SEMI-AUTOMATIC ONTOLOGY-BASED KNOWLEDGE EXTRACTION AND VERIFICATION FROM UNSTRUCTURED DOCUMENT

By

JIONGHUA JI

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

UNIVERSITY OF FLORIDA

2000

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to Dr. Stanley Su and Dr. Abdelsalam Helal for serving on my committee. I would particularly like to thank my advisor, Dr. Li-Min Fu, for his guidance and encouragement throughout my research, without which this work would not have been possible.

My love and gratitude are extended to the members of my family: my parents and my brother, without whose unconditional love and support on a daily basis, I would not have become the person I am today or achieved whatever I have accomplished. I am forever in their debt for helping me realize my dream and giving me the opportunity to reach my potential. It is to them that I dedicate this work.

TABLE OF CONTENTS

page

LIST OF FIGURES

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Science

SEMI-AUTOMATIC ONTOLOGY-BASED KNOWLEDGE EXTRACTION AND VERIFICATION FROM UNSTRUCTURED DOCUMENT

By

Jionghua Ji

December 2000

Chairman: Dr. Li-Min Fu
Major Department: Computer and Information Science and Engineering

With the arrival of the information age, world knowledge is accumulating at an astronomical rate. Information available on-line stored in the form of unstructured documents that are hard to classify and search for. Without making the information efficiently accessible, the ratio of attainable knowledge over existing information will be extremely low. This thesis presents a semi-automatic approach to extracting and verifying knowledge from unstructured documents based on the ontology inherited in the domain of interest. Medline summaries as a result of search on keyword, hepatitis causation, are chosen as the unstructured documents.

Not only retrieving knowledge from the vast information source is needed for effective knowledge management, but also the ability to validate or verify the knowledge extracted against the current knowledge base has become crucial in achieving a consistent and reliable knowledge base. Under the context of this research, knowledge verification

implies the detection of knowledge repetition and conflict. The thesis introduces a new approach to knowledge verification by employing a domain ontology, which imposes a structure on the concepts and thus limits the interpretation among them.

CHAPTER 1
INTRODUCTION

Background and Motivation

With the arrival of the information age, world knowledge is accumulating at an astronomical rate. As Picht and Draskau noted (1985, 24),

> The 19th century was remarkable for the giant strides with which scientific progress advanced and found practical applications. This situation led to a vast need for terminology, and it soon came to be realized that these explosive developments likewise called for the organization of knowledge.

Moreover, Skuce (2000, 96) says,

> today's systems for storing and sharing knowledge on a large scale are undergoing rapid change, driven mainly by the success of the World Wide Web and its search engines. Soon a vast array of material will be come available…. We seek better ways of dealing with this impending revolution.

Since on-line information is stored in the form of unstructured documents that are hard to classify and search on, without making the information accessible and efficiently accessible, the ratio of attainable knowledge over existing information will be extremely low.

One example of on-line information growing unmanageable is the electronic form of academic papers, especially medical papers. On the search of the keyword, hepatitis causation, 3756 papers are returned from Medline (www.ncbi.nlm.nih.gov). If one spends 30 minutes to read a paper, it will take more than eight months to finish all of them, and in 8 months, 400 more new papers on the subject will appear. If one only reads

the abstracts which takes five minutes each, it still requires 39 days to complete. Fortunately, there are Medline summaries available that are usually composed of one sentence. However, after reading 3756 or more sentences, can a human being remember how many times a knowledge has been repeated or if there is knowledge conflict either directly or indirectly with the knowledge already encountered ?

Not only retrieving knowledge from the vast source of information is needed for efficient knowledge management, but also the abilility to validate the knowledge extracted against the current knowledge base has become crucial in achieving a consistent and reliable knowledge base.

This situation motivates the research of this thesis. The goal of this project is to take in Medline summaries and turn them into knowledge and populate the knowledge base. Each time a new piece of knowledge is added to the knowledge base, the system should be able to check if it already exists, which is a knowledge repetition, and if it violates any existing knowledge, which is a knowledge conflict.

Organization of the Thesis

The remainder of the thesis is organized as follows. Chapter 2 discusses the term *ontology* and its various applications in the information processing field and its implications in this project. Current ontologies in use are presented and related work introduced. In Chapter 3, the author provides the rationale behind the choices that are made during the design process of this knowledge extraction and verification system. Chapter 4 gives the actual implementation, result, and performance evaluation. Chapter 6 concludes the research and gives remarks on future direction and improvements.

CHAPTER 2
ONTOLOGY-BASED APPROACH TO KNOWLEDGE ACQUISITION

What is Ontology ?

According to Webster's Dictionary, ontology is a) a branch of metaphysics relating to the nature and relations of being b) a particular theory about the nature of being or the kinds of existence. In other words, ontology is a part of metaphysics that specifies the most fundamental categories of existence, the elementary substances or structures out of which the world is made.

There are two aspects to the term *ontology*. One is from the philosophical point of view, which sees ontology as a theoretical enterprise whose content remains the same independently of the language, which is used to express it.

The other aspect of the term ontology, formal ontology, as used by researchers in artificial intelligence or information processing community, is used to define the models of the world or a specification of a conceptualization (Gruber, 1993). In essence, only the objects defined in the ontology can exist or have meaning in that model -- the ontology has become the basic level of knowledge representation scheme. For this reason, ontology is designed with a specific use and computational environment in mind. It is certainly a different sense of the word than its use in philosophy.

Moreover, the philosopher-ontologist will attempt to establish the truth about the world and reflect it in the identities (the objects, properties and relations, the states) of the ontology, while the information engineer is more concerned about the application of such knowledge structure. The absolute truthfulness of such structures is not important as long

as they hold correct in the domain of interest. Essentially, ontology becomes a computational tool, which has been used to solve a variety of problems.

Ontology Specification and Representation

Formally, ontology is represented by carefully defined terms with formal interpretations. It consists of a set of entities and the relationships among them, which are described by a representational vocabulary with formal axioms to constrain their meanings. While the domain knowledge is represented in a declarative formalism, the definitions of the terms provide human-understandable text describing their meaning.

In its simplest form, ontology is a type hierarchy, specifying classes and their subsumption relationships.

Current Ontology Applications

As noted by Guarino, there are six major ontology application areas (Guarino, 1998).

1. Knowledge engineering, representation, management, sharing, integration
2. Information retrieval and extraction
3. Natural language translation
4. Database design, conceptual modeling, and information system design
5. Enterprise integration
6. Military applications

In addition, from the fundamental roles an ontology plays, Uschold and Gruninger have divided its use space into three categories (Uschold & Gruninger, 1996) :

1. Communication
2. Inter-operability
3. Specification, reliability, and reusability

Since ontology is defined as a representation of a shared conceptualization of a particular domain, it is inheritantly language independent and is embedded with structures and imposes constraints among the concepts. These two characteristics about ontologies have found a variety of applications in the information processing and knowledge management fields.

Being purely conceptual and language independent, ontologies serve successfully as a communication medium between and among people, organization and computer systems by providing a uniform communication framework that otherwise is impossible or much more costly. This area of ontology application promotes shared understanding, knowledge reuse, knowledge integration and knowledge interchange. Furthermore, ontology also plays a roll in standardizing knowledge representation among tools.

By having limited interpretation between concepts in the ontology, which is achieved through a structure of concepts and axioms among concepts, ontology is able to maintain the knowledge integrity and provides reliability and knowledge validation.

Independent of application domain and industry, some manner of communication seems to be the fundamental functionality of ontologies. According to how ontologies are used, there are five categories.

1. Information retrieval and extraction
2. Knowledge reuse
3. Knowledge interchange
4. Knowledge integration
5. Knowledge representation standard

Information Retrieval and Extraction

The goal of information retrieval is to find relevant documents in a large collection, in response to users' queries expressed in natural language, for example, internet navigation which derives information from the immense diversity of sources of internet. A common ontology that structures and unifies the information on the Internet is able to provide efficient and precise information retrieval. This process can be viewed as a form of communication between end users and the data sources. That is with the help of ontologies, the vast data sources are refined, classified, properly related and presented as a uniform source of information rather than a diversity of unrelated and unstructured documents.

In addition, by providing a uniform conceptual source of information, the traditional approach of keyword search, which is solely based on syntax matching, is replaced with much more accurate and efficient content matching.

OntoSeek is an online yellow page search tool that takes this approach. However in order to provide the uniform knowledge framework, both the online information producer and the web surfer are involved to achieve a shared understanding of the sources and usage (Guarino, Masolo & Vetere, 1999).

The database of ThoughtTreasure which consists of 25,000 concepts and 50,000 assertions enables computers to parse natural languages and extract information from text and answer questions based on common sense reasoning (Mueller, 1998).

Knowledge Reuse

Various ontologies have been constructed with specific purposes for specific domains. When one ontology works well in one domain, it may not perform as expected in other domains as the underlining assumptions normally change with different domains.

Without letting the human effort of building individual ontologies go to waste and aiming at achieving a universal ontology that incorporates and reuses the existing ontologies, the Knowledge Systems Laboratory is constructing a library of ontologies that can be reused and adapted to different classes of problems and contexts (Farquhar et al., 1995).

Knowledge Interchange

Intuitively by acting as the medium of knowledge communication, ontologies make knowledge interchange possible.

