Ontology Enrichment from Free-text Clinical Documents: A Comparison of Alternative Approaches

by

Kaihong Liu

Doctor in Medicine, Beijing University of Chinese Medicine, 1987

Master of Science in Biomedical Informatics, University of Pittsburgh, 2007

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Kaihong Liu

It was defended on

July 21th, 2011

and approved by

Dr. Wendy W. Chapman, Associate Professor, Associate Professor, Division of Biomedical Informatics, University of California San Diego

Dr. Rebecca Hwa, Assistant Professor, Department of Computer Science

Dr. William Hogan, Associate Professor and Chief, Division of Biomedical Informatics University of Arkansas for Medical Sciences

Thesis Director/Dissertation Advisor: Dr. Rebecca Crowley, Associate Professor, Department of Biomedical Informatics

Ontology Enrichment from Free-text Clinical Documents: A Comparison of Alternative Approaches

Abstract

Kaihong Liu, MD, MS

While the biomedical informatics community widely acknowledges the utility of domain ontologies, there remain many barriers to their effective use. One important requirement of domain ontologies is that they achieve a high degree of coverage of the domain concepts and concept relationships. However, the development of these ontologies is typically a manual, time-consuming, and often error-prone process. Limited resources result in missing concepts and relationships, as well as difficulty in updating the ontology as domain knowledge changes. Methodologies developed in the fields of Natural Language Processing (NLP), Information Extraction (IE), Information Retrieval (IR), and Machine Learning (ML) provide techniques for automating the enrichment of ontology from free-text documents. In this dissertation, I extended these methodologies into biomedical ontology development. First, I reviewed existing methodologies and systems developed in the fields of NLP, IR, and IE, and discussed how existing methods can benefit the development of biomedical ontologies. This previously unconducted review was published in the Journal of Biomedical Informatics. Second, I compared the effectiveness of three methods from two different approaches, the symbolic (the Hearst method) and the statistical (the Church and Lin methods), using clinical free-text documents. Third, I developed a methodological framework for Ontology Learning (OL) evaluation and comparison. This framework permits evaluation of the two types of OL approaches that include three OL methods. The significance of this work is as follows: 1) The results from the comparative study showed the potential of these methods for biomedical ontology enrichment. For the two targeted domains (NCIT and RadLex), the Hearst method revealed an average of 21% and 11% new concept acceptance rates, respectively. The Lin method produced a 74% acceptance rate for NCIT; the Church method, 53%. As a result of this study (published in the Journal of Methods of Information in Medicine), many suggested candidates have been

incorporated into the NCIT; 2) The evaluation framework is flexible and general enough that it can analyze the performance of ontology enrichment methods for many domains, thus expediting the process of automation and minimizing the likelihood that key concepts and relationships would be missed as domain knowledge evolves.

TABLE OF CONTENTS

DEDICATION.. XII

ACKNOWLEDGEMENTS XIII

1.0 INTRODUCTION.. 15

1.1 WHAT IS ONTOLOGY? 15

1.2 APPLICATION OF ONTOLOGY 17

1.3 ONTOLOGY DEVELOPMENT AND CURRENT PROBLEMS............... 18

1.4 KNOWLEDGE ACQUISITION AND ONTOLOGY LEARNING 20

1.5 NATURAL LANGUAGE PROCESSING APPROACH FOR ONTOLOGY LEARNING.. 22

2.0 NATURAL LANGUAGE PROCESSING METHODS AND SYSTEMS FOR BIOMEDICAL ONTOLOGY LEARNING.. 26

2.1 NATURAL LANGUAGE PROCESSING METHODS 29

2.1.1 Synonyms and concepts extraction.. 29

2.1.1.1 Symbolic methods 29

2.1.1.2 Statistical methods 32

(a) Methods that use clustering approaches 32

(b) Methods that use machine learning approach.................................... 36

2.1.2 Taxonomic relationship extraction.. 41

2.1.2.1 Symbolic methods ... 41

2.1.2.2 Statistical methods ... 49

2.1.3 Non-taxonomic relationships extraction ... 50

2.1.3.1 Symbolic methods ... 50

2.1.3.2 Statistical methods ... 51

2.1.4 Denovo generation of ontologies ... 55

2.2 NATURAL LANGUAGE PROCESSING SYSTEMS ... 57

3.0 DEVELOPMENT OF RESEARCH QUESTIONS ... 67

RESEARCH QUESTIONS ... 70

4.0 RESEARCH DESIGN AND METHODS ... 73

4.1 STUDY METHOD FOR EVALUATING THE LSP MATCHING METHOD FOR ONTOLOGY ENRICHMENT USING CLINICAL DOCUMENTS 76

4.1.1 Lexico-Syntactic Pattern (LSP) matching method ... 76

4.1.2 Extraction of sentences containing LSPs ... 76

4.1.3 Calculating LSP frequencies and distributions ... 78

4.1.4 Evaluation of ontology suggestions ... 78

4.1.5 Defining evaluation metrics ... 82

4.2 STUDY METHOD FOR EVALUATING THE STATISTIC METHODS FOR ONTOLOGY ENRICHMENT USING CLINICAL DOCUMENTS ... 85

4.2.1 The statistical methods ... 85

4.2.1.1 Church method ... 85

4.2.1.2 Lin method ... 86

4.2.2 Named Entity Recognition system used in this study ... 87

4.2.3 Generating new concept suggestions 88

4.2.4 Evaluation study.. 90

4.2.4.1 Methodology development 91

(a) Establishing a reference standard for methodology development........ 92

(b) Generating optimal parameters for the statistic methods using precision and recall curve 93

4.2.4.2 Final evaluation by domain experts 95

5.0 RESULTS 96

5.1 RESULTS FROM SYMBOLIC METHOD STUDY 96

5.1.1 Lexico-Syntactic Pattern (LSP) matching method 96

5.1.2 How many medically meaningful terms (MMTs) could be identified for each LSP presented in the corpus... 98

5.1.3 New Concept Suggestion Rate (CSR) and new Concept Acceptance Rate (CAR) ……………………………………………………………………….99

5.1.4 Distribution of semantic relationships that being extracted by each type of LSP ……………………………………………………………………..100

5.1.5 New Relationship Suggestion Rate (RSR) and new Relationship Acceptance Rate (RAR)... 102

5.1.6 Estimated Concept Yield (ECY) and Estimated Relationship Yield (ERY) ……………………………………………………………………..103

5.1.7 Reasons for why some of the suggested relationships would not be added in the corresponding resource.. 103

5.2 RESULTS FROM STATISTIC METHOD STUDY.................................... 104

5.2.1 Reference standards 104

5.2.2 Precision and recall curves 104

5.2.3 Parameters generated for Lin method 107

5.2.4 Parameters generated for Church method 108

5.2.5 Summarization of the best parameters generated for both Lin and Church methods 108

5.2.6 New concept suggestion rate and acceptance rate 111

6.0 DISCUSSION 113

6.1 THE EFFECTIVENESS OF THE SYMBOLIC-BASED APPROACH FOR ONTOLOGY ENRICHMENT USING CLINICAL DOCUMENTS 114

6.2 THE FFECTIVENESS OF THE STATISTICAL-BASED APPROACH FOR ONTOLOGY ENRICHMENT USING CLINICAL DOCUMENTS 118

6.3 COMPARISON OF THE EFFECTIVENESS OF THE OL METHODS FOR BIOMEDICAL ONTOLOGY DEVELOPMENT USING THE EVALUATION FRAMEWORK AND METRICS I DEVELOPED 120

7.0 LIMITATIONS 125

8.0 CONCLUSION AND FUTURE WORK 127

APPENDIX A 131

APPENDIX B 133

APPENDIX C 135

BIBLIOGRAPHY 156

LIST OF TABLES

Table 1. Ontology learning task and their corresponding learning methods 28

Table 2. Characteristics of existing ontology learning systems.. 58

Table 3 . Lexico-Syntactic Patterns with examples from corpora 77

Table 4. Syntactic relationships experimented with Lin’s method... 95

Table 5. Frequencies of various LSPs in the pathology and radiology corpora 97

Table 6. Number of medically meaningful terms (MMTs) extracted by the LSP method........... 99

Table 7. Comparison of new concept suggestion rate and acceptance rate 100

Table 8. Distribution of semantic relationships extracted using the LSP matching method 101

Table 9. Comparison of new concept relationship suggestion rate and acceptance rate 102

Table 10. Precisions and recalls of Lin method with different syntactic relationships combination 107

Table 11. Precisions and recalls of Church method using different window sizes..................... 108

Table 12. Best parameters for Church and Lin methods for each domain 110

Table 13. Suggestion and acceptance rates for the Lin and Church methods............................. 112

Table 14. Comparison of acceptance rates of suggested terms extracted by different OL methods 123

LIST OF FIGURES

Figure 1. Overview of the study design 75

Figure 2. Domain expert training process 81

Figure 4. New Concept Suggestion Generated using Statistical Methods.................................... 89

Figure 5. Precision and recall curve of Lin method for NCIT enrichment using pathology reports for all syntactic relationships (Set 1) 105

Figure 6. Precision and recall curve of Lin OL method for RadLex enrichment using radiology reports for all syntactic relationships. 106

DEDICATION

This thesis is dedicated to my mother

Qiying Hang

who introduced me to the joy of learning;

who inspired and enabled my work;

who is unable to see it finished today.

ACKNOWLEDGEMENTS

I wish to express my profound gratitude to my advisor, Professor Rebecca Crowley. This thesis could not have been completed without her expert advice and unfailing patience. I am also most grateful for her emotional support and encouragement, especially through the sometimes-difficult events and circumstances that have unfolded throughout these years.

I am deeply indebted to my many colleagues and members of my committee as well, for their unstinting support during the completion of this project.

I want to thank my daughter, Isabella, for putting a smile on my face every day, and George, my life partner and best friend, who supported me throughout this journey, patiently plying me with words of assurance and many much-needed cups of coffee.

I would like to express a special word of thanks to my friends Eileen, Clark, Laura, Liz, Bob, Heather, Jim, and the countless others who have believed in me, encouraged me, and provided babysitting services when they were most needed. It would take a document as long as this thesis to list them all and the multitude of ways they helped me to achieve my life goal. I know for certain that this work could not have been accomplished without their support. My heartfelt thanks to all!!

Finally, a special thanks to Heather Aronson, Karma Lisa Edward, and Lucy Cafeo for editorial support, and to the National Library of Medicine Medical Informatics Training

Program, and the National Institute of Health Ontology Development and Information Extraction Grant for their financial support.

1.0 INTRODUCTION

1.1 WHAT IS ONTOLOGY?

The term “ontology” has been used for centuries. However, because the determination of what constitutes an ontology varies based on assumed perspectives, philosophers, lexicographers, librarians, and computer scientists have defined “ontology” in many different ways [1-4]. From a computer scientist perspective, Gruber [3] stated that “ontology is an explicit, formal specification of a shared conceptualization of a domain of interest.” Here, the word “conceptualization” refers to an abstract model of some phenomenon in the world that identifies the relevant concepts of that phenomenon. Gruber uses the term “explicit” to make the point that the types of concepts and the constraints on their use are clearly defined. Gruber’s requirement of “formal[ity]” refers to the fact that the ontology should be machine understandable, and his specification that the conceptualization be “shared” reflects the notion that an ontology captures consensual knowledge that is not private, but accepted by a group people who have a common interest. Sowa [5] defined ontology as “the study of the categories of things that exist or may exist in some domain.” “The product of such study,” he wrote, “is a catalog of the types of things that are assumed to exist in a domain of interest from the perspective of a person who uses a language for the purpose of talking about the domain.” A more practical perspective and technical definition dictates that an ontology is a standardized classification system that enables data from different sources to be combined, accessed, and manipulated.

What these perspectives have in common is that they define an ontology as a representation of entities and their relationships in a particular domain. Debates as to whether the ‘entities’ represented are concepts [6] or real-world things [7] continue. Nevertheless, a key requirement is that each entity has one unique reference in the ontology (typically a meaningless identifier to avoid confusion among natural language terms). Each identifier is linked to at least one natural language term, and is often linked to greater than one natural language term to capture the synonymy inherent in human language. A standard ontology facilitates aggregation of data from multiple data sources if each source uses the identifiers from the ontology. Interoperability is one of the primary reasons, if not *the* primary reason, that groups have been engaged in the development of ontologies.

Ontology developers usually capture the relationships among entities as formal, logical relationships. To do so, they frequently use one type of logic from a family of logics known as “description logics.” Description logics constitute a family of fragments of first-order logic (nearly all of which are decidable), in which members of this family are primarily differentiated based on the set of allowed logical operators. For example, some logics exclude negation and universal quantification, which in turn determine the computational complexity of inference with the language. The Web Ontology Language (OWL) is a standard ontology language that captures the semantics of many description logic languages.

1.2 APPLICATION OF ONTOLOGY

Natural Language Processing (NLP) and text mining are research fields aimed at exploiting rich knowledge resources with the goal of understanding, extraction, and retrieval from unstructured text. Knowledge resources that have been used for these purposes include the entire range of terminologies, including lexicons, controlled vocabularies, thesauri, and ontologies. For the purposes of this description, I followed the framework for describing terminologies and terminological systems defined by de Keizer [8, 9] and Cornet [10]. The authors define **concepts** as “cognitive constructs” of objects that are built using the “characteristics of the objects,” **terms** as “language labels” for concepts, and **synonyms** as two or more terms that designate “a unique concept.”

For simple NLP tasks, such as named entity recognition, almost any type of terminology can be used. Slightly more complex tasks, such as identification of concepts, require the representation of synonyms, and therefore limit the resources to terminological systems such as controlled vocabularies, thesauri, and ontologies that encode multiple lexical representations in natural language [11]. For example, “liver cell” and “hepatocyte” would be represented in the vocabulary or ontology as synonyms. Therefore, during Named Entity Recognition (NER) they would be classified as the same concept.

In contrast, some NLP tasks require more complex relationships between concepts and limit the types of terminological systems that may be used. Examples include word sense disambiguation [12], co-reference resolution [13-15], discourse reasoning, and extraction of attributes and values

[16]. For example, if “hepatocellular carcinoma” and “liver neoplasm” are both used in a document to refer to the same entity, then these terms can be determined to co-refer if a relationship is represented in the terminology [17].

Ontologies can be used to make even more complex inferences and to derive rules necessary for semantic interpretation [18, 19] and question-and-answering systems [20]. For this reason, ontologies have been of particular interest to researchers developing NLP systems. For example, to answer the question: “What role do infectious organisms play in liver cancer?” an ontology can be used to perform the query expansion and retrieve related textual information, if the ontology contains the following information: 1) a synonym relationship between ‘liver cancer’ and ‘hepatocellular carcinoma’; 2) a hierarchical relationship between various hepatitis viruses and ‘infectious organisms’; and 3) an etiologic relationship between hepatitis viruses and hepatocellular carcinoma.

1.3 ONTOLOGY DEVELOPMENT AND CURRENT PROBLEMS

At present, the process of ontology development is largely manual. One by one, humans must add identifiers, their synonyms and relationships. The financial investment in labor-intensive ontology development is huge. The National Human Genome Research Institute has funded the Gene Ontology (GO) Consortium since 2001 [21], when the GO was already enjoying widespread success. In 2009, this funding was $3.4 million plus a $1 million supplement [21]. In 2005, the National Center for Biomedical Ontology (NCBO) received $18.8 million over five years [22]. An effort to build the infectious disease ontology just received $1.25 million over four years [23]. The National Science Foundation recently invested >$900,000 over two years in an ontology of *Hymenoptera* [24]. The National Library of Medicine has paid approximately $6 million per year for the ongoing development of SNOMED-CT since 2007 [25], after an initial investment of $32.4 million in 2003 [26].

One approach to accelerating the manual process of ontology development is to use informatics tools to improve and facilitate the interactions among domain experts and ontologists. An important recent development is the NCBO's BioPotal. BioPortal enables the biomedical community to find, comment on, and contribute to biomedical ontologies, thereby facilitating interactions among ontology users and developers to increase the value of the ontologies [27]. Stanford has developed Collaborative Protégé to allow synergistic ontology development in real time by users in different locations [28]. The earliest examples of such technologies date to the mid-1990s, with work done by Campbell et al. to facilitate geographically distributed development of SNOMED-RT and its successor, SNOMED-CT [29] .

Another approach to reducing resources required in ontology development is the division of labor. The goal is to avoid the wastefulness of recreating multiple representations of the same entity, and its synonyms and relationships, in multiple ontologies. Multiple ontologies result in multiple identifiers for entities, one per ontology. The Open Biological and Biomedical Ontologies (OBO) Foundry seeks to alleviate duplication of efforts and thereby facilitate ontology development by mandating orthogonality of ontologies. That is, it has a well-defined goal of having only one representation of an entity in all of the ontologies in the Foundry [30].

Already, per Smith et al., this principle has resulted in the consolidation of several ontologies [30]. This project also has the goal of increasing interoperability by avoiding the necessity for 'mapping' identifiers among ontologies that represent the same entities (i.e., asserting that identifiers from multiple ontologies refer to the same entity).

Lastly, there is a large body of research that describes automating the development and maintenance of ontologies using NLP. Because literature and text documents are major mechanisms for reporting new knowledge about a domain, ontological knowledge is usually stated explicitly or implicitly within the text. These reference documents serve as important, knowledge-rich resources for ontology learning. Since the NLP often uses ontological knowledge to interpret the texts (see Section 1.1), it can also help to enrich and enhance the linguistic representations of an ontology. Many researchers have been utilizing methods from fields of Natural Language Processing (NLP), Computational Linguistics (CL), and Artificial Intelligence (AI) to partially or fully automate semantic knowledge extraction. This approach has been termed "ontology learning," and represents a sub-field of Knowledge Acquisition (KA), which is the focus of this dissertation.

1.4 KNOWLEDGE ACQUISITION AND ONTOLOGY LEARNING

Knowledge Acquisition (KA) is a broad field that encompasses the processes of extracting, creating, and structuring knowledge from experts and heterogeneous resources [31]. Semi-automated and automated approaches to KA utilize data derived from structured, semi-structured, or unstructured data sources, and result in many different types of knowledge representation [32]. Ontology learning (OL), however, is limited to the extraction of ontological elements from knowledge-rich resources. A further delineation is made for ontology learning from text, which builds on a large body of work within the fields of NLP, CL, and AI [33, 34]. Biomedicine text resources for ontology learning from text include scientific literature and clinical documents, many of which are already available in electronic format. Additionally, ontology learning from text can be further subdivided by task based on the ontological element learned from the resources [33]. These tasks include term extraction, synonym extraction, concept extraction (both taxonomic and non-taxonomic), relationship extraction, and axiom extraction. An axiom is often defined as a set of logical assertions (including rules) used to constrain information in an ontology.

Although there continues to be dissent over whether instances (individuals) should be included in biomedical ontologies at all, many NLP tasks cannot be accomplished without knowledge of instances and their relationship to the corresponding ontology classes. These tasks include information extraction, co-reference resolution, and question answering. Many researchers in KA and OL consider learning of new instances represented in the ontology to be part of ontology learning [33], as encompassed by some combination of term extraction, synonym extraction, and concept extraction, depending upon the way knowledge is modeled in the ontology.

For this dissertation, I choose to define instance learning as a task of ontology learning. I recognize that this task may not be relevant to all ontology engineering efforts. As previously described, the broader field of KA includes research that is easily applied to some of these tasks (particularly term and synonym extraction). Therefore, for these tasks, I have not strictly limited our review to those methods that could be labeled as “ontology learning.” For a more complete treatment of the general field of KA and automated approaches, recent review articles and book chapters [31, 35-37] can provide background information for a better understanding of this field.

I have chosen to exclude axiom learning from the ontology learning tasks reviewed, because there has been so little relevant work done in this area.

1.5 NATURAL LANGUAGE PROCESSING APPROACH FOR ONTOLOGY LEARNING

For several decades, fields of studies such as CL, NLP, AI, and Machine Learning (ML) have been developing methods and algorithms for information retrieval and extraction from free-text knowledge resources. Some of these methods have been used and tested for ontology learning from text and have shown promising results. In general, these methods can be categorized into symbolic, statistical, and hybrid approaches (Table 1).

The symbolic approach utilizes linguistic components to extract information from text. For instance, noun phrases are considered to be linguistic representations of concepts and are often used to represent concepts in an ontology. Linguistic rules describing the relationships between terms in the text can also be used to identify conceptual relationships within an ontology. As first explored by Hearst [38], the most common symbolic approach is to use lexico-syntactic pattern (LSP) matching. LSPs are surface relational markers that exist in a natural language. In the phrase “systemic granulomatous diseases such as Crohn's disease or sarcoidosis,” the words “such as” can help us infer that “systemic granulomatous diseases” is a hypernym of “Crohn's disease” and “sarcoidosis.” Another symbolic approach is to use the internal syntactic structure of component terms. Concepts are often represented using compound or multi-word terms. In general, a compound term is more specific than a single compositional term. The basis of this method is the assumption that a compound term is likely a hyponym of a single term. Using this approach, the term “prostatic carcinoma” can be considered to be a hyponym of “carcinoma.” It is also possible to use multiple symbolic approaches at the same time. For example, the LSP method can be combined with information from compound terms.

The statistical approach, which has also been labeled as the “corpus-based approach,” utilizes large corpora of text data. Harris [39] popularized this approach with his distributional hypothesis, advancing Firth’s notion that “a word is characterized by the company it keeps” [40]. Building on Harris’s theory, it became common practice to classify words not only on the basis of their meaning, but also on the basis of their co-occurrence with other words. The advantage of this method is that it requires minimal prior knowledge of the corpus and can be generalized to other domains. However, a large corpus of text is needed for reliable information to be obtained. Statistical techniques often utilize different linguistic principles and features for statistical measurements to extract semantic information. One of these linguistic principles is known as selectional restriction [41], a limitation on what words can logically accompany particular words in a sentence. For example, the problem with the sentence “A rooster laid an egg” is that a rooster is a male, and thus cannot lay an egg. Therefore, it is incorrect to use the word “rooster” with the word “egg” in this sentence.

Statistical methods (machine learning) can be categorized into two major subsets: unsupervised clustering and supervised classification. The clustering technique for extraction is based on a similarity measure, while the supervised classification attempts to treat the knowledge extraction problem as a classification process. The following paragraphs describe the characteristics of these two techniques.

Clustering is useful for two purposes: first, the similarity measurements can provide information about the hierarchical relationships of concepts for relationship extraction; second, the identification of distinct clusters of similar terms can aid in identifying concepts and their synonyms. The extraction techniques for clustering similar terms are based on definitions of a context within a given corpus. In general, the context of the target word refers to the surrounding linguistic elements. The precise definition of context can vary somewhat depending on the scope. For example, the "first order word context" defined by Grefenstette [42] utilizes information only in the immediate vicinity of the target word [43, 44]. In contrast, the "second order word context" utilizes syntactic information, such as noun-modifiers [45], dependency triples [46], verb-arguments [47], and preposition structures [42, 48]. When utilizing second order context similarity to cluster similar words, semantically similar words are expected to cluster even though they would not typically appear next to each other. For example, the synonyms 'tumor' and 'tumour' would cluster together because they are likely to appear in similar contexts, even though they would not be found together. Context can be further defined as the entire document. In this approach, concepts are represented by a concept signature, which is a vector of co-occurring terms within a set of domain-specific documents [46, 49]. Similarity between concepts can then be calculated by comparing concept signatures. Another approach that utilizes the context of the entire document is the association rule mining technique for concept relationship discovery [50-52]. This technique will be described later in section 2.1.3.2.

Supervised classification, a subcategory of machine learning, is another technique used for many NLP tasks, such as POS tagging, chunking, and co-reference resolution. Most applications of this type of machine learning to ontology learning from text focus on the relatively simpler task of new concept identification, and use supervised methods [53-57]. Using machine learning methods to identify the precise taxonomic location for a concept is a much more difficult task for fully automated systems [58-61].

There are limitations to both symbolic and statistical approaches. Despite the widely accepted belief that statistical methods for ontology learning provide better coverage and scalability than symbolic methods, Basili [62] points out that statistical methods only provide a probability. The output is often represented by words, word strings, or word clusters with associated probabilities. The conceptual explanation of the results is not provided. Ultimately, a human analyst must make sense of this data, because, at present, full automation seems unachievable. Therefore, many researchers have explored a hybrid approach that has the potential to combine the statistical and the symbolic approaches for knowledge extraction. The following section, 2.0, includes my review of existing NLP methods and systems on knowledge acquisition and ontology learning.

2.0 NATURAL LANGUAGE PROCESSING METHODS AND SYSTEMS FOR BIOMEDICAL ONTOLOGY LEARNING

This section is devoted to the review of prior research on NLP approaches and systems that apply to ontology learning from text and are based on associated learning tasks: synonym and concept extraction (Section 2.1); taxonomic relationship extraction (Section 2.2); non-taxonomic relationship extraction (Section 2.3); and generation of ontologies de novo (Section 2.4). The task of term extraction (instance extraction) is encompassed by concept or synonym extraction and not separately considered. In many cases, a particular method, especially the statistical method, can be used for more than one task. For the purposes of this review, I have classified each paper by the task considered most salient, and other tasks that may be accomplished when are relevant to this topic. Because the focus is on describing approaches and algorithms, I have further defined approaches by primary methodology type (e.g., LSP and clustering) and distinguished approaches that are primarily symbolic from those that are primarily statistical, noting those cases in which the approaches overlap.

