# KNOWLEDGE REPRESENTATION AND INDEXING USING THE UNIFIED MEDICAL LANGUAGE SYSTEM

KENNETH BACLAWSKI
*Jarg Corporation*
*Waltham, MA 02453*
*and*
*College of Computer Science*
*Northeastern University*
*Boston, MA 02115*

JOSEPH CIGNA
*College of Health Science*
*Northeastern University*
*Boston, MA 02115*

MIECZYSLAW M. KOKAR
*Department of Electrical and Computer Engineering*
*Northeastern University*
*Boston, MA 02115*

PETER MAGER
*Jarg Corporation*
*Waltham, MA 02453*

BIPIN INDURKHYA
*Tokyo University of Agriculture and Technology*
*Tokyo, Japan*

Ontologies and semantic frameworks can be used to improve the accuracy and expressiveness of natural language processing for the purpose of extracting meaning from technical documents. This is especially true when a rich ontology such as the Unified Medical Language System (UMLS) is available. This paper reports on some tools being developed to make this possible and on some experience with a user interface based on ontologies and semantic networks that allows for interactive knowledge exploration.

## 1 Introduction

Standard Natural Language Processing (NLP) techniques such as lexical scanning (i.e., splitting the text into discrete words), morphological analysis (also known as "stemming") and parsing are important tasks in the process of automated understanding of natural language text. However, general purpose

NLP techniques can only process a document syntactically. To understand a document, it is necessary to have a semantic framework for understanding and means for representing the knowledge contained in the document. Such a framework can be supported with an *ontology*, consisting of a vocabulary and theories of various kinds expressing the meaning of the vocabulary terms within the community using the vocabulary. Given an ontology, knowledge can be represented using semantic networks and the vocabulary of the ontology.

In this paper we report on the use of a large biomedical ontology, namely the Unified Medical Language System (UMLS),[1,2] for the purpose of constructing and indexing knowledge representations of biomedical documents. In Section 3 we describe the elements of the UMLS that are useful for NLP. While the UMLS is a very rich and well structured ontology, some effort was required to adapt the UMLS for use as a basis for understanding biomedical text. We discuss in Section 4 our approach to NLP in which the processing steps use both domain-independent and domain-specific background knowledge.

To evaluate the effectiveness of our natural language processing tools, we conducted a survey of biomedical personnel. This survey is discussed in Section 6. In this survey, we showed several examples of biomedical knowledge representations produced by our NLP tools to a biomedical subject. The subject was asked to evaluate the knowledge representations and to compare them with traditional keyword representations. The results of the survey are presented in Section 7. The knowledge representation diagrams produced by our NLP tools and used in the survey are called *keynets*. The keynet representation language is founded on well-established principles from the data and knowledge representation communities as well as the object-oriented programming community. For more details about this foundation, see the Resource Description Framework (RDF) [3]. Keynets are closely related to the RDF knowledge representation. Although our survey used knowledge representations produced by our NLP tools, the results should be more generally applicable to any NLP tool that represents knowledge using either RDF or the UMLS.

## 2 Related Work

In the past, we have developed ontology-based techniques for representing, indexing and retrieving biological documents, materials and methods [4]. This early work used a relatively small ontology. Since that time we have been researching the problems of scaling up these techniques to large ontologies, especially the UMLS.

Most of the existing work on ontologies deals with reasoning and knowledge sharing. We recognized already in 1992 that ontologies can be used effectively

for information retrieval [4,5,6]. With the dramatic increase in interest in the World Wide Web since 1996, ontologies are now beginning to be used for information retrieval in a number of other research systems such as OntoSeek [7] and commercial ventures such as InQuizit [8,9,10]. These systems are based on ontologies such as WordNet [11] that are not specific to any particular domain, as in the case of the UMLS for biotechnology.

A review of many existing approaches to ontologies can be found in Fridman [12], and a discussion of research issues related to developing an ontology for biological knowledge can be found in Hafner and Fridman [13]. A more recent review of ontologies specifically for molecular biology can be found in Schulze-Kremer [14] which also outlines a prospective ontology for molecular biology.

Our NLP tools construct the knowledge representation by processing the text through a series of stages. This technique and the stages that we use are typical of NLP systems. See Cowie and Lehnert [15] for a survey of information extraction technology. The NLP tools that we have developed differ from standard tools in using not only domain-dependent background knowledge but also domain-dependent knowledge. In addition, we use a much larger and more richly connected ontology than those used in other systems. The UMLS is over 10 times as large as the ontology used by OntoSeek [7,1,2].

