Towards Applying Text Mining and Natural Language Processing for Biomedical Ontology Acquisition

Tasha R. Inniss
Spelman College
Department of Mathematics
Atlanta, GA 30314-4399
tinniss@spelman.edu

Marc Light
Thomson Legal and Regulatory
610 Opperman Drive
Eagan, MN 55123
marc.light@thomson.com

George Thomas
University of Iowa
Department of Computer Science
Iowa City, IA 52242
george-thomas@uiowa.edu

John R. Lee
Assistive Intelligence, Inc.
Iowa City, IA
52245-3210
AIResearcher@gmail.com

Michael A. Grassi
University of Chicago
Department of Ophthalmology
Chicago, IL 60637
mgrassi@uchicago.edu

Andrew B. Williams
Spelman College
Computer Science Department
Atlanta, GA 30314-4399
williams@spelman.edu

ABSTRACT
The use of text mining and natural language processing can extend into the realm of knowledge acquisition and management for biomedical applications. In this paper, we describe how we implemented natural language processing and text mining techniques on the transcribed verbal descriptions from retinal experts of biomedical disease features. The feature-attribute pairs generated were then incorporated within a user interface for a collaborative ontology development tool. This tool, IDOCS, is being used in the biomedical domain to help retinal specialists reach a consensus on a common ontology for describing age-related macular degeneration (AMD). We compare the use of traditional text mining and natural language processing techniques with that of a retinal specialist's analysis and discuss how we might integrate these techniques for future biomedical ontology and user interface development.

Categories and Subject Descriptors
I.2.6 [Artificial Intelligence]: Learning - *Knowledge acquisition.* I.2.7 [Artificial Intelligence]: Natural Language Processing - *Text analysis.* I.2.11 [Distributed Artificial Intelligence]: Multiagent Systems. I.5.4 [Pattern Recognition]: Applications - *Text processing.* J.3 [Life and Medical Sciences]: Biology and Genetics

General Terms: Design, Experimentation, Human Factors, Standardization

Keywords: Text mining, information extraction, natural language processing (NLP), ontology

1. INTRODUCTION
Acquiring the vocabulary for a biomedical knowledge domain from human experts may benefit from the use of text mining and natural language processing techniques. In certain domains, such as some eye diseases, there is no existing standardized (uniform) vocabulary and classification of these disease variations, or subtypes. That is, several clinicians may observe variations in a particular disease, such as age-related macular degeneration (AMD), but the group of clinicians, often dispersed geographically, does not share a common, agreed upon vocabulary for these diseases. What is needed is a knowledge acquisition method to generate new standardized vocabularies for these subtypes. In this paper, we describe how we developed a user interface by performing a manual, ad-hoc knowledge acquisition and then compare how this approach might be improved by using text mining and natural language processing.

The goal of this research was to extract the feature and attribute descriptions for the vocabulary of AMD, or, more precisely, to produce an ontology specification that could be integrated in a user interface for a collaborative, biomedical ontology development tool called IDOCS (Intelligent Distributed Ontology Consensus System). See [14] for more detailed information about ontologies.

Section 2 will provide a motivation for our work and present related work. Section 3 will discuss the three methodologies (Human Experts, Natural Language Processing, and Text Mining) that were used to determine a vocabulary that could be used to describe AMD. Section 4 will present the results of the three methodologies and will include a comparison of the methods. Section 5 will outline our proposed methodology for a semi-automated ontology generation system. In Section 6, we provide our plans for future research in this area.

2. MOTIVATION AND RELATED WORK
Since the mid-90s, a vast amount of research has been conducted on applying natural language processing techniques in the area of medicine, biomedicine, and molecular biology [12, 24, 28, 30, 33]. It was recognized that natural language processing and text data mining is effective for information extraction. Most of this work focuses on extracting information and knowledge from research literature and abstracts [4, 5, 7, 13, 17, 19, 27, 34] from such online repositories as Medline and PubMed. In some domains, this extraction is referred to as literature mining or web mining [23]. In addition, natural language processing systems such as MedLEE, UMLS, and GENIES, have been developed to