First, ontology learning methods and algorithms are reviewed (Section 2.1) and categorized by ontology learning task and by approach (Table 1). For each of these categories, I reviewed related papers that are prominent in the field of ontology learning, focusing on algorithmic methods and the advantages and disadvantages of each method. Second, I provided examples of several state-of-the art systems that use these various approaches to support ontology learning from text (Section 2.2).

Task	Primary Method	Secondary Method	Authors
Synonym and Concept extraction	Symbolic	Compound noun information	Hamon [63]
		Lexico-syntactic patterns (LSP)	Downey [64]
		LSP + compound noun information	Moldovan, Girju [65]
	Statistical	Clustering	Church [43]
			Smadia [44]
			Grefenstette [66], Hindel [47]
			Geffet, Dagan [67]
			Agirre [49]
			Faatz, Steimetz [68]
		Hidden Markov Model (HMM)	Collier [53]
			Bikel [69]
			Morgan [54]
		Support Vector Machine (SVM)	Shen [55]
			Kazama [56]
			Yamamoto [57]
		Conditional Random Fields (CRFs)	Chanlekha [70]
Taxonomic relationship extraction	Symbolic	LSP	Hearst [38]
			Caraballo [71]
			Cederberg, Widdow [72]
			Fiszman [73]
			Snow [74]
			Riloff [75]
		Compound noun information	Velardi [76] Cimiano [77]
			Rinaldi [78]
			Morin [79]
			Bodenreider [80]
			Ryu [81]
	Statistical	Clustering	Alfonseca, Manandler [59]
		Machine learning	Witschel [82]
Non-taxonomic relationship extraction	Symbolic	LSP	Berland [83]
			Sundblad [84]
			Girju [85]
			Nenadić, Ananiadou [86]
	Statistical	Co-occurring information	Kavalec [87]
		Association rule mining	Gulla [50]
			Chefi [51]
			Bodenreider [52]
Ontology generation (combining all tasks)	Statistical	Dependency triples	Lin [46]
		Nearest neighbor clustering	Blaschke, Valencia [88]

Table 1. Ontology learning task and their corresponding learning methods

2.1 NATURAL LANGUAGE PROCESSING METHODS

2.1.1 Synonyms and concepts extraction

Extraction of synonyms and concepts has been approached using a variety of methods, because, in many cases, a single method alone cannot distinguish between these ontological elements. In other cases, a particular method that has been used for one of these tasks could easily be used for another learning task. Thus, I consider approaches in this category along a spectrum of complexity, starting with symbolic methods designed primarily to extract synonyms.

2.1.1.1 Symbolic methods

Compound noun information provides a simple symbolic method for synonym identification. Hamon et al. [63] used a general purpose thesaurus as the knowledge resource, along with the following three heuristics: (1) *IF two compound terms' noun heads are identical and have modifiers which are synonyms; or (2) IF two noun heads are synonyms and have modifiers which are identical; or (3) IF two noun heads are synonyms and have modifiers which are also synonyms, THEN the two compound terms are synonyms.* To use a biomedical example: the terms "hepatic tumor" and "hepatic tumour" can be considered synonyms because the modifiers are identical and the head nouns "tumor" and "tumour" are synonyms. Working with a corpus of engineering documents, Hamon et al. evaluated this method and found that 37% of the extracted synonym pairs were correct. The first two heuristics were most effective, producing 95% of the total of correct synonyms.

Another approach for extracting synonyms and concepts relies on the application of LSP, often using a bootstrap method. In this case, a set of seed concepts or patterns is used to extract new concepts or patterns, initiating a cycle of discovery and extraction. An important problem is to control the quality of the extraction, using some discriminating performance metric. Downey and colleagues [64] illustrated this approach, which they defined as the Pattern Learning algorithm (PL). The algorithm started with a set of seed instances generated by domain-independent patterns (e.g. Hearst patterns). For each seed word in the set, they retrieved more instances that contained the seed word from the WWW. Patterns were obtained by creating a window of *w* words around the seed word (*w* was set to 4 in their experiment), which acted as a threshold for selecting pattern candidates. In the first step, patterns with relatively high estimated recall and precision were selected, and these patterns were used to extract new concept candidates from the WWW in order to improve the recall. Use of the selected patterns boosted recall from 50% to 80%. In the second step, they used Turney's [89] Pointwise Mutual Information (PMI) in order to improve the precision. PMI is a statistical measure of the strength of association between an extraction and discriminator (pattern). PMI is calculated as Counts (*D*+*E*) / Counts (*E*) where *D* is the pattern, *E* is the extraction and *D*+*E* is the pattern with the extraction as the instance placeholder. Downey and colleagues used the PMI scores for a given extraction as features in a Naïve Bayes classifier, to determine whether the pattern should be used as an extractor. For example, in the pattern "city of <CITY>" *D* represents the pattern "City of <*X*>", while *E* represents the various instances extracted as "<CITY>". This pattern has a high PMI because "City of" rarely extracts instances that are not cities, and the cities extracted are frequently associated with this pattern. In contrast, the pattern "<CITY> hotels" has a low PMI because many other terms (such as "budget") are also extracted. The classification step is performed to improve accuracy because a single threshold will not work for every domain. Using this method of discrimination, Downey increased precision from 70% to 87%. This method seems highly amenable to applications in the biomedical domain, as we often observe patterns that have high PMIs. For example, "<protein> activates <*X*>" will extract either a "Protein" or "Process" in the biomedical domain (e.g. "Fyn activates Cbl,"; "Bcl-2 activates apoptosis"). The method could be used to extract terms which may be either new synonyms or new concepts, but it is unlikely to distinguish between them.

Combining both compound noun information and LSP matching, Moldovan and Girju [65] developed an approach to enrich domain-specific concepts and relationships in WordNet. The source for acquiring new knowledge was a general English corpus, augmented by other lexical resources such as domain-specific corpora and general dictionaries. The user provided domain-specific seed concepts, which were used to discover new concepts and relationships from the source. The method was tested on five seed concepts selected from the financial domain: "interest rate"; "stock market"; "inflation"; "economic growth"; and "employment." Queries were formed with each of these concepts, and a corpus of 5,000 sentences was extracted automatically from the Internet and TREC-8 corpora. From these extractions, they discovered a total of 264 new concepts not defined in WordNet, of which 221 contained the seed concepts and 43 are other related concepts. Compound noun information and LSP can also be used to extract taxonomic relationships. Moldovan and Girju used this combined method to discover 64 new relationships that linked these concepts with each other or with other WordNet concepts.

2.1.1.2 Statistical methods

(a) Methods that use clustering approaches

Clustering methods have been commonly applied to concept and synonym extraction, because text corpora provide a great deal of data for computing similarity measures. These methods may be able to distinguish synonyms from new concepts based on the degree of statistical similarity. Because these measures can be compared to the existing ontology, these methods can also be used to suggest placement of the concept in the hierarchy.

One of the first to suggest the clustering approach was Church [43], who proposed methods to measure word association based on the information theoretic notion of mutual information. In this approach, the association ratio of two words (x, y) is calculated as the probability of observing x and y in the same context (the joint probability) divided by the probability of observing x and y independently (the product of the marginal probabilities). If there is a genuine association between x and y, then the joint probability P(x, y) should be larger than chance P(x) P(y). In this case, context is the immediate vicinity of a given word in a window. Church suggested that smaller window sizes might identify fixed expressions (idioms) and other relationships that hold over short ranges, while larger window sizes might highlight semantic concepts and other relationships over a larger scale.

Smadia [44] further extended Church's proposal by using Church's method as the first stage and adding two more stages to raise the precision. The two added stages are both filtering functions. One of them calculated the histogram of the frequency of the target word (x) relative to position of the collocated word (y) with a five word window before and after the target. If the histogram was flat, the association between x and y was rejected. The other filter calculated which spike to pick if more than one spike appeared in the histogram. These two additional functions eliminated the noise introduced by non-specific associations.

A similar approach is used in Grefenstette [66] and Hindle [47], both of whom describe the clustering of terms according to the verb-argument structures they display in the text corpus. The approach termed “selectional restriction” exploits the restrictions on what words can appear in a specific structure. For example, wine might be “drunk,” “produced,” or “sold,” but not “pruned.” Using 6 million words in the 1987 AP news corpus, Hindle extracted a set of Subject-Verb-Object triples and calculated the mutual information between verb-noun pairs. Using this approach, Hindle found that the nouns with the highest associations as objects of the verb “drink” were “beer,” “tea,” “Pepsi,” “wine,” and “water,” etc. Then, he calculated the similarity between nouns by considering how much mutual information these nouns shared with the verbs in the corpus. The similarity between nouns with high associations with the same verb may be even more pronounced in biomedical domains, in keeping with Harris’s sublanguage theory [90, 91], as meanings of a term and vocabularies are further restricted. For example, in the biomolecular domain, the predicate “interaction” includes two subcategories, “activate” and “attach.” For the semantic groups "protein" and “process,” “protein” is constrained to co-occur with the “activates process” pattern, rather than the "attaches process" pattern. Therefore, the Subject-Verb-Object triple approach may prove to be very effective for similar-term extraction. Examples of the effective use of this technique in biomedical domains include Friedman’s MedLEE [92] and Sager’s Linguistic String Project (LSP) system [93].

Geffet and Dagan [67] further explore the relationship between the distributional characterization of words, proposing two new hypotheses as refinements to the distributional similarity hypothesis. They claimed that distributional similarity captures a somewhat loose notion of semantic similarity, but in the case of tight semantic relationships (for example, synonym relationships), the distributional similarity measure may not be sufficient. Particular attention is paid to this type of semantic relationship. They describe a “lexical entailment relationship” as a relationship between a pair of words such that the meaning of one word sense can be inferred through substitution with the paired word. The refined versions of the distributional similarity hypothesis for lexical entailment inference are as follows: First, let “v_i” and “w_j” be two word senses of the words v and w, correspondingly, and let “v_i => w_j” denote the (directional) entailment relation between the two words senses; further, assume that they have a measure that determines the set of characteristic features for the meaning of each word sense. From this step, 1) if “v_i => w_j”, then all the (syntactic) characteristics of “v_i” are expected to appear with “w_j”, or 2) if all the syntactic characteristic features of “v_i” appear with “w_j”, then we expect that “v_i => w_j”. An empirical analysis performed on a selected sample to test the validity of the two distributional inclusion hypotheses revealed that the first hypothesis completely fit the data, while the second hypothesis held 70% of the time. Geffet and Dagan further employed the inclusion hypotheses as a method to filter out non-entailing word pairs, with the result that precision was improved by 17% and F1 was improved by 15% over the baseline.

By incorporating information from an entire document, Agirre [49] exploited a topic signature approach for concept clustering to enrich WordNet. He showed that topic signatures could be used to disambiguate word senses, a common problem in using text corpora for ontology learning. His work followed Lin and Hovy [46], who originally developed this approach for text summarization. First, Agirre composed a query using the WordNet concept with its synset to extract documents from the WWW. Each document collection was used to build a topic signature for each concept sense. The topic signature for a concept sense, derived from WordNet, was a set of words from a collection of selected documents that revealed a higher frequency of the concept sense when compared with the the frequency of concept sense in the remaining documents. For a given new concept candidate, the topic signature was obtained and compared to the signature calculated for the concept sense, using the chi-square statistic. The word sense with the highest chi-square score was the chosen sense for that concept candidate.

Faatz and Steinmetz [68] developed a sophisticated method to utilize distances inherent to an existing ontology in order to optimize enrichment. The method utilized a comparison between the statistical information of word usage in a corpus and the structure of the ontology itself, based on the Kullback-Leibler divergence measure. Although Faatz and Steinmetz also used collocation information for the similarity measure, their method differed in that they defined a parameterization by assigning weights to each word collocation feature so that the parameters used in the calculation could be optimized. One interesting advantage of this approach is that it might preferentially select candidates which approximate the level of abstraction for a given ontology.

(b) Methods that use a supervised machine learning approach

Supervised classification machine learning methods can also be used for concept and synonym extraction. Collier et al. [53] described how to extract Molecular Biology terminology from MEDLINE abstracts and texts using Hidden Markov models (HMM). They trained the HMM with bigrams based on lexical and character features in a relatively small corpus of 100 MEDLINE abstracts that had been marked up by domain experts with eleven term classes, such as “proteins” and “DNA.” Word features used for their HMM were based on Bikel [69] and included 23 features, such as *Digital Number*, *Single Capitalized Letter, Greek Letter, Capitalized and Digits*, *Hyphen* etc. The testing data consisted of 3,300 MEDLINE abstracts from a subdomain of molecular biology, retrieved using the query terms *human*, *blood cell,* and *transcription factor*. Using the HMM classifier, they extracted named entities related to the eleven classes, and determined the accuracy of classification of the named entities, using F-score as their metric. The method performed adequately, with an average F-score of 73%.

Morgan [54] further extended Collier’s approach, developing a method appropriate for learning new instances without human-annotated training data. Considering such hand-annotation to be a limitation of Collier’s method, Morgan leveraged an existing FlyBase resource to provide supervision. The FlyBase model database was created by human curation of published experimental findings and relations in the Drosophila literature. The resource contains a list of genes, related articles from which the gene entries were drawn, and a synonym lexicon. Morgan applied a simple pattern matching method to identify gene names in the associated abstracts, and filtered these entities using the list of curated entries for that article. This process created a large quantity of imperfect training data in a very short time. Using a process similar to Collier, an HMM was trained and used to extract relevant terminology. The resulting F-score was 75%, making it quite comparable to that of Collier's report. This method has the advantage of being rapidly transferable to new domains wherever similar resources exist.

Shen et al. [55] used feature selection to identify lexical features that can capture the characteristics of a biomedical domain. Using HMM, they determined the additive benefit of (1) simple deterministic features such as capitalization and digitalization, (2) morphological features such as prefix/suffix, and (3) part-of-speech features, and compared these features alone as compared to adding (4) semantic trigger features such as head nouns and special verbs. Head noun trigger features enable classification of n-grams. For example, the n-gram "activated human B-cells" would be classified as "B-cells." Similarly, special verb trigger features are verbs such as "bind" and "inhibit" that prove useful in biomedical documents for extracting interactions between entities. The GENIA corpus (Ver. 1.1) [94], a human-annotated corpus of 670 biomedical journal abstracts taken from the MEDLINE database, and which includes annotations of 24 biomedical classes by domain experts, was used as the training and evaluation corpus.. The overall F-score was 66.1%, which is 8% higher than that of Kazama's work [56], which used the identical data set. Simple deterministic features only achieved an F-score of 29.4%. Addition of morphological features increased the F-score to 31.8%. Addition of POS features provided the largest boost, increasing the F-score to 54.3%. Head nouns provided an additional improvement, leading to an F-Score of 63.0%, but special verb trigger features did not increase the F-score at all. They speculated that past and present participles of some special verbs often play adjectival roles within the biomedical terms, and may have influenced the classification. For example, in the phrase “IL-10 inhibited lymphocytes,” the term “inhibited” is a past participle, linking two terms that are not taxonomically related. This may limit the accuracy of this method for taxonomic classification, but suggests that other kinds of ontological relationships could be derived using this method.

Support Vector Machines (SVM) have also been utilized for biomedical Named Entity Recognition (NER) and subsequent classification. Both Kazama [56] and Yamamoto [57] used the GENIA corpus as training data for their work with SVM. Kazama formulated the named entity recognition as a classification task, representing each word with its context as three simple features (termed “*BIO*”) to facilitate the SVM training. *B* indicates that the word is the first word in the named entity, *I* indicates that the word is in another position in the named entity, and *O* indicates that the word is not a part of the named entity. *B* and *I* can be further differentiated by the named entity class annotated within GENIA. Thus, there can be a total of 49 (2N+1) classes when the *BIO* representation is used. For example, in the sentence fragment “Number of **glucocorticoid receptors** in **lymphocytes** and ...,” where “glucocorticoid receptors” has been human-annotated as a member of the class PROTEIN and “lymphocytes” has been human-annotated as a member of the class CELL-TYPE, the sentence fragment can be represented the following way:

Number	of	**glucocorticoid**	**receptors**	in	**lymphocytes**	and ...
O	O	B_{protein}	I_{protein}	O	$B_{\text{cell-type}}$	O

Because the GENIA corpus has a skewed distribution of classes, with the majority of words belonging to the O class, Kazama used a splitting technique to subclass all words in the O class based on POS information. This technique not only made training feasible, but had the added benefit of improving accuracy, because in NER we need to distinguish between nouns in the named entities and nouns in ordinal noun phrases that do not participate in named entities. Kazama achieved an average F-score of 54.4% using these techniques.

Yamamoto [57] explored the use of morphological analysis as preprocessing for protein name annotation using SVM. He noted that Kazama's work ignored the fact that biomedical entities have boundary ambiguities that are unlike those found in general English. For example, in general English, it may be assumed that the space character is a token delimiter. In contrast, named entities in biomedical domains are often compound nouns, where the space character is not a token delimiter. Consequently, simple tokenization and POS-tagging developed for general English may not be adequate for biomedical domains. Kazama proposed a new morphological analysis method that identifies protein names by chunking based on morphemes (the smallest units determined by morphological analysis) as well as by word features. This method can avoid the under-segmentation problem that often exists with traditional word chunking. Thus, if a named entity appeared as a substring of a noun phrase, chunking based on noun phrase only would fail to identify the named entity because of coarse segmentation. For example, for the noun phrase "SLP-76-associated substrate," use of a traditional chunking method would only tokenize "SLP-76-associated substrate." In contrast, Yamamoto's morpheme-based chunking method would tokenize both "SLP-76" and "SLP-76-associated substrate." Using the GENIA corpus 3.01, Yamamoto achieved an F-score of 70% for protein names and an F-score of 75% for protein names including molecules, families, and domains. These results suggest that this preprocessing method can be easily adapted to any biomedical domain and improve language processing.

Another machine-learning algorithm, the Conditional Random Fields (CRFs) model, has become popular for term extraction due to its advantages over Hidden Markov Models (HMMs) and Maximum Entropy Markov Models (MEMMs) [95]. Like HMMs and MEMMs, CRFs are discriminative probabilistic models that have been applied to a wide range of problems in text and speech processing. However, CRFs permit relaxed independence assumptions about the random variables and use undirected graphic representations that avoid bias toward states with fewer successor states, the major shortfall of HMMs and MEMMs. For example, Chanlekha and Collier [70] used a CRFs-based NER module to learn new concepts of a specific semantic type, namely the spatial information of an event (the predicate that describes the states or circumstances in which something changes or holds true). They treated spatial terms as attributes to each event, and tried to identify the spatial location of an event based on three sets of features. First, they studied the kinds of textual features that people often use to perceive the place where an event in a news report occurred and found eleven, such as: “location of the subject” and “location of the object.” Such features can be used to train the CRFs model; for example, the location of the subject can often indicate the location of an event. In this sentence, “Head of South Halmahera district health office, Dr. Abdurrahman Yusuf, confirmed the spread of diarrhea and malaria in the villages,”.the phrase “South Halmahera district” indicates the location of the subject “Dr Abdurrahman Yusuf,” and serves as a clue for the location of the event embodied in the phrase “confirmed the spread of diarrhea and malaria.” Second, the researchers discovered that the type of event could also be a beneficial feature for spatial term extraction. Using an automatic classifier that they developed, Chanlekha and Collier categorized the events into three groups: spatially locatable event; generic informational event; and hypothetical event. Third, they incorporated the subject type (disease, pathogen, symptom, government or medical officers, person, organization, and location) into the feature set. For evaluation, they compared the CRFs for spatial term recognition with two other methods (a simple heuristic approach and a probabilistic based approach), from a set of 100 manually annotated outbreak news articles from the BioCaster corpus. Using n-fold cross validation, they found that the CRFs approach achieved the highest performance (precision 86.3%, recall 84.7%, and F-score 85.5%) when compared with both a probabilistic approach (precision 69%, recall 74.3%, and F-score 71.6%) and a simple heuristic approach (precision 52.8%, recall 51.2%, and F-score 52%).

2.1.2 Taxonomic relationship extraction

Extraction of taxonomic relationships has been extensively studied, using both symbolic and statistical methods.

2.1.2.1 Symbolic methods

One of the earliest attempts to derive relationships from text corpora was described by Hearst [38], who used LSPs for semantic knowledge extraction. She hypothesized those linguistic regularities such as LSPs within a corpus can permit identification of the syntactic relationship between terms of interest, and therefore can be used for semantic knowledge acquisition. To prove this hypothesis, Hearst searched for a set of predefined LSPs that indicated general relationships, such as hyponym/hypernym, in Grolier's American Academic Encyclopedia text. Out of 8.6 million words in the encyclopedia, she found 7,067 sentences that contain the pattern 'such as,' from which 330 unique relationships were identified. Of these, 152 relationships involved unmodified nouns for both hypernym and hyponym, comprising a total of 226 unique words. Using WordNet as a validation resource, Hearst found that 180 of these 226 words were present in the WordNet hierarchy, suggesting that these linguistic rules extract meaningful information. She concluded that the LSP matching method could be an effective approach for finding semantically related phrases in a corpus because a) the method does not require an extensive knowledge base; b) a single, specially expressed instance of a relationship is all that is required to extract meaningful information; and c) the method can be applied to a wide range of texts. Hearst acknowledged low recall as an inherent problem with this method.

Other researchers have applied the LSP matching approach to other domains and investigated methods to increase recall and precision of the LSP approach. Caraballo [71] addressed the low recall problem by applying noun coordination information to the LSP method. Coordination is a complex syntactic structure that links two or more elements, known as conjuncts or conjoins. The conjuncts generally have similar grammatical features (e.g. syntactic category, semantic function). He assumed that nouns in a coordination structure, such as conjunction and appositives, are generally related, as has been discussed previously by Riloff and Shepherd [96] and Roark and Charniak [97]. For example, in the sentence "Sugar, honey, nutmeg, and cloves can increase the flavor of a dish," the words "nutmeg" and "cloves" share a conjunction structure, and are therefore considered to be semantically similar. If "spice" is known to be a hypernym to "nutmeg," then from the sentence above, it can be inferred that "spice" is also a hypernym to “cloves.” This linguistic structure can be observed often in biomedical corpora, as in the sentence "In the ovine brain, GnRH neurons do not contain type II glucocorticoid (GR), progesterone (PR), or α estrogen (ERα) receptors." Thus, if α estrogen receptor (ERα) is a steroid receptor in the ontology, we can define GR and PR as steroid receptors as well.

Cederberg and Widdows [72] described two other methods that can be added to the extraction process to increase recall and precision: a graph-based model method and latent semantic analysis. In the first method, they used a graph-based model of noun-noun similarity learned automatically from coordination structures. This method is very similar to Caraballo’s method using coordination information. But in contrast to Caraballo’s hierarchy-building method, Cedeberg used an alternative graphic-based clustering method developed by Widdows [98], in which nouns are represented as nodes and noun-noun relationships are represented as edges. In Cederberg’s graph, the edges between two nouns are connected if they appear in a coordination structure. The algorithm extracts similar words when a seed word is provided by the user, where the seed word is normally a known hyponym of one category. For example, if “clove” is the seed word and is a hyponym of “spice,” then all the words that appear in the coordination structure will be hyponyms of “spice” as well. This method obtained additional hypernym-hyponym pairs extracted by LSPs and improved recall five-fold.

In the second method, Cederberg and Widdows used latent semantic analysis [99, 100] to filter the LSP-extracted hyponyms. Latent semantic analysis is a statistical method that can measure the similarity of two terms based on the context in which they appear. Each term’s context is represented by a vector of words that co-occur most frequently with the target term. Similarity between two terms was calculated using the cosine of the angle between the two vectors. A hyponym and its hypernym extracted with the LSP matching method should be very similar. Therefore, by establishing a threshold, term pairs with low scores can be filtered and excluded from further consideration. Using this method, the researchers increased precision of LSP matching from 40% to 58%.

Within the biomedical domain, Fiszman et al. [73] have shown that the Hearst LSPs can be used for hypernymic propositions to improve the overall accuracy of the SemRep semantic processor developed by Rindflesch and Fiszman [101, 102]. SemRep uses syntactic analysis and structured domain knowledge such as the SPECIALIST lexicon and UMLS Semantic Network to capture semantic associations in free-text biomedical documents such as MEDLINE. For example, given the sentence “Alfuzosin is effective in the treatment of benign prostatic hyperplasia,” SemRep produces the semantic predication: “Alfuzosin-TREAT-Prostactic Hypertrophy, Benign.” SemSpec is an extension to SemRep that utilizes LSPs such as appositive structures and Hearst patterns (e.g. “including,” “such as,” and “especially”) to identify hypernymic propositions. Once a hypernymic proposition is established, the more specific term can replace the more general terms in a semantic association that has been captured by SemRep. For example, for the sentence “Market authorization has been granted in France for **pilocarpine**, an old **parasympathomimetric agent**, in the treatment of xerostomia,” SemRep produces “Parasympathomimetric Agents-TREATS-Xerostomia” and captures the hypernymic position “Pilocrapine-ISA-Parasymphomimetic Agents.” From this extraction, a more accurate semantic association, “Pilocrapine-TREATS-Xerostomia,” can be inferred. Using a manually tagged set of 340 sentences from MEDLINE citations that were limited to the UMLS Semantic Network predicate TREATMENT, Fizman et al. found that SemSpec increased SemRep's recall by 7% (39% to 46%) and precision by 1% (77% to 78%).