For example, ontologies can act as inter lingua in natural language translation. Instead of creating ad hoc translators for each pair of languages, a common ontology is constructed. Translation is performed between local languages and the common target. The advantage of this approach is that the time complexity is lowered from $O(n^2)$ to $O(n)$ as illustrated in Figure 2.1.

Process Interchange Format (PIF) developed by Lee and associates (Lee et al., 1998) aims to support the exchange of business process models among different process representations. Tools communicate by translating between their local format and PIF. Automating the translation process is another goal of this project with minimum loss of meaning. When needed, some minimal form of human effort may be involved to assist the translation.

A similar application, Process Specification Language (PSL) a joint effort by Schlenoff and associates is an interchange format designed to help exchange process information automatically among a wide variety of manufacturing applications such as process modeling, process planning, scheduling, simulation, workflow, project management, and business process re-engineering tools. These tools would communicate by translating between their native format and PSL. Noticeably, any system would be

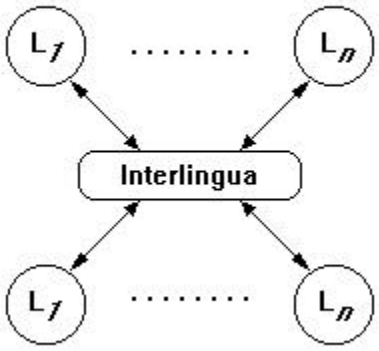

Figure 2.1. Ontology as Inter-lingua

able to automatically exchange process information with any other system via PSL (Schlenoff et al., 2000).

However, as noted by Uschold and Gruninger, there are three obstacles or limitations for ontologies to function as inter lingua (Uschold & Gruninger, 1996).

1. The common language should be declarative

2. The native language has to be less expressive than the common languages

3. The common language should be easily extendable for future addition

Knowledge Integration

Knowledge sharing and interchange among disparate systems enables a further application, knowledge integration.

Several noticeable applications areas in knowledge integration are military application, enterprise integration and database integration.

Military application. In military domains, different amies need to be integrated together into a single force, and military information from different regional and international sources needs to be coded within a single framework for effective communication.

Enterprise integration. International industries often face communication problems among its subsidiaries in different countries due to cultural differences in addition to language differences. A common ontology that provides a framework of communication will solve this problem. Even within one local enterprise, personnel at different level may view the organization issues and structures from different point of views, which in turn may result miscommunication.

The major role of the Enterprise Ontology (EO) is to act as a communication medium between different people and between people and implemented computational systems. It is motivated to adapt to the fast changing business environment and to achieve greater flexibility, more effective communication and integration for business enterprises by providing a collection of terms and formal definitions and an enterprise modeling framework, which is manifested in the Enterprise Tool Set. The tool set helps user to capture aspects of a business and identify business requirements, problem and solutions (Uschold et al., 1998).

Database integration. Database constructions of divergent data derived from different sources bearing various tasks in mind can be unified into a single system by developing common ontologies. With databases designed with a specific need for a specific organization, terms, categories and relations are used differently when storing data objects. For example, different databases may use the same categories but with different semantic meanings; alternatively, the same concept may be expressed by different terms. Again, by providing a unifying specification of terms and relations, ontology can act as a communication medium among divergent data and solve this problem.

Knowledge Representation Standard

Acting as a communication medium among disparate entities, ontologies naturally grow into playing the role of standardizing knowledge representation.

Knowledge Interchange Format (KIF). KIF developed by Genesereth and associates (Genesereth & Fikes, 1991) is a good example of a knowledge representation language designed for knowledge interchange among disparate programs. Ontolingua, a language based on KIF, developed by Gruber and associates (Gruber, 1995) and its Ontology Editor by the Stanford KSL Network Services has facilitated the process of converting the natural language specification of the ontology into a formal language in EO. It has also been used as a working tool for the construction of a medicine ontology by the Ontology group of the Medical Informatics Unit in Rome.

Ontology Interchange Language (OIL). OIL is proposed to provide a joint standard for integrating ontologies with existing and arising web standards. The Ontology Inference Layer is a Web-based representation and inference layer for ontologies, which combines the widely used modeling primitives from frame-based languages with the formal semantics and reasoning services provided by description logics (Fensel et al., 2000).

Common Object Request Broker Architecture (CORBA). COBRA networks developed and endorsed by the Object Management Group (OMG) is a standard for retrieving objects and invoking operations on objects across. It incorporates informal notions of ontologies by including a glossary of terms to be used in the object model, thus informally providing a framework for shared understanding (Mowbray & Zahavi, 1995).

International Standard for the Exchange of Product Data (STEP). STEP is a noticeable effort in providing an inter lingua for defining and specifying products. It

achieves inter-operability and enables product data to be exchanged among different computer systems throughout the complete product life cycle. STEP data files are built on the ground of specific standardized languages or format. These files can then be read directly by STEP processors or applications or through STEP access interfaces, allowing better integration of applications based on companies disparate systems. Its ultimate goal is an integrated product information database to support a product over its life cycle (STEPTools, 1999).

Knowledge Validation

As representations of a shared conceptualization of a particular domain, ontologies have been used in various areas of applications mainly as a communication medium among disparate systems.

Divergent sources of information that are otherwise unmanageable can be manipulated, retrieved and extracted by relying on a unifying information framework supported by ontologies as shown in the field of information retrieval and extraction. Knowledge can also be interchanged via a common ontology; furthermore, knowledge integration across systems becomes possible as a result. This enforces knowledge sharing and reuse as another major application area of ontologies. Acting as a interchange format or interchange language, ontologies have successfully played the role in standardizing the knowledge representation as an effort to further improve consistency among domains.

Besides working successfully as a communication medium, another important aspect of ontology is the structure or limited meaning it imposes on the entities in the application domain. By doing so, relationships between concepts are restricted, inferences can be made, and knowledge can be validated. The actual knowledge validation normally occurs parallel with the knowledge construction. For example, in the

field of information retrieval, before information can be extracted, a knowledge base is built. During the construction process, each knowledge inserted is checked with the ontology to ensure the proper structure or relationships among the concepts in the knowledge is observed. In addition, when an extraction query is made against a knowledge base, that query is turned into knowledge and its structure validated.

In summary, ontology has proven a useful tool in knowledge engineering, which organizes and presents knowledge so that the knowledge system responds to the user requests and forces a consistent understanding on a subject rather than a fragmented understanding of pieces of knowledge in isolation. It acquires and formalizes knowledge and builds knowledge bases to provide logical, intelligent navigation through information.

Major Implemented Ontologies

CYC. CYC is an effort by the Microelectronics and Computer Technology Corporation (MCC) since 1984 which later spun-off Cycorp, Inc., based in Austin, Texas, in 1995. It is the leading supplier of formalized common sense by developing ontologies for a variety of domain-specific applications.

Aiming at filling the gap of common sense reasoning between human and machines, CYC's knowledge base has over 1,000,000 hand-entered assertions (or "rules") designed to capture a large portion of what is considered consensus knowledge about the world. By using an efficient inference engine on simple assertions with inference rules and control rules for inference, new assertions can be derived. For example, CYC knows a 9-year-old girl can not be a grandmother (Lenat & Guha, 1990).

Knowledge About Complex Technical systems for multiple USe (KACTUS). KACTUS is an European ESPRIT-iii project aiming at the developing a methodology for

the reuse of knowledge for technical systems during their life cycle. The project started on January 1994 and ran for 30 months using about 40 men-years.

By employing domain ontologies and reusing them for different applications, KACTUS uses the same knowledge base for design, diagnosis, operation, maintenance, redesign, instruction, etc. thus supports an integrated approach for reuse knowledge across different applications.

In addition, KACTUS provides a tool kit, VOID, to support an interactive environment for browsing, editing and managing ontologies. It supports the theoretical and application oriented work packages by providing an environment in which one can experiment with theoretical issues and also perform practical work (e.g. browse, edit and query ontologies in various formalisms) (Schreiber, Wielinga & Jansweijer, 1995).

Ontology-based Approach to Knowledge Extraction

As noted above, ontologies have been applied extensively and successfully in conceptual modeling and knowledge representation to enable communication, sharing, and reusing among different entities i.e. organizations, people, software tools, information process, knowledge bases and ontologies. However, knowledge extraction via ontology is a recent development that has demonstrated to be promising. This approach relies on a domain ontology, which imposes a structure on the knowledge, to classify retrieved information and provide a framework for further manipulation.

Following are the some of the current projects on ontology-base approach to knowledge extraction that have influenced some of the design decisions in this research project.

Plinius Ontology

The goal of the Plinius project is to achieve semi-automatic knowledge extraction from short natural-language texts that are the title and abstract fields of bibliographic document descriptions on material properties of ceramic materials. As the text cover a wide range of subjects, ontologies are constructed to capture materials and their properties, processes to make them, processes operating on samples, etc.

One of its design concern is the cost. Realizing full manual acquisition would be too expensive and total automation unrealistic, it takes a human-aided knowledge acquisition approach by employing a lexicon that translates natural language entities into their conceptual counterparts and a knowledge base with background knowledge.

The output of the system is a knowledge base that contains a subset of the knowledge and data conveyed by the texts.

Human experts are involved in various stages of the process as shown in Figure 2.2 to achieve more accuracy.

OntoSeek

OntoSeek is developed by Nicola Guarino and associates for searching online yellow pages and product catalogs based on content matching.