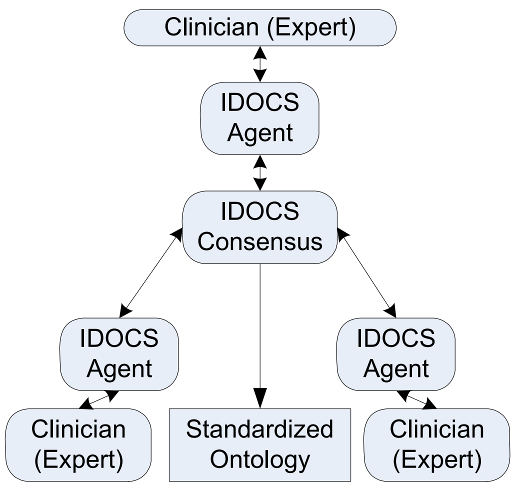
TMBIO'06, November 10, 2006, Arlington, Virginia, USA.

assist in the extraction of specific clinical information [5, 9, 10, 11, 13, 18, 27, 34].

Because of the vast amount of new biological and medical data that is generated daily, current research has focused on the development of biomedical ontologies [2, 29] as well as the development of methodologies for connecting new information to currently existing ontologies [1, 22, 25].

The goal of our paper is to initially determine a common vocabulary that can be used to describe age-related macular degeneration (AMD) via the use of retinal experts, text mining, and natural language processing. In [8], comparisons are made between automated methods and an internal medicine resident for identifying pneumonia. Our research methodology is similar; but differs in that we are describing AMD features. The ultimate goal is to develop an ontology for AMD, which will be the subject of future research.

Figure1. IDOCS Dataflow Diagram



3. METHODOLOGY

3.1 Human Experts

The manual, ad-hoc approach for acquiring a new biomedical ontology involved interviewing experts, transcribing their text, and manually mining the text for feature-attribute pairs that could be incorporated in the user interface for a collaborative, biomedical ontology development tool, IDOCS [35]. We enlisted the participation of four clinical experts in retinal diseases to view 100 sample eye images containing variations, or disease subtypes, of AMD. These retinal experts, who were in different geographic locations, described their observations of the features informally using digital voice recorders. The rationale behind having the experts use dictaphones to record their observations was to allow them to freely associate a description with the analysis of a funduscopic image without the constraints of a pre-ordained vocabulary or knowledge elicitation paradigm. Their verbal descriptions were then transcribed into text. Another retinal clinician then manually parsed the text and extracted all key words which were then organized, using the clinician's domain knowledge, into a structured vocabulary for AMD, with candidate feature names, attribute names for those features and the possible values for those attributes. These feature attributes and values were then incorporated into the user interface of our collaborative biomedical ontology development tool, Intelligent Distributed Ontology Consensus System (IDOCS). Figure 1 provides an overview of the proposed IDOCS dataflow. The overall goal of IDOCS is to involve agents that will assist the human experts in the collaborative ontology process, possibly via a proxy server.

3.2 Natural Language Processing (NLP)

Natural language processing (NLP) is the study of computer processing of human language [21]. Some tasks of NLP that are relevant to our research project are information extraction and automatic summarization. Both of these tasks benefit from being able to identify short sequences of words that have meaning over and above a meaning composed directly from their parts. For example, the sequence of words, "extreme programming," does have something to do with "programming" and "extreme" but also has meaning over and above the simple combination of an adjective and noun as in "extreme cold." Extreme programming refers to a method of programming. Identifying such sequences, known as collocations and idioms, is an established topic of research ([21], See chapter 5 for an overview). For our project, we conducted experiments on the same feature description text generated by the original interviews of the clinicians (eye experts) using a number of collocation discovery methods from NLP. The "Ngram Statistics Package" (NSP) [3] was used to identify two-word (bigram) sequences of text that occur more often than expected. NSP, developed by Santanjeev Banerjee and Ted Pederson, is a "general purpose software tool that allows users to define Ngrams as they wish and then utilize standard methods from statistics and information theory to identify interesting or significant instances of Ngrams in large corpora of text".