The LSP matching method can be further improved by using machine learning methods to learn LSP patterns. Snow [103] represented the Hearst patterns using a dependency parse tree, and found all features along the path for each LSP. These features were used to train a classifier. Snow not only rediscovered the Hearst patterns, but also identified several new patterns. Riloff [75] developed the Autoslog-TS system, which uses a bootstrapping method for generating LSPs from untagged text. This system is an extension of her earlier Autoslog work [104] and has been further extended in Thelen and Riloff's [105] Basilisk system for semantic lexicon extraction. The input for Autoslog-TS was a text corpus and a set of seed words that belonged to six semantic categories (building, event, human, location, time, and weapon). The seed words were generated by sorting all words in the corpus based on frequency, and then manually assigning high frequency words to a category. For example, "bomb," "dynamite," "guns," "explosives," and "rifles" are seed words for "weapon." Seed words were then used to extract contiguous LSPs, and the resulting patterns were ranked based on their tendency to extract known category terms. The top patterns were used to extract other terms, which were then scored; those with high scores were added into the semantic lexicon. Using a bootstrapping method, this process was then repeated multiple times. The MUC-4 corpus was used to evaluate performance of both Autoslog and Autoslog-TS pattern extraction for aiding semantic information extraction. Autoslog achieved 62% recall and 27% precision, while Autoslog-TS achieved 53% recall and 30% precision. The difference between Autoslog and Autoslog-TS is that Autoslog-TS creates a pattern dictionary with an unannotated training text, whereas Autoslog uses an annotated text and a set of heuristic rules. This method has some specific advantages in biomedicine, because of the breadth of resources available for obtaining seed words for a particular semantic category. For example, “ATP,” “kinase,” “gene transcription,” and “binding site” are seed words for “cell activation,” and can be obtained from the UMLS or existing biomedical ontologies. Markó and Hahn [106] have developed a methodology for automatic acquisition and augmentation of a multilingual medical subword thesaurus using seed terms from the UMLS Methathesaurus.

Another linguistic technique for relationship extraction uses compound noun information. For example, Velardi [76] and Cimiano [77] used the following head matching heuristic for hyponym term discovery: *IF term A and term B head nouns are the same and term A has an additional modifier THEN term A is a hyponym of term B.* Using a tourism domain corpus, Velardi achieved 82% precision while, Cimiano achieved 50% precision. However, the precisions obtained from the different studies are not directly comparable due to the different corpora used.

Rinaldi [78] further expanded Hamon’s work, by using Hamon’s method to extract all the synsets for each concept and adding the following simple heuristic to organize these synsets into a taxomomic hierarchy: *IF term A is composed of more individual terms than term B, THEN term A is a hyponym of term B.* A manual expert evaluation found 99% accuracy for synonym discovery and 100% accuracy for hyponym links. Morin et al. [79] explored to add a hypernym relationship by mapping one word terms to multiword terms. For example, given a link between “fruit” and “apple,” a relationship between the multi-word terms “fruit juice” and “apple juice” can be added. This is often the case in biomedical domains as well; for example, given a relationship between "nucleotide" and "ATP," relationship between the multiword terms "nucleotide transport" and "ATP transport" can be added. Morin et al. based their work on several heuristics: *IF (1) two multi-terms share the same head noun (juice); and (2) the substituted words have the same grammatical function (modifiers); and (3) the substituted words are semantically similar ("apple" and "fruit"), THEN the two terms are related*. For the third clause of the heuristic, semantic information would come from an existing semantic resource, such as an ontology. For their knowledge resource, Morin et al. used the Agrovoc Thesaurus, a multilingual thesaurus in the agriculture domain managed by the Food and Agriculture Organization of the United Nations. This method could potentially be very effective in the medical domain, where multiword terms like "diabetes mellitus" and "insulin-dependent diabetes mellitus" are quite common and are likely to express taxonomic relationships.

Bodenreider and colleagues [80] explored ways to use modifier information to establish groups of similar terms. First, a group of compound nouns was collected from MEDLINE citations. Terms extracted from MEDLINE were compared to current UMLS concepts in an attempt to discover concept candidates for the UMLS Methathesaurus. Each component noun was then parsed into a modifier and head noun using an underspecified syntactic analysis [101] and the SPECIALIST Lexicon. The component noun became a concept candidate *if: 1) the head noun of the component noun was found in the Methathesaurus; and 2) concepts existing in the Methathesaurus had the same modifier.* The concept candidate was incorporated into the Methathesaurus based on the head noun's position in the hierarchy. From three million randomly selected MEDLINE component nouns, 125,464 were captured as concept candidates for Methathesaurus. Evaluation of a sample of randomly selected concept candidates determined how well these candidates could be incorporated into the Methathesaurus using head noun matching. The authors defined three levels of relevance: the highest level, “relevant,” was used for cases where the addition of the candidate to the terminology was relevant even if there was a more specific concept available. The intermediate level, “less relevant,” was used for cases where the parent selected for the candidate was too general to be informative. The lowest level, “not relevant,” was used for cases where the parent selected for the concept was irrelevant. Of the 1,000 randomly selected candidates, 834 were classified as “relevant,” 28 were classified as “less relevant,” and 138 were classified as “irrelevant.”

Investigating an alternative approach to heuristics, Ryu [81] explored a mathematical method for determining hierarchical position using ‘specificity’ as defined in the field of information theory [107], where specificity of a term is a measure of the quantity of domain-specific information contained in the term. In this method, the higher the specificity of the term, the more specific the information it contained (further details regarding this measure are discussed in section 3.2). A weighting scheme excluded terms that appeared frequently as modifiers but provided no additional information. The taxonomic position of the term was then determined based on the specificity. For example, “insulin-dependent diabetes mellitus” had a higher specificity, and thus could be positioned as a child of “diabetes mellitus.” Using a flat collection of terms obtained from a sub-tree of MeSH and a set of journal abstracts retrieved from the MEDLINE database, the authors generated a hierarchy for the MeSH terms, and compared it to the MeSH hierarchy. The precision for ontological hierarchy placements was increased from 69% (for a word frequency baseline method) to 82%.

2.1.2.2 Statistical methods

Clustering and machine learning methods have also been applied to the extraction of relationships, albeit less frequently and with less success than the extraction of concepts. Alfonseca and Manandhar [59] followed Agirre's [49] topic signature technique. Through a top-down search, starting with the most general concept in the hierarchy, the new concept was added to the existing concept whose topic signature was the closest to its own. Several experiments with seven general target words were conducted. The task was to place these words into the right category in the ontology; the best result was an 86% accuracy. The researchers concluded that it was better, for this task, to consider a smaller context of highly related words to build the signature rather than a larger context that included more words.

Another group led by Witschel [82] extended a decision tree model for taxonomy enrichment. They first identified potential new concepts using a combination of statistical and linguistic methods [108], termed "semantic description," that is based on co-occurrence within German language texts (such as newspapers, fiction, etc.). Witschel's 'semantic description' method is similar to Alfonseca's 'distributional signature' [58]. Witschel's group evaluated its method using a general-German language text to enrich a sub-tree of GermaNet (the German equivalent to WordNet). Two measures were computed: accuracy of the decisions (the percentage of nodes that were correctly classified as hypernyms); and learning accuracy [61], which takes into consideration the distance of the automated placement from the expected location in the tree. The accuracies for enriching a furniture sub-tree and a building sub-tree were 11% to 14% respectively, findings comparable to Alfoneca's results. The learning accuracy reached 59%, a result that was significantly better than Alfoneca’s. Again, the absence of a common reference standard for testing makes it difficult to directly compare these results.

2.1.3 Non-taxonomic relationships extraction

Extraction of non-taxonomic relationships, i.e. non-IS-A relationships, has also been studied, and has been considered to be the most difficult ontology learning task. Both symbolic and statistical methods have been employed.

2.1.3.1 Symbolic methods

The LSP method has been used by Berland [83], Sundblad [84], and Girju [85] for part-whole (meronymic) relationship discovery. Berland combined both the LSP method and statistical methods and used them on a very large corpus. The output of the system was an ordered list of possible parts for a set of six seed “whole” objects. Berland achieved 55% accuracy.

Nenadić and Ananiadou [86] used three symbolic approaches to discover terms from MEDLINE abstracts: 1) an LSP-based similarity measure (SS) using Hearst patterns, coordination patterns, apposition patterns, and anaphora; 2) a component noun-based similarity measure (called the “lexical similarity measure,” or LS); and 3) a contextual pattern-based similarity measure. The third approach, which was considered novel by the author, learns contextual patterns by discovering significant term features. The procedure is performed as follows and illustrated using a ATP example: First, for each target term, its context constituents are tagged with POS tags and grammatical tags. These tags became the context pattern for the target term. For example, in the phrase “ATP binds heterodimers with ***high affinity,***” “high affinity” is the target term, and the left context pattern (CP) is “*V: bind TERM: rxr_heterodimers PREP:with.*” Second, all the CPs for each term are collected, and a normalized CP-value is calculated in order to measure the importance of the CP. The CP-value is calculated based on the length and the frequency of the pattern. The similarity between two terms based on CP is termed “CS (t_1, t_2),” and is calculated based on the number of common and distinctive CPs of the two terms. Since none of the three similarity measures is sufficient on its own, the researchers introduced a hybrid term similarity measure called “Contextual Lexical Similarity” (CLS), which is a linear combination of the three similarity measures with three parameters: $CLS(t_1, t_2) = \alpha\ CS(t_1, t_2) + \beta\ LS(t_1, t_2) + \gamma\ SS(t_1, t_2)$. In the final step, the three parameters (α, β, γ) were adjusted automatically by supervised learning methods. Nenadić and Ananiadou tested the CLS measure on a corpus of 2008 abstracts retrieved from MEDLINE. Random samples of results were evaluated by a domain expert to see whether the two similar terms based on CLS measure were indeed similar. They also used the CLS measure for term clustering, and achieved a precision of 70%.

2.1.3.2 Statistical methods

Kavalec [87] used a statistical approach, supplemented with some linguistic information to extract non-taxonomic relationships. In this case, the linguistic feature used was based on the assumption that relational information is typically conveyed by verbs at the sentence level. For example, the verb “induce” defines a non-taxonomic (associational) relationship between a gene and a protein. Therefore, Kavalec first selected verb v and a pair of concepts that co-occur within a certain window of verb v. Second, the concept-concept-verb triples were ordered by frequency.

The highest frequency triples were candidates for relationship labels of the given concept association. The association measure was a simple statistical measure based on a verb and a concept pair's conditional frequency (co-occurrence), $P(c_1, c_2|v)$. However, the conditional frequency of a pair of concepts, given a verb, could be high even in the absence of a relationship between the concepts and the verb, because a verb may occur separately with each of the concepts at high frequency, even though it has nothing to do with any of the mutual relationships between the two concepts. Therefore, the authors defined an "above expectation" (AE) measure (see equation 1 below), which was a measure of the increased frequency when compared to the frequency expected under the assumption of independence of association of each of the concepts with the verb. This measure is very similar to the "interest measure" suggested by Kodratoff [109] for knowledge discovery in text, and is also similar to the Church mutual information metric [43].

$$\text{AE (c1, c2|v)} = \frac{P(c_1, c_2 \mid v)}{P(c_1 \mid v) \cdot P(c_2 \mid v)} \qquad \text{Equation 1}$$

The authors performed several experiments to evaluate this approach. In one of the experiments, an ad hoc tourist document collection was used as input for the method. In another experiment, the SemCor corpus that had been semantically tagged with the WordNet senses was used. The results were promising: at AE=1.5 (1 is equal to expectation value), the recall was 54% and precision was 82% for the tourist corpus (measured against a human annotated reference standard). For the SemCor corpus, expert judges evaluated the output, yielding a precision of 72%. Recall could not be measured.

An alternative statistical approach uses association rule mining methods to extract relationships between concepts [50-52]. This method was first introduced by Agrawal et al. [110] as a technique for market analysis using a large database of transaction data. The rules extracted can be exemplified as “90% of the transactions that purchased bread and butter also purchased milk.” The advantage of this method, which has been adapted to mine domain text for concept relationships, is that it does not require deep textual analysis. However, it does tend to generate a large number of association rules. Statistical indices such as support and confidence are then used to select the most meaningful and significant rules. Although the method does not distinguish among types of relationships, it could easily be used as a starting point for human curation.

Gulla [50] evaluated and compared this method with traditional similarity measure methods that utilize vector space models. The output was judged by four human experts who separated extracted relationships into three categories: “not related”; “related”; and “highly related.” The results revealed that more than half of the relationships found by association rule methods were also identified by the similarity measure method. However, the distribution of mined rules differed for these two methods. A further experiment combining the methods produced much improved results. The authors concluded that these two methods might be complementary when combined for relationship extraction. Cherfi, et al, did another work on association rule method. [51]. They investigated how the characteristics of several statistical indices such as support, confidence, interest, conviction, and novelty influence the performance of association rule mining and how a combination of different indices ensures that a subset of valid rules will be extracted.

In the biomedical domain, Bodenreider, et al. [52] evaluated and compared the association rule method (ARM) with two other statistical methods that use similarity measures: the vector space model (VSM); and co-occurrence information (COC), for identifying associations of GO terms between three GO sub-ontologies (molecular function, cellular components, and biological processes). The authors took advantage of several existing databases of human annotations, using GO terms that were publicly available. For the VSM method, gene products that associated with the GO term in the databases were used to form a vector, and the similarity of two GO terms was calculated as the cosine of the two vectors. For the COC method, the frequencies of co-occurring GO terms in the database was represented as a contingency table (number of gene products annotated with both term A and B, number of gene products annotated with term A only, number of gene products annotated with term B only, number of gene products annotated with neither term A or B), and a chi-square test was used to test the independence of the two GO terms. If the terms were not dependent, they were considered to be associated. For the ARM method, each annotation of gene products with GO terms was treated as a transaction. Association rules were extracted using the Apriori algorithm [111]. Bodenreider, et al. evaluated the validity of the extraction by comparing the overlap between the statistical methods, and by comparing statistical methods to another set of methods that were non-statistical and not based on a document corpus. These non-statistical methods included extracting relationships between GO terms existing in UMLS or MeSH (where the relationship is not also included in GO), and determining lexical relationships based on composition between existing between GO terms (where the relationship is not also included in GO). A total of 7,665 associations between GO terms was identified by at least one of the three statistical methods (VSM, COC, and ARM). Among the 7,665 associations extracted by these statistical methods, 936 (12%) were identified by at least two of the three statistical methods, and 201 (3%) were identified by all three statistical methods. When the non-statistical methods were employed, 5,963 associations were identified. However, the authors note that when comparing the relationships extracted by statistical methods to those obtained using the non-statistical methods, only 230 overlapping associations were found. They conclude that multi-method approaches may be necessary to extract a more complete set of relationships.

2.1.4 Denovo generation of ontologies

In contrast to the process of ontology enrichment (which seeks to add or modify existing ontologies), a few researchers have explored the possibility of learning the entire ontology by combining methods for multiple tasks.

Lin [46] explored the distributional pattern of dependency triples as the word context to measure word similarity. Lin's work is very similar to Grefenstette's approach [112], in which dependency triples were treated as features. A dependency triple consists of two words and the grammatical relationship between them in the input sentence. As an example in our own domain, the triples extracted from the sentence "The patient has a mild headache" would be "(has subj patient), (patient subj-of has), (headache obj-of has), (headache adj-mod mild), (mild adj-mod-of headache), (headache det a), (a det-of headache). The description of a word w consists of the frequency counts of all the dependency triples that matched the pattern (w, *, *). Therefore, the similarity between two words is calculated based on the count of dependency triples for each word. Using this similarity measure, Lin created a thesaurus and evaluated it against WordNet and Roget's Thesaurus. He found that his thesaurus was more similar to WordNet than it was to Roget's Thesaurus, and that using all types of dependency triples was better than using only subject and object triples, as Hindle did [47].

Blaschke and Valencia [88] explored the statistical clustering method for building an ontology-like structured knowledge base using the biomolecular literature. They adapted Wilbur's method [113] by clustering the key terms that have been derived from the documents associated with each individual gene. They first retrieved over 6,000 gene names associated with *Saccharomyces cerevisiae* from SWISS-PROT and SGD. 63,131 MEDLINE abstracts were obtained with the search terms "saccharomyces" and/or "cerevisiae." The authors then grouped the documents based on each gene name with which they were associated, and created a fingerprint for each group that could describe the specific content of the documents. The fingerprint consisted of a list of key terms (including bi-grams) and the scores (calculated by comparing frequencies between groups of documents) for each term. This fingerprint was used to calculate the similarity between two genes, a and b (SimScore $_{a,b}$), as the sum of the scores for all significant terms that appear in both fingerprints.

$$\text{SimScore}_{a,b} = \frac{\sum (score_i^a + score_i^b)}{2} \qquad \text{Equation 2}$$

To construct the ontology, a distance matrix for all pairs of genes was created by calculating the similarity score for each pair of genes. Two genes with the highest score were clustered together and removed from the distance matrix, and the two groups of documents for these two genes were merged. A new fingerprint for the merged documents was created. This process was repeated until none of the clusters shared more significant terms. The final output was a gene tree, which was compared with the hand-curated GO ontology by domain experts and found by them to be compatible. Some relationships in the tree that were not in the GO could be added. The authors concluded that this automatic clustering method can be utilized as an instrument to assist human expert ontology building, and could be particularly useful for domains experiencing rapid growth. For example, in genomics, many new genes have been discovered as a result of the advances in genomic sequencing. The number of potential relationships among these genes and proteins is quite large, and therefore could be amenable to a semi-automated approach.

2.2 NATURAL LANGUAGE PROCESSING SYSTEMS

In recent years, a number of ontology learning systems have been developed using one or more of the algorithms described above, with the goal of reducing the human effort required for ontology development. In this section, I compare eleven state-of-the-art ontology learning systems. Three of these systems were developed primarily for the biomedical domain, and the remaining eight systems were developed for general language or other domains. I examine and compare the elements learned from the text, as well as the different approaches employed and the different evaluations performed. Table 2 summarizes these comparisons. All eleven systems are able to learn concepts and taxonomic relationships. Additionally, the DOODLE II, HASTI, STRING-IE, Text-To-Onto, and Text2Onto systems can also learn non-taxonomic relationships.

	Input	Language	Ontological elements learned	Degree of automation	Resource	Ontology enrichment or De Novo generation	Learning Methods
ASIUM	Free text documents.	French	Concepts, taxonomic relations.	Semi-automated	N/A	Deno Vo	Conceptual and hierarchical clustering
DODDLE II	Dictionary, domain specific text documents	English	Concepts, taxonomic relations, non-taxonomic relations.	Semi-automated	WordNet	Enrichment	Matching and trimming against WordNet for taxonomic relations; statistical co-occurrence information.
HASTI	Free text documents	Persian	Concepts, taxonomic relations, non-taxonomic relations, axioms.	Two modes: semi-automated and fully-automated	N/A	Deno Vo	Combination of logical, linguistic, template, and heuristic
KnowItAll	Web pages	English	Concepts,	Automatic	Domain ontology	Enrichment	Combination of linguistic and statistic methods
MEDSYNDIKATE	Medical domain documents	German	Concepts, taxonomic relations.	Semi-automated	Own general and medical lexicons; Fully lexicalized dependency grammar.	Enrichment	Input text is mapped to corresponding text knowledge bases (TKB) which represent the text content; Generates concept hypothesis and ranks hypothesis based on quality
OntoLearn	Free text documents	English	Concepts, taxonomic relations	Semi-automated	WordNet; SemCor	Enrichment	Machine learning; Statistical approach
STRING-IE	Free text documents from PubMed	English	Non-taxonomic relations	Automated	SWISS-PROT, Saccharomyces Genome Database	Enrichment	Linguistic and rule based approach
Text-To-Onto **Text2Onto**	Dictionaries, Databases, Semi-structured text, Free text documents.	German	Concepts, taxonomic relations, non-taxonomic relations,	Semi-automated	Domain ontology (Tourism)	Enrichment	Combination of association rules, formal concept analysis and clustering
TIMS	Free text documents	English	Concepts, taxonomic relations	Automated	N/A	Enrichment	Automatic term recognition using both linguistic and statistical approach and automatic clustering using average mutual information
WEB→KB	Web pages	English	Concepts, taxonomic relations	Automated	Domain ontology	Enrichment	Statistical and Logical

Table 2. Characteristics of existing ontology learning systems

ASIUM [114] (Acquisition of Semantic Knowledge Using Machine Learning Methods) is a system developed to acquire ontological knowledge and case frames. The input to the system is a set of domain-specific documents in French that have been syntactically parsed. The system uses clustering methods, based on a two-step process which produces successive aggregations. The first step is conceptualization clustering, which is similar to work done by Harris [39], Grefenstette [42], and Peat [115], in which the head words associated with their frequencies of appearance in the text are used to calculate the distances among concepts. Based on the sub-categorization of verbs, the head words that occur with the same verb after the same preposition (or with the same syntactical role) are clustered into the basic cluster. The second step is a pyramidal clustering approach adopted from Diday [116], in which the basic clusters are built into the hierarchy of the ontology [117]. This approach is promising, but an evaluation with real cases and real problems has not yet been performed.

DODDLE II [118] is a domain ontology rapid development environment. The inputs to the system are a machine-readable dictionary and domain-specific texts. It supports the building of both taxonomic and non-taxonomic relationships. The taxonomic relationships come from WordNet, while the non-taxonomic relationships come from domain-specific text and from analyzing the lexical co-occurrences based on WordSpace [119], which is a multi-dimensional, real-valued vector space representing lexical items according to how semantically close they are. Evaluation in the domain of Law was done with two small-scale case studies. One study used 46 legal terms from Contract for the International Sale of Goods part II (CISG); the other study used 103 terms that included general terms from the CISG corpus. For taxonomic relationships, the precision was 30%. For non-taxonomic relationships, the precision was 59%.

HASTI [120] is a system that learns concepts, taxonomic and non-taxonomic relationships, and axioms. It is the only system that also learns axioms from text documents (in Persian). HASTI employs a combination of symbolic approaches, such as Hearst patterns [38], logic, and template, as well as semantic analyses and heuristic approaches. It has two modes for conceptual clustering: automatic and semi-automatic. HASTI requires only a very small kernel of an ontology containing essential meta-knowledge, such as primitive concepts, relations and operators for adding, moving, deleting, and updating ontological elements. Based on this kernel, the system can learn both lexical and ontological knowledge. The kernel is language-neutral and domain-independent. Therefore, it can be used to build both general and domain ontologies, essentially from scratch. To prove that the system can be generalized, the authors evaluated HASTI with two test cases. With a text corpus consisting of primary school textbooks and storybooks, the precision was 97% and the recall was 88%. With a text corpus consisting of computer technical reports, the precision was 78% and the recall was 80%.

KnowItAll [121] is an automatic system that extracts facts, concepts, and relationships from the WWW. There are three important differences between this system and other, similar systems. First, KnowItAll addresses the scalability issue by using weakly supervised methods and by bootstrapping learning techniques. Using a domain-independent set of generic extraction patterns, it induces a set of seed instances, thus overcoming the need for a hand-coded set of training documents that is typically required for these kinds of systems. Second, it uses Turney's PMI-IR methods [89] to assess the probability of extractions using statistics computed by treating the web as a large corpus of text (so called "web-scale statistics"). This assessment overcomes the problem of maintaining high precision, and enables the system to automatically trade recall for precision. Third, it is able to make use of the ample supply of simple sentences on the WWW that are relative easy to process, thus avoiding the extraction of information from more complex and problematic texts. Details of the algorithmic methods [64] were described earlier in section 3.1.1.

MEDSYNDIKATE [122] is an extension of the SYNDIKATE system. It is the only knowledge acquisition system aimed at acquiring medical knowledge from medical documents (in German).

MEDSYNDIKATE enables the transformation of text documents to formal representation structures. The system addresses one of the shortcomings of information extraction systems by providing a parser that is particularly sensitive to the treatment of textual reference relationships as established by various forms of anaphora [123]. It distinguishes between text at the sentence level and the text level. A deeper understanding of textual referential relationships is based on their *centering* mechanism [124]. Additionally, MEDSYNDIKATE initiates a new conceptual learning process (knowledge enrichment) while understanding the text. Domain knowledge and grammatical constructions, such as LSPs in the source document in which the unknown word occurs, are used to access the linguistic quality and conceptual evidence. This information is then used to rank the concept hypotheses. The most credible hypotheses based on ranking are selected for assimilation into the domain knowledge base. Another technique for concept generation is based on the reuse of available comprehensive knowledge sources such as UMLS. Evaluation of MEDSYNDIKATE was performed on the deep semantic understanding of the input text but not on the concept learning aspect of the system. Although this is a system developed for the medical domain, the German language basis of the system may somewhat limit its transfer to English language documents. Nevertheless, methodologies developed and insights derived from MEDSYNDIKATE are extremely valuable to researchers developing ontology enrichment systems for English language documents in biomedical domains.

