THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

A research division of the National Library of Medicine

TECHNICAL REPORT
LHNCBC-TR-2004-002

Modeling and Learning Methods;
A Report to the Board of Scientific Counselors

May 2004

Mehmet Kayaalp, M.D., Ph.D.

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

Modeling & Learning Methods

A report to the Board of Scientific Counselors

May 13, 2004

Investigator:
Mehmet Kayaalp, M.D., Ph.D.
Cognitive Science Branch

Lister Hill National Center for Biomedical Communications
National Library of Medicine
National Institutes of Health
Department of Health and Human Services

Executive Summary

Modeling is the most important and arguably the most difficult intellectual process in a scientific endeavor. Identification, conceptualization, and communication of scientific problems, as well as formation, testing, and evaluation of hypotheses are all performed utilizing models. The scientific modeling process depends on reliable, high quality, and, sometimes, high volumes of information. Advances in computing and networking technologies within the last decade and the emergence and ubiquity of the Internet have drastically increased the growth of information as well as its availability. On the other hand, we have not developed as rapidly the necessary modeling methods.

The Modeling & Learning Methods (MLM) project at the Lister Hill Center of the National Library of Medicine® (NLM®) is aimed towards addressing this very problem. Its goal is to develop new modeling methods that enable researchers to rapidly construct effective computational models from large datasets.

The objectives of the project are to develop machine learning methods that automate the process of constructing probabilistic models for (1) identifying relevant information among large datasets and corpora, (2) mapping identified information to networks of ontologies, (3) accessing queried information accurately, and (4) answering user queries through mining the data located in heterogeneous information sources. Interest in probabilistic models ranges over a wide spectrum of biomedical fields, including computational biology; bio-, clinical, and healthcare informatics; and epidemiology.

The objectives of the project will be evaluated with a set of suitable metrics such as receiver operating characteristics (ROC) that measure the performance of prospective models in terms of sensitivity and specificity in reaching their target functions. Depending on the domain of the models and the problems of interest, domain subjects and/or experts might be needed to determine the gold standards or the target functions for the performance evaluations of the models if such gold standards or target functions are not readily available.

1 Introduction

For thousands of years, from antiquity to the end of the twentieth century, libraries have kept their identities as being physical locations of book collections. We are now witnessing a rapid change in that characteristic of libraries as well as in our modes of interaction with information.

What awaits us in the near future? Practically all written material will be digitally stored and most will be accessible. Numerical data are being collected at increasingly many points of social, scientific, and business processes. Soon, measuring and recording almost everything numerically will be technologically, financially, and logistically feasible.[1] Libraries will become virtual locations, connected to a network of physical data repositories, providing information services to consumers through intelligent software agents.

Today, at this early stage of the digital information revolution (DIR), we struggle with problems related to rapid changes in our professional endeavors. Without having proper methods and tools, we might be overwhelmed with staggering volumes of data when we are in need of extracting critical information. This sentiment is shared most strongly

[1] Today many private healthcare companies measure a number of physiological parameters of their customers in real-time 24 hours a day.

among groups (such as US intelligent agencies and US Department of Defense) that collect data more intensely and depend on timely information more urgently than others.

Scientists in the US are at the frontline of DIR. The National Institutes of Health (NIH) began the Biomedical Science and Information Technology Initiative (BISTI)[2] and then reaffirmed its commitment with a set of roadmap initiatives through which intra- and extramural scientific activities are going to be transformed from their conventional mode of operations, which were suitable before DIR, into a new set of interactions (Zerhouni, 2003).

Digital information is at the center of the new scientific endeavor and, if managed carefully, it may bridge scientists and clinicians across disciplines. Biological data collected by scientists from a particular discipline are practically useless for other scientists if data are not properly interpretable. The same is true for corpora of scientific articles. Scientists need enabling tools to overcome cross-disciplinary communication barriers. The desirable property of such tools is the transferability of data from one model to another based on the questions and terminology of the scientist at the other end of the communication line. Such tools require new methods for learning models from different datasets with heterogeneous data models, for which we need a thorough understanding of the fundamentals of modeling.

2 Project Objectives

The main mission and goals of the National Library of Medicine (NLM) may be characterized in part as acquiring, organizing, and disseminating health-related information to researchers and the general public (National Library of Medicine (U.S.) Board of Regents, 2000). Accordingly, the aim of the MLM project is to develop methods as the bases of new technologies that would support users of NLM services in their pursuit of biomedical knowledge acquisition.

The ultimate goals of any intelligent digital library system may be summarized in four steps: (1) interpret queries and the needs of users accurately, (2) identify all relevant information among a comprehensive set of sources, (3) evaluate and combine relevant information, and (4) compose answers according to the needs of the users.

---

[2] The fundamental objective of BISTI is: "To make optimal use of information technology, biomedical researchers need, first of all, the expertise to marry information technology to biology in a productive way. New hardware and software will be needed, together with deepened support and collaboration from experts in allied fields. Inevitably, those needs will grow as biology moves increasingly from a bench-based to a computer-based science, as models replace some experiments and complement others, as lone researchers are supplemented by interdisciplinary teams. The overarching need is for an intellectual fusion of biomedicine and information technology" (BISTI, 2000).

These goals have led the formation of the MLM project objectives, which are to develop methods to

(1) identify information blocks in the existing datasets and corpora,

(2) integrate heterogeneous datasets by mapping identified information to the concepts of relevant ontologies,[3]

(3) access the requested information with high accuracy, and

(4) learn inferential models from data based on user queries and information obtained in Step (3), and draw inferences that are necessary to synthesize answers to user queries.

2.1 Identifying Information Blocks

In order to transform data from one model to another, one must identify relevant information embedded in the model. Information identification is followed by labeling or tagging data and forming a well-defined variable set.

The importance of information identification may be best understood when the dataset is a corpus of text. For analyzing textual information, the information unit is reduced from documents to sentences and sentence substructures. Necessary tools in this process include tokenizers, lexicons, spell-checkers, part-of-speech (POS) taggers, parsers, and tools for structural and semantic disambiguation. Given the fact that the research in computational linguistics is quite involved and requires specific expertise, we intend to resolve such problems with existing tools whenever they are available.

2.2 Mapping Information to Ontologies

The high volume of information prohibits manual accounting and organization of the relationship between pairs of concepts. We need an ontological approach to integrate heterogeneous information located in various information sources so that we can connect conceptually related information. Instead of committing to a single ontology, we intend to use all available relevant ontologies in parallel. Further details on our approach can be found in the Section "Methods and Procedures."

Because the Unified Medical Language System® (UMLS®) contains a map of concepts linked to a number of information sources, it will be our starting point, but additional biomedical information sources are also needed. For example, an identified genetic information block should also be mapped to concepts and records located in LocusLink, SwissProt, online databases about gene-expression, and relevant entries of dictionaries and established knowledge sources such as the Genetics Home Reference.

[3] Here, any conceptual organization (i.e., pairs of concepts connected through well-defined relations) is considered as an ontology.

Our objective is to build a progressively growing map of (1) online information sources, (2) their ontologies (conceptual organizations), (3) concepts found in those sources, and (4) data associated with those concepts.