Two phases are involved when storing a natural sentence into OntoSeek's database. One is lexical disambiguation, and the other semantic validation.

In the lexical disambiguation phase, the system takes in a resource description in natural text and converts it into an lexical conceptual graphs (LCGs), which is done with the help of a user interface. Upon receiving a word, the user interface asks the user to input its correct sense. For example, "motor repair" may refer to automobile engine rebuilding or any general motor rebuilding, which is done by two different kinds of

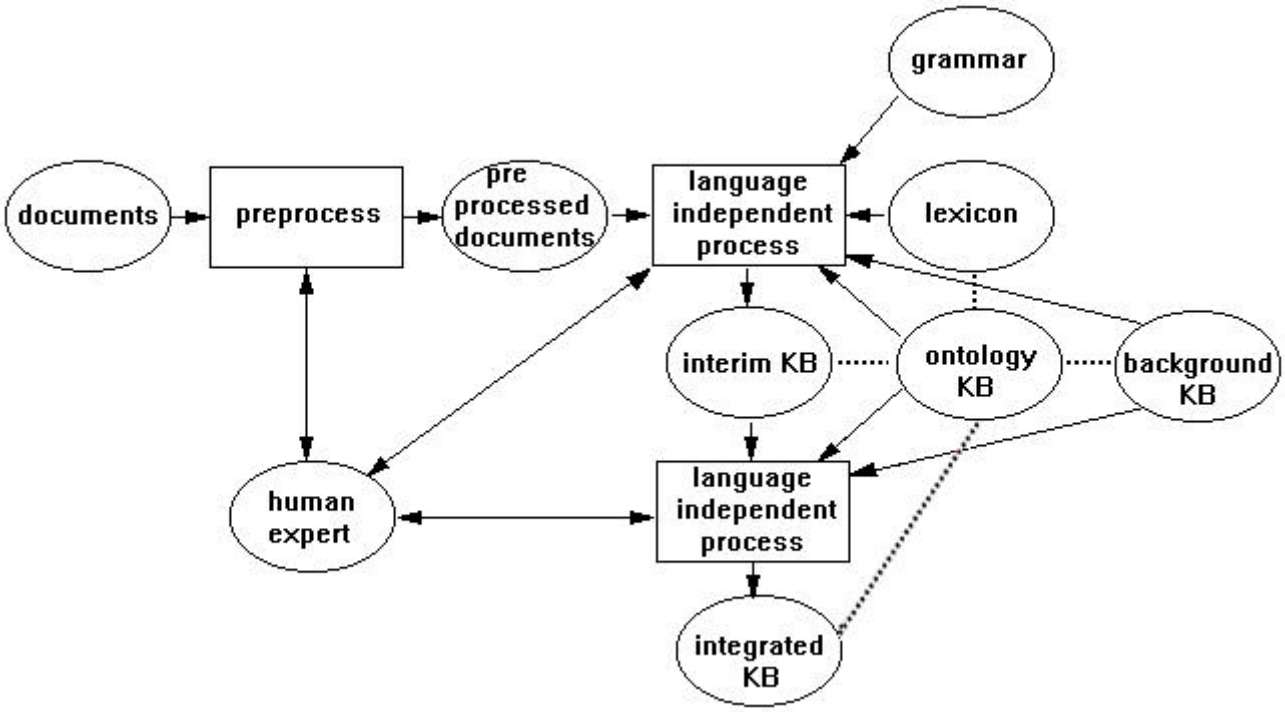

Figure 2.2. The Plinius Processes and Knowledge Sources

workshops. This process takes advantage of available linguistic resource of WordNet (Miller, 1995). After this phase, a natural sentence is converted into a graph of senses, with each one corresponding to a node in the ontology.

The semantic validation phase checks a LCG and ensures its compliancy with the ontology. The Sensus (Knight & Luk, 1994) ontology was chosen as the ontology used in OntoSeek.

The user query process works in the dual way with the addition of a database search to match the encoded query with all items that it subsumes.

OntoSeek is a powerful search engine since it returns all subsumption resources in the ontology, and it is accurate because of the phase of sense disambiguation. The architecture is shown in Figure 2.3. Solid arrows indicate flow of information; dashed arrows connect among the three main data structures.

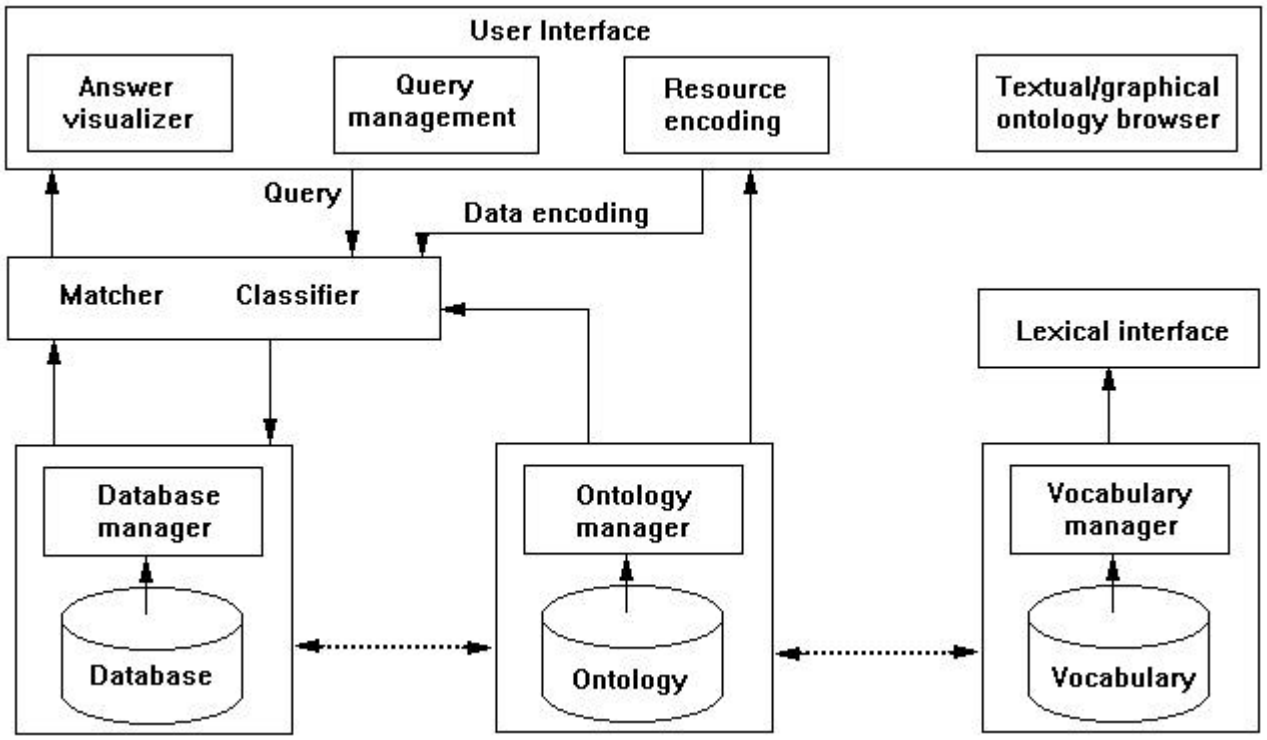

Figure 2.3. OntoSeek's Function Architecture.

Knowledge-based Discovery Tool

Another system that extracts and retrieves relevant document is called "knowledge-based Discovery Tool" (KBDT). It is being developed by Lossau and associates (Lossau et al., 1999). KBDT extracts content from military doctrine and builds knowledge indexes for later search.

The content extracted or knowledge assertion is represented as a group of words and phases. One sentence produces one knowledge assertion that has components of actor, process, other, time, and location, and each component is a collection of words and phases. The words and phases are extracted by breaking down the sentence and ignoring extraneous and irrelevant ones. When encountering a new phase or word, the system relies on a knowledge engineer or user to determine whether to breakdown further or to

add it to the system. The knowledge assertions consist of keywords or key phases from the actual doctrine but take on appropriate roles.

When searching on a user query, the system again parses the user input into assertions and then finds the list of knowledge assertions that overlap the user assertions. Even though only keywords are extracted, this approach is still more powerful than the conventional approach of search on keywords since it groups the keywords into roles which ensures context thus accuracy.

The role of ontology plays in this system is to improve search flexibility. That is user can specify more powerful search by including searching on relationships like parent/child, composition, and membership.

The ontology used in KBDT was initially derived from the WordNet database (Fellbaum, 1998) and revised to suit the military domain.

Information Extraction from Unstructured Documents

A simple yet effective system of extracting information from unstructured documents is developed by David Embley and associates (Embley et al., 1998). The extraction from two different types of documents based on the domain ontology reaches recall ratios in the 80% and 90% range and precision ratios near 98%. The system converts an unstructured document into tuples of attribute values to populate a generated database schema, thus transforming unstructured documents into structured documents that can be further manipulated and searched upon.

Newspaper advertisements for automobiles and newspaper job listings for computer-related jobs were chosen as the input document and as the ontology domain. Both types of documents are data rich and narrow in the ontological breadth, which means those documents has many identifiable constants like names, dates and etc., but

the size of the domain ontology is relative small. The ontology for the automobiles ad consists of nine nodes, one is the car, the other eight are attributes of the car, such as year, make, etc.