OntoLearn [76] is a very sophisticated system that uses a combination of symbolic and statistical methods. Domain-specific terms are extracted and related to corresponding concepts in a general purpose ontology, and relationships between the concepts are examined. First, statistical comparative analysis is done on the target domains and the contrasting corpora to identify terminology that is used in the former but not the latter. Second, lexical knowledge of WordNet is used to interpret the semantic meaning of the terms. OntoLearn then organizes the concepts, based on taxonomic and non-taxonomic relationships, into a forest, using WordNet and a rule-based inductive learning method. Finally, it integrates the domain concept forest with WordNet to create a pruned and specialized view of the domain ontology. The validation of the process is performed by an expert. The system has been evaluated by two human judges, across a variety of ontology learning algorithms, with encouraging results. Across several different domains (art, tourism, economy, and computer network), the authors achieved recall ranging from 46% to 96% and precision ranging from 65% to 97%.

STRING-IE [125] is a system designed to extract non-taxonomic relationships between concepts in the biomedical domain using symbolic features and rules (heuristic). More specifically, it extracts regulation of gene expression and (de-)phosphorylation related to yeast *S.cerevisae*. Although the language rules reated are specific for *S.cerevisae* organism, the algorithm has been tested on three other organisms (*Escherichia coli*, *Bacillus subtilis* and *Mus musculus*) and has achieved equally good results. Therefore, the method appears to be generalizable. The input to the system is a set of abstracts and full text papers from PubMed Central, retrieved with the terms ‘Saccharomyces cerevisiae,’ ‘S.cerevisiae,’ ‘Baker’s yeast,’ ‘Brewer’s yeast,’ and ‘Budding yeast.’ The documents were POS-tagged, and a name-entity recognition was used to identify names of genes and proteins. The NER module uses syntactic-semantic chunking. For example, the text “the ArcB senory kinase in Escherichia coli” would be chunked as “[$_{nx\_kinase}$ [$_{dt}$ the] [$_{nnpg}$ ArcB] [$_{jj}$ senory] [$_{kinase}$ kinase] [$_{in}$ in] [$_{org}$ Escherichia coli]]. The label $_{nx\_kinase}$ indicates that this is a noun chunk (*nx*) semantically denoting a *kinase*. After NER, two types of relationships were extracted using heuristics to identify verbs related to these relationships, as well as other symbolic features, such as the pattern “x but not y” and pre-defined information about linguistic restriction. A set of rules over groups of verbs and relational nouns, triggered by key words related to the regulation of gene expression, such as “phosphorylate,” “induce,” “decrease,” “regulate,” and “mediate,” was also created. For evaluation, one million PubMed abstracts that related to the organisms above were used. A total of 3,319 regulatory network and phosphorylation relationships were extracted, with an accuracy of 83-90% for regulation of gene expression and 86-95% for phosphorylation.

Text-To-Onto [126] is a semi-automatic ontology learning system that employs a shallow parser (in German) to pre-process text documents coming from the WWW. The advantage of this system is that it has a built-in algorithm library that supports several distinct ontological engineering tasks. The library includes several algorithms for ontology extraction and several algorithms for ontology maintenance, such as ontology pruning and refinement. It gives the user the ability to pick extraction and maintenance algorithms for various inputs and tasks. For ontology concept and concept relationship extraction, Text-To-Onto utilizes a combination of statistical methods, such as Srikant’s [127] generalized association rule discovery, and symbolic methods, such as Hearst’s LSP method. Details of extraction algorithms are described in other manuscripts [128-131]. The Text2Onto system which was developed later [132] was distinguished from the earlier system in three important ways. First, the learned knowledge is represented at a meta-level, termed the Probabilistic Ontology Model (POM), in the form of instantiated model primitives. In this way, learned knowledge remains independent of a concrete target language and can be translated into any knowledge representation formalism (e.g. RDFS, OWL, and F-Logic). Second, to facilitate user interaction, the POM is employed to calculate a confidence for each new learned object. Users can thus filter the POM, selecting only a number of relevant instances of modeling primitives that fit Text2Onto's interests. Third, changes to the ontology since the last change in the document collection are explicitly tracked so that users can trace the evolution of the ontology over time as new documents are processed. An obvious benefit is that there is no longer the need to process the entire document collection when additional documents are added later. Such transparency into the working of the system over time could also enable greater human supervision of the enrichment process.

Both taxonomic relationship discovery using Hearst's pattern match method, and non-taxonomic relationship discovery using Srikant's generalized association rule discovery method [127], were evaluated in a tourism domain. For taxonomic relationship (IS-A) discovery, 76% accuracy was achieved. For non-taxonomic relationship discovery, a small ontology, with 284 concepts and 88 non-taxonomic relationships as the gold standard, was manually developed. Because the traditional evaluation metrics, precision and recall, cannot measure the real quality of automatic relationship discovery if the relationships are of varying degrees of accuracy, four categories of relationship matches against the gold standard were defined as "utterly wrong," "rather bad," "near miss," and "direct hit." A new metric called Generic Relation Learning Accuracy (RLA) was then defined in order to measure the average accuracy of an instance of a relationship discovered against the best counterpart from the gold standard. The best RLA earned when experimenting with different parameters (support and confidence) was 67%.

TIMS (Tag Information Management System) [133] is a terminology-based knowledge acquisition and integration system in the domain of molecular biology. The system is very comprehensive and can support ontology population using automatic term recognition and clustering, and knowledge integration and management, using XML-data management technology, as well as information retrieval. For knowledge acquisition, TIMS uses automatic term recognition (ATR) and automatic term clustering (ATC) modules. The ATR module is based on the C/NC –value method [134], which uses both symbolic information, such as POS tagging, and statistical information, such as the frequency of occurrence of a term. The C/NC method is specifically adapted to multi-word term recognition. The ATC module is based on Ushioda’s AMI (Average Mutual Information) hierarchical clustering method [135], and is built on the C/NC results. The output of ATC is a dendrogram of hierarchical term clusters. Preliminary evaluation of ATR showed precision from 93% to 98% for the top 100 terms taken from a NACSIS AI-domain corpus and a set of MEDLINE abstracts.

Focusing on the vast quantity of information available on the WWW, WEB → KB [136] is an ontology learning system that uses a machine-learning approach for trainable information extraction. The system takes two inputs: 1) a knowledge base consisting of ontology-defined classes and relationships; and 2) training examples from the Web that describe instances of these classes and relationships. Based on these inputs, the system determines general procedures capable of extracting additional instances of these classes, as well as rules for extracting new instances, rules for classifying pages, and rules for recognizing relationships among several pages. WEB → KB uses mainly statistical, machine-learning approaches to accomplish these tasks. For evaluation, the authors attempted to learn information about faculty, student, course, and departments from Web pages, creating an organizational knowledge base. The average accuracy was over 70%, at a coverage level of approximately 30%. The authors also explored and compared a variety of learning methods, including statistical bag-of-words classifier, first-order rule learner, and multi-strategy learning methods. More complex methods, such as first-order rule learning, tended to generate greater accuracy than the simple bag-of-word classifier, at the expense of lower coverage.

3.0 DEVELOPMENT OF RESEARCH QUESTIONS

Based on the literature review described in the previous chapter, I believe that methodologies developed in the fields of Natural Language Processing, Information Retrieval, Information Extraction, and Artificial Intelligence can be utilized for ontology enrichment to alleviate the knowledge acquisition bottleneck in Biomedicine. However, there are many issues that must be addressed before we can completely realize the potential benefits of these methods for fully automated or even semi-automatic ontology enrichment in biomedical domains.

Although current methods can be applied to ontology learning in biomedical domains, some methods may be more useful than others, due to the constraints of medical and biological language. Some features utilized by the various linguistic approaches are quite prevalent in medical and biological text, and make it particularly appealing to attempt ontology enrichment using these methods. For example, compound nouns are common in the biomedical domain, because many biomedical terms are composed by adding additional modifiers to the existing terms. A number of researchers have explored this phenomenon in detail [137-139], especially because of its implication for post-coordination and compositional models [139]. Methods that utilize such component information could be effective for hyponym placement [63, 79].

Our field boasts many well-developed knowledge and lexical resources, such as existing ontologies and terminologies, domain-specific corpora, and general dictionaries that are necessary for knowledge extraction. WordNet provides an important resource for ontology learning of general English domains [49], and could be utilized in ontology learning in biomedical domains. Combined approaches that leverage both WordNet and biomedical ontologies and vocabularies could be particularly interesting. With wide recognition of the importance of sound and complete ontologies in the field of biomedical informatics, endeavors such as the NCI's Enterprise Vocabulary Services, Open Biomedical Ontologies Consortium, and the Gene Ontologies Consortium provide ample opportunities to explore the benefit of the enrichment of existing biomedical ontologies. However, many of these techniques have never been tested and evaluated in biomedical domains. Nearly all systems built for ontology knowledge learning and extractions have been developed specifically for domains other than the biomedical, and often in languages other than English.

There are significant barriers to the immediate translation of previous research to the biomedical domain. First, biomedical language is very different from that used in these other domains [140-143]. Sources of biomedical text, such as clinical and biomedical texts, also differ in their characteristics. Many clinical reports are structured in such a way that the header or sections provide context that must be used to make inferences regarding further content. For example, in pathology reports, text such as "PROSTATECTOMY: Adenocarcinoma," requires an inference about the origin of the disease from knowledge regarding the procedure. Few algorithms have specifically addressed the issues related to section segmentation and inference. Most of patterns used in the symbolic approach were discovered from general English domain. They may miss some domain specific patterns existed in the biomedical domain. Researchers have suggested that this difficulty might be alleviated by the discovery of domain-specific patterns from domain corpus using a pattern-learning approach, such as Hearst's pattern matching method [103, 144, 145].

Second, because sublanguages differ widely from one another, many general English-based algorithms may not be effective when applied to more specific sublanguages. The performance of existing methods is likely to vary by domain and task, and little research into a systematic evaluation of these methods in the biomedical domain has been conducted. In general, investigators in this area would benefit from systematically testing and extending existing approaches that can best explore the characteristics of biomedical and clinical text, and directly comparing the performance of these methods.

Third, because of the lack of an evaluation framework and reference standards for methods used in the biomedical domain, the direct application of previous work is rendered more challenging. Furthermore, researchers working in the same area may be evaluating different aspects of ontology enrichment; thus, their work cannot be compared. Only a few researchers have dedicated significant work to developing appropriate evaluation methods. The OntoClean methodology [146] developed by Guarino's group describes a set of rules that can be applied systematically to taxonomies to remove the erroneous subclass (in the is-a relationship) and that may be useful for ontology pruning and refinement. Another group led by Faatz and Steinmetz [147] studied an evaluation framework for ontology enrichment, describing a quality measurement framework for ontology enrichment methods with relevance and overlap heuristics.

More research is needed in this area to develop robust performance metrics, and to move the field towards more standardized approaches that permit meta-analysis.

The goals of this thesis study are to address two of the barriers mentioned above. First, I sought to design a comparative study in which the existing NLP approaches will be evaluated for ontology enrichment in the biomedical domain. I intended to explore the extent to which these techniques can be used to discover new concepts, and to determine how the techniques can be incorporated into the existing ontologies: in what ways can the existing biomedical domain knowledge and resources be utilized for ontology enrichment? Second, while comparing the different approaches for ontology learning, I intended to explore and develop a framework that can be used for the evaluation and development of ontology learning methods in the biomedical domain. I especially sought to develop metrics for rapid and repeatable evaluation. With such metrics and a defined set of evaluation strategies, one can compare different learning methods for certain domains more rapidly than the current methods allow. Comparing and analyzing these metrics can provide complementary insights for ontology learning from texts. With these goals in mind, I developed the following research questions for this study.

RESEARCH QUESTIONS

1. How do certain characteristics of biomedical texts have an impact on the effectiveness of NLP methods for biomedical ontology enrichment?

2. How effective are various NLP approaches for biomedical ontology enrichment using free-text medical documents as a learning resource?
3. How do the structure and contents of the biomedical ontology itself affect the efficacy of these NLP approaches for ontology enrichment?
4. How can the existing biomedical ontology be utilized for ontology enrichment method evaluation?
5. Are traditional evaluation metrics, such as recall and precision, suitable for ontology enrichment methods? If not, what kind of metrics should be used?

To answer these questions, I studied three commonly used NLP methods for ontology information extraction and determined their suitability and efficacy for ontology concept discovery. These three methods represent two fundamentally different methodological approaches--the symbolic and the statistical--and have been highly regarded in their respective fields. The three methods include the following: 1) the LSP matching method: 2) Church's mutual information method; and 3) Lin's combination of the mutual information method and the syntactic context information method.

I also used two domain free-text learning resources, a large corpus of pathology reports and a large corpus of radiology reports, which represent two different subdomains of medical language, as well as two biomedical ontologies: the National Cancer Institute Thesauri (NCIT) and RADLex, both of which serve as target ontologies as well as knowledge resources.

The focus of this thesis is two-fold: to study the interactions between the three variables expressed in the research questions – learning methods, domain corpus, and domain ontology -- and to explore and develop an evaluation framework that can be used for researchers to compare and develop different OL methods for ontology enrichment for different domains. These two goals are not mutually exclusive; in fact, they are actually synergistic, as I cannot compare and evaluate the OL methods without evaluation strategies and metrics.

4.0 RESEARCH DESIGN AND METHODS

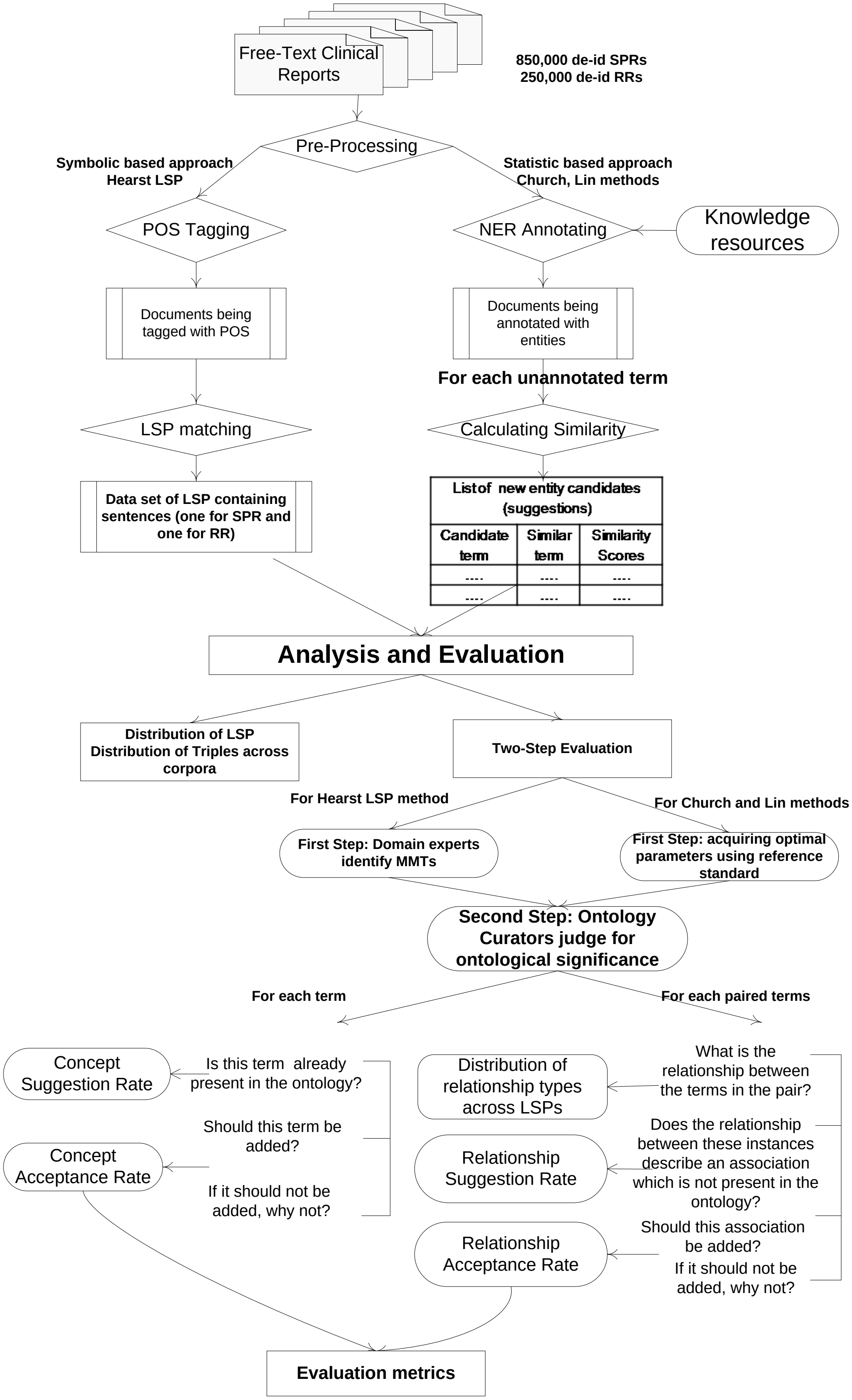
Figure 1 provides an overview of the study's design, and includes some of the results of this dissertation work. To fulfill one of the goals of this study, I developed a methodology framework and metrics that enabled me to evaluate and compare several OL methods for ontology enrichment in the biomedical domain. The clinical documents and knowledge resources described below were used for both OL method studies. In the following sections (4.1 and 4.2), I will describe the studies conducted on each type of OL method, as well as the development of an evaluation framework and evaluation metrics in the context of method study in the corresponding sections (4.1 and 4.2).

Clinical corpora used in this study

I used two clinical document types as ontology learning resources: surgical pathology reports and radiology reports. The corpus of surgical pathology reports included a total of 852,764 documents; the corpus of radiology reports included a total of 209,997 documents. Both corpora were obtained from the clinical information systems of the University of Pittsburgh Medical Center (UPMC), which includes a total of 18 hospitals. Both corpora were de-indentified to meet the requirements of HIPAA "safe harbor" [45]. Use of the clinical corpora was approved by the University of Pittsburgh Institutional Review Board (IRB# PRO07070252).

Targeted biomedical knowledge resources

I selected two biomedical knowledge resources in active development that had the potential to benefit from ontology enrichment using clinical text. The National Cancer Institute Thesaurus (NCIT) [148] is a description logic-based ontology sponsored by the National Cancer Institute. It includes more than 75,000 key biomedical concepts in over 20 categories, including Disease, Abnormal Cell, Molecular Abnormality, Organism, Biological Process, etc. RadLex [149] is a lexicon for the uniform indexing and retrieval of radiology information resources, sponsored by the Radiology Society of North American (RSNA). It includes over 11,000 concepts in 12 categories, including Imaging Observation, Procedure, Characteristic, Treatment, etc. RadLex has previously been used to derive an application ontology for radiologic reporting, and seems likely to evolve into a formal ontology.


Free-Text Clinical Reports
850,000 de-id SPRs
250,000 de-id RRs
Pre-Processing
Symbolic based approach
Hearst LSP
Statistic based approach
Church, Lin methods
POS Tagging
NER Annotating
Knowledge resources
Documents being tagged with POS
Documents being annotated with entities
For each unannotated term
LSP matching
Calculating Similarity
Data set of LSP containing sentences (one for SPR and one for RR)
List of new entity candidates (suggestions)
Candidate term
Similar term
Similarity Scores

Analysis and Evaluation
Distribution of LSP
Distribution of Triples across corpora
Two-Step Evaluation
For Hearst LSP method
For Church and Lin methods
First Step: Domain experts identify MMTs
First Step: acquiring optimal parameters using reference standard
Second Step: Ontology Curators judge for ontological significance
For each term
For each paired terms
Concept Suggestion Rate
Is this term already present in the ontology?
Should this term be added?
Concept Acceptance Rate
If it should not be added, why not?
Distribution of relationship types across LSPs
What is the relationship between the terms in the pair?
Does the relationship between these instances describe an association which is not present in the ontology?
Relationship Suggestion Rate
Relationship Acceptance Rate
Should this association be added?
If it should not be added, why not?
Evaluation metrics

Figure 1. Overview of the study design

4.1 STUDY METHOD FOR EVALUATING THE LSP MATCHING METHOD FOR ONTOLOGY ENRICHMENT USING CLINICAL DOCUMENTS

4.1.1 Lexico-Syntactic Pattern (LSP) matching method

LSPs are surface markers that exist in natural language and often indicate a semantic relationship between terms in the text. For example, in the phrase “systemic granulomatous diseases **such as** Crohn’s disease or sarcoidosis,” the LSP “**such as**” can help us infer that “systemic granulomatous diseases” is a hypernym of “Crohn’s disease” and “sarcoidosis.” I identified a set of LSPs for use in this study, including those LSPs identified by Hearst [150] and Berland [151], and supplemented by some from our own manual inspection of clinical documents. Table 1 lists LSPs used in this study and provides example sentences that contain patterns observed in the corpora.

4.1.2 Extraction of sentences containing LSPs

Free-text pathology and radiology reports were processed in two steps (Figure 1). First, I tagged Parts Of Speech (POS) using a maximum entropy POS tagger that I had previously retrained with pathology reports [152]. Second, I used regular expressions over POS tags to extract all of the LSPs shown in Table I. For example, in the phrase “Compatible with benign eccrine neoplasia, ***such as*** nodular hidroadenoma,” the terms “benign eccrine neoplasia” and “nodular

LSP Category	LSP	Examples
Hearst	NP$_0$ such as {NP$_1$, NP$_{2\ldots}$, and|or} NP$_n$ Such NP$_0$ as {NP$_1$,}* { or | and} NP$_n$	Compatible with benign eccrine neoplasia, **such as** nodular hidroadenoma
		Such atypical pneumonia **as** mycoplasma or viral pneumonitis
	NP$_1$ {, NP$_2$} * {,} or other NP$_0$ NP$_0${, NP$_1$}*{,} and other NP$_2$	Residual basal cell carcinoma **or other** malignancy
		Pneumoconiosis **and other** chronic process
	NP$_0${,} including {NP$_1$,}*{or | and} NP$_2$	Peripheral blood pancytopenia **including** macrocytic anemia and rare nucleated red blood cells
		Chronic obstructive pulmonary disease **including** bronchial wall thickening
Other	NP$_0$ [a.k.a.|aka | also known as] NP$_1$ *{or | and} NP$_n$	Sebaceoma (**aka** sebaceous epithelioma)
	NP$_0$ so called {NP$_1$,}*{or | and} NP$_n$	Pleomorphic adenoma (**so called** hybrid adenoma)
Berland	Part NN in PREP {the | a} DET mods [JJ|NN]* whole NN Parts NN-PL in PREP wholes NN-PL	Phospholipids **in** the cell membrane…
	Part NN-PL of PREP {the | a} DET mods [JJ|NN]* whole NN Parts NN-PL of PREP wholes NN-PL	Membrane **of** a cell

Table 3. Lexico-Syntactic Patterns with examples from corpora

NP: Noun Phrase; NN: Noun; PREP: preposition; DET: determiner; JJ: adjective; NN-PL: Noun plural form; mods: modifiers.

Regular expression notation: {x }: x is optional; x|y: either x or y; x*: zero or more instances of repetition of x hidroadenoma" are Noun Phrases (NPs) and will match the LSP "NP_0 ***such as*** NP_1." This phrase will then be extracted for presentation to the domain experts. The output is a list of all sentences containing LSPs for each corpus. Processing was performed using the GATE platform [153].

4.1.3 Calculating LSP frequencies and distributions

For each LSP, I calculated the number of documents and sentences in which it was contained. Because many sentences contained the same terms and LSPs, I also calculated the number of sentences containing unique LSPs. Frequency data enabled us to compute the potential yield of concepts and relationships within a corpus if the rate at which LSPs provide useful information for ontology or lexicon curators is known. Additionally, I used frequency data to determine the sample number for each LSP that was provided to human judges. For LSPs with more than 50 unique instances, I sampled 50 instances. For LSPs with 25 to 50 unique instances, I included all instances. I excluded LSPs with fewer than 25 unique instances.

4.1.4 Evaluation of ontology suggestions

I developed a two-step process to determine the value of suggestions generated with the LSP approach. The evaluation approach relies on manual annotations, assuming that automated methods using POS and noun-phrase identification can later be used to approximate the results of the human annotation.

In Step 1, domain experts examined each sentence containing an LSP and identified the Medically Meaningful Terms (MMTs) preceding and following the LSP. From the manual annotations, I could evaluate the maximum yield expected from applying LSPs to each corpus when the assumption is that all the MMTs have been correctly extracted. Manually identified MMTs from Step 1 were used as input to Step 2, in which I evaluated the value of the MMTs for ontology enrichment. The use of human annotations of MMTs for the evaluation of Step 2 permitted us to more accurately determine the true value of ontology enrichment without confounding the evaluation with possibly incorrect MMTs.