2.3 Accessing Information Accurately

Conventional ad hoc information retrieval (IR) systems rely on inverted indexing and vector space model, where the tokens usually are words with no conceptual relationship between them (Salton, 1989). This approach is efficient for locating well-specified data points across documents; however, it also is conceptually shallow and does not comprise information about the data.

Empirical evidence suggests that information about data can make a significant difference in accuracy of information access (Srinivasan & Rindflesch, 2002; Kayaalp et al., 2003). The Google experience might be another good example—the number of external links pointing to a particular website has been found to be valuable information in ranking websites.

Information access cannot be as accurate as it should be, if the tools and their underlying methods are limited to the representation of raw data and ignore information about data. Indexing conceptual information on an ontological backbone is arguably a better architecture for information access than conventional IR methods, as it would enable us to collect and locate information based on conceptual neighborhood, i.e. conceptual similarity and distance.

Our objective is to develop probabilistic methods and metrics for accurate information access that effectively utilize ontological and information theoretic evidence and extend the capabilities of conventional ad hoc search strategies.

2.4 Learning Models from Data

Accessing information accurately is a necessary objective but it is not an end goal. In many cases, available information in its original composition may not be adequate. Our fourth objective in the MLM project is to develop machine learning and data mining methods for extracting new or implicit information that is sought by users.

Given the fact that today's scientists are highly specialized in their domains using distinct methods and vocabularies, information exchange between scientists across disciplines becomes a major communication problem. We believe that the learning methods that we plan to develop would serve scientists (and the general public) as tools for extracting necessary information in their vocabulary from data and documents created in other fields of research.

3 Significance

The digital information revolution confronts us with an information explosion that is currently managed poorly due to a lack of necessary methods and tools. The objectives of this project represent a plan for developing methods that would help us manage information through conceptual organizations and with intelligent agents dedicated to data analysis, modeling, data mining, and inference.

As the leading biomedical library in the world, NLM will inevitably be required to provide analytical data mining and comprehensive knowledge support in the future. The outcomes of this project will form the methodological bases for some of the intelligent services of tomorrow's digital libraries.

4 Background

Modeling and learning are two central concepts of this project; thus, this section is devoted to defining essential terms related to modeling and learning and to provide a high-level perspective of the domain and its processes.

4.1 Modeling

> *The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have* ***meaning****; that is they refer to or are correlated according to some system with certain physical or conceptual entities.* (Shannon, 1948)

Meanings emerge when they are represented in one form or another and become an object of discourse. Modeling refers to this representation process, of which outcomes are called *models*. A scientific model consists of a structure and, frequently, a set of parameters inferred from data.[4] The structure of a mathematical model usually comprises a set of variables and a set of relations defined on the variable set. The structure provides an orientation and context to the data, which then becomes interpretable and may be called information. Hence, models are central objects of scientific communications (see Figure 1).

Models[5] are representations of *systems*, which may constitute tangible entities (e.g., cells, bricks, and buildings), intangible entities (e.g., idea, time, and dimension), or both. Here,

[4] Axiomatic mathematical models may not contain external data associated with the structure until they are applied.

[5] "To an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A" (Minsky, 1965). The object A can be generically called a "system".

we generically call Subject A in Figure 1 the *designer*, and Subject B the *interpreter* or the *user*. An interpreter may also be the designer of the interpreted model, as in the case of editing the model that he/she has created at the first place (e.g., writing and editing a letter).[6] Interpreters and designers are collectively called "agents". An agent may be a living organism (e.g., a person), a device (e.g., a radio), or software.

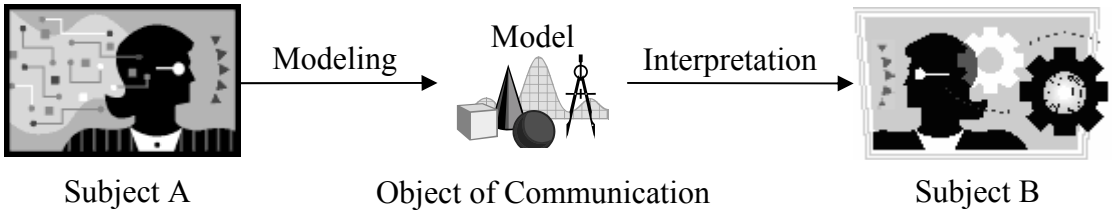

Figure 1: The Essential Role of Models in the Scientific Communication

Bidirectional information exchange is a particularly important mode of communication using models in various forms, such as biological signals in Nature, or spoken languages in human discourse (see Figure 2). In biological cases, models M_A and M_B may be two different sets of biochemical compounds such that M_A produced by the agent $\mathcal{S}_A$ may be interpreted by the agent $\mathcal{S}_B$ through its dedicated receptors, and may cause $\mathcal{S}_B$ to produce M_B. In human discourse, an agreement between subjects $\mathcal{S}_A$ and $\mathcal{S}_B$ may be reached when the semantic gap between M_A and M_B is narrowed to the extent that both $\mathcal{S}_A$ and $\mathcal{S}_B$ are satisfied with. In knowledge acquisition, a user $\mathcal{S}_A$ who submits a query M_A might be satisfied with the response M_B produced by a digital library system $\mathcal{S}_B$, if M_B is comprehensive enough and the semantic gap between the intent of the query and the interpretation of the context of M_B is negligibly small.

Although models are not necessarily formal in nature, we here limit our discussion to constructing formal computational models, but we should be able to interpret informal models such as natural language sentences. A model is formal if it is well-defined such that it is interpreted uniformly by all intended agents that are knowledgeable about the discipline and its standards based on which the model is formulated.

[6] Thought formation may be modeled as an iterative feedback loop of revising a model through continuous interpretation, where the capabilities of representation tools (e.g., the language as in Whorfian hypothesis (Whorf, 1940)) and modeling methods (as seen in the evolutions of many art forms such as painting, music or dance) might be the limiting factors of the intellectual process. An obvious corollary of this hypothesis is that we may improve our intellectual capacity (or its realization) by improving our modeling tools and methods. Accordingly, science, society, and our conceptualization of the world are being transformed (once again (Kuhn, 1996)) through the rapid adoptions of computers and computational methods.

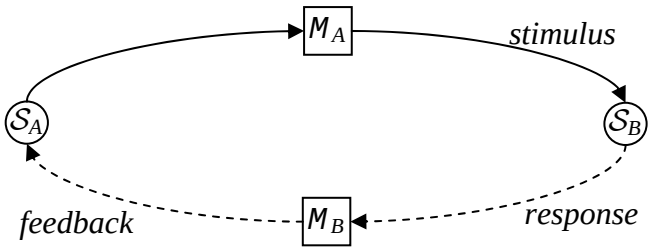

Figure 2: Bidirectional Communication and Control in Nature

Modeling involves a conceptualization process, which generally constitutes at least one of the following three simplification methods: abstraction, aggregation, and idealization (Carson, Cobelli, & Finkelstein, 1983).

Abstraction is the elimination of system components that are nonessential to the model and selection of those components that enable the model to best convey the intent of its designer. In computational modeling, the abstraction process may also be called reduction, variable elimination, or feature selection. For example, a patient may be associated with hundreds of variables but usually only a small portion of them is needed to model a patient in the context of the underlying illness.