## 3 Using the UMLS to Express Document Semantics

The ontology for knowledge representation in a given domain has a number of components. A component is a class of objects with associated attributes, relationships, and behaviors such as inference rules. These affect the process of constructing knowledge representations and the subsequent management of the knowledge representation.

The main components of an ontology relevant to constructing knowledge representations from technical articles in a domain are as follows:

1. *Semantic categories or types*: The UMLS currently has over 130 semantic categories. "Organism", "Anatomical Structure" and "Mental or Behavioral Dysfunction" are examples of UMLS semantic categories. The definition of "Organism" is "Generally, a living individual, including all plants and animals".

2. *Semantic relationships*: A semantic relationship associates two semantic categories. The most important semantic relationship is the **isa** relationship, which specifies that one category or concept is a special case of another. The UMLS currently has over 50 semantic relationships. Two examples of UMLS relationships are **physically_related_to** and

**part_of**. The definition of **physically_related_to** is "Related by virtue of some physical attribute or characteristic".

3. *Categorical links*: A categorical link is specified by two semantic categories and a semantic relationship. Such a specification asserts that the semantic relationship can hold between the first semantic category and the second semantic category. The existence of a link between two semantic categories does not automatically imply that there is a link between specializations of the two semantic categories. The UMLS currently has over 7,000 categorical links. As an example of a categorical link, "Anatomical Structure" is related by the **part_of** relationship to "Organism".

4. *Semantic concepts*: Semantic concepts form the vocabulary of an ontology. The UMLS currently has over 475,000 semantic concepts. For example, "Chronic fictitious illness with physical symptoms" and "Munchausen Syndrome" are two different concepts in the UMLS. Both of these are categorized by the semantic category "Mental or Behavioral Dysfunction".

5. *Concept Maps*: A concept map defines a "mapping" from source data to a semantic concept. The source data is usually text in a technical article. To assist the NLP, the concept map is annotated with a part of speech, a semantic category and other syntactic and semantic relationships. The UMLS currently has over 1,000,000 concept maps.

6. *Categorizations*: A categorization relates a semantic concept to a semantic category. A single semantic concept can be associated with more than one semantic category. There are over 600,000 categorizations currently in the UMLS. For example, both "Chronic fictitious illness with physical symptoms" and "Munchausen Syndrome" are categorized by the semantic category "Mental or Behavioral Dysfunction".

7. *Conceptual links*: A conceptual link is specified by two semantic concepts and a semantic relationship. The most important semantic relationship is the **isa** relationship. Other conceptual links can usually be inferred from the categorical links of the categories to which the concepts belong. The UMLS has 372,808 conceptual links. After eliminating the inverse links, there are 186,669 conceptual links. For example, "Munchausen Syndrome" and "Chronic fictitious illness with physical symptoms" are related by the relationship **mapped_to**.

8. *Explanations*: A concept may be explained in a variety of ways such as the preferred textual representation, full form of an abbreviation or acronym, generic name for a trademark name, definition and so on. For example, "COLD" is an acronym for "Chronic obstructive lung disease".

9. *Knowledge Representation Rules*: A knowledge representation rule consists of an identifier, a grammar rule, a predicate that determines whether the knowledge representation rule applies and a method that builds the knowledge representation when the knowledge representation rule applies. The knowledge representation rules define a mapping from parse trees to knowledge representations.

## 4 Constructing Knowledge Representations

Diagrammatic representations are ubiquitous in molecular biology and medicine. For example, in a textbook such as Cox and Sinclair [16] most of the figures are informal knowledge representations. Our contention is that NLP, combined with keynets, can be used to automate the process of organizing and visually representing retrieved biomedical information.

The tools we developed for constructing knowledge representations from biomedical natural language text are based on two kinds of background knowledge:

1. *Domain-independent knowledge.* This includes NLP processing knowledge such as parts of speech and grammar rules. This kind of background knowledge will be called *syntactic* knowledge.

2. *Domain-dependent knowledge.* This includes the terminology of the domain as well as other components to be introduced below. This kind of knowledge will be called *semantic* knowledge.

The process of constructing knowledge representations takes as input a document and produces as output a knowledge representation conforming to the ontology. The knowledge representation is constructed by processing the text through a series of stages known as the Knowledge Representation Pipeline. See Figure 1.