3.3 Text Mining

3.3.1 Background and Definition

In the mid-90s, data mining became a prominent and important field for both practitioners and researchers. Data mining can be defined as the process of analyzing large data sets using statistical, pattern recognition, and knowledge discovery techniques to determine meaningful and sometimes subtle trends and information. The term "text mining" was coined since it is a natural extension of data mining and is the extraction (or mining) of patterns, useful information or knowledge from natural language text. The process of text mining is not a new development as it is used in statistical natural language processing and information extraction. According to Hearst [15], text data mining tasks can be classified as (i) question answering (information retrieval), (ii) information extraction, or (iii) thesaurus generation. Text mining can be used to discover prevalent concepts in a collection of documents, to summarize documents, or to classify documents into categories. In this paper, we focus on discovering prevalent concepts with the goal of information extraction and potentially thesaurus generation.

3.3.2 Text Mining Methodology

A collection of documents, called a corpus, is used as input into any text mining algorithm. The corpus is then parsed into tokens ("contiguous string of characters delimited by spaces, punctuation or other character separators [32]) or terms (tokens in a particular language). The unstructured text in the corpus becomes a structured data object via the creation of a term-by-document frequency matrix. Numerical measures can be used to weight certain terms depending on the goal of text mining for a particular project. To address the "curse of dimensionality", mathematical dimension reduction techniques can be used to transform the data.

For our project, the unstructured text is the transcribed interviews of the retinal experts. Since the overall goal is to develop a biomedical ontology, we wanted to discover those concepts that occur most often in the most number of documents. Thus, we simply used "counts" as the frequency weights. In SAS' Text Miner, this frequency weight measure is called "none." Frequency weights are called the local weights. To take into consideration that some documents may be longer than others and thus may have larger local weights, we compute "global weights." Global weights, also called "term weights", are used "to adjust the frequency weights to account for the distribution of terms across documents" [31].

3.3.3 SAS Text Miner

In SAS' Text Miner, there are five different term weighting measures: ***Entropy***, ***Inverse Document Frequency*** (IDF), ***Global Frequency (GF)-IDF***, ***Normal***, and ***None***. ***Entropy*** is the default measure and is most useful if the goal is to discriminate between documents. ***None*** simply assigns each term a global weight of 1. The "normal term weight" is most appropriate for our project since our goal is to determine the most prevalent concepts. The normal term weight is defined as follows:

$$G_i = \frac{1}{\sqrt{\sum_j a_{i,j}^2}}$$

where G_i is the global weight of term *i* and $a_{i,j}$ is the frequency of term *i* in document *j*.

The weighted term-by- document frequency matrix is very sparse when there are more terms than documents and not all terms exist in all of the documents. To reduce dimension yet preserve information, the method of *Singular Value Decomposition* is used.

4. RESULTS AND EVALUATION

4.1 Human Expert Results

A retinal specialist selected 100 representative images from patients who had been examined by an ophthalmologist and found to have signs consistent with the clinical diagnosis of AMD. (The recruitment and research protocols for human subjects were reviewed and approved by the University of Iowa institutional review board and informed consent was obtained from all study participants.) The images were displayed on a user-friendly computer interface for easy navigation and visual assessment by the four retinal experts. Using digital dictaphones, they described in detail the ophthalmoscopic appearance of each image. This method was generally well received by the retinal experts and the results indicate that a broader, more expansive range of data was captured than would have been possible through slower, more cumbersome interfaces. The data from the dictaphones was then transcribed and analyzed. Table 1 below lists the most frequent feature names provided by the four retinal experts.

Our retinal specialist was asked to manually parse the transcripts of the interviews with the retinal experts and all key terms and phrases were extracted. The vocabulary was grouped into a hierarchical structure with candidate feature descriptions, attributes and associated values that appeared to be extensive and descriptive enough to cover the aggregate vocabulary of the clinicians. The goal of the human expert (our retinal specialist) in analyzing the collected data was to be as inclusive as possible over the transcribed dictations and capture every key descriptor such that the final vocabulary was a superset of the key terms used by the retinal experts. A single human expert was deemed sufficient because this extraction phase required only a basic knowledge of the domain, given the overall philosophy described. In addition, this structured vocabulary was finally distributed to the four retinal experts for their review and approval as representative of the spectrum of AMD vocabulary.

