

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)
www.pnl.gov