In Step 2, NCIT and RadLex curators determined whether the MMT was already present in the knowledge resource and, if not, whether it should be added. Next, they judged whether there was a relationship between the paired MMTs. If there was a relationship, the curators annotated the type of relationship and indicated whether it already existed in the ontology. If it did not exist, they determined whether it should be added. Finally, if the curators determined that the relationship should not be added, they provided a reason for their decision. I restricted the relationship types to synonym, hypernym, meronym, and other (if the relationship did not fall into any of the three predetermined relationships). These judgments required not only domain knowledge, but also an in-depth understanding about a knowledge resource’s structure and content.

Step 1: Identify the medically meaningful terms from extracted sentences

Domain experts included two resident pathologists (second and third year) and two resident radiologists (second and fourth year). Each group was presented with a sample of LSP-containing sentences from the pathology or radiology corpus, respectively. Domain experts were asked to annotate the MMTs, before and after the LSP, that could stand alone. For example, in the following text, “Abnormal slightly high T2 signal seen in the porta hepatis which may be secondary to an underlying malignancy ***such as*** Klatskin tumor or gall bladder carcinoma,”, the bold and italic term ***such as*** is the LSP. Domain experts would annotate “malignancy” as the MMT before the LSP and “Klatskin tumor” and “gall bladder carcinoma” as the MMTs after the LSP. The final product of the annotation was a table of paired MMTs from each sentence. When multiple terms were annotated before or after the LSP, I created a separate term-pair for each combination. All annotation was performed using Microsoft Excel. Domain experts were given a spreadsheet containing the sentences extracted with bolded LSPs. They annotated the MMTs before and after the LSP by copying and pasting them into a second and third column.

Domain experts were trained to perform the annotation using a modification of an existing annotation guideline for manual annotation of clinical conditions from emergency department reports developed by Chapman et al [50]. On a development set, I used a Delphi method with repeated training until the F measure exceeded the threshold of 0.9, as depicted in Figure 2. Subsequently, expert annotators were given the final sample, which consisted of 50 unique sentences for each LSP, to annotate.

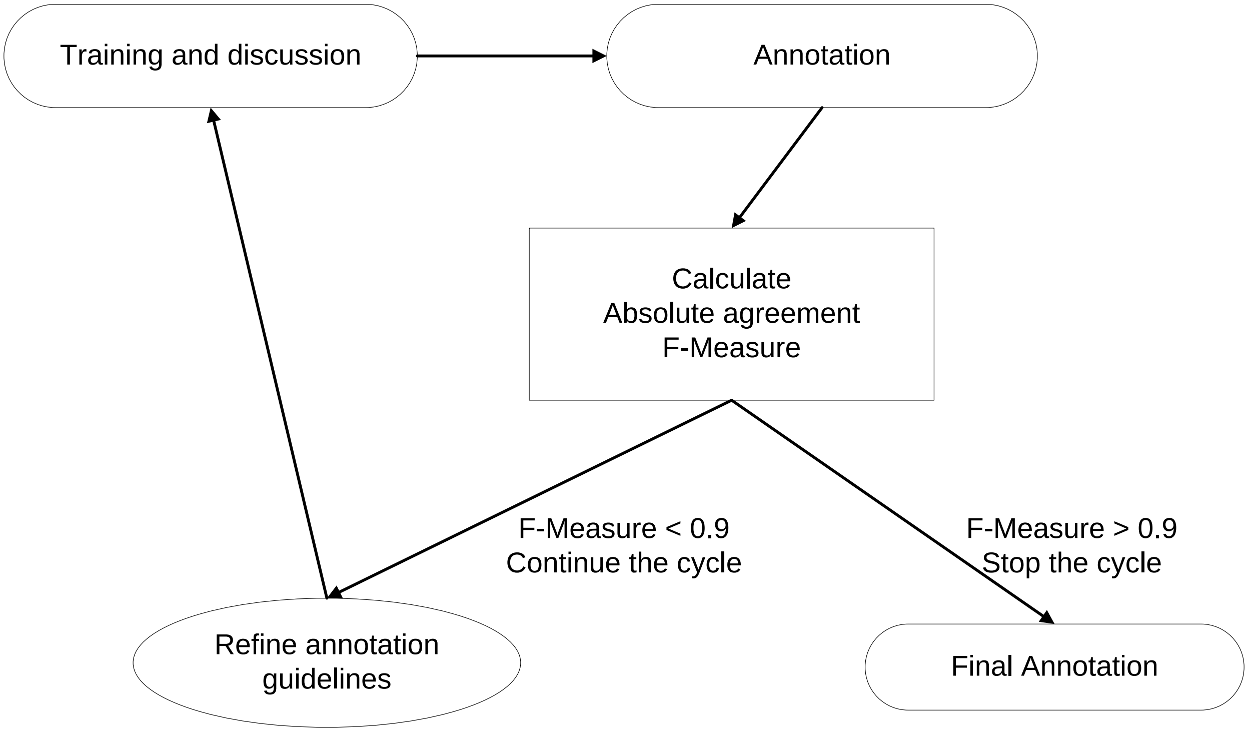


Figure 2. Domain expert training process

Step 2: Determine the value of concepts and conceptual relationships obtained from MMTs

Domain expert annotations resulted in a list of paired MMTs for pathology and a similar list for radiology. I then invited two experienced curators to judge the MMTs produced by the domain experts in Step 1. One ontology curator, a pathologist who is currently curating the National Cancer Institute Thesaurus, evaluated the term list obtained from the surgical pathology corpus. The other curator, a radiologist who is currently curating RadLex, evaluated the term list obtained from the radiology corpus.

For each term in a term-pair, curators asked the following questions:

1) Is the term already represented in the resource (possibly as a synonym)?

2) If not, should a new concept based on this term be added to the resource?

3) If not, what is the reason for the determination that it should not be added?

For each pair of terms, ontology curators also explored the following questions:

4) If there is a relationship between the two terms, what is the relationship?

(Relationship choices were restricted to synonym, hypernym/hyponym, meronym, and other.)

5) Does this relationship exist in the resource?

6) If not, should the relationship be added to the resource?

7) If no new relationship should be added, what is the reason for this determination?

4.1.5 Defining evaluation metrics

The classic measure of precision is not entirely adequate in summarizing the resulting data, since it does not capture the two-step process I anticipate using for suggesting new ontological elements. Therefore, I defined more specific evaluation metrics to quantify efficacy for the two discrete steps.

Concept Suggestion Rate (CSR):

$$\text{CSR} = \frac{\text{\# of MMTs that were not in the ontology}}{\text{Total \# of MMTs extracted by the method}} \quad \text{Equation 3}$$

This metric indicates the percentage of terms, extracted using the enrichment method, that are new concept candidates and would be presented to the curator for a given target ontology.

Concept Acceptance Rate (CAR):

$$\text{CAR} = \frac{\text{\# of MMTs that should be included as new concept, instance, or synonym in the ontology}}{\text{Total \# of MMTs extracted by the method}}$$

Equation 4

This metric indicates the percentage of terms, extracted using the enrichment method, that would be added to the relevant ontology (these may represent new concepts or new instances).

Relationship Suggestion Rate (RSR):

$$\text{RSR} = \frac{\text{\# of relationships that were not in the ontology}}{\text{Total \# of relationships extracted by the method}}$$

Equation 5

This metric indicates the percentage of concept relationships, extracted using the enrichment method, that are candidates for a new concept relationship and would be presented to the curator for a given target ontology.

Relationship Acceptance rate (RAR):

$$\text{RAR} = \frac{\text{\# of relationships that should be included in the ontology}}{\text{Total \# of relationships extracted by the method}}$$

Equation 6

This metric indicates the percentage of concept relationships, extracted using the enrichment method, that would be added to the relevant ontology.

Additionally, I defined two measures that combine this information to provide an estimate of the total number of concepts or relationships extracted from a given corpus using the LSP matching method.

Estimated Concept Yield (ECY) $_{\mathbf{LSP}}$:

$$ECY_{\mathbf{LSP}} = N * R * CAR \qquad \text{Equation 7}$$

N: Total number of unique LSPs in the corpus

R: Average number of MMTs that can be extracted per LSP, which is equal to the total number of MMTs divided by the total number of LSPs

CAR: Concept Acceptance Rate

Estimated Relationship Yield (ERY) $_{\mathbf{LSP}}$:

$$ERY_{\mathbf{LSP}} = N * P * RAR \qquad \text{Equation 8}$$

N: Total number of unique LSPs in the corpus

P: Prevalence of a single relationship that is equal to the percentage of a single type of relationship among all of the relationships being extracted

RAR: Relationship Acceptance Rate

4.2 STUDY METHOD FOR EVALUATING THE STATISTIC METHODS FOR ONTOLOGY ENRICHMENT USING CLINICAL DOCUMENTS

4.2.1 The statistical methods

Two statistical methods are being evaluated and compared in this study: Church's mutual information method [43] and Lin's similarity measure method [46]. I selected these two methods because they differed in their utilization of syntactic features for similarity measures, allowing a comparison of efficacy for ontology enrichment, and because each of the methods was well respected in its particular field.

4.2.1.1 Church method

Church [43] used word co-occurrence information to measure the degree of similarity between two words. The mutual information (I) between two words in a corpus, x and y, is defined as follows:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x) * P(y)} \qquad \text{Equation 9}$$

Where P(x) is the probability of x, P(y) is the probability of y, and P(x, y) is the joint probability of x and y. If there is a genuine association between x and y, then the joint probability P(x, y) will be much higher than chance, P(x)* P(y), and consequently I(x, y) >> 0. If there is no relationship between x and y, then P(x, y) $\approx$ P(x)* P(y), and thus I(x,

y) = 0. If x and y have complementary distribution, P(x, y) will be much less than P(x)*P(y) and I(x, y) <<0.

The P(x) and P(y) can be estimated by calculating the frequency of the word in a corpus. For x and y to be considered highly associative, there are two variables that can influence the results: one is the window (W) of x and y; and the other is the threshold of I. The output is given in a table with three columns. The first column contains the target word x. The second column contains y words that are similar to x based on the mutual information measure I. The third column contains the I score.

4.2.1.2 Lin method

Lin's method [46] uses the syntactic information between two words for similarity measure in addition to word co-occurrence. The co-occurrence information between two words (w_1, w_2) and the grammatical relationship r between them are collected as the dependency triples (w_1, r, w_2). For example, in the sentence "Patient had a high fever," the following set of dependency triples can be obtained: (had – subject – Patient); (had – object – fever); (fever – adjective-modifier – high); and (fever – determiner – a). The similarity between the two words, w_1 and w_2, can be calculated as follows:

$$\text{Sim }(w_1, w_2) = \frac{\sum (r,w) \in T(w_1) \cap T(w_2)(I(w_1,r,w) + I(w_2,r,w))}{\sum (r,w) \in T(w_1) I(w_1,r,w) + \sum (r,w) \in T(w_2) I(w_2,r,w)} \quad \text{Equation 10}$$

Where $I(w_1, r, w_2) = \log \frac{\| w_1, r, w_2 \| * \| *, r, * \|}{\| w_1, r, * \| * \| *, r, w_2 \|}$ Equation 11

There are two parameters in this formula: one is the threshold of I; the other is the type of syntactic relationship I that needs to be included in the calculation. The output is given in a table similar to that for Church’s method, as detailed in the previous paragraph.

4.2.2 Named Entity Recognition system used in this study

A Named Entity Recognition (NER) system, based on IndexFinder [154], was developed in our research group in order to help us identify ontology concepts present in the clinical documents. This NER engine is built in such a way that the knowledge resource it used to annotate the class concept in the corpus can be changed based on user selection. Given a text corpus as input, the NER system can identify named entities in a free-text corpus based on any knowledge resource provided by the user (Fig. 3). The clinical documents annotated with the knowledge resource by NER constitute the system’s output, from which I determined how many terms in the corpus were present in the knowledge resource.

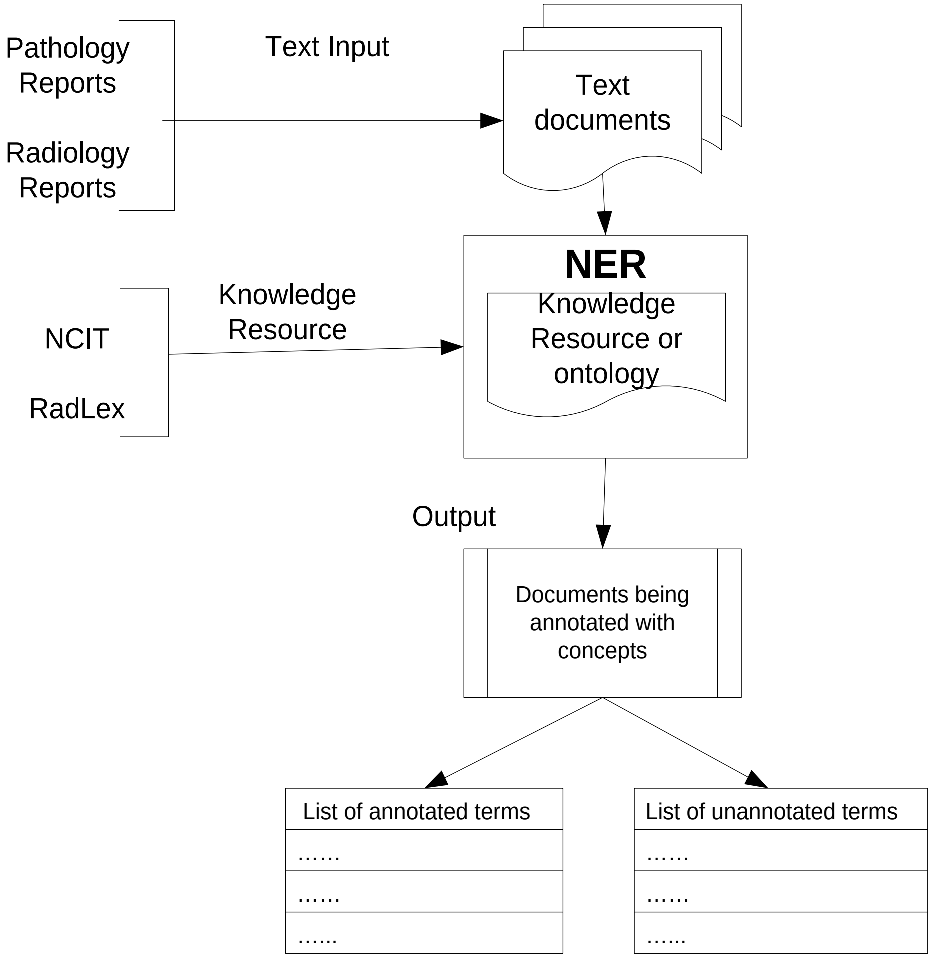
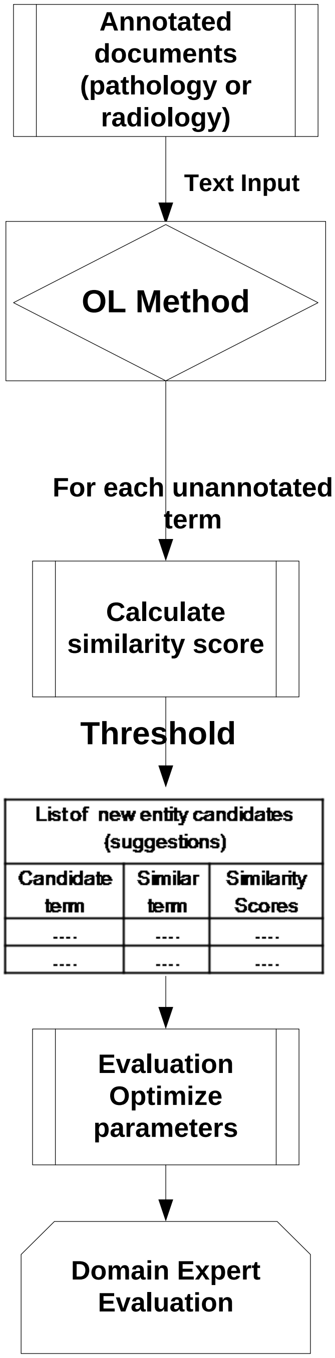


Figure 3. NER system

4.2.3 Generating new concept suggestions

The clinical documents were divided into two equal sets: the development set and the evaluation set. I used the development set to experiment with the statistical methods in order to determine the optimal parameters; the evaluation set was used for the final evaluation of the efficacy of these statistical methods for ontology enrichment. Both data sets served as document input for statistical methods to produce new concept suggestions (Fig. 4).


Annotated documents (pathology or radiology)
Text Input
OL Method
For each unannotated term
Calculate similarity score
Threshold
Evaluation Optimize parameters
Domain Expert Evaluation

List of new entity candidates (suggestions)		
Candidate term	Similar term	Similarity Scores
----	----	----
----	----	----

Figure 4. New Concept Suggestion Generated using Statistical Methods

As noted in Section 4.2.1), both the Church and Lin methods use word co-occurrence and/or other syntactic features to measure the similarity of two terms. For an ontology enrichment task, I sought to discover new concepts that are similar to existing concepts in the target ontology. If a term in the concept was found to be similar to an existing concept, as determined by the statistical methods, this term was suggested as a new concept candidate. The benefit of this approach is twofold: 1) any new concept suggestions are more likely to fall within the scope of the target ontology; and 2) the information that a suggested concept is associated with an existing concept may help an ontology developer to appropriately place it.

I first used NER (see Fig.1) to annotate the clinical corpora (pathology or radiology) with the ontologies (NCIT 0609d or RadLex v2.00). These annotated clinical reports were then run through the two statistical enrichment methods. For each of the annotated terms, the methods returned a list of similar terms, with similarity scores given in descending order. The output of the methods consisted of lists of paired terms that were considered similar based on determinations from the statistical methods that one of the two terms was already in the ontology. Terms were added to the list of new concept suggestions when similarity scores exceeded the threshold. Different thresholds produced different sets of concept suggestions.

4.2.4 Evaluation study

The conventional evaluation study for a statistical method involves the use of a pre-established reference standard, against which the output is compared; evaluation metrics, such as recall and precision, are also employed. In most cases, the reference standard is divided into two subsets: a training data set and a testing data set. The training data set is used to train the statistical method in order to obtain the optimal parameters used in the statistical method. Because the optimal parameters vary for different domains, this training process allows the optimal parameters for the targeted domain to be more easily obtained. The evaluation data set is used for the final evaluation, in which the method is tuned and optimal parameters are discerned. Therefore, I divided the study into a methodology development phase and a method evaluation phase.

4.2.4.1 Methodology development

I first tested the statistical methods for specific domains (pathology and radiology) and task (new concept suggestion). As noted earlier, two parameters, window size and similarity score threshold in the Church method and syntactic triples and similarity score threshold in the Lin method, can influence a similarity score. For example, In the Church method, the window size (w) of two terms appearing together in text can influence the similarity score, I; the threshold I is used to determine to what degree the two terms are similar enough to be considered a new concept suggestion. The optimal parameters are those that can capture all the new concept suggestions without introducing a lot of noise (e.g., false positives). In this phase of the study, I utilized the existing ontologies to create our reference standards. The following section describes the process of this methodology development for both statistical methods.

(a) Establishing a reference standard for methodology development

Our task was to enrich an existing ontology by adding new concepts through use of statistical methods on free-text clinical reports. There were no reference standards readily available for this study, and it was not possible to create them due to the extensive labor, resource, and time such a project would require. Manual evaluations by human experts were very time consuming and expensive. Therefore, I sought to utilize the existing ontologies to generate the reference standards. Although the target ontologies used in this study (NCIT and RadLex) have shortcomings, they were developed manually under the supervision of a group of domain experts. I believe for the purpose of methodology development, these ontologies could be used as the reference standard.

For each targeted ontology, I retrieved early, published ontology: for the NCIT, I used NCIT 0909c (published in 2009), and for the RadLex, I used version v.3.03 (also published in 2009). I first used the NER system to identify all the concepts of these two ontologies that could be found in the respective clinical corpus (NCIT for the pathology corpus, and RadLex for the radiology reports). I then extracted concepts and compiled a list of those from each ontology that also appeared in the clinical corpus. These lists became our reference standard data sets. The assumption was that if these concepts did not already exist in the ontologies, and we were going to discover them from the corpus using statistical methods, the entire list would constitute the upper-bond, or the majority, of the concepts we could discover. Once the reference standard has been established, traditional evaluation metrics, such as recall and precision, can be applied to the concepts.

The traditional precision and recall considers the matching of two concepts at the lexical level. In this definition, the reference standard set consists of all the concepts that presented in the corpus. When I judge whether a suggested term has achieved a positive result, I look at both terms in the pairs. If both can be found in the reference standards, then the result is positive; otherwise, the result is a false positive. I also calculate the F-measure, which is the harmonic mean of recall and precision.

$$\text{Recall (R)} = \frac{\text{Total \# of positives}}{\text{Total \# of reference standards}} \qquad \text{Equation 12}$$

$$\text{Precision (P)} = \frac{\text{Total \# of positives}}{\text{Total \# of suggestions}} \qquad \text{Equation 13}$$

$$\text{F-measure (F)} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad \text{Equation 14}$$

(b) Generating optimal parameters for the statistic methods using precision and recall curve

I performed the same study for both the pathology and radiology corpora, as described below.

For each set of suggestions, I calculated recall and precision; a precision and recall curve based on these metrics was then drawn.

For the Church method, I used seven window sizes (2, 3, 4, 5, 6, 7, and 15) and a series of incremental thresholds for I, and generated a precision and recall curve for each window size. After examining all the precision and recall curves, the best combination of window size and threshold was determined.

For the Lin method, the parameters are threshold and type of syntactic relationship. I experimented with several sets of syntactic relationships and summarized them in Table 2. I then generated a precision and recall curve for each syntactic relationship; the best combination of syntactic relationship and threshold was determined after examining all the precision and recall curves.

Set	**Syntactic Relationship (SR)**
1	All syntactic relationships
2	Top ten most frequent syntactic relationships
3	obj
4	mod
5	nn
6	conj
7	obj, mod, conj, nn
8	obj, mod, conj, nn, det
9	subj

10	obj, mod, conj, nn, det, s
11	lex-mod

Table 4. Syntactic relationships experimented with Lin's method

(obj: objective; mod: modifier; nn: noun; conj: conjounction; det: determiner; s: sentence; lexmod: lexical modifier).

4.2.4.2 Final evaluation by domain experts

After I obtained suitable parameters for each method and domain, I proceeded to the final stage: evaluation by domain experts. A set of 30,000 clinical reports for each domain (pathology and radiology) was used as the learning resource. OL methods were run separately on each set of clinical documents. Although human expert evaluation provides the best estimated value of suggestions generated by statistical approaches, it isn't possible to have domain experts evaluating every single suggestion; the time and resources required for such a tedious process are daunting. Consequently, I randomly selected a subset of 100 pairs of terms from each output, and gave these data to the domain experts for final evaluation.

The domain expert evaluation method and evaluation metrics such as concept suggestive rate and concept acceptance rate are the same as those used in the symbolic method evaluation study (see 4.1.5 and 4.1.6).

5.0 RESULTS

5.1 RESULTS FROM SYMBOLIC METHOD STUDY

5.1.1 Lexico-Syntactic Pattern (LSP) matching method

Table 5 shows the frequency of seven LSPs across the radiology and pathology corpora. Sentences that contained any LSPs were extracted. The data are shown as LSPs per sentence and per unique sentence. The overall frequency of patterns appearing in the corpora was low. Although it is not possible to determine how accurate the LSPs are in extracting all relevant instances, the method is expected to perform well in this regard because it is based on string matching. I have not observed false negatives during manual inspection of sample documents from the corpus. Nevertheless, there are factors that could affect the accuracy of results: a POS tagging error could occur, and it is possible that some instances of LSPs could be missed due to misspellings and other typographical errors. The POS tagger was trained with a pathology corpus, and achieved 93% POS tagging accuracy; the accuracy was 91% when the radiology reports were tagged.

LSP	Surgical pathology reports			Radiology reports		
	852,764 reports 16,157,608 sentences			209,997 reports 4,057,228 sentences		
	# sentences	# sentences containing unique LSP	# randomly selected sentences	# sentences	# sentences containing unique LSP	# randomly selected sentences
NP such as NP	98	95	50	906	251	50
NP including NP	6291	4952	50	1403	747	50
NP other NP	6940	2251	50	10622	1407	50
NP also called NP	48	37	37	29	22	0
NP aka NP	5396	460	50	2	2	0
NP in the NP	47124	23178	50	64044	29285	50
NP of the NP	246798	70735	50	173016	54895	50
total	312695	101708	337	250022	86609	250

Table 5. Frequencies of various LSPs in the pathology and radiology corpora

5.1.2 How many medically meaningful terms (MMTs) could be identified for each LSP presented in the corpus

Table 6 shows the number of medically meaningful terms (MMTs) that could be identified by domain experts in a sample of sentences obtained from each corpus. The total number of sentences used for each LSP is shown in Table 3. For each LSP, there was at least one MMT preceding the LSP and more than one MMT following the LSP. Thus, multiple MMTs could be extracted from a sentence that contained a single LSP.