The benefit of this approach is that the retinal expert, while not requiring extensive expert knowledge, still has the basic domain knowledge that enables her/him to make informed decisions on categorization of various terms. The results of this manual parsing are based on the categories *Drusen*, *Retinal Pigmented Epithelium*, and *Optic Disk* and presented in Table 1.

Table 1. Results of Expert Manual Parsing of Interviews

Drusen	**Retinal Pigmented Epithelium**	**Optic Disk**
Hard, **Soft****, Calcified, Reticular, Dystrophic, Cuticular/Basal Laminar	granularity, mottling, disruption, figures, clumping, irregularity, intraretinal migration, Hypertrophy, Hypo-pigmentation, Hyper-Pigmentation, Pigmentation	glaucomatous
coalescent, discrete, fine, punctate, distinct, indistinct, homogenous, inhomogenous	Superior, inferior, nasal, temporal	Peripapillary atrophy
Grouped, Clustered, Scattered, Radial, Linear	Drusenoid, serious, chronic	Scleral crescent
Large**, medium, **small****	Geographic, Non-geographic, focal, coalescent, discrete	
few, many, moderate, **confluent****, none, **extensive****, sub-confluent	Peripapillary	
Foveal	Loss of choriocapillaris	

4.2 Natural Language Processing Results

The Ngram Statistics Package (NSP) was implemented on transcribed text to determine word pair associations (bi-grams) measured using log-likelihood ratio and pointwise mutual information (PMI). Log-likelihood is a statistical test of association between two random variables (words), whereas PMI is simply a measure of association since it does not give a value of statistical significance. PMI between two words (bigram or collocation) x and y is defined as

$$PMI(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

where *P(x,y)* is the joint probability of *x* and *y* together and each of *P(x)* and *P(y)* are marginal probabilities. PMI essentially measures the degree of association (statistical dependence) between two words [6]. If there is an association, greater than what would be expected, then PMI would be large.

The results in Table 2 are based on log-likelihood ratios. The null hypothesis of a log-likelihood ratio test is that there is no association between the words beyond chance. Essentially, like a chi-squared test, observed frequencies are compared to expected frequencies. If there is a large difference (statistically significant) between these frequencies, then we can make an inference about the words occurring together more than what is expected.

Table 2. NSP Results Based On Log Likelihood Ratios

Bigram	Log-likelihood Value	Bigram	Log-likelihood Value
Foveal center	471.8231	Geographic atrophy	440.7184
Non geographic	371.0657	Retinal pigment	278.9330
Optic nerve	273.3231	Pigment epithelium	262.6582
Nerve head	230.3569	Drusen throughout	193.9924
Left eye	159.5534	Eye Shows	142.6221
Shows extensive	135.2535	Right eye	123.2690
Poor quality	100.2583	Medium sized	87.8320
Quality image	79.1112	Central macula	77.2148
Confluent drusen**	71.9671	Reticular pattern	71.7646
Small drusen**	66.4513	Large confluent	55.7267
Soft drusen**	44.6217	RPE hyperpigmenta tion	36.7882
Pigment epithelial	33.9625	Dystrophic drusen	32.8880
Extensive large	30.4726	Shows large	30.4041
Large drusen**	27.4051	Sized drusen	27.3609
Extensive small	26.7291	Pigment migration	26.0272
Atrophy between	20.3224	Central geographic	18.6789
Extensive drusen**	17.8943	Atrophy temporal	12.2694

4.3 Text Miner Results

SAS' data mining software, "Enterprise Miner" with the Text Miner node was implemented on the transcribed interviews from the retinal experts.

Recall that the frequency weight used is "none" (counts) and the term weight used is "normal." These measures were chosen since the goal of our analysis is to discover those terms or concepts that are most frequently occurring in the corpus of interviews. Table 3 lists the concepts or "noun groups" that were discovered.