Since documents used is narrow in its ontological breath, the application ontology is relatively small and can be built beforehand and stays unchanged during the process of the extraction, unlike the previous two approaches in which the ontology grows as information being extracted and knowledge added. In addition, the nodes in the ontology serve to be attributes in the generated SQL tables thus creating an SQL schema. For the data that does not have a corresponding concept node in the ontology, it simply ignores it.

The contribution of this approach is its straightforwardness and flexibility. It uses syntactic constant recognition instead of semantic interpretation, which enables full automatic wrapper generation. A constant/keyword recognizor takes the input of an unstructured documents and outputs all meaningful combination of the constants and keywords. Then, the keyword proximity is used to determine the final mapping between constants and keywords. For example, when seeing a number, the constant/keyword recognizor will give three plausible mappings to the keywords, year, mileage and price. The distance between the position of the number and the position of the keyword is then used to reject or accept the mappings. Relying on the syntax instead of the meaning of the sentence greatly simplifies the extraction process.

The same process readily adapts to any other unstructured, data rich, and ontologically simple documents via an independent ontology parser. It has been shown that good precision and recall ratios on the different documents it tested on. However, the limitation to this approach lies in the fact that only data rich and ontologically simple

documents can be successfully parsed and turn into useful structured documents, the database entry in this case. The architecture is illustrated in Figure 2.4.

Conclusion

The greatest advantage of ontologies is that by providing a knowledge structure, ontology becomes independent of representation languages while its essential structure enforces knowledge constraints and safeguards its interpretation, meaning, and usage, thus the ultimate integrity of knowledge. We have seen various goals can be achieved by using ontology where impossible or costly otherwise.

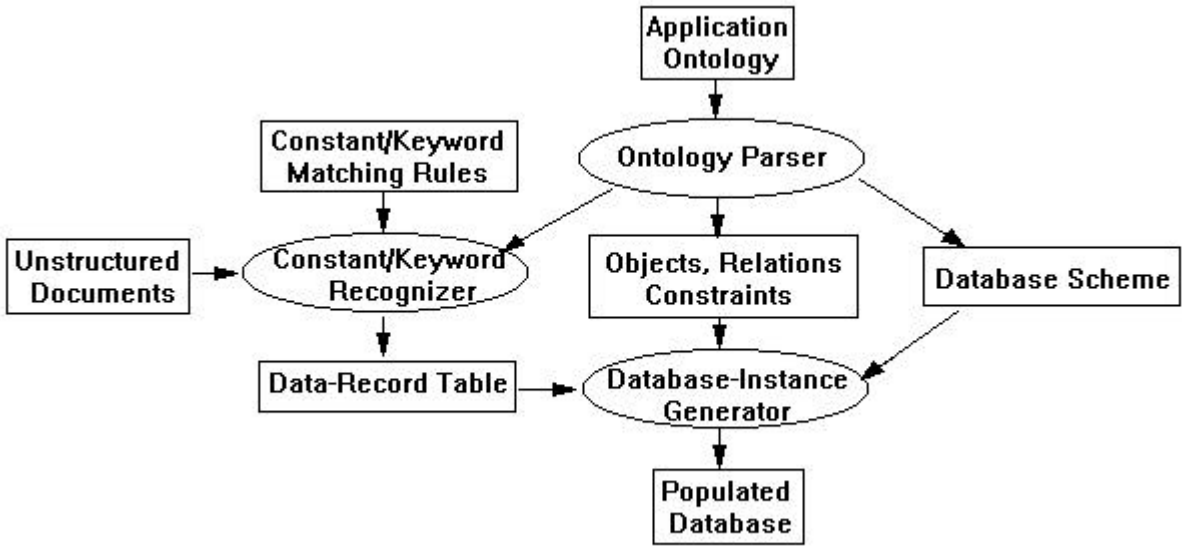

Figure 2.4. Architecture of the Newspaper Advertisement Project

CHAPTER 3
DESIGN DECISIONS

Introduction

The goal of this research is to develop an efficient and accurate way of understanding academic papers. Normally there are two aspects involved in understanding natural texts. First is knowledge extraction. Second is the use of the knowledge extracted. Under the context of this research and taking into consideration of the nature of academic papers, we would like to discover strong academic statements which are supported by more than one paper and controversial ones which are the results of opposite opinions given by different papers.

Document Types

Before discuss the details of design choices, we would like to introduce some background on the nature of the document used in this research.

Structured vs. Unstructured Documents

The type of documents in the form of natural text is called unstructured documents, which are not machine understandable. In order for computers to be able to perform tasks like extracting knowledge and translating from one language to another, the documents have to be put into a machine readable format. Structured document is one of the machine readable texts. One example of structured document is a database record represented as tuples of attributes.

Academic papers naturally fall into the category of unstructured document. In order to convert academic papers into machine readable format, some natural language processing (NLP) is required. NLP is a major branch of Artificial Intelligence that has been described as "AI-hard", which means to solve the NLP problem requires solving the ultimate AI problem (Nilsson, 1998). Many approaches have been developed to process natural text automatically to some extent but the full automatic NLP still awaits. In this research, we would like to present a semi-automatic approach to the NLP that is not completely automatic but does relieve some human effort during the process.

Data rich vs. Concept Rich Documents

Data rich documents have a number of identifiable constants such as dates, places, prices, names and so forth. Newspaper advertisements for automobiles is a good example of such documents. We have seen from the previous chapter, such documents can be used to populate a product database which in turn can be searched upon by users.

On the contrary, concept rich documents can contain a large even infinite number of concepts and the relationships exhibited among the concepts are more sophisticated and interesting. Academic paper is a form of concept rich document that covers more conceptual knowledge than mere data.

Domain

Two factors determined the final choice of using Medline summaries as the unstructured documents used in this research.

Availability. Even though academic papers are available in large quantity, after narrowing down to a specific subject area, there are few academic fields that warrant more than 500 papers. Medline becomes a good choice for its huge database and having

an independent search engine. When searched on keyword, hepatitis causation, 3756 abstracts were returned.

Format. Medline also provides the options of displaying the search results as abstracts, titles, or summaries. Originally, the system was to extract summaries from the abstracts. Since summaries of abstracts are readily available, this part of information retrieval is deleted from the design and implementation.

Relationships

Relationships Embedded in the Summaries

After deciding on Medical summaries as the input to the system, we need to determine the type of summaries to be used. Since we would like to discover repeating academic statements and detect conflicting ones, an obvious choice is to select those summaries that contain causal relationships so that if one summary declares phenomenon A causes phenomenon B, and another summary denies such causality, there is a knowledge conflict.

According to Y. Iwasaki, “It is clear that causality plays an essential role in our understanding of the world … to understand a situation means to have a causal explanation of the situation” (Iwasaki, 1998, 314). For example, phenomenon A causes phenomenon B to occur is more interesting than phenomenon A appears in place B. The first statement is embedded with a causal relationship, and it can be further tested with other causal statements about phenomenon A. The second statement is more about the attributes of phenomenon A, i.e. the location of phenomenon A. We can still test the statement further but it should be relatively simply. Thus, we focus on the statements that

contain causal relationships only. This is the reason that hepatitis causation is the keywords searched in Medline.

Relationships Embedded in the Ontology

An ontology can have more than one relationship among concepts, such as part-of, membership, and parent/child, and axioms have been used to introduce more sophisticated relationship. However, the central form of inference lies in the usage of the subsumption relationships (Brachman et al., 1991). It is found that the concepts extracted from the Medline summaries display subsumption relationships only. As a result, the ontology implemented in this research will impose subsumption relationship only among its concepts.

Efficient information retrieval based solely on subsumption hierarchy has been developed successfully as demonstrated by the ontology-aware database management system (Andrade & Saltz, 1999). However in their approach, named relationships that directly link two concepts are differentiated from complex relationships that are capable of reasoning along the hierarchy structure. In this research, every relationship is a complex relationship in that sense.

Ontology Construction

The decision is to build the domain ontology from scratch, that is the ontology is empty prior to the system execution. This approach differs from all the previous approaches discussed in Chapter 2; all ontologies were built beforehand, and some of them never change throughout the execution. Even though the OntoSeek ontology is evolving and incrementing as more knowledge is inserted, the basic ontology structure is fixed before the system startup.

The unavailability of a medical ontology influences the choice of building the ontology from scratch. If a medical ontology were available, it would still be too broad or too detailed to be manageable for the mere scope of hepatitis causation.

The advantage of building ontology from scratch is that only the concepts encountered during the knowledge extraction process will be present in the ontology. There is no extraneous or irrelevant concepts crowding the ontology. Limiting the ontology as essential and as small as possible makes extending and editing an easier process.

The ontology is also made flexible in a sense that the concepts can be moved, deleted, inserted, and copied as situation arises. The added flexibility ensures the most essential structure of the concepts. For example, there are several ways of viewing the concept, hepatitis. According to Sensus ontology, hepatitis is an infectious disease and a liver disease. At the beginning, we can simply put hepatitis as a subclass of disease and ignore its role as an infectious disease and a liver disease. Only when the system encounters the concepts related to a liver disease or an infectious disease, is a more precise classification of disease necessary. However before then, those subtle classification is not required and will only create an ontology too large and complicated than necessary. Even though the burden of building and maintaining the ontology is passed onto the end user, this approach has the advantage that it works for those domains that do not have an ontology available and the cost is kept minimum.