	Surgical pathology reports		Radiology reports	
LSP	Preceding the LSP	Following the LSP	Preceding the LSP	Following the LSP
	Ratio (# of MMTs/ # of instances of LSP)	Ratio (# of MMTs/ # of instances of LSP)	Ratio (# of MMTs/ # of instances of LSP)	Ratio (# of MMTs/ # of instances of LSP)
NP such as NP$_1$, NP$_2$	1.04 (52/50)	1.88 (94/50)	1.0 (50/50)	1.9 (95/50)
NP including NP$_1$, NP$_2$	0.98 (49/50)	1.62 (81/50)	1.0 (50/50)	1.72 (86/50)
NP other NP$_1$, NP$_2$	1.0 (50/50)	1.06 (53/50)	1.0 (43/43)	1.0 (43/43)
NP also called NP$_1$, NP$_2$	0.95 (35/37)	0.97 (36/37)	NA	NA
NP aka NP$_1$, NP$_2$	0.96 (47/50)	1.18 (59/50)	NA	NA
NP in NP$_1$	1.0 (50/50)	1.0 (50/50)	0.94 (47/50)	0.76 (39/50)
NP of NP$_1$	1.0 (50/50)	1.0 (50/50)	0.8 (40/50)	0.68 (34/50)
Average	0.99 (333/337)	1.26 (423/337)	0.95 (230/243)	1.22 (296/243)
Average # MMT per LSP	2.25		2.21	

Table 6. Number of medically meaningful terms (MMTs) extracted by the LSP method

5.1.3 New Concept Suggestion Rate (CSR) and new Concept Acceptance Rate (CAR)

Table 7 shows the new concept suggestion rate and the new concept acceptance rate as determined by the curators. For NCIT, the concept suggestion rates ranged from 37% for the pattern “NP such as NP1, NP2” to 11% for the pattern “NP of NP1,” with an average of 24% over seven patterns. For RadLex, the suggestion rates were higher, ranging from 52% for the pattern “NP such as NP1, NP2” to 18% for the pattern “NP in NP,” with an average of 37% over five patterns. However, nearly all the terms suggested would be accepted into the NCIT. The concept suggestion rate and concept acceptance rate were nearly equal. In contrast, the majority of terms suggested for RadLex would not be accepted into the terminology by the curator.

LSP	Surgical pathology reports		Radiology reports	
	CSR	CAR	CSR	CAR
NP such as NP$_1$, NP$_2$	37% (52/140)	31% (43/140)	52% (75/145)	10% (14/145)
NP including NP$_1$, NP$_2$	32% (61/189)	32% (60/189)	39% (54/138)	14% (19/138)
NP other NP$_1$, NP$_2$	16% (18/113)	16% (18/113)	33% (28/86)	8% (7/86)
NP also called NP$_1$, NP$_2$	14% (10/74)	10% (7/74)	NA	NA
NP aka NP$_1$, NP$_2$	31% (37/119)	31% (37/119)	NA	NA
NP in NP$_1$	12% (12/100)	6% (6/100)	18% (13/74)	8% (6/74)
NP of NP$_1$	11% (11/98)	6% (6/98)	26% (21/80)	14% (11/80)
Average	24% (201/833)	21% (177/833)	37% (191/523)	11% (57/523)

Table 7. Comparison of new concept suggestion rate and acceptance rate

5.1.4 Distribution of semantic relationships that being extracted by each type of LSP

One of the advantages of the LSP matching method is that the extracted terms preceding and following the LSP are expected to be semantically related. In our study, curators evaluated the semantic relationships between the pairs of MMTs and I calculated the distribution of each type of relationship based on the curator annotation (Table 8).

Corpus	**LSP**	**Semantic Relationship**				
		Hyponym	**Synonym**	**Meronym**	**Other**	**None**
Surgical Pathology Reports	**NP such as NP$_{1,}$ NP$_2$**	**37% (24/65)**	**0%**	**2% (1/65)**	**57% (37/65)**	**5% (3/65)**
	NP including NP$_{1,}$ NP$_2$	**10% (11/114)**	**1% (1/114)**	**6% (7/114)**	**78% (89/114)**	**5% (6/114)**
	NP other NP$_{1,}$ NP$_2$	**39% (24/61)**	**0% (0/61)**	**2% (1/61)**	**46% (28/61)**	**8% (5/61)**
	NP also called NP$_{1,}$ NP$_2$	**22% (9/41)**	**20% (8/41)**	**0%**	**37% (15/41)**	**10% (4/41)**
	NP aka NP$_{1,}$ NP$_2$	**10% (6/59)**	**44% (26/59)**	**0%**	**39% (23/59)**	**5% (3/59)**
	NP in NP$_1$	**0%**	**0%**	**0%**	**100% (45/45)**	**0%**
	NP of NP$_1$	**5% (2/44)**	**0%**	**18% (8/44)**	**61% (27/44)**	**16% (7/44)**
	Average	**18% (76/429)**	**8% (35/429)**	**4% (17/429)**	**62% (264/429)**	**7% (28/429)**
Radiology Reports	**NP such as NP$_{1,}$ NP$_2$**	**72% (26/36)**	**0%**	**0%**	**28% (10/36)**	**0%**
	NP including NP$_{1,}$ NP$_2$	**39% (7/18)**	**0%**	**11% (2/19)**	**33% (6/18)**	**17% (3/18)**
	NP other NP$_{1,}$ NP$_2$	**76% (16/21)**	**0%**	**0%**	**0%**	**24% (5/21)**
	NP in NP$_1$	**4% (1/26)**	**0%**	**12% (3/26)**	**42% (11/26)**	**42% (11/26)**
	NP of NP$_1$	**4% (4/27)**	**0%**	**0%**	**0%**	**96% (26/27)**
	Average	**40% (51/128)**	**0%**	**4% (5/128)**	**21% (27/128)**	**35% (45/128)**

Table 8. Distribution of semantic relationships extracted using the LSP matching method

5.1.5 New Relationship Suggestion Rate (RSR) and new Relationship Acceptance Rate (RAR)

Table 9 shows the new relationship suggestion rate and the new relationship acceptance rate as determined by the curators. For NCIT, on average, the relationship suggestion rate was 64%, and the relationship acceptance rate was 14%. For RadLex, on average, the relationship suggestion rate was 55%, and the relationship acceptance rate was 44%.

	Pathology reports (Enrich NCIT)		**Radiology reports (Enrich RADLex)**	
LSP	**RSR**	**RAR**	**RSR**	**RAR**
NP such as NP_1, NP_2	**55% (36/65)**	**26% (17/65)**	**94% (34/36)**	**94% (34/36)**
NP including NP_1, NP_2	**78% (89/114)**	**15% (17/114)**	**61% (11/18)**	**39% (7/18)**
NP other NP_1, NP_2	**51% (31/61)**	**8% (5/61)**	**57% (12/21)**	**57% (12/21)**
NP also called NP_1, NP_2	**29% (12/41)**	**10% (4/41)**	NA	NA
NP aka NP_1, NP_2	**73% (43/59)**	**24% (14/59)**	NA	NA
NP in NP_1	**84% (38/45)**	**0% (0/45)**	**50% (13/26)**	**12% (3/26)**
NP of NP_1	**64% (28/44)**	**5% (2/44)**	**0% (0/27)**	**0% (0/27)**
Average	**64% (277/429)**	**14% (59/429)**	**55% (70/128)**	**44% (56/128)**

Table 9. Comparison of new concept relationship suggestion rate and acceptance rate

5.1.6 Estimated Concept Yield (ECY) and Estimated Relationship Yield (ERY)

Using the metrics Estimated Concept Yield (ECY) and Estimated Relationship Yield (ERY) for both pathology corpus and radiology corpus, I estimated that as many as 15,000 (for radiology corpus) to 16,000 (for pathology corpus) new concepts, instances, or synonyms could be added, and perhaps as many as 2,000 (for pathology corpus) to 5,000 (for radiology corpus) new relationships could be added.

5.1.7 Reasons for why some of the suggested relationships would not be added in the corresponding resource

I also explored reasons why some of the suggested relationships would not be added into the corresponding resource. The top three reasons were: 1) that the relationships between classes of concepts are not modeled in the ontology (60%; e.g., the NCIT does not support relationships between anatomic concepts, procedure concepts, and findings); 2) that the relationship between two concepts is too general or vague to be included (20%; e.g., the relationship between "complication" and "Primary biliary cirrhosis" was considered to be too general); and 3) that there is no relationship between the two extracted concepts (10%).

5.2 RESULTS FROM STATISTIC METHOD STUDY

5.2.1 Reference standards

I obtained a total of 5,281 NCIT entities that were identified in the corpus of pathology reports, and a total of 660 RadLex entities that were identified in the corpus of radiology reports. These two sets of entities served as our reference standards for ontology learning algorithms for each respective domain.

5.2.2 Precision and recall curves

The following Figures show the precision and recall curves I obtained while tuning OL methods for ontology enrichment using clinical documents. Figure 5 shows the precision and recall for the Lin Method using all syntactic relationships (Set 1) for NCIT enrichment using pathology reports, while Figure 6 shows the precision and recall for the Lin method using all syntactic relationships (set 1) for RadLex enrichment using radiology reports. In general, with increased threshold, recall increases and precision decreases.

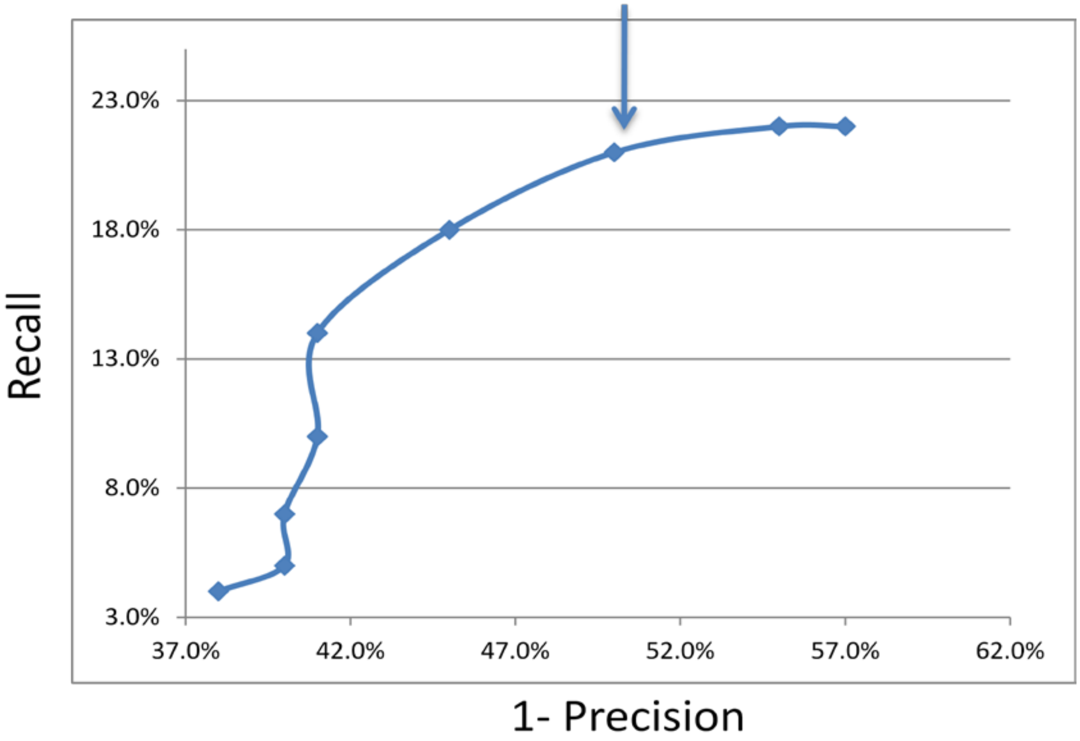


Figure 5. Precision and recall curve of Lin method for NCIT enrichment using pathology reports for all syntactic relationships (Set 1)

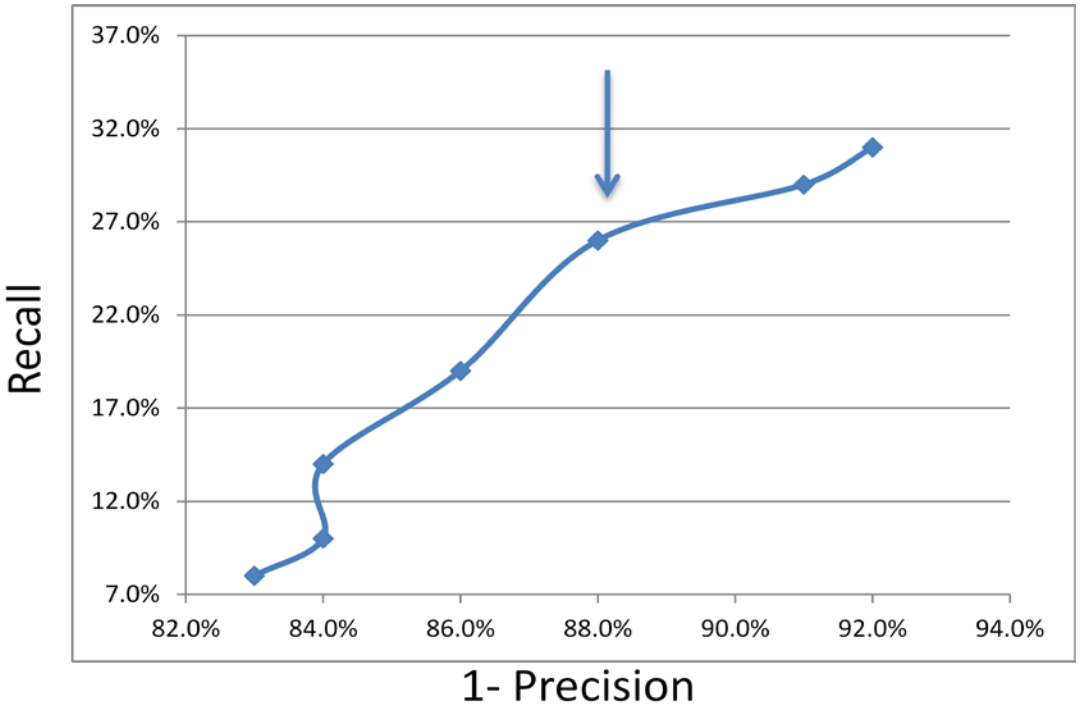


Figure 6. Precision and recall curve of Lin OL method for RadLex enrichment using radiology reports for all syntactic relationships.

For each values of the parameter (e.g. syntactic sets), I selected the best threshold based on the operating point on the curve that is the best combination of recall and precision. In Figure 2, the arrow designates the operating point, which has 22% recall and 50% precision. The threshold for that point is 0.07. The operating point for Figure 3 is at threshold 0.09, with 27% recall and 12% precision.

5.2.3 Parameters generated for Lin method

Table 10 shows the different combinations of syntactic relationships, with the operating point for enrichment of both NCIT and RadLex. It is evident that the top ten most frequent relationships (Set 2) is the best combination of syntactic relationships for the Lin OL method using NCIT and pathology reports. With a threshold of 5.46, Lin method achieved 22% recall and 51% precision. In contrast, the combination of object, modifier, conjunctive, noun phrase, and determiner (set 8) is the best combination of syntactic relationships for the Lin OL method using RadLex and radiology reports. This combination achieved 25% recall and 13% precision at 7.0 thresholds.

Syntactic Relationship Set (description)	**NCIT**				**RadLex**			
	T	**Recall**	**Precision**	**F**	**T**	**Recall**	**Precision**	**F**
1 (all relationships)	**0.07**	**0.22**	**0.48**	**0.30**	**0.09**	**0.27**	**0.12**	**0.16**
2 (top ten frequent relationships)	***5.46***	***0.22***	***0.51***	***0.30***	**-2.63**	**0.05**	**0.03**	**0.08**
7(obj, mod, conj, nn)	**0**	**0.22**	**0.43**	**0.30**	**3.4**	**0.28**	**0.11**	**0.16**
8 (obj, mod, conj, nn, det)	**0.1**	**0.22**	**0.43**	**0.30**	***7.00***	***0.25***	***0.13***	***0.18***

Table 10. Precisions and recalls of Lin method with different syntactic relationships combination

5.2.4 Parameters generated for Church method

The same approach described above was followed for the investigation of window size through use of the Church method; however, only the final results are detailed below. Table 11 shows the different window sizes, with the operating point for enrichment of both NCIT and RadLex. From this table, it is evident that window sizes 3 and 5 generate the best evaluation results and are nearly identical. In this case, I selected window size 3 over window size 5 for computational efficiency.

Window	NCIT				RadLex			
Size	T	Recall	Precision	F	T	Recall	Precision	F
2	**15.74**	**0.36**	**0.54**	**0.22**	**16.38**	**0.61**	**0.04**	**0.08**
3		**0.46**	**0.48**	**0.23**		**0.64**	**0.04**	**0.08**
4		**0.42**	**0.49**	**0.23**		**0.53**	**0.04**	**0.08**
5		**0.46**	**0.48**	**0.23**		**0.64**	**0.04**	**0.08**
15		**0.52**	**0.41**	**0.23**		**0.52**	**0.04**	**0.08**

Table 11. Precisions and recalls of Church method using different window sizes

5.2.5 Summarization of the best parameters generated for both Lin and Church methods

Table 12 summarizes the best parameters selected for each method and domain, along with the recalls and precisions obtained with these parameters. For the Church OL method, window size 3 and threshold 14.2 are most effective for NCIT, and window size 3 and threshold 15.9 for

RadLex. For the Lin OL method, the best results are attained with a syntactic relationship set 8 and threshold 7 for NCIT domain and syntactic relationship set 2 and threshold 5.46 for a RadLex domain. Both methods performed better for NCIT than for RadLex. Overall, the Church OL method generated better recall than the Lin method (46% vs 22% for NCIT and 64% vs 25% for RadLex).

	Church OL Method					Lin OL Method				
	Parameter		Evaluation Metrics			Parameter		Evaluation Metrics		
Domain	WS	T	Recall	Precision	F	SR	T	Recall	Precision	F
Pathology	3	14.0	46%	48%	0.46	8 (obj, mod, conj, nn, det)	7	22%	51%	0.30
Radiology	3	15.9	64%	4%	0.08	2 (top 10 most frequent relationships)	5.46	25%	13%	0.18

Table 12. Best parameters for Church and Lin methods for each domain

5.2.6 New concept suggestion rate and acceptance rate

From the thresholds obtained during the development of the OL methods, I was able to extract a total of 2,249 suggestions for NCIT and a total of 1,300 for RadLex using Lin’s method. Among these, 49% of the 2,249 suggestions had not been found in the NCIT, and 87% of the 1,300 suggestions had not been found in the RadLex. Using Church’s method, I attained a total of 4,529 suggestions for NCIT and 12,174 for RadLex. 52% of the 4,529 had not been found in the NCIT and 96% of the 12,174 were not in the RadLex. These results are represented in Table 11 as Suggestion Rates (SRs). The Acceptance Rates (ARs) were based on the domain experts’ evaluations on 100 of sampled terms from all the suggestions. For NCIT, among the 100 sampled terms from Lin’s extractions, 38 had not been found in the ontology; domain expert determined that 28 of these 38 terms should be added, for an AR was 74%. Among the 100 sampled terms from Church’s extraction, 73 were not in the ontology; domain expert determined that 39 of the 73 terms should be added, for an AR of 53% (Table 13). For RadLex, among the 100 sampled from Lin’s extractions, 38 had not been found in the ontology; domain expert determined that 9 of the 38 terms should be added, for an AR of 9% (Table 13). I also manually examined the two lists of suggested terms for NCIT and two lists of suggested terms for RadLex by each method. I found there were no overlapping terms. These lists have been attached to this thesis as Appendix C.

	Pathology reports (Enrich NCIT)		Radiology reports (Enrich RADLex)	
OL Method	**SR**	**AR**	**SR**	**AR**
Lin Method	**49%(1,100/2,249)**	**28%(28/100)**	**87%(1,135/1,300)**	**9% (9/100)**
Church Method	**52%(2,159/4,529)**	**39%(39/100)**	**96%(11,743/12,174)**	**16% (16/100)**

Table 13. Suggestion and acceptance rates for the Lin and Church methods

6.0 DISCUSSION

The goal of this dissertation study was threefold. The first goal was to systematically review past research work on NLP methods and systems used for ontology learning and provide an overview of current advances and problems related to biomedical ontology learning. I sought to answer the following question: What could be the potential benefits of NLP methods for biomedical ontology learning? The second goal was to design a comparative study in which the existing NLP approaches were evaluated and compared for ontology enrichment in the biomedical domain. Finally, while comparing the different approaches for ontology learning, I explored and developed a methodology that can be utilized to extend and evaluate ontology learning. With this methodology and the metrics I developed for it, I was able to compare different learning methods for the biomedical domain. Comparing and analyzing these metrics provided useful insights for ontology learning from text. In the following paragraphs, I will discuss what has been learned regarding these goals.

6.1 THE EFFECTIVENESS OF THE SYMBOLIC-BASED APPROACH FOR ONTOLOGY ENRICHMENT USING CLINICAL DOCUMENTS

The LSP matching method is a symbolic method that has been studied by many researchers for domain knowledge extraction in the past; therefore, it was the first method I selected for my research. In this study, I evaluated the LSP matching method for ontological knowledge extraction using two types of clinical documents, pathology and radiology, that represent two sub-medical domains. My results indicated that the LSP matching method is an effective tool for semantic information extraction from clinical documents, with some limitations.

First, the LSP matching method can be expected to produce many suggestions for new concepts, instances, synonyms, and relationships. Several factors contribute to this expectation. Each instance of a pattern that appeared in the text resulted in the extraction of more than two Meaningful Medical Terms (MMTs) per sentence. Second, for both corpora tested, at least one quarter of the terms that could be extracted were not associated with corresponding concepts in the existing knowledge resource. With regard to acceptance, the results were mixed. For the pathology corpus, nearly all of these terms were accepted by the curator as useful concepts for the ontology. For the radiology corpus, however, less than one third of the suggested concepts were accepted by the curator as useful. In many cases, the scope and structure of the knowledge resource was the limiting factor in concept acceptance. Using this LSP approach, more than half of the relationships identified in the text corpora were not found to be present in either resource. However, curators rated these relationships for acceptance quite differently between NCIT and RadLex, with a much lower overall acceptance rate for NCIT. The low relationship acceptance rate for NCIT was mainly due to the fact that many relationships in the text were not within the scope of the ontology. For example, relationships between findings and disease are not defined in the NCIT, but these relationships are plentiful in the corpus. In future work, syntactic information derived from concepts and conceptual relationships in the ontology could be used to further constrain suggestions. Selecting candidate concepts based on the type of relationships modeled in the ontology might increase the acceptance rate by limiting suggestions that are clearly not modeled in the ontology.

The value of the LSP matching method also depends on how frequently these patterns occur in a domain corpus, given that these patterns are likely to extract more meaningful medical terms. Quantity is not the only measure. Our study showed that distributions of LSPs are heterogeneous. Some LSPs have higher frequencies than others; these proportions differ across the two corpora studied. Some of the patterns can be highly effective because a single specifically expressed instance of a relationship is all that is required for new semantic knowledge extraction. For example, even though the "NP_aka_NP" pattern in "Schwannoma (aka neurilemoma)" has a low frequency of occurrence, we can extract from a single instance a correct synonym relationship between schwannoma and neurilemoma. The frequencies of pattern occurrence in two different types of clinical documents varies; some of the patterns (e.g. "NP_aka_NP" and "NP_so called_NP" in radiology reports) either do not occur or occur at a very low frequency. Thus, the frequency of the patterns in the corpus does not guarantee a high suggestion rate. Based on suggestion rates, the top three patterns are "NP such as NP$_1$, NP$_2$,” “NP including NP$_1$, NP$_{2,}$” and “NP other NP$_1$, NP$_2$” for both corpora. However, these three patterns have relatively low frequencies in both corpora.

To determine the overall value of the LSP pattern set as a method of semantic extraction, I computed an estimate of the yield of concepts and relationships for each corpus. For a large corpus, the yield of concepts and relationships could be quite substantial. However, this method carries with it several limitations. In contrast to the findings in Hearst’s paper, there is little information in the LSP that accurately predicts the semantic relationship between concepts in the LSP, a finding borne out in both of the domains I studied. The distribution of relationships extracted with the LSP method was heterogeneous. Of the three named relationship types that curators evaluated (hyponym, meronym, and synonym), the most frequent relationship extracted with “NP_such as_NP,” “NP_including_NP,” or “NP_other_NP” patterns was hypernym/hyponym, and the most frequent relationship extracted using “NP_aka_NP’ was synonym. In some cases, there was no identifiable relationship between the Meaningful Medical Terms extracted. In many cases, the relationship was determined to be of some other type. Thus, the LSP cannot be used as an indicator of the type of relationship expressed between the entities. Because the LSP extracts terms in pairs, if one of the terms extracted by the LSP method is already in the ontology, a general position in the hierarchy for a concept based on the complementary term can be determined. The ability to assign one of the terms to the ontology based on the other term’s position can also become a very useful feature for a semi-automated ontology learning platform where a human curator is required to determine the type of relationship between two terms.