Aggregation is about the level of detail or granularity that is appropriate to the modeling task. In biomedical modeling, for instance, questions such as "should we consider a single cell, a cluster of cells with the same type, a tissue, an organ, a physiological system, an organism, or a population of organisms?" must be answered before constructing a functional model. In computational modeling, given a variable, the level of aggregation determines the range of the variable.

Idealization is a set of assumptions on the system that are known to be slightly inaccurate or that do not always hold. Assuming our planet is geometrically a sphere (instead of an ellipsoid) is a typical idealization that is frequently made in our conversations.[7] Assumptions about random variables, such as they are independent and identically distributed, or about time series, such as they are stationary, are typical examples of idealization in statistics.

In the context of computational models, the necessity of simplification techniques can be analyzed in three dimensions:

1. Cognitive load. Simplification techniques may alleviate cognitive load of the designer and may prevent some complexity related errors. By yielding clarity in the

[7] The sphere model of earth obviously abstracts out anything below the surface and aggregates all types of terrain structures into a uniformly smooth surface.

communication, the simplicity reduces cognitive load of the consumer of the model.
2. Computational complexity. Model size (e.g., number of variables and variable interactions) usually determines the computational load in terms of execution time and storage capacity.
3. Sample size. Small models with sparse interactions do not require large samples for reliable parameterization.

Everything should be made as simple as possible, but not simpler (Albert Einstein).

There is always tension between simplicity and complexity. Complex problems generally require complex models that cannot (and ought not to) be simplified further. Such problems are usually problems that we do not understand in depth and detail. Unlike engineering problems, in which we frequently deal with human artifacts for which complete blueprints are available, biomedical problems are defined on complex systems about which we know much less compared to what we know about engineering systems. In other words, biomedical problems frequently necessitate complex computational models (Kayaalp & Sullins, 1992), which usually are hard to construct, hard to compute, and require a large number of observations (i.e. large samples). These arduous prerequisites hinder the biomedical modeling process.

4.2 Learning

How can we overcome these barriers that face the modeling process? Arguably, the most important barrier of all is the cost associated with expertise and the manual nature of model construction. Unlike computational power and the growth of our data collection capabilities, expertise does not increase exponentially and constitutes the major bottleneck of the modeling problem. Fortunately, recent advances in machine learning, shortly referred to as *learning*, enable us to automate the process of model construction from data.

Contemporary learning techniques extensively use mathematical logic, statistics, information theory and other disciplines of mathematics. So, we define learning as an *algo-*

rithmic process of constructing models from data and metadata[8] *using methods of mathematics and artificial intelligence* (see Figure 3).[9]

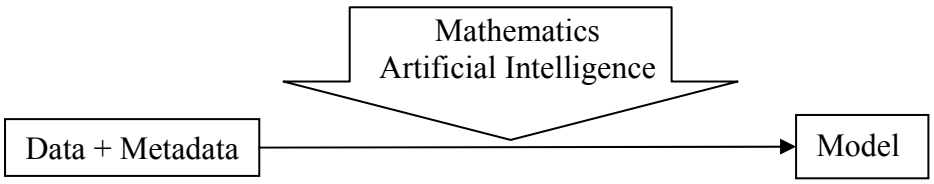

Figure 3: The Process of Learning (Model 1)

The model of the learning process in Figure 3 has three explicit components:

1. inputs (an aggregation of data and metadata),
2. output (model), and
3. methods (implicitly extracted from methods of mathematics and artificial intelligence (AI)).

It also has a number of implicit components, three of which are

1. a designer $\mathcal{D}$ with domain knowledge,
2. a set of learning algorithms $\mathcal{A}$, and
3. a machine $\mathcal{C}$ (e.g., computers or robots) executing $\mathcal{A}$.

The model of the learning process in Figure 3 presumes (idealizes) that the data and metadata were separately available to both the designer and to the learner as inputs. In reality, the designer needs to either collect that data and metadata through direct observation of the data-generating system (e.g., measurements of a specimen or its processes by a biologist) or to extract them from another set of models in which both data and metadata are embedded (e.g., measurements of a specimen presented in a scientific article). We here call such models *prior models*. A model of the learning process that represents the modeling problem—the interpretation of the prior model along with explicit accounts of $\mathcal{D}$, $\mathcal{A}$, $\mathcal{C}$ —is illustrated in Figure 4.

[8] Metadata constitutes information about the data such as names, descriptions, types, and range of variables, about the method and conditions of the data collection process as well as the characteristics of the model in which data are represented, and any other pertinent information such as prior knowledge about the domain and nature of the data-generating system.

[9] Other definitions of learning focus on particular aspects of learning such as the mechanical (search (Turing, 1950)) and the behavioral (improving the performance of the learner (Buchanan, Mitchell, Smith, & Johnson, 1978; Simon, 1989; Russell & Norvig, 1995; Langley, 1996; Mitchell, 1997), or increasing knowledge and improving skills of the learner (Cohen & Feigenbaum, 1982; Buchanan & Shortliffe, 1984)).

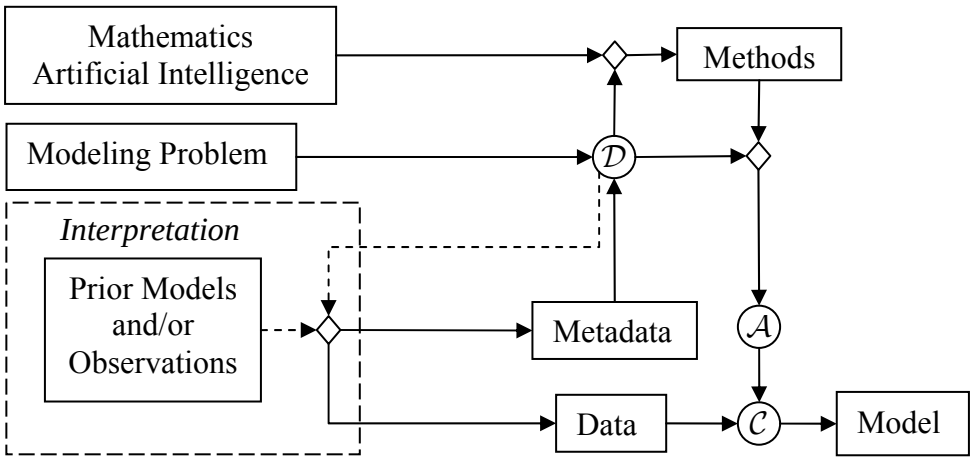

Figure 4: The Process of Learning (Model 2)

Given a well-defined modeling problem, a performance metric, and a sufficiently large random sample of labeled examples (data), intelligent methods may be capable of learning abstraction, aggregation, idealization (AAI), and relations (interactions) between variables; that is, a learner can search a well-defined hypothesis space of AAI and variable relations by evaluating each hypothesis based on the given performance measure and conclude with the hypothesis that maximizes the performance metric within the allotted time frame. In real world problems, the hypothesis space is usually so large that it cannot be searched exhaustively; thus, a proper heuristic search must be devised.