### *4.1 Scanning*

The first step in the process of constructing knowledge representations is scanning. The data source is examined and a set of discrete elementary units is produced. These elementary units are called "lexemes", which are objects that

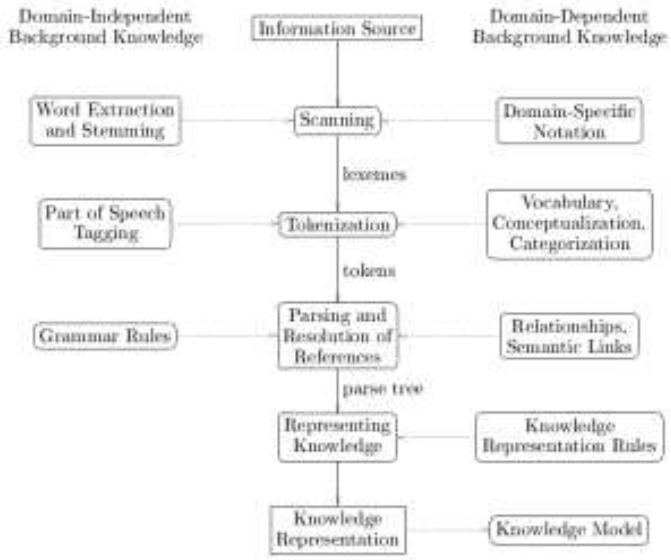



Figure 1: Knowledge Representation Pipeline

are related to one another by low-level relationships such as physical proximity in the data source.

In our NLP tools, the scanning process is dependent on both domain-independent and domain-dependent background knowledge. Domain-specific notation is very common in biology and medicine, and it is important for NLP to be able to recognize commonly used notations.

The lexical scanner can also perform stemming, which extracts the base (uninflected) form of a lexeme from the lexeme that actually occurs in the source document. Stemming reduces the size and complexity of the conceptualization mapping needed for the tokenization step that is discussed next.

Specialized nomenclature such as protein and chemical names are extracted using domain specific extracting rules such as those in Fukuda et al. [17], except that we combine this technique with the specialized vocabulary of the UMLS ontology. This allows us to identify more frequently names that differ in syntax yet refer to the same material.

### *4.2 Tokenization*

Tokenization converts lexemes to conceptual units, called "tokens", which relate a piece of the original source data to a semantic concept of the ontology. Tokenization also annotates the semantic concept using a part of speech and a semantic category. The relationship between lexemes and tokens is not one-to-one. For example, "in vivo" consists of a pair of lexemes, but represents a single semantic concept and therefore a single token. In general, a token corresponds to a sequence of lexemes. The conceptualization mapping discussed above defines the mapping from sequences of lexemes to semantic concepts. There can be many sequences of lexemes that map to the same semantic concept.

Another important responsibility of tokenization is determining the part of speech of a lexeme. For example, the word "store" can be either a noun or a verb. The semantic analog of a part of speech is the *semantic category*. Like a part of speech, a semantic category specifies the context of a semantic concept. For example, "fibrin" can belong to the semantic category "Protein," the semantic category "Biologically Active Substance," or the semantic category "Chemical," depending on the context.

### *4.3 Parsing and Resolution of References*

Parsing takes as input a sequence of tokens, and produces a parse tree using grammar rules defined by the language of the data sources. These grammar rules are not domain-specific, so they are not part of the ontology.

Pronouns and other anaphoric constructions are converted to references from one node in a parse tree to another node in a (possibly different) parse tree. Resolution of references is also not domain-specific, so it is not part of the ontology.

Both parsing and anaphoric references are highly ambiguous in most textual documents. Disambiguation is an important part of this step of natural language processing. The ontology can be used to reduce the ambiguity of the text by determining which of several possibilities has the most reasonable meaning or even has any meaning at all.

### *4.4 Knowledge Representation*

The parse tree and anaphoric references are converted to a knowledge representation using knowledge representation rules. Each rule consists of a condition and an action to be performed if the condition is satisfied. For example, one rule might state that if an Immunologic Factor occurs as the subject of a clause

while a Cell occurs as the object, then by default the subject of the clause is linked to the object of the clause by the UMLS relationship **disrupts**.

### *4.5 Indexing of Knowledge Representations*

The extracted knowledge representations are indexed using a proprietary, patented indexing technology [18,19]. This technology indexes knowledge representations by fragmenting them into smaller pieces which are then indexed in a distributed high-performance indexing engine. The indexing technique was inspired by the use of probes for genetic mapping, except that our technology performs high-resolution semantic indexing and classification, while genetic probes are associated with low-resolution genetic mapping.