However, in order to be able to built such a hepatitis ontology from scratch some medical background is needed. In this case Sensus ontology is consulted. Since Sensus ontology is not a medical ontology, a number of medical terms still have to be looked up

from other medical sources. It is very likely that during this process some mistakes are made and as a result misplacement of the terms in the ontology structure. It will be shown later that by detecting knowledge conflicts and correcting the ontology structure the errors can be kept to minimum.

Natural Language Processing

It is decided to employ a user interface to determine the precise meaning of a word. Several factors influences this decision.

Accuracy. For example, hepatitis can mean both the virus causing the hepatitis disease or the disease itself. Even though context can be consulted to disambiguate the sense, since we are using summaries as the input, the contextual information available is minimum, thus human interaction is necessary for the sense disambiguation.

Simplicity. It is also a tradeoff between accuracy and complexity. To achieve more accurate natural language processing, more resources will be used such as a lexicon and a grammar, and neither of them are trivial implementation. Also in order to work well for a medical domain like hepatitis, any general lexicon and grammar would need tremendous extension. This effort though achieves more accuracy and automation is deemed not necessary in this case.

The OntoSeek project takes on a similar approach; it relies on a user interface for sense disambiguation. Plinilus however employs a much more aggressive approach, which uses specialized lexicon and grammar to parse the documents and seeks contextual information to determine the precise meaning of a word. On the other extreme, the newspaper advertisement project uses syntactic constant recognition solely; however, it does depend on the position of a word to decide the meaning.

CHAPTER 4
IMPLEMENTATION AND PERFORMANCE EVALUATION

Training Phase

The purpose of the training phase is to establish a set of relation keywords or key phrases for selecting summaries that are embedded with causal relationships.

The set of 3,756 summaries that is the result of the search on keyword, hepatitis causation, is divided into two sets. One set consisting of 500 summaries is reserved exclusively for training purpose, and the rest for the actual input to the system.

Manual check is performed on each of the training summaries to determine if it is embedded with or without causal relationships. In particular, three types of causal relationships are to be extracted. They are *causality* as embedded in a sentence like A causes B, *association* as in A is associated with B, and *disassociation* as in A is not associated with B. For the causality relationship, the link between A and B is directional which means A causes B but not vice versa. However, both association and disassociation are bi-directional. That is A and B are either associated with each other or not.

For each summary that is embedded with such causal relationships, the keyword or key phrase that designates the relationship is extracted. For example, from the summary, *tattoo application is not associated with an increased risk for chronic viral hepatitis,* the key phrase, *not associated with,* is extracted. The result is a set of 65 keywords and key phrases.

However, the mere appearance of a keyword or a key phrase does not guarantee that the sentence is embedded with causal relationships. For example the summary, *a case-controlled study of the factors associated with spontaneous resolution of hepatitis C viremia*, is equipped with the key phrase, *associated with*, but the summary does not contain an association relationship between two concepts -- there are some factors leading to the *spontaneous resolution of hepatitis C viremia*, but the actual factors are not explicitly stated --, thus this summary has be to discarded.

In order to efficiently retrieve summaries embedded with causal relationships, an effective frequency of a keyword or key phrase is used. The frequency is calculated by dividing the number of times a particular keyword appears in the summaries that contain causal relationships by the total number of times it appears in the training summaries. For example, if a keyword appears in two summaries that are embedded with causal relationships but also appears in ninety eight other summaries that are not embedded with causality, that keyword will receive a 2% effective frequency rating and is very likely to be ignored.

Then every summary is assigned a weight in the range between 0.0 and 1.0. The weight is the sum of the effective frequencies of the keywords appearing in the summary. In the case where the sum exceeds 1.0, the weight is set to 1.0. A minimum weight threshold is then applied to the summaries. In order for a summary to be selected, its weight needs to exceed the minimum weight threshold. A more conservative approach is to assign 0.0 as the minimum weight threshold, which produces every summary that contains causal relationhips however among the selected summaries, 47.5% of them do not contain causal relationships. On the other hand, assigning 1.0 to the threshold

guarantees every summary chosen contains causal relationships but it also results in 43% of the relevant summaries not being retrieved. Balancing between the two extremes produces the optimal result -- when setting the threshold at 60%, 96% of the causal summaries are detected while 13% of the summaries retrieved are useless. This threshold has shown to work well on the actual data -- tested on 500 non-training summaries, 90% of causal summaries are hit with 18% of useless summaries.

This keyword extraction step is crucial since every domain employs its own vocabulary. Certain words are used more often than the others are. In this case, the word, *associated*, is the most frequently used verb or almost the exclusive verb when describing an association relationship between two concepts.

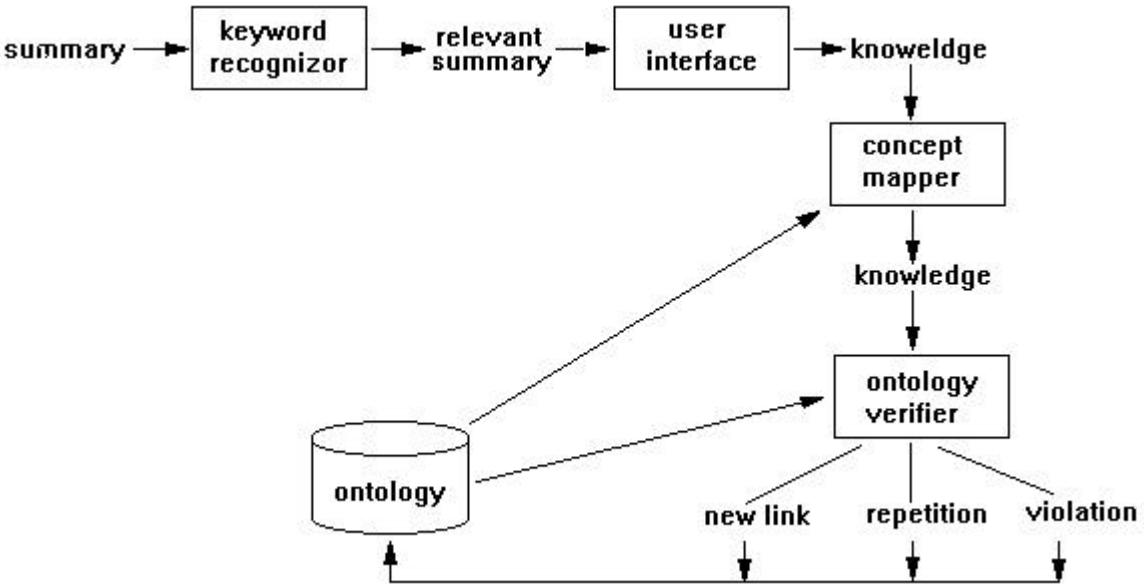

Figure 4.1. System Architecture

Architecture

The architecture of the system is illustrated in Figure 4.1. First, the summaries are processed by a keyword recognizor, which selects and outputs relevant summaries. Then, a user interface is invoked to transform the selected summaries into knowledge. Concept mapper helps mapping new concepts to existing concepts in the ontology. Finally, the ontology verifier determines the extracted knowledge being a new or a repeating or a conflicting knowledge. Both concept mapper and ontology verifier consult the domain ontology.

Keyword Recognizor

Based on the keywords and key phrases extracted during the training phase, the keyword recognizor outputs the summaries that exceed the minimum weight threshold. Although most of the summaries outputted contain causal relationships, a small number of them may not. Thus, the selected summaries need to be further filtered by the user interface to be determined finally whether they contain causality or not. Human effort is thus mainly relieved from the manual selection process.

User Interface

In addition to filtering summaries, the user interface also performs knowledge extraction and sense disambiguation. Each time a summary is chosen, the user is asked to identify the two concepts in the summary that are causally related and to identify the relationship between them.

Several relationships can be embedded in the one summary. From the summary, *hepatitis A, B is related to hepatitis C*, two causal links should be extracted, namely *hepatitis A is related to hepatitis C* and *hepatitis B is related to hepatitis C*. Upon

receiving a summary, the user interface will keep prompting the user until every causal relationship has been extracted.

As discussed earlier, more than one meaning or concept can be associated with one word, and more than one word can describe the same meaning or concept. The user interface is responsible for recognizing the correct sense of the word. The process is integrated in the ontology construction and editing. Therefore, semantic concepts rather than syntactic words are extracted in this step.

Concept Mapper

Concept mapper maps the newly extracted concepts to the existing concepts in the ontology thus saves some human effort. Upon receiving the knowledge or the causal link, the concept mapper first checks the current ontology to determine if the concepts already exist.

There are three scenarios. One is the concepts do exist in the ontology. Then, the system continues on to the next step. The other is when the ontology does not contain the concepts. The user will be asked to insert the new concepts to the ontology. The last scenario happens when the concepts do exist in the ontology but use different terminology. For example, the new concept, liver cancer, refers to the same liver disease as hepatocellular carcinoma, and hepatocellular carcinoma already exists in the ontology. It is the user's responsibility to know the different terminology and to remember if the ontology already has hepatocellular carcinoma as a concept node. If the user is aware, he or she will map liver cancer to hepatocellular carcinoma, and the concept mapper is updated with this new mapping. When the concept mapper consults the ontology again, it will find hepatocellular carcinoma already exists and continues. If the user is not aware, the concept mapper tries to add liver cancer as a new concept and prompts user to specify

a position to insert liver cancer in the ontology, i.e. to specify the proper superclass and subclass of the concept. Ultimately when the user looks up the ontology, he or she will notice the existence of hepatocellular carcinoma thus abandons the insert operation and updates the concept mapper with the new mapping. Any subsequent encounter of liver cancer will then be mapped to hepatocellular carcinoma without user awareness. To some extent, this concept mapper relieves human effort of memorizing the different terminology for the same concept. However, this step requires the user to be domain literate.