5 Methods and Procedures

The study area of modeling and learning is vast. We here can address only a small portion of the methodology and focus our attention on methods that are most relevant to the objectives of this project. The following subsections, therefore, describe the research direction of the MLM project. The prospective research is described with references to a number of other relevant methods, some of which promise to be the foundational bases of the new methods to be developed in the MLM project.

5.1 Methods for Identifying Information in Data

A major portion of useful scientific knowledge is found in scientific articles, which are difficult to analyze through computational methods. Fortunately, computational linguistics is becoming an increasingly mature field offering a wide range of methods and tools for syntactic and semantic analysis of scientific corpora.

As the initial step toward reaching the first objective of the project, we plan to apply existing methods and tools to extract as much information as possible from MEDLINE®

and other available biomedical corpora. Given the history and expertise of the Cognitive Science Branch (CgSB) in UMLS (McCray, 1991) and that in the Indexing Initiative (Aronson et al., 1999), the methods and tools developed at CgSB will be used in the first phase of the project. These tools include the SPECIALIST lexicon (McCray, Srinivasan, & Browne, 1994), the SPECIALIST parser (Rindflesch, Rajan, & Hunter, 2000), the semantic network (McCray & Nelson, 1995), other UMLS resources (McCray et al., 1993; Rindflesch & Fiszman, 2003), and the MetaMap (Aronson et al., 1999). We will also utilize Xerox's part-of-speech tagger (Cutting, Kupiec, Pedersen, & Sibun, 1992) . For identifying information blocks in MEDLINE, we currently use MetaMap, which coordinates inputs and outputs of various software packages and tools and provides the desired information. For example, given the raw sentence from a MEDLINE abstract, "Specific steroid antibodies, by the immunofluorescence technique, regularly reveal fluorescent centrioles and cilia-bearing basal bodies in target and nontarget cells", here below is the first portion of a simplified version of the processed output in which identified information blocks were tagged.

<phrase>
<NP>**Specific steroid antibodies**</NP>
<POS>adj adj noun</POS>
<MetaMap>
<Concept>*Specific antibody*</Concept>
<UMLSconceptID>C0443640</UMLSconceptID>
<SemanticType=T116>Amino Acid, Peptide, or Protein</SemanticType>
<SemanticType=T129>Immunologic Factor</SemanticType>
</MetaMap>
<MetaMap>
<Concept>*Steroids*</Concept>
<UMLSconceptID>C0038317</UMLSconceptID>
<SemanticType=T110>Steroid</SemanticType>
</MetaMap>
</phrase>
<phrase>
<Conj>**by**</Conj>
<POS>conj</POS>
</phrase>
<phrase>
<NP>**the immunofluorescence technique,**</NP>
<POS>det noun noun punc</POS>
<MetaMap>
<Concept>*Fluorescent Antibody Technique*</Concept>
<UMLSconceptID>C0016318</UMLSconceptID>
<SemanticType= T059>Laboratory Procedure </SemanticType>
</MetaMap>
</phrase>

We will further process the output with additional methods and tools such as structural and semantic disambiguation methods. Earlier, we had developed a prepositional phrase

attachment (PPA) disambiguation method, which we plan to further improve, test and apply (Kayaalp, Pedersen, & Bruce, 1997). It is a probabilistic approach to PPA resolution based on decomposable graphical models called Markov fields and decision lists (Rivest, 1987). Its performance was comparable to the back-off model (Collins & Brooks, 1995) and it was more accurate than any other previously developed method tested on the same datasets.

After completion of the information identification with in-house natural language processing (NLP) systems, we plan to augment that output with external resources and tools that are reliable, stable, and readily available (Friedman, Kra, & Rzhetsky, 2002; Hirschman, Morgan, & Yeh, 2002; Hobbs, 2002; Grishman, 2003).

Identifying information blocks in data is a task that can and should be progressively improved over time by using additional tools, some of which (e.g., probabilistic parsers) discover new information blocks whereas others (e.g., different POS taggers) preprocess the data with different methods. Our metadata description language, XML, enables us to keep outputs of different preprocessors of the same type (e.g., two different word sense disambiguation systems) at the same physical location. Both agreements and disagreements of multiple NLP system outputs stored on the same physical location of the text are expected to facilitate a more robust interpretation of the data.

The depth of information that needs to be identified from text would form a long list of desiderata. Below are a number of questions that would interest us most:

- Given a sentence *S* with a set of qualifiers, relative clauses, a list of objects or subjects, how can we split *S* into a sequence of simpler sentences (s_1,…,s_n)?
- What would be the dependency structures between such sentences? What types of dependencies are they?
- Can we identify subject, verb, object structures?
- Can we identify pronouns and other referential structures from discourse and replace them with their actual references?
- Does the sentence state a factual observation or a belief? Can we assign a value to the author's belief of that statement that we can perceive?
- What are the other possible ways to dissect or interpret a given sentence? Can we assign a confidence value to each one of a possible set of NLP outcomes?

Given the complexity of each linguistic problem in the above list, we cannot expect any single tool to provide a reliable answer. However, we might improve the robustness of the linguistic decision making by combining outcomes of all established tools that are available.

5.2 Methods for Mapping Information to Ontologies

Ontological approaches usually are conceptualizations of the world in a single uniform structure. The resulting systems such as CYC (Lenat & Guha, 1989) might be called common or unified ontologies. Such ontological systems are not only difficult to construct, maintain, and scale, but they are also destined to be inadequate for many people, since people do not have uniform views on any given concept. As every design reflects the view of its designer, every ontology represents a particularly biased view of a small section of the world, which we here call a *facet* (Kayaalp & Sullins, 1993).

Our approach is to include many relevant ontologies in an ontological network, in which ontologies are loosely connected to each other (see Figure 5). The only essential bridge between concepts of two ontologies is the identity relationship, which may be labeled as an *is* relation. This process is also known as ontology alignment (Burgun & Bodenreider, 2001). Although establishing such relationships is desirable, an *is* relation may not be established between two given ontologies. In such cases, or in other cases where it is deemed to be necessary, other relationships such as *is-a* and *part-of* may be used to bridge concepts of two ontologies directly or indirectly. In the latter case, a new parent node is introduced; e.g., two distinct entities $X \in Ontology_1$ and $Y \in Ontology_2$ may be found in the same anatomic location or in a physiological compartment $Z \notin \{Ontology_1, Ontology_2\}$, where *Z* may be represented as the new parent node of *X* and *Y* through *part-of* relations.

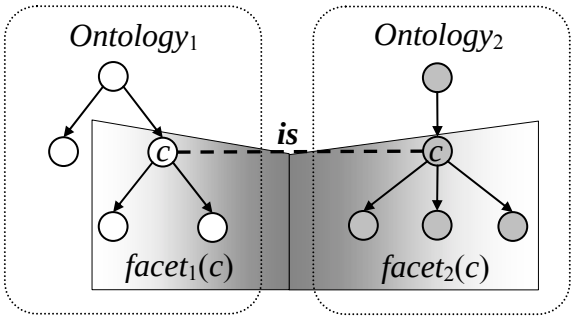

Figure 5: A Set of Loosely Connected Ontologies

Since different ontologies may constitute different facets, the resulting system may be called a *multifaceted ontological network* (Kayaalp, 1993). The underlying method of multifaceted ontological networks, muON, is based on an altruistic design philosophy. With minimal restrictions, muON incorporates all relevant views. Although each facet ought to be consistent within its perimeter, muON does not prohibit inconsistencies between facets. The method is pragmatic as well, since the only requirement of adding a new ontology to the network is identifying identical concepts between the new and an

existing ontology. Even that requirement can be relaxed by establishing other relations as described above.