Table 3. SAS Text Miner Results for None-Normal Weights

Noun Groups	Freq/ NumDocs	Weight
atrophy superior, epithelial changes, foveal sparing, hard drusen, large size, mild pigmentary changes, possible geographic atrophy, very large area	2/2	0.7071068
confluent soft drusen**, small areas	3/3	0.5773503
epithelial detachment, nasal aspect, retinal stria	3/2	0.4472136
extensive soft drusen**, few fine	4/2	0.3535534
distinct drusen, poor quality photograph, superior arcade	4/2	0.3162278
discrete drusen, photographic artifact	5/2	0.2425356
few small	6/2	0.2357023
multiple areas	6/2	0.2236068
large number, small number	6/2	0.1961161
large areas	7/2	0.164399
large area	12/4	0.1543034
pigmentary changes	10/3	0.147442
small area	9/2	0.1240347
small drusen**	10/2	0.1104315
central fovea	11/3	0.1097643
scleral crescent	13/2	0.0894427
large drusen**	21/4	0.0755929
temporal aspect	15/2	0.0712471
large soft drusen	18/3	0.0622573
fovea area	18/2	0.058722
peripapillary atrophy	30/4	0.0478913
optic nerve head	24/2	0.0434372
retinal pigment epithelium	33/2	0.0361551

soft drusen**	40/3	0.0335578
pigment epithelium detachment	35/2	0.0293991
retinal pigment	44/2	0.0243252
pigment epithelium	50/2	0.0216574
optic nerve	57/2	0.0191425
right eye	105/3	0.0152623
left eye	113/3	0.0144533
central macula	93/3	0.0137348
geographic atrophy	159/4	0.0099597
posterior pole	111/2	0.00994

4.4 Comparison

Each of the three methods was carried out by a different co-author and was done independently of the other methods to eliminate any bias. We manually parsed the data for EXACT term matches to determine those words that are common to each pair of methods. Note that the list from the retinal specialist was used as the basis for comparison and matching. Words that are common to each pairwise group of methods are listed in Table 4. Terms that are common to ALL three methods are indicated by bolded print and ** symbol.

Table 4. Common Terms for Three Methods

Expert-Retinal Specialist	**NLP-NSP**	**Text Mining-SAS Text Miner**
Hard (drusen)		Hard drusen
Soft (drusen)	**Soft drusen**	**Soft drusen** Large soft drusen Extensive soft drusen Confluent soft drusen
Reticular (drusen)	Reticular pattern	
Dystrophic (drusen)	Dystrophic drusen	
Discrete (drusen)		Discrete drusen
Fine (drusen)		Few fine
Distinct (drusen)		Distinct drusen
Large (drusen)	**Large drusen** Large confluent Extensive large Shows large	**Large drusen** Large size Large area(s) Large number Large soft drusen Very large area
Medium (drusen)	Medium sized	
Small (drusen)	**Small drusen** Extensive small	**Small drusen** Small area(s) Small number
		Few small
Few (drusen)		Few fine Few small
Confluent (drusen)	**Confluent drusen** Large confluent	**Confluent soft drusen**
Extensive (drusen)	**Extensive drusen** Extensive large Extensive small Shows extensive	**Extensive soft drusen**
Foveal (drusen)	Foveal center	Foveral sparing
Temporal	Atrophy temporal	Temporal aspect
Superior		Superior arcade Atrophy superior
Nasal		Nasal aspect
Geographic	Geographic atrophy Central geographic	Geographic atrophy Possible geographic atrophy
Non-geographic	Non-geographic	
Peripapillary Peripapillary atrophy		Peripapillary atrophy
Scleral crescent		Scleral crescent
	Optic nerve	Optic nerve
	Left eye	Left eye
	Right eye	Right eye
	Central macula	Central macula

Table 5. Frequency of Common Terms for Three Methods

Human Expert & NLP	**Human Expert & Text Mining**	**NLP & Text Mining**	**All Methods**
7	10	10	5

Table 5 above gives the numbers of words in common for each pair of methods. Based on the results stated in Table 5, it is evident that text mining is a viable and effective method for determining vocabulary that could be used to describe a particular disease. It should also be noted that text mining found some of the terms that natural language processing found, though there are more when compared to the Human Expert Method.

5. PROPOSED SEMI-AUTOMATED ONTOLOGY GENERATION

Based on the results of this paper, we propose a methodology to generate an ontology in a semi-automated manner using human experts, natural language processing, text mining and a user-interface. (See Figure 2 below)

We conjecture that this approach would result in the generation of a more uniform, standardized vocabulary that could be used to describe attribute features of any given disease.