## 5 Semantic Based User Interfaces

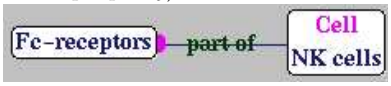
We are developing a user interface that provides insight into the underlying knowledge representation of queries and retrieved data. The user interface displays a knowledge representation, called a *keynet*, constructed from the query or data. The keynet shows the semantic context of the query or response by incorporating information about the category of concepts occurring in the query or retrieved data and relationships among these concepts. The keynet can be used as the basis for an expanded investigation that can include information about other concepts in the same general category that participate in semantic relationships to the concept originally being investigated.

## 6 Usability Survey

The purpose of the usability study was to explore the reactions of users to different representations of biomedical knowledge. To do this we compared two techniques for representing the biomedical knowledge expressed by biomedical text: keywords and keynets. The Keywords approach is used by Medline and other biomedical classification systems. Keynets are a diagrammatic knowledge representation technique intended to describe and to categorize biomedical knowledge.

A *keynet* consists of a collection of boxes (vertices) linked by lines (edges). A box contains a word or short phrase from the text. A box represents a semantic concept. If the word or short phrase is found in the UMLS, then the box also has a semantic category appearing on the top line in the box. For example, "fibrin" is categorized as "Protein" in the following: 

. A line joining

two boxes is a relationship between the two semantic concepts. A label on the line identifies the *relationship type* (also called a *property*). A dot indicates the source of the relationship. For example, 

means that Fc-receptors are part of NK cells. In RDF terminology, this is a *statement* in which "Fc-receptors" is the *subject*, "NK cells" is the *object* and "part of" is the *predicate*.

The survey was conducted by interviewing clinicians and researchers who would typically use the Internet to retrieve biomedical information. A total of eleven subjects representing the fields of medicine, biology, biochemistry, pharmacology and biomedical engineering were used in this study. Each of the subjects looked at three example pages, each of which contained

1. Some biomedical text, such as a question or a description of a pharmacological product,

2. The biomedical keywords in the text that appear in the UMLS, and

3. The corresponding biomedical keynet representation of the biomedical text.

The first example is shown in Figure 2.

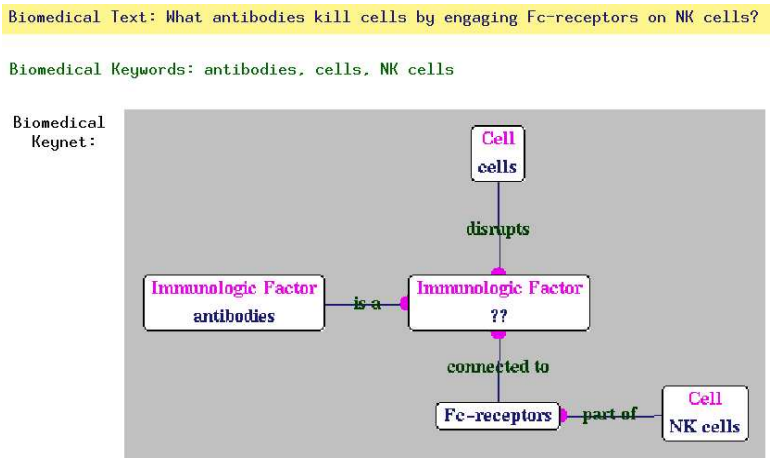



Figure 2: Survey Example 1

The survey was organized in three sections. Section one collected demographics of the subjects. Section two consisted of nine questions that were interpreted using the technique of semantic differential [20] to gauge the subject's overall reaction to keywords and keynets. A scale of 0 to 9 was used to compare examples in the ranges: confusing/clear, terrible/wonderful, frustrating/satisfying, dull/stimulating, difficult/easy, weak/powerful, rigid/flexible, inconsistent/consistent and ambiguous/precise [21]. Subjects also supplied written feedback to each question. Section 3 also uses semantic differential questions and written feedback to explore the subjects reactions to biomedical keynets themselves. This section included questions on learning to use the keynet, the overall presentation of the keynet and a comparison of the three keynets to each other.

## 7 Survey Results

The following are noteworthy findings of the usability survey:

1. **Ease of Understanding**. The level of understanding by the survey participants of the keynet representations was remarkably high given the very brief period used to complete the survey, the diversity of the population studied and the examples used in the survey. For instance, without prompting, 64% (7 of 11) of the subjects detected a missing relationship between two components in one of the keynets. We had inadvertently omitted this relationship when we constructed our keynet representations.