There are situations that when inserting a new concept, the ontology needs to be updated by not only one concept node but also several nodes. For example, initially the ontology contains the nodes, chronic hepatitis and acute hepatitis, as the only two subclasses of hepatitis. After introducing hepatitis B infection, the user will discover some overlapping between these concepts; a chronic hepatitis B infection is a hepatitis B infection and a chronic hepatitis, and the same is true for acute hepatitis B infection. Since the concepts intersect, it is necessary to introduce a new concept, chronic hepatitis B infection, to be the subclass of both hepatitis B infection and chronic hepatitis to ensure the integrity of the concepts.

Ontology Verifier

The principle behind the ontology verifier

The ontology verifier is based on the principle of inheritance that is children automatically inherit the properties -- in this case, cause-effect links -- of their parents. To illustrate, hepatitis A is a subclass of hepatitis, and if there is some medicine that can cure hepatitis, that medicine should be able to cure hepatitis A. The dual of the inheritance works the opposite direction; if a concept is associated with another concept then it

should also be associated with the subclasses or the children of that concept. This relationship is illustrated in Figure 4.2.

Upon linking node A to node B, the children of B, B1, B2 and B3 become linked to node A, and in the other direction node A is also linked to B1,B2 and B3. The link in the graph is directional implying a causal relationship. When the edge is bidirectional, the link indicates a mutual relationship between the concepts. For example, hepatitis C virus is associated with hepatitis C infection, which can be interpreted as hepatitis C virus causes hepatitis C infection and hepatitis C infection will carry hepatitis C virus.

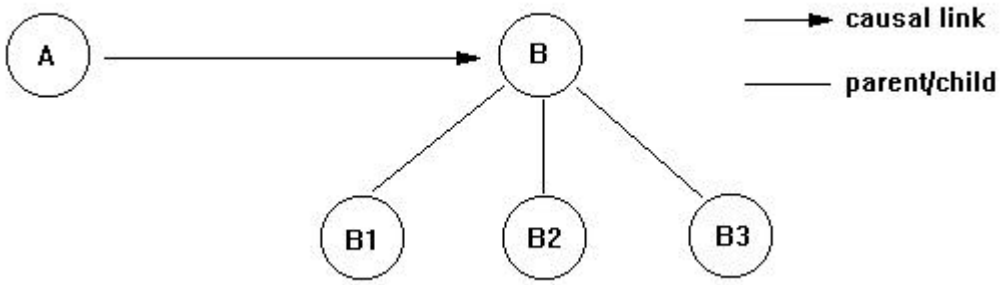

Figure 4.2. Inheritance and Relationship

Interestingly even though node A is casually connected to the children of B, it is not causally connected to its own children. The parent/child relationship should be differentiated from the cause-effect relationships. This restriction plays an important role when deciding the path between two concepts.

Reinforcement and conflict detection

The main purpose of the ontology verifier is to discover the interesting aspects about a summary that is if that summary or more appropriately the causal link extracted from the summary repeats or contradicts with what is already known. When the causal link repeats, clearly more than one Medline paper supports that knowledge thus a

reinforcement should be made. When the causal link contradicts, the summary must present an idea that disagrees with previous ideas thus a conflict should be detected.

Repetition and conflict detection can be reduced to finding a causal path between two concepts. The algorithm for finding the path is similar to the ones in the graph theory.

To detect a repetition, upon adding a new summary or causal link to the system, the ontology verifier tries to locate a path between the two concepts in the current knowledge base. When found, that knowledge is said to be repeating with the existing knowledge. The system differentiates between association path and disassociation path. To detect a disassociation repetition, a disassociation path is searched for, and an association repetition, an association path. The algorithm is the same for the two types of paths.

Conflict detection requires more attention. When an association link is added, the system validates every disassociation link in the knowledge base to ensure the newly added association does not create a new association path among the disassociated concepts. Also, when a disassociation link is added, the system searches for any association path between the two concepts. If a path is found, a conflict is detected.

Shortest Path

Because the ontology has a number of parent/child relationship and a quite few of concepts have the same child, the path between two concepts can extend quite far. In addition, multiple paths can exist between two concepts. For example, there are two paths from hepatitis C virus infection to lichen planus as shown in Figure 4.3.

One path is the direct causal link between the two. Another path follows the reasoning: hepatitis C virus infection is associated with hepatitis C virus, and in term

hepatitis C virus causes cutaneous disorder, thus hepatitis C virus infection will cause cutaneous disorder, since lichen planus is a form of cutaneous disorder, hepatitis C virus infection also causes lichen planus.

However, the most succinct knowledge is usually the most powerful. Therefore, the system selects the shortest path among those existed, in this example, the direct link from hepatitis C virus infection to lichen planus. The algorithm for selecting the shortest path is based on stack operations and is carried out at the same time as path detection. Please refer to the appendix for the algorithm.

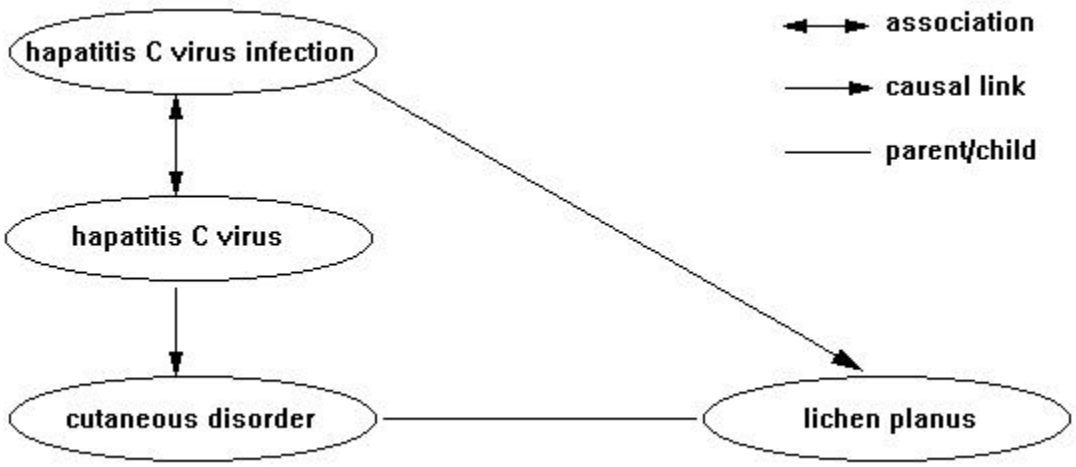

Figure 4.3. Two Paths from Hepatitis C Virus Infection to Lichen Planus

Minimum Link Weight and Maximum Path Length

In order for the path detection to work well, each link is associated with a link weight, and each path is associated with a path length, both of which are under the constraint of a minimum link weight threshold and a maximum path length threshold.

Minimum link weight

The weight on a link indicates how strong that link is. The user is asked for a minimum link weight threshold to ensure the links along the path is of some strength. Every link has its weight set to the default maximum value, 1.0, at the system startup. As the system executes, the weight of the links can be adjusted by the user, and the weight of a path is the minimum link weight of the links along the path.

When a knowledge repetition is detected, user can have two choices, either to strengthen all the link weights along the path or to modify individual link weight. Similar choices are available when a knowledge conflict is detected except that the user can choose to weaken all the link weights.

To illustrate, if a path is detected between two concepts, but the user realizes some link on that path is not strong enough for the whole path to be valid, the user can lower the weight of that particular link so that it will not be a factor in the future. Figure 4.4 demonstrates this situation. There are two paths between hepatitis C virus and hepatocellular carcinoma. One is the direct link from hepatitis C virus and hepatocellular carcinoma, and the other follows the path from hepatitis C virus to liver cirrhosis to cirrhotic hepatocellular carcinoma then to hepatocellular carcinoma. The link weights are initialized at 1.0 thus a knowledge repetition will be detected when the link from hepatitis C virus and hepatocellular carcinoma is inserted, however, the user notices while hepatocellular carcinoma is definitely a superclass of cirrhotic hepatocellular carcinoma, liver cirrhosis is not a superclass. If the user chooses to lower the weight between cirrhotic hepatocellular carcinoma and liver cirrhosis to be under the minimum link

weight threshold, 0.5 in this case, then no knowledge repetition will be detected in Figure 4.5 when the link from alcohol to hepatocellular carcinoma is introduced to the system.

It is shown that by allowing the users to change the link weight, the ontology can evolve to be more accurate and precise.

The minimum link weight threshold allowed in the system is 0.0, which will output all possible paths between two concepts disregarding the strength of the links. The threshold can also be set at the maximum value of 1.0, under which all links along the path have to be 100% strong; thus the ontology verifier will output the strongest links or the most acknowledged knowledge.