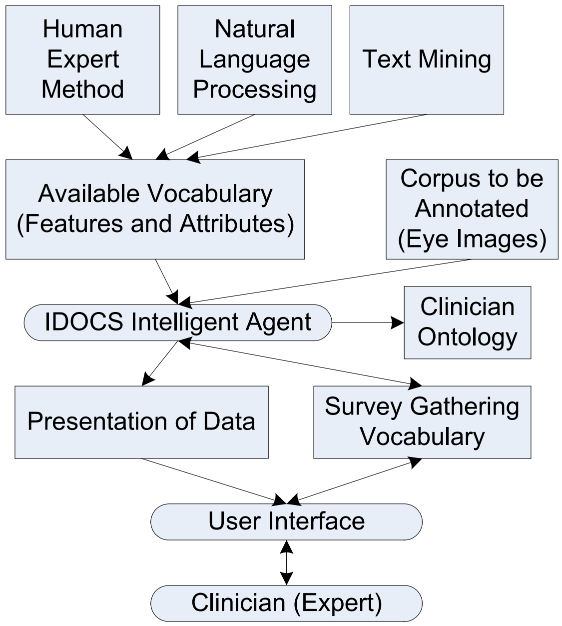


Figure 2. Semi-Automated Ontology Generation Dataflow

6. CONCLUSION AND FUTURE WORK

The Human Expert results were the best but we plan on comparing how the text mining and natural language processing results might enhance the analysis and generation of feature descriptions. It is anticipated that a more robust vocabulary can be generated. Based on analysis of our current IDOCS tool, it is believed that some key vocabulary descriptors were missed in the human analysis of the feature description text. Complementing the human analysis with text mining and natural language processing may prevent this from happening in the future.

Our ultimate goal is to develop an ontology of feature descriptions of AMD using the three methodologies implemented in this paper. An extension of this work is to evaluate the effectiveness of the automated tools, NLP and text mining. For information retrieval, the performance measures that are used are precision and recall. To determine if resulting sets of terms from text mining are the most effective for describing a disease, several similarity measures such as cosine similarity can be used. In [20], Inniss developed a clustering technique in which she evaluated several measures of (dis)similarity using an effectiveness measure she developed. It should be noted that results of SAS text miner will differ based on the weighting schemes that are used. Future work will focus on comparing and evaluating these different results for the one that is most effective.

5. ACKNOWLEDGMENTS

This work was supported by NIH grant EY013688-03. Dr. Grassi was a Heed Fellow from 2003-2005. Thanks to Kendra Maxwell and Zijan Ren for technical assistance. This investigation is currently supported by the Spelman College Award from MCMHD grant no 5P20MD000215.

7. REFERENCES

[1] Aronson, A.R. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proc AMIA Symp*, (2001), 17-21.

[2] Ashburner, M., et al. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*, 25 (2000), 25-29.

[3] Banerjee, S. and Pederson, T. The Design, Implementation and Use of the Ngram Statistics Package. *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. (Feb 2003)

[4] Bruijn, B. and Martin J. Getting to the (c)ore of Knowledge: Mining Biomedical Literature. *International Journal of Medical Informatics*, 67, 1-3, (2002), 7-18.

[5] Chen, L. and Friedman, C. Extracting Phenotypic Information from the Literature via Natural Language Processing. In *MEDINFO 2004* (M. Fieschi et al., eds), 758-762.

[6] Church, K. and Hanks, P. Word Association Norms, Mutual Information, and Lexicography. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics,* 16, 1 (1990).

[7] Eom, J.-H. and Zhang, B.-T. PubMiner: Machine Learning-based Text Mining for Biomedical Information and Analysis. *Genomics and Informatics*, 2, 2, (2004), 99-106.

[8] Fiszman, M., Chapman, W.W., Aronsky, D., Evans, R.S., and Haug, P.J. Automatic Detection of Acute Bacterial Pneumonia from Chest X-ray Reports, *Journal of the American Medical Informatics Association*, 7, (2000), 593-604.

[9] Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., and Johnson, S.B. A General Natural-Language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, 1, 2, (1994), 161-174.