These properties make muON a practical approach for co-representing and bridging multiple ontologies. As muON is developed further and becomes operational, it would help us to translate data from one model into another and facilitate communication across disciplines.

Although co-representing or aligning ontologies is relatively easy to achieve, establishing reliable communications across ontologies is not and requires formal specifications of concepts. Each concept node in the network may be represented in a frame of attributes, which should be specified using a controlled vocabulary. Each concept is a variable and may be associated with a set of values that may be distributed over a number of databases. In order to utilize the data, one needs to know precisely the levels of abstraction and aggregation of the variable. Furthermore, the particular conditions and assumptions under which the values of the variable are collected need to be specified in details. Perhaps the most challenging portion of this research objective is to establish or adopt[10] a formal specification language that is expressive enough to define a wide-range of idealizations that are frequently made in biomedical data collection processes.

5.3 Methods for Accurate Access to Information

Ad hoc information retrieval (IR) systems are based on an efficient conventional method for accessing pre-indexed data points, but they are not aware of the conceptual content of the data that they access. Preprocessing the data with lexical methods, such as obtaining lexical roots of terms, can boost the performance of an IR system, but empirical evidence suggests that such methods alone are not nearly as powerful as using conceptual information structured in ontologies such as MeSH® (Kayaalp et al., 2003).

In this project, we plan to utilize conceptual information to improve information access accuracy. One common method of utilizing ontologies in information access is called query expansion (Voorhees & Hou, 1992; Srinivasan, 1996; Aronson & Rindflesch, 1997). For example, suppose the analysis of a user query yields a target concept *C* (e.g., psychotropic drugs). A simple hierarchical ontology containing only is-a relations might indicate that there is a set of concepts *children*(*C*) (e.g., antidepressive agents, tranquilizing agents, and hallucinogens), each of which is a type of *C* but differs from each other.

[10] Formal specification methods (Woodcock & Loomes, 1988) constitute a discipline of software engineering that has had a relatively long history, during which a number of specification languages have been developed; e.g. VDM (Jones, 1986), Eiffel (Meyer, 1988), Larch (Guttag, Horning, & Modet, 1990), Z (Spivey, 1992), UML (Rumbaugh, Jacobson, & Booch, 1999), Alloy (Jackson, 2002). For a high-level introduction to the discipline, see Wing (1990).

An information access tool supported with conceptual information can access not only all instances of data points that are lexical derivatives of *C*, i.e. psychotropic drugs in the above example, but also other data points derived from *children*(*C*) (e.g., antidepressive agents and tranquilizing agents) as well as from their children recursively (e.g., an is-a path under the antidepressive agents consists of the following entities: anti-anxiety agents, benzodiazepine, diazepam, and Valium®).

Furthermore, the user who seeks information about psychotropic drugs most probably would not be interested in this report; thus, the information access tool should rank this document with a very low interest point, had this report been a part of the indexed corpus. Given the number of occurrences of the term "psychotropic drugs" in this document, conventional IR systems would probably regard this document as a very likely candidate. However, a context-aware information access tool may recognize that although "psychotropic drugs" is used frequently in this document, neither the context of those sentences nor the context of the document is about pharmaceuticals.

Analysis of the user query and translating it into a formal probabilistic model is as important as the corpus analysis portion of the information access task. We are in the process of developing such a probabilistic method, which we call *collocation networks*, early prototypes of which have already been tested (Kayaalp et al., 2003). In this method, we represent all tokens of a user query in a collocation network. A collocation network is a Bayesian network with a number of dimensions of relevance. One such relevancy dimension is the length of a matching portion of the query phrase to the document; e.g., given a query segment is *slowpoke binding protein*, a collocation network would evaluate the relevancy of each document in terms of posterior probabilities. The posterior probabilities would be expected to be highest when the matching sequence is identical to the query; i.e., *slowpoke binding protein*. It would decrease as the matching sequence gets shorter, i.e., *binding protein* or just *protein*.

Currently, we are studying combining other dimensions of relevance such as the information theoretic distance between the query and a matching sequence in a document sentence, the cumulative value of such matches within a paragraph or document, and the value of semantic distance (Hirst, 1987) due to effects of lexical and semantic variations or due to the difference in conceptual granularity.

5.4 Methods for Learning Models from Data

The muON as the co-representation of a set of ontologies is a conceptual model of which variables (concepts) initially are not associated with data. If muON concepts are linked to actual sentences of corpora, they together would become a large database.

Conventional methods that use unprocessed text usually fall short of extracting all relevant information, since the raw text cannot be parameterized well, partly due to the characteristics of natural languages, as defined in Zipf's law. Zipf's law indicates that more than half of all distinct words of a corpus occurs with very low-frequency counts. By regarding text as a prior model and transforming the search space from a lexical space to a concept space, we would have a richer set of options in selecting our variables and their parameterization. For example, we might need to learn a psychiatric model and one of the provided query terms might be diazepam. Suppose the term *diazepam* occurs in the available corpus with a very low-frequency count; thus, we cannot rely on any diazepam-related probability estimation derived from the associated sparse frequency count. Since diazepam is already an element of our ontology (MeSH in this case), we might traverse down and collect the frequency counts of all brand name products of diazepam, and, if needed, we can traverse up the concept hierarchy one level and check for frequency counts of all instances of agents in the benzodiazepine family and their brand name products.

The rich conceptual structure of such a database, together with information theoretic metrics, would facilitate resolving many structural and semantic ambiguities in individual sentences, yielding a more robust interpretation to the informally represented information and it would enable us to translate such an interpretation into formal models. For example, a discourse model represented as a sequence of concepts in an n^{th}-order Markov model may provide us a posterior probability distribution of possible contexts of a sentence or a sentence substructure. Such a probability distribution may facilitate resolving the structural or semantic ambiguities within the sentence or sentence substructure in question, because the terms used in a sentence are usually influenced by the contexts of the preceding sentences.

We plan to process informal user queries as we preprocess corpora, identify concepts represented in queries, and form a well-defined variable set. Given a set of variables associated with data, a learner can identify variable interactions that maximize the score of a chosen metric. Various metrics were proposed in the literature; e.g., frequentist metrics (Bishop, Fienberg, & Holland, 1975) such as likelihood ratio (G^2) and chi-square (χ^2) statistic, information theoretic metrics such as AIC (Akaike, 1973) and MDL (Rissanen, 1978), and Bayesian metrics such as BDe (Cooper & Herskovits, 1992; Heckerman, Geiger, & Chickering, 1995) and GU (Kayaalp & Cooper, 2002). For further details and examples, see (Kayaalp, Cooper, & Clermont, 2000; Kayaalp, Cooper, & Clermont, 2001).