2. **Limit complexity**. The larger diagrams which were used in our survey were more difficult to interpret. The subjects suggested that keynet representations should be concise with only a limited level of detail. More detailed views should be presented only at the user's request.

3. **Keyword versus Keynet**. Statistical analysis of the data revealed no difference between the two methods of representing biomedical text. The brief time period with which the subjects were exposed to keynets may have contributed to this. Nevertheless, the comments indicated that the subjects did believe that the keynet representations were at least as useful and in some cases more useful than keyword representations.

## 8 Conclusion

In this paper we have introduced the role and requirements of an ontology for the purpose of constructing knowledge representations. Specific details are

given for the construction of NLP tools for using an ontology in the special case of the UMLS. We have also tested the usefulness of representing biomedical knowledge using diagrammatic knowledge representations with generally favorable results.

## 9 Future Directions

Feedback from this usability survey is being used as input to the development of a new class of semantics-based search engine. A new usability survey has just been completed that provides further evidence of the usefulness of representing biomedical knowledge using diagrammatic knowledge representations. The long term goal is to provide the capability for users to interact dynamically with the user interface, and to use the knowledge representation of queries and answers as the basis for an expanded investigation of a subject area.

## Acknowledgments

This work was performed as part of the "Biomedical Science Information Retrieval and Management" project. The project described was supported by grant number 1 R43 LM06665-01 from the NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

## References

1. B. Humphreys and D. Lindberg. The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81(2):170–177, 1993.
2. B. Humphreys, D. Lindberg, and A. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281, August 1, 1993.
3. O. Lassila and R. Swick. Resource description framework (RDF) model and syntax specification, Feburary 1999. `www.w3.org/TR/REC-rdf-syntax`.
4. K. Baclawski, R. Futrelle, N. Fridman, and M. Pescitelli. Database techniques for biological materials & methods. In *First Int. Conf. Intell. Sys. Molecular Biology*, pages 21–28, 1993.
5. K. Baclawski, R. Futrelle, C. Hafner, M. Pescitelli, N. Fridman, B. Li, and C. Zou. Data/knowledge bases for biological papers and techniques.

In *Proc. Sympos. Adv. Data Management for the Scientist and Engineer*, pages 23–28, 1993.

6. C. Hafner, K. Baclawski, R. Futrelle, N. Fridman, and S. Sampath. Creating a knowledge base of biological research papers. In *Proc. Second Int. Conf. Intell. Sys. Molecular Biology*, pages 147–155, 1994.
7. N. Guarino. Semantic matching: Formal ontological distinctions for information organization, extraction, and integration. In M. Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, pages 139–170. Springer-Verlag, 1997.
8. K. Dahlgren. A linguistic ontology. *Int. J. Human-Computer Studies*, 43:809–818, 1995.
9. K. Dahlgren. Natural language understanding system, August 1998. United States Patent No. 5,794,050.
10. InQuizit. Website, 1999. `www.inquizit.com`.
11. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.
12. N. Fridman. *Knowledge Representation for Intelligent Information Retrieval in Experimental Sciences*. PhD thesis, College of Computer Science, Northeastern University, Boston, MA, 1997.
13. C. Hafner and N. Fridman. Ontological foundations for biology knowledge models. In *Proc. 4th International Conference on Intelligence Systems for Molecular Biology (ISMB-96)*, pages 78–87. AAAI Press, Menlo Park, CA, 1996.
14. S. Schulze-Kremer. Ontologies for molecular biology. In *Pacific Symposium on Biocomputing*, volume 3, pages 603–704, 1998.
15. J. Cowie and W. Lehnert. Information extraction. *Comm. of the ACM*, 39(1):80–91, 1996.
16. T. Cox and J. Sinclair, editors. *Molecular Biology in Medicine*. Blackwell Science, Ltd., Oxford, UK, 1997.
17. K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *Pacific Symposium on Biocomputing*, volume 3, pages 705–716, 1998.
18. K. Baclawski. Distributed computer database system and method, December 2 1997. United States Patent No. 5,694,593. Assigned to Northeastern University, Boston.
19. Jarg Corporation. Website, 1999. `www.jarg.com`.
20. C. Osgood, G. Suci, and R. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, Urbana, 1957.
21. B. Schneiderman. *Designing the User Interface*. Addison-Wesley, Reading, MA, 1998.