[10] Friedman, C., Cimino, J.J., and Johnson, S.B. A Schema for Representing Medical Language Applied to Clinical Radiology. *Journal of the American Medical Informatics Association*, 1, 3, (1994), 233-248.

[11] Friedman, C. Towards a Comprehensive Medical Language Processing System: Methods and Issues. *Proc AMIA Symp*, (1997), 595-599.

[12] Friedman, C. A Broad-Coverage Natural Language Processing System. *Proc AMIA Symp*, 19, 19, (2000), 270-274.

[13] Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles. *Bioinformatics*, 17, Suppl 1, (2001), S74-S82.

[14] Gruber, T.R. A Translation Approach to Portable Ontologies. *Knowledge Acquisition*, 5 (1993), 199-220.

[15] Hearst, M. Untangling Text Data Mining. *Proceedings of ACL'99: the 37th Annual Meeting of the Association of Computational Linguistics* (1999).

[16] Hearst, M. *What is Text Mining?* http://www.sims.berkeley.edu/~hearst/text-mining.html

[17] Hobbs, J.R. Information Extraction from Biomedical Text. *Journal of Biomedical Informatics,* 35, (2002), 260-264.

[18] Humphreys, B.L., Lindberg, D.A., Schoolman, H.M., and Barnett, G.O. The Unified Medical Language System: an Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 5, (1998), 1-11.

[19] Humphreys, K., Demetriou, G., and Gaizauskas, R. Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures. *Pacific Symposium on Biocomputing*, (2000), 505-516.

[20] Inniss, T.R. Seasonal Clustering Technique for Time Series Data. *European Journal of Operational Research*, In Press. Available online 10 August 2005 at Science Direct.

[21] Jurafsky, D. and Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, New Jersey, 2000.

[22] Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. GENIA corpus- a Semantically Annotated Corpus for Bio-textmining. *Bioinformatics*, 19, Suppl 1, (2003), i180-i182.

[23] Kosala, R. and Blockeel, H. Web Mining Research: A Survey. *Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD Explorations)*, 2, 1 (2000), 1-15.

[24] Krallinger, M., Erhardt, R. A.-A., and Valencia, A. Text-Mining Approaches in Molecular Biology and Biomedicine. *Drug Discovery Today*, 10, 6 (March 1995).

[25] Lussier, Y., Borlawsky, T., Rappaport, L.Y., and Friedman, C. PhenoGO: Assigning Phenotypic Context to Gene Ontology Annotations with Natural Language Processing. *Pacific Symposium on Biocomputing*, (2006), 64-75.

[26] Manning, C.D. and Schutze, H. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, Massachusetts, 2002.

[27] Novichkova, S., Egorov, S., and Daraselia, N. MedScan, a Natural Language Processing Engine for MEDLINE Abstracts. *Bioinformatics*, 19, 13, (2003), 1699-1706.

[28] Raychaudhuri, S., Schutze, H., and Altman, R.B. Using Text Analysis to Identify Functionally Coherent Gene Groups. *Genome Research*, 12, (2002), 1582-1590.

[29] Rosse, C. and Mejino, J.L. A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36, 6, (2003), 478-500.

[30] Sager, N., et al. Natural Language Processing and Representation of Clinical Data. *Journal of the American Medical Informatics Association*, 1, (1994), 142-160.

[31] SAS Institute Inc. *Mining Textual Data Using SAS Text Miner for SAS® 9*. SAS Institute Inc., Cary, North Carolina, 2004.

[32] SAS Institute Inc. *Getting Started with SAS® 9.1 Text Miner*. SAS Institute Inc., Cary, North Carolina, 2004.

[33] Spyns, P. Natural Language Processing in Medicine: an Overview. *Methods Inf Med*, 35, 4-5, (1996), 285-301.

[34] Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. Automatic Extraction of Protein Interactions from Scientific Abstracts. *Pacific Symposium on Biocomputing*, (2000), 541-552.

[35] Williams, A.B., Krygowski, T., and Casavant, T. I-DOCS: Distributed Agent-Assisted Knowledge Fusion for Disease Gene Discovery. *Proceeding of the Eighth International Conference on Parallel and Distributed Systems* (Kyongju, ,Korea, June 26-29 2001). IEEE Computer Society Press, 698-70.