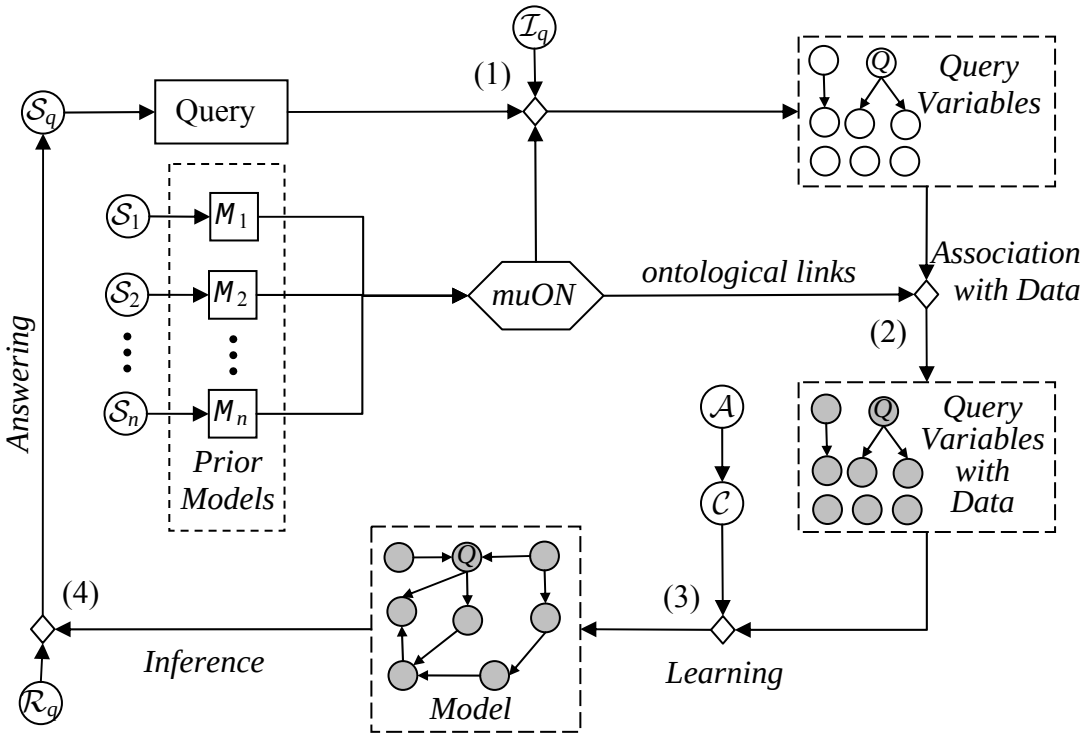

Figure 6: Query Driven Learning and Information Elicitation

The learning process driven by user queries can be conceptualized in four steps (see Figure 6): (1) Queries are interpreted by an interpretation software $\mathcal{I}_q$, which identifies concepts in the query and translates the query into a set of variables. Depending upon its level of sophistication, $\mathcal{I}_q$ may interpret implicit needs and the ontology of the user and include corresponding variables into the final set of variables. (2) Variables are associated with the existing corpora and databases (prior models M_1 through M_n) through ontological links provided by muON. A prior model M_i can simply be a scientific article produced by a scientist $\mathcal{S}_i$. (3) A learner ($\mathcal{A}$+$\mathcal{C}$) constructs a model such that the structure and parameters of the model maximizes the score of the metric in use. (4) Statistical inference is obtained through the model. The inference contains the prospective answer to the query. The inference may be formatted by a natural language synthesizer $\mathcal{R}_q$, which we expect to obtain from external sources.

6 Evaluation Plan

We plan to evaluate each new method both separately and in combination with others regarding their individual contribution to the global solution. To this end, one of the tasks

that we need to accomplish is to identify proper gold standards to evaluate the performance of our solutions.

In the information identification problem, we may use existing information created manually by the MeSH indexers and information that can be produced automatically through the tools of the Indexing Initiative. Certainly, these information resources must be excluded in the evaluation version of the system. In cases where such exclusions are not feasible, we have to develop our gold standard through a panel of experts.

Methods related to the second and third project objectives will be evaluated using established metrics, such as ROC and F-measure.

Evaluations of learning new models and their contributions to reliable scientific communication are expected to be more challenging. The learning method should be evaluated indirectly through the performance of the learned models on the particular task for which the model has been developed.

7 Project Schedule

During the first year of the project, we plan to (1) identify information blocks in MEDLINE using the NLP tools that have been developed in NLM, (2) build an initial version of muON using the UMLS concepts and associated ontological resources, (3) index information blocks identified in MEDLINE on muON, (4) test the accuracy of information access of the system against PubMed® using Boolean queries, and (5) start building the framework of learning models from textual data and identified information.

In the second year, we plan to (1) extend the identified information using some external resources, (2) complete implementation and testing the method on collocation networks, (3) update muON using additional ontological resources, (4) test the accuracy of information access of the resulting system and evaluate its performance against the results obtained in the first year, and (5) complete and test the alpha version of the model learning system.

In the following three years, we plan to improve the depth and breadth of the system.

In the third year, we plan to extend our focus from textual corpora to structured biomedical databases and, if possible, to clinical databases. Learning probabilistic models from those databases based on user queries would be an important step to automate. In the third year, we also plan to start developing a formal specification language that would accommodate requirements of actual biomedical and clinical data models and to investigate translating datasets from one model to another based on our experience in muON that we will have acquired in the earlier years of the project.

Years four and five will be dedicated to improving our understanding, methods, and the performance of our systems.

8 Conclusions

Producing information in a natural language such as English has been the most efficient and effective method for scientific communication between scientists, although interpreting such informal models algorithmically is a very challenging task. We perhaps need new computational methods for scientific modeling such that models may exchange information through some standard formal language of scientific communication (Sowa, 1984; Lenat & Guha, 1989; Genesereth, 1991; Kayaalp, 1993; Berners-Lee, 1997–2002). On the other hand, it is unrealistic to expect scientists to dramatically change their operation modes for meeting the current information demands of others.

Our task in the MLM project is to develop methods that facilitate learning formal models based on user queries from prior models, which may or may not be formal. Such formal models, when learned effectively, would serve as a bridge across disciplines and may mediate a reliable communication between agents of different disciplines.

This report states the direction of the MLM project, the problems of interest, the types of models that may lead to new useful technological solutions, and some promising methods that may be improved further. At this stage of the MLM project, this report does not attempt to describe how to solve these problems; rather it merely states what should be solved, what can be solved, and which promising methods may be used to develop new methods.

Our task is quite challenging, but the potential rewards are very high. We are certain that new scientific modeling and computational communication methods will shift the paradigm of creating and disseminating scientific knowledge and revolutionize science (Kuhn, 1996), and our hope is to contribute to this effort to the full extent of our capabilities.

Acknowledgements

The author thanks Drs. McCray, Bodenreider, and Rindflesch for their constructive comments on an earlier version of this report.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *Proceedings of the Second International Symposium on Information Theory* (p. 267–281).

Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G., Nelson, S. J. et al. (1999). *The Indexing Initiative: A Report to the Board of Scientific Counselors.* Bethesda, MD: Lister Hill National Center for Biomedical Communications, National Library of Medicine.

Aronson, A. R., & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. *Proceedings of the Annual American Medical Informatics Association (AMIA) Symposium*, 485–489.

Berners-Lee, T. (1997–2002). *Design Issues for the World Wide Web.* Web site: URL http://www.w3.org/DesignIssues/Overview.html.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

BISTI. (2000). *Recommendations of The Biomedical Information Science and Technology Initiative Implementation Group.* Web site: URL http://www.bisti.nih.gov/bisti_recommendations.cfm.

Buchanan, B. G., Mitchell, T. M., Smith, R. G., & Johnson, C. R. Jr. (1978). Models of Learning Systems. J. Belzer, A. G. Holzman, & A. Kent (Eds.), *Encyclopedia of Computer Science and Technology* (Vol. 11p. 24–51). Pittsburgh, PA: Marcel Dekker.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley.

Burgun, A., & Bodenreider, O. (2001). Mapping the UMLS Semantic Network into general ontologies. *Proceedings of the Annual American Medical Informatics Association (AMIA) Symposium*, 81–85.fd

Carson, E. R., Cobelli, C., & Finkelstein, L. (1983). *The Mathematical Modeling of Metabolic and Endocrine Systems*. New York, NY: John Wiley & Sons.

Cohen, P. R., & Feigenbaum, E. A. (1982). *The Handbook of Artificial Intelligence. Volume III*. Reading, MA: Addison-Wesley.

Collins, M., & Brooks, J. (1995). Prepositional Phrase Attachment through a Backed-off Model. *Proceedings of the Third Workshop on Very Large Corpora*.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, *9*, 309–347.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A Practical Part-of-Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*.

Friedman, C., Kra, P., & Rzhetsky, A. (2002). Two Biomedical Sublanguages: a Description Based on the Theories of Zellig Harris. *Journal of Biomedical Informatics*, *35*(4), 222–235.

Genesereth, M. R. (1991). Knowledge Interchange Format. *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning* (p. 238–249). San Mateo, CA: Morgan Kaufmann.

Grishman, R. (2003). Information Extraction. R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (p. 545–559). Oxford, UK: Oxford University Press.

Guttag, J. V., Horning, J. J., & Modet, A. (1990). (Report No. SRC Research Report 58). Digital Equipment Corporation.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, *20*(3), 197–243.

Hirschman, L., Morgan, A. A., & Yeh, A. S. (2002). Rutabaga by Any Other Name: Extracting Biological Names. *Journal of Biomedical Informatics*, *35*(4), 247–259.

Hirst, G. (1987). *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge, UK: Cambridge University Press.

Hobbs, J. R. (2002). Information Extraction from Biomedical Text. *Journal of Biomedical Informatics*, *35*(4), 260–264.

Jackson, D. (2002). Alloy: A Lightweight Object Modelling Notation. *ACM Transactions on Software Engineering and Methodology*, *11*(2), 259–290.

Jones, C. B. (1986). *Program Specification and Verification in VDM.* (Report No. UMCS-86-10-5). Manchester, UK: Computer Science Department, University of Manchester.

Kayaalp, M. (1993). *Multifaceted Ontological Networks: Methodological Studies toward Formal Knowledge Representation.* Master thesis, Southern Methodist University, Dallas, TX.

Kayaalp, M., Aronson, A. R., Humphrey, S. M., Ide, N. C., Tanabe, L. K., Smith, L. H. et al. (2003). Methods for accurate retrieval of MEDLINE citations in functional genomics. *Proceedings of the 12th Annual Text Retrieval Conference.* (p. 175–184). Gaithersburg, MD: National Institute of Standards and Technologies.

Kayaalp, M., Cooper, G. F., & Clermont, G. (2000). Predicting ICU Mortality: A Comparison of Stationary and Nonstationary Temporal Models. *Proceedings of the Annual American Medical Informatics Association (AMIA) Symposium* (p. 418–422). Los Angeles, CA.

Kayaalp, M., Cooper, G. F., & Clermont, G. (2001). Predicting with Variables Constructed from Temporal Sequences. *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics 2001* (p. 220–225). Morgan Kaufmann.

Kayaalp, M., & Cooper, G. F. (2002). A Bayesian Network Scoring Metric That Is Based on Globally Uniform Parameter Priors. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-2002)* (p. 251–259).

Kayaalp, M., Pedersen, T., & Bruce, R. (1997). A Statistical Decision Making Method: A Case Study on Prepositional Phrase Attachment. *Proceedings of the 1997 Meeting of the ACL SIG in Computational Natural Language Learning (CoNLL97)* (p. 33–42).

Kayaalp, M., & Sullins, J. (1992). Representation Issues in Medical Knowledge. *Proceedings of the First European Joint Conference on Engineering Systems Design and Analysis: Bioengineering* (p. 161–166).

Kayaalp, M., & Sullins, J. (1993). System Design with Multiple Abstraction Facets. *Proceedings of the Energy-Sources Technology Conference and Exhibition: Computer Applications and Design Abstraction* (p. 145–149).

Kuhn, T. S. (1996). *The Structure of the Scientific Revolution* (Third ed.): University of Chicago Press.

Langley, P. (1996). *Elements of Machine Learning*. San Francisco, CA: Morgan Kaufmann.

Lenat, D. B., & Guha, R. V. (1989). *Building Large Knowledge-Based Systems. Representation and Inference in the Cyc Project*. Reading, MA: Addison-Wesley.

McCray, A. T. (1991). Extending a natural language parser with UMLS knowledge. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, 194–198.

McCray, A. T., Aronson, A. R., Browne, A. C., Rindflesch, T. C., Razi, A., & Srinivasan, S. (1993). UMLS knowledge for biomedical language processing. *The Bulletin of the Medical Library Association*, *81*(2), 184–194.

McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods of Information in Medicine*, *34*(1–2), 193–201.

McCray, A. T., Srinivasan, S., & Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, 235–239.

Meyer, B. (1988). Eiffel: A Language and Environment for Software Engineering. *Journal of Systems and Software*, *8*(3), 199–246.

Minsky, M. L. (1965). Matter, Mind and Models. *Proceedings of the International Federation for Information Processing (IFIP) Congress* (p. 45–49). Washington, DC: Spartan Books.

Mitchell, T. M. (1997). *Machine Learning*. New York, NY: McGraw-Hill.

National Library of Medicine (U.S.) Board of Regents. (2000). *National Library of Medicine Long Range Plan 2000–2005/Report of the Board of Regents.* Bethesda, MD: National Library of Medicine.

Rindflesch, T. C., & Fiszman, M. (2003). The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text. *Journal of Biomedical Informatics*, *36*(6), 462–77.

Rindflesch, T. C., Rajan, J. V., & Hunter, L. (2000). Extracting Molecular Binding Relationships from Biomedical Text. *Proceedings of the Sixth Applied Natural Language Processing Conference* (p. 188–195).

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471.

Rivest, R. L. (1987). Learning Decision Lists. *Machine Learning*, *2*, 229–246.