A second limitation of this method, as noted in previous research [71, 72, 75], is low recall. Many candidate concepts in the corpus never appear in any pattern. Attempts to improve the recall of the LSP method have focused on three major approaches. The first approach is to use additional syntactic features, such as noun coordination information, in combination with LSPs. For example, consider the following sentence containing a coordination structure: "In the ovine brain, GnRH neurons do not contain Type II glucocorticoid (GR), progesterone (PR), or α estrogen (ERα) receptors." If "ERα" is a steroid receptor in the ontology, the assumption that coordinated concepts are related permits defining "GR" and "PR" as steroid receptors as well. Caraballo [155] and Cederberg [156] used this approach to obtain additional related pairs of terms. The second approach is to use the machine learning method to acquire new patterns with either seed terms (Riloff [145] and Downey [157]) or seed patterns (Xu [158]), in an iterative bootstrapping process. A third approach combines pattern and co-occurrence information to learn new patterns. To illustrate this approach, Pantel et al. [159] used the minimal edit distance algorithm for pattern learning.

A final limitation of the LSP method is that it focuses on the use of simple English patterns, rather than on domain-specific patterns. The pattern learning approaches discussed above have been applied to specific corpora to learn domain-specific extraction patterns [145, 160]. Embarek and Ferret [144] discovered many medically related patterns using Pantel's algorithm, using them to discover semantic relationships in the medical domain. These patterns showed good results when evaluated via a medical corpus of the EQueR evaluation campaign for question-answering systems in French. Future enhancements that build on the work of Riloff [145], Snow [160], and other investigators [155, 156] could further reduce the limitations of the LSP method for ontology enrichment in biomedicine.

6.2 THE FFECTIVENESS OF THE STATISTICAL-BASED APPROACH FOR ONTOLOGY ENRICHMENT USING CLINICAL DOCUMENTS

The statistical algorithm, or the corpus-based approach for knowledge extraction, has been regarded as having the benefits of broad coverage, speed, and scalability, utilizing large corpora to generate statistics of features that capture the characteristics of a domain. Despite significant work in the area of OL, these methods have rarely been applied to biomedical ontologies [56, 88, 161]. One of the major barriers to progress in using these methods is the significant difficulty in evaluating OL methods [162]. Currently, there are no systematic evaluation methods or reference standards. Consequently, it is extremely hard to compare performance between algorithms, among domains, or for different knowledge resources. The absence of human reference standards for this task makes it particularly difficult to select parameters for statistical methods. In my dissertation study, I sought to address this problem by utilizing existing ontologies. In my experiments, I derived a set of reference standards from two existing ontologies: NCIT and RadLex. I limited the extraction task to new entities only, as the reference standards for these are the easiest to obtain. For other extraction tasks, such as relationship extraction, establishing reference standards is not easily accomplished without further research. Given the reference standards I was able to obtain, I tested, tuned, and directly compared, through recall, precision, and F-score, two statistical OL methods for ontology enrichment using clinical documents. I found that the Church mutual information method performed better than Lin’s similarity measure for both domains, with 46% vs. 22% recall for NCIT, and 64% vs. 25% recall for RadLex using these reference standards. Further examinations of the results of Lin’s method suggest that the lower performance may have been partly due to parsing errors. For Lin’s method to work, the texts have to be fully parsed so that all the dependency triples can be extracted. I selected the Minipar parser for this task, because it was created by the Lin method’s creator, Dekang Lin, and is available free of charge for noncommercial use. It is a broad-coverage parser for general English language and gives an output of dependency triples from inputted text documents. An evaluation with a general English corpus, SUSANNE Corpus, showed 88% precision and 80% recall with respect to dependency triples [163]. As there is no annotated medical domain corpus available, I was unable to determine the performance of the Minipar on medical corpora. Therefore, it is difficult to estimate how greatly the parsing errors may have contributed to the low performance of the new entity extraction task - a difficulty that will need to be addressed in future studies. In contrast, the Church method doesn’t require texts to be fully parsed, as only noun phrases are used for similarity measures. I was able, however, to use the medical domain corpus to train the POS tagger used in the Church method, and made the preliminary determination that the occurrence of preprocessing errors was greatly minimized.

6.3 ESTABLISHING AN EVALUATION FRAMEWORK AND METRICS THAT CAN BE USED FOR COMPARISON OF THE EFFECTIVENESS OF THE OL METHODS FOR BIOMEDICAL ONTOLOGY DEVELOPMENT

The establishment of an evaluation framework for OL methods is essential for fully realizing their potential for biomedical ontology development; as such, it serves as the most important contribution of my thesis to the field. The framework I established has helped me to evaluate and compare different OL methods across different domains.

The evaluation of ontology learning methods is a difficult task because traditional evaluation metrics, such as recall and precision, which are often used in Information Retrieval and Natural Language Processing, cannot easily be used in this context, because they aren't as clearly defined. Take Part-of-Speech (POS) tagging as an example: in this task, a set of POS tags is often predefined. The performance of the POS tagging algorithm can be evaluated based on how many POS tags can be identified among all the POS tags that exist in the corpus (recall), and, of these, how many have been correctly identified (precision). In ontology learning, one needs to acquire such ontological knowledge as the discovery of new entities and new relationships between entities. Often, there is no clear definition of what constitutes correct knowledge, nor is there a clearly predefined set of knowledge that needs to be acquired. Therefore, the evaluation framework and metrics need to be modified to reflect the more complicated characteristics of ontology learning.

In the symbolic method study, I described a two-step method for evaluating purely symbolic/syntactic OL for biomedical domains. In the first step, domain judges identified the meaningful medical terms on either side of the lexical pattern in text documents. In the second step, ontology developers evaluated the accuracy of the entity candidate and relationships. However, for OL methods that are either statistical or hybrid, the two-step human evaluation approach is inefficient. Given the wide range of possible parameters for any of these methods, the effort required for human evaluation could be enormous. I have therefore devised a variation of a previous approach that begins with automated parameter selection using the target ontology as a reference standard. This method would be considered as the first step of a two-step evaluation approach for OL method development. In the second step of the evaluation, the performance of each OL method can be determined through judgments by ontologists, based on metrics that include the Concept Suggestion Rate, Concept Acceptance Rate, Relationship Suggestion Rate, and Relationship Acceptance Rate for individual OL methods (all described previously in section 4.1.5).

The method for formative evaluation described here offers the following advantages: 1) it utilizes existing ontologies; 2) it allows for the automation of formative aspects of evaluation; 3) it facilitates a comparison of the performance of OL methods used for the same domain; and 4) it ensures that OL methods developed using this approach will be more likely to learn new entities that reside within the scope of the targeted ontology. Another potential benefit of using a reference standard for OL method evaluation is that it provides a systematic method of evaluation for multiple OL tasks, including learning of entities and taxonomic relationships. Although I limited our evaluation in this study to the entity enrichment by OL method, similar methods can be used to evaluate algorithms for other OL tasks.

I believe this methodology is sufficiently general and flexible enough to permit comparison of any OL method for a specific corpus and ontology of interest. However, I also see some limitations to this approach. First, a basic assumption is that the ontology or knowledge resource used to generate the reference standard is adequate and correctly represents the domain knowledge. But sometimes this is not the case. A second assumption is that the corpus used for ontology enrichment has some overlap with the target ontology. When overlap is absent, however, the size of the reference standard may not be adequate for OL method evaluation. For example, I was only able to identify a total of 660 RadLex entities in the radiology corpus. The poor overlapping between corpuses and targeting ontology is perhaps the most important contribution to the poor precision results for RadLex enrichment (4% for Church's method and 13% for Lin's method). Though, valid comparisons can be made across OL methods for a single corpus and ontology, but cannot be made across corpora and multiple ontologies.

As a second step, suggested entities that do not exist in the current ontologies have to be judged by domain ontologists to determine whether they can be added to an ontology; this evaluation must be performed by ontologists, as only they have the knowledge, experience and authority necessary to make these judgments. However, this process is tedious and time consuming. We decided to give the ontologists a randomly selected subset of suggestions from the output of OL methods for evaluation. The acceptance rate (AR), therefore, is defined in the section 4.1.5 to be an estimated value of the suggested terms generated by the OL methods. A comparison of the ARs of the methods revealed that both are quite good for NCIT enrichment (39% for Church method and 28% for Lin method). For Radlex, the AR is 16% for Church method and 9% for Lin method.

Using this framework, I found that the Church method engendered the best acceptance rate (39%) for NCIT compared to that produced by the two other methods (Lin's 28% and Hearst's 21%; Table 14).

OL method	Acceptance Rate	
	NCIT	RadLex
Hearst method	**21%**	**11%**
Church method	**39%**	**16%**
Lin method	**28%**	**9%**

Table 14. Comparison of acceptance rates of suggested terms extracted by different OL methods

Because the acceptance rate is the percentage of suggested terms that can be added to the target ontology, as determined by the ontologist, it is a good measure of the value of an OL method. In general, the higher the acceptance rate, the better the method. The preference is for a method whose suggested terms can be included in an ontology as many times as is possible.

However, the acceptance rate should not be used as the only indicator of the effectiveness of an OL method. Other factors, including the following, may be influencing the results. First, computational resources required for an OL method could influence how quickly an algorithm can run on a set of corpus. Some statistical methods require longer running times when the size of the corpus increases. In such cases, the Lin method is more computer-intensive than the Church method; the generation of new entity suggestions takes a lot longer to complete. The LSP matching method, however, uses a simple string-matching algorithm that renders it a more swift and efficient process.

Preprocessing is another factor that can influence the results of ontology learning. Because text corpora are learning resources, all OL methods will require some degree of text preprocessing. However, text preprocessing is less crucial to some methods than to others. For example, POS tagging is the preprocessing step for the LSP matching method. It has minimal effect on the LSP matching method because of its simple string-matching algorithm and the high accuracy of its POS tagging. Preprocessing for the Lin and Church methods is more complex, as it consists of sentence boundary detection, chunking, and named entity recognition. These preprocessing events have been packed into the NER system we developed. A shortcoming of my study is that the new NER system has not been formally evaluated. Manual examination of the output of the NER system reveals that some terms were not properly chunked and some terms were misrecognized as the named entities. During the development of statistical methods, the preprocessing errors will systematically present for all experiments. While these errors are acceptable for fine-tuning the methods, their presence is unacceptable in the final human evaluation.

7.0 LIMITATIONS

One of the limitations of this study is that the NER system used here was not formally evaluated. I have found that some of the terms are not properly chunked, and some of the terms in the clinical corpus have not been identified as concepts (i.e.: they are false negatives). These weaknesses have no impact on the Hearst method, as it doesn’t require the NER system for preprocessing (simple pattern matching). For the Church and Lin methods, however, the improper chunking can affect the statistical calculation for similarity scores. During the development of statistical methods, the preprocessing errors will systematically present for all experiments. While these errors are acceptable for fine-tuning the methods, their presence may not be acceptable in the final human evaluation. Thus, the extent of the effect of NER errors on ontology learning is hard to estimate.

One basic assumption for utilizing an existing ontology or knowledge resource as a reference standard is that the ontology is adequate and correctly represents the domain knowledge. Sometimes, however, such may not be the case. A second assumption is that the corpus used for ontology enrichment has some overlap with the target ontology. When this is not the case, the size of the reference standard may not prove adequate for OL method evaluation. I have found that I was only able to identify a total of 660 RadLex entities in the radiology corpus, which was, perhaps, a very important factor in the poor precision results achieved with the RadLex enrichment (4% for Church's method and 13% for Lin's method). Another limitation of this study is that I was only able to explore these OL methods in two medical domains, instead of in the three that I had originally planned to use, due to the difficulty of finding a sufficient number of domain experts. Both the Church and Lin methods performed poorly for the RadLex domain. I believe the inadequate quantity of reference standards available with the radiology domain may have contributed the lower performance of the statistical methods employed. However, as I have done only two domain studies, I was unable to draw definite conclusions when I compared the effectiveness of OL methods across the domains.

Although I believe that the evaluation framework can be used to evaluate algorithms that are aimed at other OL tasks, such as relationship extraction, I have limited my evaluation in this study to entity enrichment by OL methods, due to time constraints. Further study in this area is warranted.

8.0 CONCLUSION AND FUTURE WORK

Ontology development is a very challenging task because it requires grappling with many unresolved issues, such as the knowledge acquisition bottleneck, the difficulty of ontology learning evaluation, and the extent of an ontology’s scope and cycle. I have argued the notion that extending OL methods into the biomedical domain is a powerful approach for alleviating the knowledge acquisition bottleneck. To that end, I have evaluated three methods for ontology enrichment from two types of OL approaches (the symbolic and the statistical) in two medical domains (pathology and radiology). Overall, these methods proved very effective for new entity extraction from clinical corpora for both the pathology and radiology domains, with some limitations. I have found that the Hearst method, because of its simplicity, is the superior method: it can be applied easily to a wide range of domain corpora and doesn’t require a language or knowledge resource; thus, its implementation can be easy and swift. It does, however, suffer from the shortfall of low recall. Both the Church and Lin methods have the advantage of producing higher recall than the Hearst method, but their precisions are low. One important finding I can attest from my study is that the statistical method necessitates the use of high-quality preprocessing tools that require time and resources to develop. Further research on reducing preprocessing errors is greatly needed.

The lack of a formal evaluation method and reference standards is another major impediment to ontology development. I believe a systematic evaluation methodology is required before we can fully realize the potential of NLP methods for ontology learning. In this study, I have established a framework and metrics that have enabled me to evaluate and compare the performance of three NLP methods for ontology enrichment across two medical domains. The new entity suggestion rate and acceptance rate metrics that I created along with this framework allow for a subjective comparison of the performance of several OL methods as well. I believe this framework offers the following advantages: 1) it utilizes existing ontologies; 2) it allows for the automation of formative aspects of evaluation; 3) it facilitates a comparison of the performance of OL methods used for the same domain; and 4) it ensures that OL methods developed using this approach will be more likely to learn new entities that reside within the scope of the targeted ontology. Another potential benefit of using a reference standard for OL method evaluation is that it provides a systematic method of evaluation for multiple OL tasks including learning of entities and taxonomic relationships. The framework is flexible and carries with it the potential for other ontology developers to test and evaluate NLP methods for other domains.

The effectiveness of the use of these OL methods in the medical domains of my study has pointed to several directions for future research in this field. First among these is the testing and evaluation of alternative NLP methods for ontology learning. In this study, I have tested only one symbolic method and two statistical methods. As I described in section 2.0., there are many alternative methods with different approaches that have been well studied in the fields of NLP, A, IE, and IR. The comparison of the performance of these methods under the same framework and metrics would provide a beneficial contribution to the biomedical ontology development community.

Second, while many alternative methods should be tested, I believe the symbolic approach embodied in Hearst's pattern matching method has the potential for superior performance of new entity discovery, due to its simplicity and precision. This method can be easily and quickly implemented across different domains; therefore, it can be very beneficial for the ontology developer who does not have the time and resources for method development and implementation. I would like to continue to study this method and focus on the improvement of its performance by increasing its recall. In the past, many researchers have explored a variety of techniques for boosting performance, such as the utilization of additional noun coordination information [71, 72], bootstrapping in order to learn new patterns [64, 75] [164, 165], and using a hybrid of these two processes. The pattern-learning technique is particularly interesting, because it can assimilate patterns that are domain-specific.

Third, ontology relationship learning from text represents another important subject for research in the field of ontology learning. Because of time constraints, my study has mainly focused on new entity discovery. Although I have found that the Hearst method is the superior method for discovery of new entities that are related in some way, the relationships between the paired entities is not uniformly associated with one particular pattern. Further, the other two statistical methods give little support to relationship discovery. I would like to address these weaknesses by testing other OL methods, especially those that focus on relationship discovery, such as the association rule learning method.

Fourth, the development of evaluation methodology is crucial for ontology learning. I would like to further the study by contrasting the framework I developed here against other standard methods, such as data-driven ontology evaluation [166]. With objective metrics as evaluation measures, such study could provide a valuable assessment of the efficacy of this framework.

Finally, I would like to address an even larger research question about biomedical ontology learning: what is the most effective way to utilize ontology learning methods for biomedical ontology development? While automated ontology learning from texts using the NLP method is an effective approach for new entity and/or relationship discovery, semi-automatic ontology development platform is considered to be the more practical approach at this current stage. While OL methods can effectively generate many new entity candidates from text corpora, human judgment is required when deciding whether a suggested new entity should be incorporated to an ontology. Therefore, human-computer interaction is another compelling research area that holds the promise of making a positive contribution to the biomedical ontology development community. Research on ways to effectively present new entity candidates to domain experts is paramount in helping domain experts with the decision-making process.

APPENDIX A

PUBLICATION 1:

Natural Language Processing methods and systems for biomedical ontology learning.

Liu K, Hogan WR, Crowley RS. J Biomed Inform. 2011 Feb;44(1):163-79.

Journal of Biomedical Informatics 44 (2011) 163–179

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

Natural Language Processing methods and systems for biomedical ontology learning

Kaihong Liu[a], William R. Hogan[b], Rebecca S. Crowley[a,c,*]

[a] *Department of Biomedical Informatics, University of Pittsburgh School of Medicine, USA, USA*
[b] *Division of Biomedical Informatics, University of Arkansas for Medical Sciences, USA*
[c] *Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA*

ARTICLE INFO

Article history:
Received 21 April 2009
Available online 18 July 2010

Keywords:
Ontology
Ontology learning from text
Ontology enrichment
Information extraction
Natural Language Processing

ABSTRACT

While the biomedical informatics community widely acknowledges the utility of domain ontologies, there remain many barriers to their effective use. One important requirement of domain ontologies is that they must achieve a high degree of coverage of the domain concepts and concept relationships. However, the development of these ontologies is typically a manual, time-consuming, and often error-prone process. Limited resources result in missing concepts and relationships as well as difficulty in updating the ontology as knowledge changes. Methodologies developed in the fields of Natural Language Processing, information extraction, information retrieval and machine learning provide techniques for automating the enrichment of an ontology from free-text documents. In this article, we review existing methodologies and developed systems, and discuss how existing methods can benefit the development of biomedical ontologies.

Published by Elsevier Inc.

1. Background

1.1. Knowledge resources used in Natural Language Processing

Natural Language Processing (NLP) and text mining are research fields aimed at exploiting rich knowledge resources with the goal of understanding, extraction and retrieval from unstructured text. Knowledge resources that have been used for these purposes include the entire range of terminologies, including lexicons, controlled vocabularies, thesauri, and ontologies. For the purposes of this description we follow the framework for describing terminologies and terminological systems defined by de Keizer [1,2] and Cornet [3]. The authors define concepts as "cognitive constructs" of objects that are built using the "characteristics of the objects", terms as "language labels" for concepts, and synonyms as two or more terms that designate "a unique concept."

For simple NLP tasks, such as named entity recognition, almost any type of terminology can be used. Slightly more complex tasks such as identification of concepts, requires the representation of synonyms, and therefore limits the resources to terminological systems such as controlled vocabularies and ontologies that encode multiple lexical representations in natural language [4]. For example, "liver cell" and "hepatocyte" would be represented in the vocabulary or ontology as synonyms, and therefore during NER they would be classified as the same concept.

In contrast, some NLP tasks require more complex relationships between concepts, and therefore limit the types of terminological systems that may be used. Examples include word sense disambiguation [5], co-reference resolution [6–8], and discourse reasoning and extraction of attributes and values [9]. For example, if "hepatocellular carcinoma" and "liver neoplasm" are both used in a document to refer to the same entity, then these terms can be determined to co-refer if a relationship is represented in the terminology [10].

Ontologies can be used to make even more complex inferences and to derive rules necessary for semantic interpretation [11,12] and question and answering systems [13]. For this reason, ontologies have been of particular interest to researchers developing NLP systems. For example, to answer the question: "What role do infectious organisms play in liver cancer?" an ontology can be used to perform the query expansion and retrieve related textual information, if it contains the following information: (1) a synonym relationship between 'liver cancer' and 'hepatocellular carcinoma', (2) a hierarchical relationship between various hepatitis viruses and 'infectious organism', (3) an etiologic relationship between hepatitis viruses and hepatocellular carcinoma.

1.2. Ontologies and ontology development

Researchers define 'ontology' in different ways [14–17], but these definitions have in common that an ontology is a representation of entities and their relationships in a particular domain,

[*] Corresponding author. Address: Department of Biomedical Informatics, UPMC Cancer Pavilion, Suite 301, 5150 Centre Avenue, Pittsburgh, PA 15232, USA. Fax: +1 412 647 5380.

E-mail address: mailto: kaihong@pitt.edu (R.S. Crowley).

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2010.07.006

APPENDIX B

PUBLICATION 2:

Effectiveness of Lexico-syntactic Pattern Matching for Ontology Enrichment with Clinical Documents.

Liu K, Chapman WW, Savova G, Chute CG, Sioutos N, Crowley RS. Methods Inf Med. 2010 Nov 8;49(6).

Effectiveness of Lexico-syntactic Pattern Matching for Ontology Enrichment with Clinical Documents

K. Liu[1]; W. W. Chapman[1,2]; G. Savova[3]; C. G. Chute[3]; N. Sioutos[4]; R. S. Crowley[1,2,5]
[1]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA;
[2]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA;
[3]Department of Health Services Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA;
[4]Lockheed Martin Corporation, Fairfax, Virginia, USA;
[5]Department of Pathology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Keywords
Ontology learning from text, knowledge acquisition, ontology enrichment, natural language processing, lexico-syntactic pattern

Summary

Objective: To evaluate the effectiveness of a lexico-syntactic pattern (LSP) matching method for ontology enrichment using clinical documents.

Methods: Two domains were separately studied using the same methodology. We used radiology documents to enrich RadLex and pathology documents to enrich National Cancer Institute Thesaurus (NCIT). Several known LSPs were used for semantic knowledge extraction. We first retrieved all sentences that contained LSPs across two large clinical repositories, and examined the frequency of the LSPs. From this set, we randomly sampled LSP instances which were examined by human judges. We used a two-step method to determine the utility of these patterns for enrichment. In the first step, domain experts annotated medically meaningful terms (MMTs) from each sentence within the LSP. In the second step, RadLex and NCIT curators evaluated how many of these MMTs could be added to the resource. To quantify the utility of this LSP method, we defined two evaluation metrics: suggestion rate (SR) and acceptance rate (AR). We used these measures to estimate the yield of concepts and relationships, for each of the two domains.

Results: For NCIT, the concept SR was 24%, and the relationship SR was 65%. The concept AR was 21%, and the relationship AR was 14%. For RadLex, the concept SR was 37%, and the relationship SR was 55%. The concept AR was 11%, and the relationship AR was 44%.

Conclusion: The LSP matching method is an effective method for concept and concept relationship discovery in biomedical domains.

Correspondence to:
Kaihong Liu, MD, MS
Department of Biomedical Informatics
UPMC Cancer Pavilion, Suite 301
5150 Centre Avenue
Pittsburgh, PA 15232
USA
E-mail: kaihong@pitt.edu

Methods Inf Med 2010; 49: ■–■
doi: 10.3414/ME10-01-0020
received: March 3, 2010
accepted: October 6, 2010
prepublished: November 8, 2010

1. Introduction

The development of biomedical ontologies represents a key advance of biomedical informatics during the past two decades [1–3]. Biomedical ontologies provide the foundation for system interoperability such as HL7 on the Reference Information Model [4]; are important elements of decision support systems [5–7], support clinical information retrieval [8, 9]; and, are needed for natural language processing (NLP) tasks such as information extraction [10], anaphora resolution [11, 12], and question answering [13]. Despite the critical role of ontologies in biomedicine, there remain many barriers to their widespread use. One well-known problem, termed the "knowledge acquisition bottleneck", is the extraordinary manual effort that is required to create and maintain these resources [14–16]. The fields of knowledge acquisition, ontology learning, and ontology learning from text provide methods for automated and semi-automated ontology enrichment, which may help reduce the burden of populating ontologies.

Knowledge acquisition is a broad field that encompasses the tasks of acquiring and structuring knowledge from a wide range of resources, including experts. Semi-automated and fully automated methods for knowledge acquisition use data that can be derived from structured data sources (e.g. databases), semi-structured sources (e.g. web pages), or completely unstructured sources (e.g. free text). Knowledge acquisition methods can be used to populate many kinds of knowledge representations. Ontology learning represents a subfield of knowledge acquisition that is specifically interested in extraction of ontological concepts and relationships from knowledge-rich resources. Ontology learning from text defines a more specific task that focuses exclusively on extraction of ontological elements from unstructured sources.