Maximum path length

The user also specifies the maximum length of the path, which is the number of links along the path, for the path detection. Since paths can extend quite long, limiting the path length produces outputs that are more meaningful. For example, there is a path from hepatitis B virus infection to porphyria cutanea tarda via the following links.

hepatitis B virus infection is associated with hepatitis B virus

hepatitis B virus is associated with HIV

HIV is associated with hepatitis E virus

hepatitis E virus is associated with hepatitis C virus

hepatitis C virus is associated with cutaneous disorder

porphyria cutanea tarda is a subclass of cutaneous disorder

In order for this path to be meaningful, only when a hepatitis B patient is also infected with three other viruses can he be having the porphyria cutanea tarda disorder. Normally, the longer the path, the less significant it is. Clearly, a successful path length threshold value depends on the particular application domain. In this case, 94% of

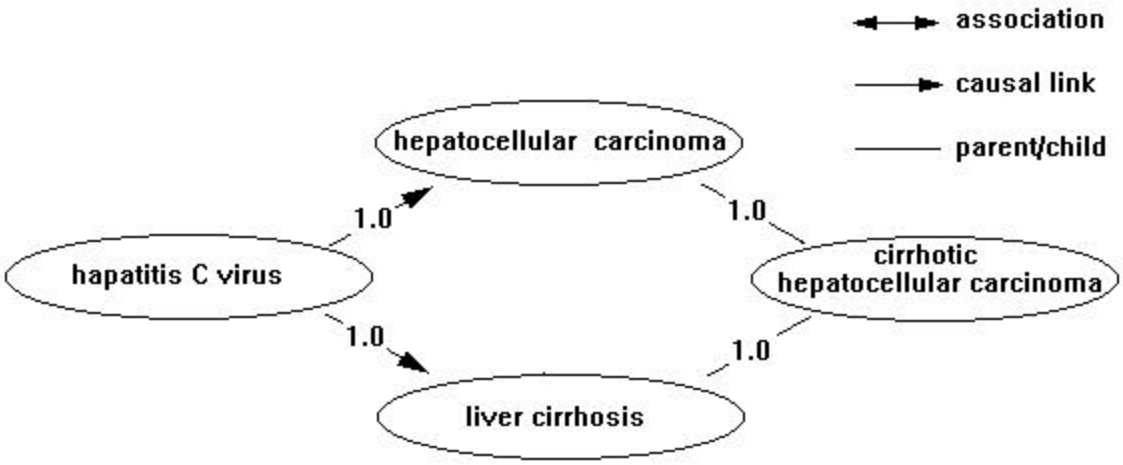

Figure 4.4. Two Paths from Hepatitis C Virus to Hepatocelluar Carcinoma

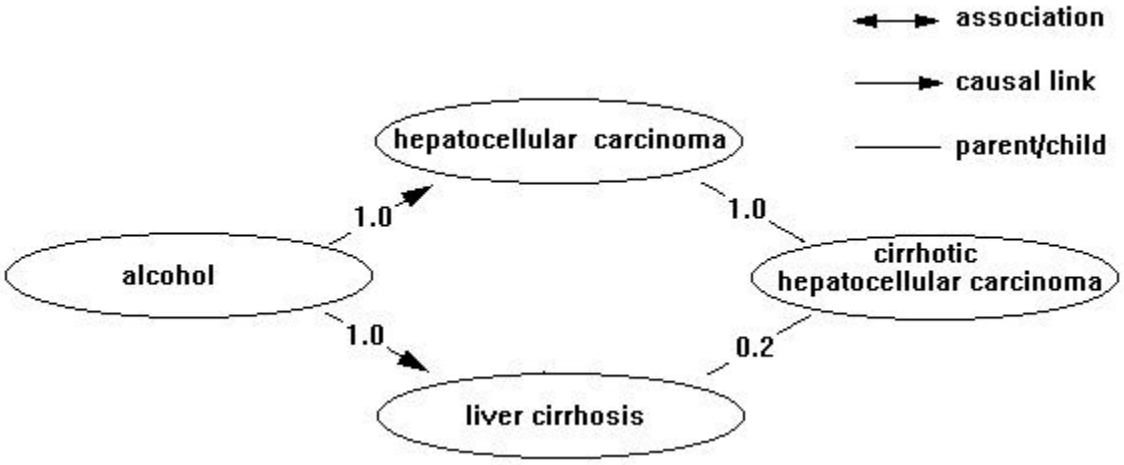

Figure 4.5. One Path from Alcohol to Hepatocelluar Carcinoma Using Minimum Weight Threshold, 0.5

plausible paths are generated when setting the threshold at six. Interestingly paths of length one, two, and three are 73% of total meaningful paths. Figure 4.6 and 4.7 illustrate the distribution.

Results

After executing on 3756 Medline summaries, the ontology composes of 21 concept nodes and 185 knowledge links. Among the 185 links, 44 knowledge repetition

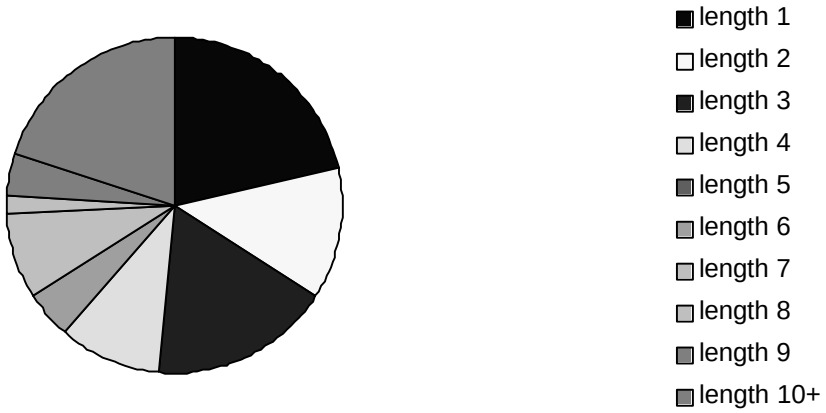

Figure 4.6. Number of Summaries at Various Path Lengths

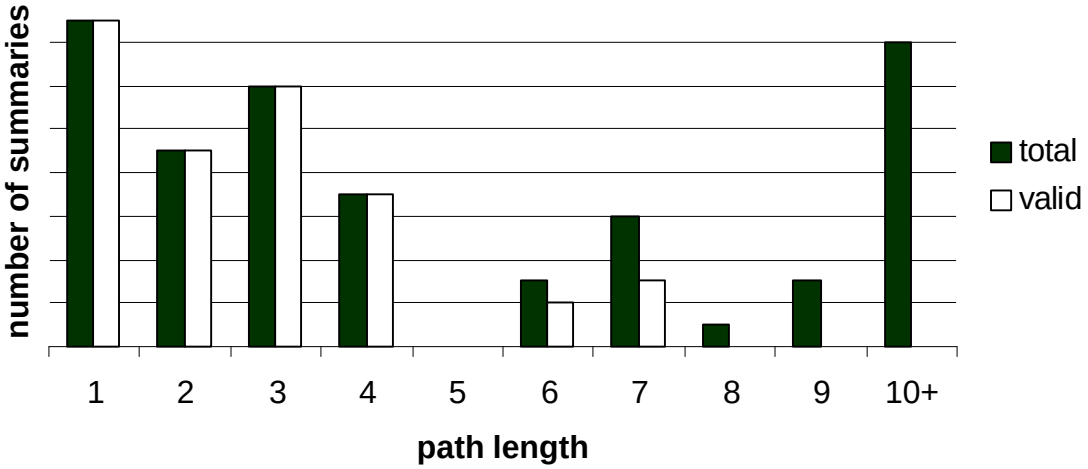

Figure 4.7. Number of Total and Valid Summaries

is detected at 23.7% and 10 knowledge conflict at 5.4%. In addition, 14 changes are made to the ontology at 7.6%. Therefore, the goal of discovering similar or conflicting ideas given by the Medline summaries thus Medline papers has been achieved.

Performance Evaluation

Five criteria are used to evaluate the performance of this ontology-based approach to knowledge extraction.

1. Whether every summary that is embedded with the causal relationship has been detected.

2. Whether the knowledge extracted from the summary is accurate and sufficiently represented.

3. Whether the knowledge repetition detected is correct.

4. Whether the knowledge conflict detected is correct.

5. Whether every knowledge repetition and conflict has been detected.

As discussed earlier in this chapter, the training phase determines the weight associated with a keyword and a minimum weight threshold, which is used to filter the summaries. There is a tradeoff between a low threshold and a high threshold; a low threshold produces more irrelevant summaries while a high threshold less relevant summaries. After experimenting with the threshold, an acceptable balance is reached between the two extremes at 0.6 which yields 90% of relevant summaries and 18% irrelevant ones on the test data. The missing 10% summaries indicate that there are keywords that have not been encountered during the training phase. Those keywords may frequently appear in the test data or may not. The first scenario can be improved if we increase the size of the training data set. However, regardless the size of the training set, there is still a possibility of some keywords only appear in the test set but not training set. On the other hand, if the missing keywords do not play a significant role in the test data set, those keywords can be discarded.

Whether the knowledge extracted from the summary is accurate and sufficiently represented has to do with the representation power of the ontology and the user interface. Under the context of this research, causal relationships between two concepts are the sole knowledge we focus upon, and the user interface is designed with that purpose in mind. Three factors affect the accuracy level of the knowledge extracted.