Rumbaugh, J., Jacobson, I., & Booch, G. (1999). *The Unified Modeling Language Reference Manual*. Reading, MA: Addison-Wesley.

Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading, MA: Addison-Wesley.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Simon, H. A. (1989). *Models of Thought. Volume II*. New Haven, CT: Yale University Press.

Sowa, J. F. (1984). *Conceptual Structures. Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.

Spivey, J. M. (1992). *The Z Notation: A Reference Manual* (Second ed.). Englewood Cliffs, NJ: Prentice-Hall.

Srinivasan, P. (1996). Retrieval feedback in MEDLINE. *Journal of the American Medical Informatics Association*, *3*(2), 157–167.

Srinivasan, P., & Rindflesch, T. (2002). Exploring Text Mining from MEDLINE. *Proceedings of the American Medical Informatics Symposium.* (p. 722–726).

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 422–460.

Voorhees, E. M., & Hou, Y.-W. (1992). Vector Expansion in a Large Collection. *First Text Retrieval Conference (TREC-1)* (p. 343–351). Gaithersburg, MD: NIST.

Whorf, B. L. (1940). Linguistics as an Exact Science. *Technology Review (M.I.T.)*, *43*, 61–63, 80–83.

Wing, J. M. (1990). A Specifier's Introduction to Formal Methods. *Computer*, *23*(9), 8–24.

Woodcock, J., & Loomes, M. (1988). *Software Engineering Mathematics*. Reading, MA: Addison-Wesley.

Zerhouni, E. (2003). The NIH Roadmap. *Science*, *302*(5642), 63–64, 72.

Curriculum Vitae

Mehmet Kayaalp, M.D., Ph.D.
Staff Scientist, NIH/NLM/LHNCBC/CgSB

Education and Training

University of Istanbul	M.D.	1989	Medicine
Southern Methodist University	M.S.	1993	Computer Science
University of Pittsburgh	Ph.D.	2003	Artificial Intelligence (specialization in Medical Informatics)

Research Areas of Interest

- Learning intelligent models from data and text
- Artificial intelligence and statistical decision making in medicine and biology

Employment

2003–today	*Staff Scientist*, Cognitive Science Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland.
2001–02	*National Library of Medicine Fellow*, Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania.
2001	*Teaching Assistant* in medical informatics, Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania.
1998–01	*Research Assistant* in an IAIMS project: Identifying Patient Subsets of Interest in Electronic Medical Record Repositories, Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania.
1995–97	*Research Assistant* in a statistical natural language learning project funded by Office of Naval Research at Southern Methodist University, Dallas, Texas.
1991–95	*Teaching Assistant* in computer science at Southern Methodist University, Dallas, Texas.
1990–91	*Research Scholar* at Southern Methodist University, Dallas, Texas.
1989–90	*Physician* and *Chief Medical Officer* of a Turkish regiment of 3000 soldiers, Turkish Army, Izmir, Turkey.
1987	*Intern* at Klinikum Grosshadern, Ludwig-Maximillian University, Munich, Germany.

Honors

2001–02	*National Library of Medicine Fellowship*, University of Pittsburgh, Pittsburgh, Pennsylvania
1998–01	*Graduate Student Research Award* funded by Center for Biomedical Informatics through an IAIMS grant, University of Pittsburgh, Pittsburgh, Pennsylvania
1997–98	*Intelligent Systems Program Scholarship* (tuition and stipend), University of Pittsburgh, Pittsburgh, Pennsylvania
1996 & 98	*American Association for Artificial Intelligence (AAAI) Scholarship* (conference and travel expenses)
1995–97	*Research Assistantship* (tuition and stipend) funded by School of Engineering and Applied Sciences, Southern Methodist University through an Office of Naval Research (ONR) grant

1991–95	*Teaching Assistantship* (tuition and stipend) funded by School of Engineering and Applied Sciences, Southern Methodist University, Dallas, Texas
1994	*Student Paper Award* (conference and travel expenses), Knowledge Acquisition Workshop, Banff, Canada
1993	*The Rick A. Barrett Memorial Award*, for institutional services to Southern Methodist University, Dallas, Texas
1987	*Exchange Fellowship* (stipend and travel expenses) by the Government of Germany.

Publications

Kayaalp, M.; Aronson, A. R.; Humphrey, S. M.; Ide, N. C.; Tanabe, L. K.; Smith, L. H.; Demner, D.; Loane, R. R.; Mork, J. G., and Bodenreider, O. (2003). Methods for Accurate Retrieval of MEDLINE Citations in Functional Genomics. *Proceedings of the Text Retrieval Conference (TREC-2003)*: 175–184.

Kayaalp, M. (2003). Learning Dynamic Bayesian Network Structures from Data. *Ph.D. Dissertation*, University of Pittsburgh.

Kayaalp, M., and Cooper, G. F. (2002). A Bayesian Network Scoring Metric That Is Based on Globally Uniform Parameter Priors. *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-2002)*: 251–259.

Kayaalp, M.; Cooper, G. F., and Clermont, G. (2001). Predicting with Variables Constructed from Temporal Sequences. *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics 2001*: 220–225.

Kayaalp, M.; Cooper, G. F., and Clermont, G. (2000). Predicting ICU Mortality: A Comparison of Stationary and Nonstationary Temporal Models. *Proceedings of the Annual American Medical Informatics Association Fall Symposium*: 418–422.

Aronis, J. M.; Cooper, G. F.; Kayaalp, M., and Buchanan, B. G. (1999). Identifying patient subgroups with simple Bayes'. *Proceedings of the Annual American Medical Informatics Association Fall Symposium*: 658–662.

Cooper, G. F.; Buchanan, B. G.; Kayaalp, M.; Saul, M., and Vries, J. K. (1998). Using computer modeling to help identify patient subgroups in clinical data repositories. *Proceedings of the Annual American Medical Informatics Association Fall Symposium*: 180–184.

Kayaalp, M.; Pedersen, T., and Bruce, R. (1997). A Statistical Decision Making Method: A Case Study on Prepositional Phrase Attachment. *Proceedings of the 1997 Meeting of the ACL SIG in Computational Natural Language Learning (CoNLL97)*: 33–42.

Pedersen, T.; Kayaalp, M., and Bruce, R. (1996). Significant Lexical Relationships. *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-1996)*: 455–460.

Kayaalp, M. and Sullins, J. (1994). Multifaceted Ontological Networks: Reorganization & Representation of Scientific Knowledge. *Proceedings of the Eighth Knowledge Acquisition for Knowledge-Based Systems Workshop*: 25.1–19.

Kayaalp, M. (1993). Multifaceted Ontological Networks: Methodological Studies toward Formal Knowledge Representation. *Master Thesis*, Southern Methodist University.

Kayaalp, M. and Sullins, J. (1993). System Design with Multiple Abstraction Facets. *Proceedings of the Energy-Sources Technology Conference and Exhibition: Computer Applications and Design Abstraction*: 145–149.

Kayaalp, M. and Sullins, J. (1992). Representation Issues in Medical Knowledge. *Proceedings of the First European Joint Conference on Engineering Systems Design and Analysis: Bioengineering*: 161–166.