1. The understanding of the summary itself. Lack of sufficient medical background can cause misunderstanding of the summary thus lead to inaccurate or even incorrect knowledge extraction by the user. This part can be compensated by domain experts.
2. Insufficient information in the summaries. Since summary is a concise statement of a document, it can be ambiguous sometimes even when a keyword is embedded. Here a further consultation of the abstract is necessary. However, to decrease human involvement in the process as much as possible, this approach is not used in this project.
3. The knowledge representation. Knowledge representation restraints the number of entities and the relationships can be imposed on them. It may seem as a drawback by not having more sophisticated relationships among the concepts. However, by limiting the possible interpretation between concepts, more control is given to the system. In addition, providing a degree associated with a relationship has greatly improved the representation power of the ontology.

Whether a knowledge repetition or conflict detected is correct, as discussed earlier, has to do with the length of the path between two concepts. Longer the path, less meaningful the detection usually is. Statistics has shown that more than half of the

detection is made where paths are shorter than three and the paths of length one being the most among them. This implies that quite a number of summaries give the same view on the subjects - paths of length one -- or closely related views - path of length two and three. The correctness of such statements is easier to prove than a long chain of reasoning. Also, statistics are kept for knowledge repetition and conflict so that the repetition and conflict can be reinforced and become stronger knowledge. For example, the link, hepatitis B virus can cause hepatocellular carcinoma, occurs six times alone in the summaries and fifteen times in the paths between other concepts, thus it should receive strong recognition.

To check whether every knowledge repetition or conflict has been detected is straightforward. Since every link is either an association or disassociation, if there is an association, a path should exist, and if there is a disassociation, no path. In addition, the algorithm for detecting path between two concepts has been tested extensively that it does captures every existing path.

In conclusion, the system has reached the goal of extracting knowledge from the Medline summaries and detecting knowledge repetition and conflict semi-automatically. The partial automation is carried out in the steps of keyword recognition and concept mapping that relieve human from reading every Medline summary and memorizing every terminology used.

CHAPTER 5
CONCLUSION AND FUTURE DIRECTION

Ontologies have proven to be a valuable tool in the field of knowledge engineering by acting as a communication medium. In addition, this research project has demonstrated ontologies can play an important role in knowledge verification and extraction.

In order to take advantage of the vast amount of electronic information, automatic or semi-automatic knowledge retrieval has become more important than ever. However, the major obstacle in the information retrieval field still largely lies in the fact that natural language processing is still not as efficient as we would like it to be. This project introduces a semi-automatic knowledge extraction that alleviates some human effort in the process of browsing through the online medical documents.

By keeping the cost low, there are several drawbacks to this approach that can be further improved with more human involvement.

1. The summaries do not necessarily reflect the content of the paper. They are short and condensed. It would be better to perform a complete analysis on the abstract rather than the summaries. This approach requires more sophisticated natural text processing and summary extraction.
2. A more sophisticated and complete medical ontology can improve the accuracy of the system.

3. Medical background is a necessity in understanding the documents and extracting the knowledge.
4. More powerful ontology reasoning can be achieved by applying more sophisticated relationships.
5. Extracting more concepts can provide more complete knowledge representation of the summaries.

APPENDIX
SHORTEST PATH ALGORITHM

Search(*concept*,*destination_concept*) detects paths between *concept* and *destination_concept*.

search(*concept*, *destination_concept*)

1. push *concept* onto stack

2. if *concept* = *destination_concept*,

    if the size of the stack < *shortest_length*

        update *shortest_path*

        update *shortest_length*

3. for each concept, *linked_concept*, that *concept* links

    search(*linked_concept*,*destination_concept*)

    for each child, *child_concept*, of concept, *linked_concept*

        search(*child_concept*,*destination_concept*)

4. get the parent of concept, *parent_concept*

    search(*parent_concept*, *destination_concept*);

5. pop one element off stack

Based on Search(*concept*, *destination_concept*), the shortest path between *start_concept* and *destination_concept* is obtained :

initialize *shortest_path*

shortest_length = MAX_VALUE

search(*start_concept*, *destination_concept*)

print *shortest_path*

LIST OF REFERENCES

Andrade, H., and J. Saltz. "Towards a Knowledge Base Management System (KBMS); An Ontology-Aware Database Management System (DBMS)." *The Proceedings of the 14th Brazilian Symposium on Databases, Florianopolis, Brazil*, October 11-13, 1999.

Brachman, R., D. McGuinness, P. Patel-Schneider, A. Borgida and L. Resnick. "Living with CLASSIC: When and How to Use a KL-ONE-Like Language." *Principles of Semantic Networks*, 401-456, May, 1991.

Embley, D., D. Campbell, R. Smith and S. Liddle. "Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents." *CIKM*, 52-59, 1998.

Farquhar, A., R. Fikes, W. Pratt and J. Rice. "Collaborative Ontology Construction for Information Integration." *Technical Report KSL-95-63*, Stanford University Knowledge Systems Laboratory, 1995.

Fellbaum, C. Ed. *WordNet: An Electronic Lexical Database*. Cambridge, Mass.:MIT Press.1998.

Fensel, D., L. Horrocks, F. Van Harmelen, S. Decker, M. Erdmann and M. Klein. "OIL in a Nutshell." *Knowledge Acquisition, Modeling, and Management, Proceedings of the European Knowledge Acquisition Conference*, 2000.

Genesereth, M., and R. Fikes. *Knowledge Interchange Format, Version 3.0 Reference Manual*. Computer Science Department, Stanford University, June 1991.

Gruber, T. "A Translation Approach to Portable Ontologies." *Knowledge Acquisition*, Vol. 5, No. 2, 199-220, 1993.

Gruber, T. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing." *International Journal of Human and Computer Studies*, Vol. 43, No. 5/6, 907-928, 1995.

Guarino, N. *Formal Ontology in Information Systems*. Amsterdam, Berlin, Oxford, Tokyo, Washington, DC: IOS, 1998.

Guarino, N. "Understanding, Building and Using Ontologies. A Commentary to *Using Explicit Ontologies in KBS Development*, by van Heijst, Schreiber, and

Wielinga." *International Journal of Human and Computer Studies*, Vol. 46, No. 2/3, 293-310, 1997.

Guarino, N., C. Masolo and G. Vetere. "OntoSeek: Content-Based Access to the Web." *IEEE Intelligent Systems,* Vol. 14, No. 3, 70-80, May/June 1999.

Knight, K., and S. Luk. "Building a Large Knowledge Base for Machine Translation." *Proc. Amer. Assoc. Artificial Intelligence Conf. (AAAI-94),* AAAI Press, Menlo Park, Calif., 773-778, 1994.

Iwasaki, Y. "Causal Ordering in a Mixed Structure." *Proceedings of AAAI '88*, 313-318, 1988.

Lee, J., M.Grunninger, Y. Jin, T. Malone, A. Tate, G. Yost and other members of the PIF Working Group. "The PIF Process Interchange Format and Framework, Version 1.2." *The Knowledge Engineering Review*, Vol. 13, No. 1, 91-120, 1998.

Lenat, D., and R. Guha. *Building Large Knowledge-based Systems: Representation and Inference in the CYC Project*. Addison Wesley, 1990.

Lossau, K., I. Mayk, C. York and E. Eilerts. *Developing A Knowledge Based Index to Distributed, Disparate Data Sources*. Austin Information System, 1999.

Miller, D. "WORDNET: A Lexical Database for English." *Comm. ACM*, Vol. 2, No. 11, 39-41, 1995.

Mowbray, T., and R. Zahavi. *The essential COBRA: System Integration Using Distributed Objects*. John Wiley and Object Management Group, 1995.

Mueller, E. *Natural Language Processing with ThoughtTreasure*. New York: Signiform. 1998.

Nilsson, N. *Artificial Intelligence: a New Synthesis*. San Francisco, Calif. : Morgan Kaufmann Publishers, 1998.

Picht, H., and J. Draskau. *Terminology: An Introduction*. Guilford: University of Surrey, 1985.

Schlenoff, G., M. Gruninger, F. Tissot, J. Valois, J. Lubell and J. Lee. *The Process Specification Language (PSL) Overview and Version 1.0 Specification*. NISTIR 6459, National Institute of Standards and Technology, Gaithersburg, MD, 2000.

Schreiber, G., B. Wielinga and W. Jansweijer. "The Kactus View on the 'O' Word." *Workshop on Basic Ontological Issues in Knowledge Sharing. International Joint Conference on Artificial Intelligence*, 1995.

Skuce, D. “Integrating Web-Based Documents, Shared Knowledge Bases, and Information Retrieval for User Help.” *Computational Intelligence,* Vol. 16, 95-113, 2000.

STEPTools. The ISO step standards. Technical report, STEP Tools Company, 1999.

Uschold, M., and M. Gruninger. "Onotologies: Principles, Methods and Applications." *Knowledge Engineering Review,* Vol. 11, No. 2, June 1996.

Uschold, M., M. King, S. Moralee and Y. Zorgios. "The Enterprise Ontology.” *Knowledge Engineering Review,* Vol.13, No.1, 31-89, March 1998.

BIOGRAPHICAL SKETCH

Jionghua Ji was born in Shanghai, P.R.China. After she completed high school, she came to the University of Florida and enrolled in undergraduate study majoring in computer and information science and engineering. After earning her Bachelor of Science with honors in 1997, she worked for Information Systems at the University of Florida for 18 months before returning to school. Currently, she is pursuing her Master of Science degree in computer and information science and engineering at the University of Florida.