There are two major advantages of ontology learning from text in biomedical domain. First, the biomedical literature is an important mechanism for reporting new discoveries in biomedical science. MEDLINE, the largest and most widely used biomedical literature repository, contains ap-

APPENDIX C

APPENDIX C1 TERMS EXTRACTED USING LIN METHOD AND PRESETNED TO NCIT ONTOLOGIST

ID	Term existed in the ontology	Proposed Term	Is this entity already present in the ontology?	If the entity is not present, should it be added to the ontology?
1	adjacent	biliary NCIT: Biliary(Code C2801)	Yes	
2	cavity	cysts NCIT: Cyst (Code C2978)	Yes	
3	cirrhotic	dense NCIT: Dense Connective Tissue (Code C32450)	Yes	
4	confirms	destruction NCIPT: Destruction (Code C62100)	Yes	
5	annular	distal NCIPT: Distal (Code C25237)	Yes	
6	cholestasis	dropout	No	No
7	crusted	dry NCIT: Dry Skin (Code C74592)	Yes	
8	chronicity	dyspoietic Myelodysplastic Syndrome (Code C3247)	Yes	
9	cellular	eosinophilic NCIT: Eosinophilic Infiltrate (Code C35981)	Yes	

10	cytological	epithelial NCIT: Epithelial (Code C13315)	Yes	
11	elongation	fibrosis NCIT: Fibrosis C30444	Yes	
12	five	fourteen	No	No
13	capsular	glandular NCIT: Glandular Cell C33923	Yes	
14	dermatopathic	granulomatous NCIT: Granulomatous C25302	Yes	
15	dry	gritty NCIT: Gritty Cut Surface C96185	No	
16	gallstones	hemorrhages NCIT: Hemorrhage C26791	Yes	
17	golden	opaque NCIT: Opaque C82125	Yes	
18	cavity	hilum NCIT: Hilum C73467	Yes	
19	eosinophilia	histiocytosis NCIT: Histiocytic Infiltrate C3106	Yes	
20	cholestasis	hyalin NCIT: Mallory Bodies Present C35381	Yes	
21	dysplasia	hyalinization NCIT: Hyalinization C96238	No	yes
22	glomerulosclerosis	hypertrophy NCIT: Glomerular Hypertrophy C96239	No	yes
23	blasts	hypogranularity NCIT: Hypogranularity C96264	No	yes
24	alveolar	intramedullary NCIT: Intramedullary C96266	No	yes
25	classic	invasive NCIT: Invasive C14159	Yes	

26	edematous	irregular NCIT: Irregular C63815	Yes	
27	bronchial	labial NCIT: Labial C96267	No	
28	inclusions	lesions NCIT: Lesion C3824	Yes	
29	circular	linear	No	No
30	iris	lipofuscin NCIT: Intracytoplasmic Lipofuscin Present C96268	No	yes
31	atelectatic	lobular NCIT: Lobular C25557	Yes	
32	adenomatous	lymphoid NCIT: Lymphoid C 13810	Yes	
33	inguinal	malar NCIT: Malar C96269	No	yes
34	desmin	margins NCIT: Margin C25563	Yes	
35	caval	mediastinal NCIT: Mediastinal C25310	Yes	
36	lesions	microabscesses NCIT: Microabscess C96272	No	yes
37	finely	mildly	No	No
38	bodies	mitoses NCIT: Mitosis C16864	Yes	
39	discolored	mottled NCIT: Mottled Skin C96273	No	yes
40	hyperplasia	neovascularization NCIT: Neovascularization C16900	Yes	
41	anisocytosis	nuclei NCIT: Nucleus C13197	Yes	
42	inclusions	organisms NCIT: Organism C14250	Yes	
43	bundles	osteoid NCIT: Osteoid C33228	Yes	
44	largest	ovoid	No	No

45	intramural	parenchymal NCIT: Parenchyma C74601	Yes	
46	cortical	patchy NCIT: Patchy C73945	Yes	
47	duodenum	pedicle NCIT: Pedicle of Vertebral Arch C96274	No	yes
48	aortic	pericolonic NCIT: Pericolic C25614	Yes	
49	mesenteric	perigastric NCIT: Perigastric C96275	No	yes
50	abundant	perimysial NCIT: Perimysial C96276	No	yes
51	anterior	peroneal NCIT: Peroneal C96277	No	yes
52	erosions	plasmacytosis NCIT: Plasmacytosis C96278	No	yes
53	discoloration	pleomorphism NCIT: Pleomorphism C17000	Yes	
54	firm	polypoidal NCIT: Polypoid C96279	No	yes
55	plasmacytoid	positive NCIT; Positive C25246	Yes	
56	intense	preponderantly	No	No
57	pmn	promyelocyte NCIT: Promyelocyte C13114	Yes	
58	edematous	red NCIT: Red C48326	Yes	
59	dense	redundant NCIT: Redundancy (Code C55286)	Yes	
60	dermoid	remnant NCIT: Remnant C96280	No	yes
61	lingular	retroperitoneal NCIT: Retroperitoneal C28256	Yes	

62	post	scar NCIT: Scar C34483	Yes	
63	forehead	segment NCIT: Segment (Code C45312)	Yes	
64	abundant	semitransparent NCIT: Semitransparent C96284	No	yes
65	obvious	serous NCIT: Serous C14168	Yes	
66	hyalinization	thrombus NCIT: Blood Clot C27083	Yes	
67	lumina	sinuses NCIT: Sinus C33556	Yes	
68	fragmented	smaller	No	No
69	ectasia	spongiosis NCIT: Spongiosis C96291	No	yes
70	crohn	sprue NCIT: Celiac Disease C26714	Yes	
71	retroperitoneal	stellate NCIT: Stellate C94437	Yes	
72	glomeruli	stones NCIT: Stone C35708	Yes	
73	bladder	stump NCIT: Stump C96294	No	yes
74	necrotizing	subacute	No	No
75	biopsies	submucosal NCIT: Submucosal C96296	No	yes
76	deposition	suture NCIT: Suture C50365	Yes	
77	cadherin	synaptophysin NCIT: Synaptophysin Staining Method C23029	Yes	
78	hemorrhagic	synthetic	No	No
79	suture	tag NCIT: Skin Tag C3374	Yes	
80	brown	tan NCIT: Tan C96298	No	yes

81	dupuytren	tenosynovium NCIT: Tendon Sheath C96299	No	yes
82	ligation	vein NCIT: Vein C12814	Yes	
83	apoptosis	thrombi NCIT: Blood Clot C27083	Yes	
84	leiomyomas	tissues NCIT: Tissue C12801	Yes	
85	membranous	translucent NCIT: Translucent C96300	No	yes
86	opaque	transparent NCIT: Transparent C94589	Yes	
87	cytokeratin	trichrome NCIT: Trichrome Staining Method C23012	Yes	
88	spongy	uniform NCIT: Uniform C73944	Yes	
89	immature	unremarkable NCIT: Unremarkable C96301	No	yes
90	occipital	upper NCIT: Upper C25355	Yes	
91	fibroelastosis	vacuolation NCIT: Cytoplasmic Vacuolation C96302	No	yes
92	identifiable	vague NCIT: Vague C96303	No	yes
93	ankle	valve NCIT: Cardiac Valve C12729	Yes	
94	hyperplastic	variegated NCIT: Variegated C96304	No	yes
95	abscesses	vasculitis NCIT: VasculitisC26912	Yes	
96	nodular	verrucous NCIT: Verrucous Lesion C5028	Yes	
97	lumen	vessel NCIT: Blood Vessel C12679	Yes	
98	cells	villi NCIT: Microvillus C33112	Yes	
99	pedunculated	wrinkled	No	No

100	grayish	yellow NCIT: Yellow C48330	Yes

APPENDIX C2 TERMS EXTRACTED USING CHURCH METHOD AND PRESETNED TO NCIT ONTOLOGIST

ID	Term existed in the ontology	Proposed Term	Is this entity already present in the ontology?	If the entity is not present, should it be added to the ontology?
1	a band-like inflammatory cell infiltrate	an occasional macrophage NCIT: Occasional Macrophages Present C96168	No	yes
2	an immunohistochemical profile	architectural disorder NCIT: Dysplastic Nevus C3694	Yes	
3	a minimal interstitial infiltrate	arteriolosclerosis NCIT: Arteriolosclerosis C35543	Yes	
4	adenocarcinoid	atypical carcinoid NCIT: Atypical Carcinoid Tumor C72074	Yes	
5	cholesterosis	chronic acalculous NCIT: Chronic Acalculous Cholecystitis C96169	No	yes
6	columnar mucosa showing intestinalization	columnar mucosa shows NCIT: Intestinal Metaplasia of Columnar Epithelium C96170	No	yes
7	bunionectomy	revision hammer toe left foot NCIT: Revision Hammer Toe Surgery C96172	No	yes
8	atypical features	darkly pigmented melanophages NCIT: Darkly Pigmented Melanophages Present C96173	no	yes
9	adventitia	deep adventitial inked margin NCIT: Deep Adventitial Inked Margin C96174	no	yes
10	a piece of soft tissue reddish	deep lobe NCIT: Parotid Gland Deep Lobe C96176	no	yes
11	deep edges	early neurotization NCIT: Nerve Regeneration C96177	no	yes
12	damaged and regenerating glands	early regeneration NCIT: Early Regeneration C96178	no	yes
13	a well differentiated hepatocellular neoplasm	effacement NCIT: Architectural Distortion	Yes	

		C82986		
14	diabetic nephropathy	efferent arteriolar hylinosis NCIT: Efferent Arteriolar Hyalinosis C96179	no	yes
15	atrophic squamous	endocervical glandular epithelium NCIT: Endocervical Glandular Epithelium C96180	no	yes
16	an inflammatory lamina propria	epithelial damage NCIT: Epithelial Gamage C 96181	no	yes
17	a superficial lymphocytic infiltrate	epithelial projections NCIT: Cilium C32318	Yes	
18	amorphous debris	fascicles NCIT: Fascicle C32586	Yes	
19	an igg	fibrillary glomerulonephropathy NCIT: Fibrillary Glomerulonephritis C96182	no	yes
20	eccentric thickening	fibroblast proliferation NCIT; Fibroblastic Proliferation Present C96183	No	yes
21	a moist	focal hemorrhage and ulceration NCIT: Localized Hemorrhagic and Ulcerated Lesion C96184	No	yes
22	any gallstones	focally greenish stained	No	No
23	focally hemorrhagic parenchyma	gritty sensation NCIT: Gritty Cut Surface C96185	No	yes
24	bilateral salpingo-oophorectomy	histerectomy and bilater salpingo-oophorectomy NCIT: Total Abdominal Hysterectomy with Bilateral Salpingo-Oophorectomy C51761	Yes	
25	beta-hcg	human chorionic gonadotropin NCIT: Human Chorionic Gonadotropin C2275	Yes	
26	cholesterol clefts	hyaline angiopathy NCIT: Hyaline Arteriolosclerosis C96186	No	yes
27	focal intraepithelial microvesicle formation	intramucosal neutrophils NCIT: Intramucosal Neutrophilic Infiltrate C96187	No	yes
28	a markedly distorted	intrathyroid parathyroid	No	yes

	total thyroidectomy	NCIT: Intrathyroidal Parathyroid C96188		
29	but not into mascularis propria	invades into muscularis mucosae NCIT: Invasion of Muscularis Mucosa Present C96189	No	yes
30	an mst stain	ischemic glomerulopathy NCIT: Ischemic Glomerulopathy C96190	No	yes
31	cmv negative	kappa immunoglobulin light chain non-contributory	No	No
32	focal stainable iron	mainly periportal	No	No
33	a hypercellular marrow	markedly decreased erythroids NCIT: Erythroid Series Cells Decreased C96203	No	yes
34	breast lobules	mastopexy NCIT: Mastopexy C96204	No	yes
35	immunoperoxidase	mature chromatin NCI PT: Heterochromatin C13241	Yes	
36	a nodular and diffuse infiltrate	medium-sized lymphoid cells NCI PT: Neoplastic Medium-Sized Lymphocyte C37004	Yes	
37	arthroscopic shavings right knee	meniscal cyst right knee NCIT: Meniscal Cyst C96205	No	yes
38	kossa stain	michalis NCIT: Michaelis-Gutmann Body C36016	Yes	
39	a red blood cells	mild-moderate anisopoikilocytosis NCIT: Mild to Moderate Anisopoikilocytosis C96207	No	yes
40	arteriolar hyalinosis	moderate arteriosclerosis NCIT: Arteriosclerosis C34398	Yes	
41	left lateral lobe	mostly soft pale tan tissue	No	No
42	mesosalpinx	multiple translucent grape	No	No
43	ganglionic tissue	negative for neoplasia NCIT: Negative for Neoplasia C96208	No	yes
44	cortical gliosis	neocortex NCIT: Cortical Cell Layer of the Cerebral Cortex C49136	Yes	
45	a femoral head specimen	eburnation or fibrillation	No	No
46	length metallic prosthesis	femoral shaft NCIT: Femoral Shaft C96209	No	yes
47	myopathic changes	neuropathic NCIT:	No	yes

		Neuropathic Pain C96210		
48	a stellate	serosal lesions NCIT: Serosal Lesion C96211	No	yes
49	a few elongated nuclei	nuclear palisading NCIT: Nuclear Palisading C49015	Yes	
50	a chondroid	osseous areas NCIT: Osseous Component Present C54171	Yes	
51	lens	other interior globe structures	No	No
52	appendiceal abscess	overall specimen dimensions	No	No
53	an interfollicular expansion	paler cells	No	No
54	abdominal apron	panniculectomy NCIT: Abdominal Panniculectomy C51597	Yes	
55	a warthin	papillary cystadenoma lymphomatosum Warthin Tumor C2854	Yes	
56	a few myofibers	perimysial and endomysial inflammatory cells NCIT: Perimysial and Endomysial Inflammatory Infiltrate C96212	No	yes
57	dyskeratotic squamous cells	perivascular lymphocytic infiltrate NCIT: Perivascular Lymphocytic Infiltrate C62777	Yes	
58	endoscopic excision	pituitary tissue NCIT: Pituitary Gland C12399	Yes	
59	pituitary adenoma	pitutary NCIT: Anterior Lobe of the Pituitary Gland C12772	Yes	
60	lined cystic structure	possible urachal remnant	No	No
61	an unoriented fibroadipose tissue	posterior renal space tumor	No	No
62	antibiotics	preparation NCIT: Preparation C25625	Yes	
63	a fibroepithelial polyp	procto NCIT: Anus C43362	Yes	
64	fewer eosinophils	prominent interface activity	No	No
65	mild mucosal atrophy	quiescent NCIT: Inactive C45422	Yes	
66	mastoid	radical tympanomastoidectomy NCIT: Radical Tympanomastoidectomy C96213	No	yes

67	fine needle aspirate biopsy	reactive lymphoid tissue showing crushed	No	No
68	re-excised tissue	right malar border stitch superior	No	No
69	a few small vessels	scattered interstitial eosinophils NCIT: Scattered Interstitial Eosinophils Present C96214	No	yes
70	sialolith	single amorphous structure identified	No	No
71	especially anterior dome	small irregular elevation	No	No
72	lul	stem bronchus NCIT: Main Bronchus C12284	Yes	
73	localized intraluminal neutrophils	stromal hemosiderin NCIT: Stromal Hemosiderin Deposition C96215	No	yes
74	a tan discoloration	subcutaneous mass NCIT: Subcutaneous Nodule C39618	Yes	
75	mild focal thickening	subepithelial collagen band NCIT: Subepithelial Collagen Band Present C96216	No	yes
76	focal severe atypia	subepithelial connective tissue melanophages	No	No
77	pseudohyphae	superficial and intermediate squamous cells	No	No
78	no fallopian tissue	surface membranous tissue and deeper smooth muscle tissue	No	No
79	a neuronal phenotype	synaptophysin NCIT: Synaptophysin Staining Method C23029	Yes	11/23/2009 m
80	a thickened area	tan prominent rugae	No	No
81	dorsal portion	the cingulate gyrus NCIT: Cingulate Gyrus C96217	No	yes
82	no definite sinus tracts	the dissected tibial and fibula	No	No
83	crushed cyst	the fibrotic band	No	No
84	the excised periprostatic tissue	the left posterior region	No	No
85	moderate amyloid angiopathy	the leptomeningeal and parenchymal blood vessels	No	No
86	bifurcated vascular tissue	the long side measures	No	No
87	mild accentuation	the pericentral sinusoidal fibrous tissue	No	No
88	fibula	the skin eschar NCIT: Skin	No	yes

		Eschar C96218		
89	the subvalvular tissue	the supravalvular wall	No	No
90	periduodenal adipose tissue	the surrounding pancreas	No	No
91	the ongoing lymphoplasmacytic interface activity	the thin fibrous bridges	No	No
92	binucleate	the total population	No	No
93	decidualized endometrium	tissue passed per vagina	No	No
94	labial mucosa	tobacco induced hyperparakeratosis NCIT: Tobacco Induced Hyperparakeratosis C96220	No	yes
95	glandularis	urethritis cystica NCIT: Urethritis Cystica C96225	No	yes
96	mixed composition	viscid yellow-green bile	No	No
97	some adjacent haversian	volkmann canals NCIT: Perforating Canal C33293	Yes	
98	pneumocytes	widespread alveolar pneumocyte damage NCIT; Widespread Alveolar Pneumocyte Damage Present C96237	No	yes
99	a remnant	wrinkled and interrupted fragments lens capsule	No	No
100	a well circumscribed tan	yellow mass	No	No

APPENDIX C3 TERMS EXTRACTED USING LIN METHOD AND PRESETNED TO RADLEX ONTOLOGIST

ID	**Term existed in the ontology**	**Proposed Term**	**Is this entity already present in the ontology?**	**If the entity is not present, should it be added to the ontology?**
1	acetabular	anterior	Yes	
2	artery	bifurcation	Yes	
3	amount	blood	Yes	
4	any joint without contrast left	ct upper extremity without contrast left	No	No
5	biopsy	cytology	Yes	
6	circumferential	diffuse	Yes	
7	canal	disc	Yes	
8	colon	diverticulum	Yes	
9	bronchoscopy	drainage	No	No
10	aspect	duodenum	No	No
11	cholecystitis	embolism	Yes	
12	colon	esophageal	Yes	
13	correlation	evaluation	No	No
14	activity	focus	No	No
15	cavernous	frontal	Yes	
16	flexure	fundus	No	Yes
17	adenoma	gallstone	Yes	
18	catheter	ganz	Yes	
19	calcaneocuboid	glenohumeral	No	No
20	cystic	globular	No	No
21	fracture	hemorrhage	Yes	
22	adrenal	hepatic	Yes	
23	esophagus	ileum	Yes	
24	hypervascular	intraluminal	Yes	
25	fracture	ischemia	Yes	
26	embolism	leak	Yes	
27	gallstones	lesions	Yes	
28	duct	loop	Yes	
29	active	malignant	No	No
30	clip	marker	No	Yes
31	area	mass	No	No
32	hilum	mediastinum	Yes	
33	mediastinum	mesentery	Yes	
34	base	middle	Yes	
35	horizontal	minor	Yes	

36	angiogram	mra	Yes	
37	duodenum	muscles	Yes	
38	mediastinum	musculature	Yes	
39	infiltration	necrosis	Yes	
40	gallstones	nodules	Yes	
41	inflamed	normal	Yes	
42	normal	oblique	Yes	
43	intraparenchymal	occipital	Yes	
44	activity	opacity	No	Yes
45	area	osteophyte	No	No
46	occipital	paratracheal	No	No
47	flank	paravertebral	Yes	
48	hemisphere	pole	Yes	
49	ganglia	pontine	No	Yes
50	location	position	No	No
51	perfusion	postcontrast	Yes	
52	patellar	posterior	Yes	
53	fixation	process	Yes	No
54	infiltration	prominence	Yes	
55	pressure	question	Yes	
56	mri	radiographs	Yes	
57	cholecystectomy	reconstruction	Yes	
58	dilatation	recurrence	Yes	
59	chronic	respiratory	Yes	
60	right breast sonogram	right breast ultrasound	No	No
61	left wrist	right shoulder	No	No
62	disease	sclerosis	No	Yes
63	disease	scoliosis	No	Yes
64	collection	segment	No	No
65	saturation	sequences	No	No
66	gallstones	sludge	Yes	
67	soft	solid	No	No
68	diffusion	spgr	Yes	
69	neck	spine	Yes	
70	progression	stenosis	No	No
71	paravertebral	subareolar	No	No
72	Body	subchondral	Yes	
73	occipital	suboccipital	No	No
74	aortocaval	subpleural	No	No
75	segment	subsegmental	No	Yes
76	subcutaneous	superficial	Yes	
77	presence	surgery	Yes	
78	apex	tail	Yes	
79	progression	tear	No	No

80	cartilage	tendons	Yes	
81	the chest	the intracranial circulation	Yes	
82	the left adrenal gland	the right hepatic lobe	No	No
83	the left lower lung	the right paratracheal region	No	No
84	the left adrenal gland	the satellite nodule	No	No
85	enhancement	thinning	Yes	
86	pancreatic	thyroid	Yes	
87	glenohumeral	tibiotalar	No	No
88	disease	tortuosity	No	Yes
89	malignancy	tumor	Yes	
90	guidance	ultrasound guidance	No	No
91	flow	uptake	Yes	
92	bronchus	ureter	No	Yes
93	contours	vasculature	Yes	
94	cortex	ventricle	Yes	
95	granulomatous	vessel	Yes	
96	subcentimeter	water	Yes	
97	noncontrasted	weighted	No	No
98	sensitive	weighted	No	No
99	pancreatectomy	whipple	Yes	
100	valve	wires	Yes	

APPENDIX C4 TERMS EXTRACTED USING CHURCH METHOD AND PRESETNED TO RADLEX ONTOLOGIST

ID	**Term existed in the ontology**	**Proposed Term**	**Is this entity already present in the ontology?**	**If the entity is not present, should it be added to the ontology?**
1	lung attenuation	vascular abnormalities	No	no
2	a left pelvic extraperitoneal infiltration	adjacent hematoma	No	no
3	an anatomic variant	aneurysm	No	no
4	a rectal stump	barrel right lower quadrant ostomy	No	no
5	abdomenfollowing oral barium	biphasic abdominal	No	no
6	a nasal tube	cervical esophagus	No	yes
7	a proximal wire	chest pacer	No	no
8	a side branch ipmn	cystic pancreatic neoplasms	No	yes
9	bilateral iliac arteries	diffuse atheromatous calcification	No	yes
10	diffuse osteopenia withmultilevel	facial ct scanning	No	no
11	left vocal cord	fdg activity	No	no
12	a contrast-enhanced ct scan	femoral canal	No	yes
13	benign enhancement	fibrocystic change	No	no
14	evaluate due	film technique	No	no
15	a normal unenhanced appearance	indicate c-diff colitis	No	no
16	a y view	internal rotation grashey view	No	no
17	fat herniates	intrathoracic lymphadenopathy	No	no
18	a complex perianal abscess	ischioanal fossa	No	no
19	compromise evaluation	kidneys secondary	No	no
20	blends	lacrimal gland	No	no
21	expected postoperative subdural and intraventricular	left occipito	No	no
22	a completely drowned	lower lobe bronchi	No	no

	right lower lobe and right			
23	a peripheral nodular	lower lobe parenchyma	No	yes
24	evaluate lung expansion	meconium aspiration	No	no
25	hyperintense signal	median nerve	No	yes
26	medical attention	medication restrictions	No	no
27	a representative node	mesenteric adenitis	No	no
28	intraluminal fluidpredominantly	middle andbasilar segments	No	no
29	bilateral ligamentum flavum hypertrophy	mild bilateral neural foramen narrowing	No	yes
30	contrast bilateral	mr breast	No	no
31	contrast left tibia	mri lower extremity	No	no
32	fluid filled small bowel	multiple dilated loops	No	yes
33	a diffuse bone marrow edema	naviculocuneiform joint	No	no
34	ct criteria	discrete cervical nodes	No	no
35	a bladder calculus	pelvic ureteral	Yes	
36	left inguinal hernia containing loops	nondilated bowel	No	no
37	a ventral abdominal hernia	nonobstructed transverse colon	Yes	
38	a thyroid ultrasound	nontoxic multinodular goiter	No	yes
39	a widely patent graft	normal amplitude	No	no
40	biceps groove	normal infraspinatus	Yes	
41	acuteischemia	normal mr angiography	No	no
42	fungal	organisms such as nocardia	No	yes
43	sonographic appearance	pap smear	No	no
44	both sca	pca vessels	No	no
45	prior pelvic ct	pelvis radiograph	No	no
46	post hysterectomy	periaortic and pelvic lymph	No	no
47	mesenous appendiceal carcinoma	peritoneal disease	No	no
48	intracranial mr angiography	phase contrast technique	No	yes

49	possible submucosal hemorrhage	post fundoplication	No	no
50	one right superior pulmonary vein	postobstructive consolidation	No	yes
51	mri findings	preoperative needle localization	No	no
52	embolism protocol	pulmonary arterial enhancement	No	no
53	minimal left basilar	residual pleural fluid	No	no
54	a subpleural	reticular opacities	Yes	
55	a myelomatous deposit	right hypoglossal canal	No	no
56	multiple other stones	right lower quadrant ileal conduit	No	no
57	choledocholithiasis withbiliary sludge	right upper quadrant ultrasound	No	no
58	residual disease	right-sided retropharyngeal metastatic node	No	yes
59	a nonobstructed appearance	scattered bowel gas	No	no
60	an enlarged and tortuous left gonadal vein	several parametrial collateral vessels	No	no
61	overt cirrhotic morphology	severe diffuse fatty infiltration	No	no
62	parietal scalp	similar soft tissue	No	no
63	bowel loops show normal caliber	sma and celiac trunk	No	no
64	anatomic impingement	small subacromial subdeltoid bursitis	No	no
65	right upper quadrant sonography	speckled doppler analysis	No	no
66	exam	spine mr	Yes	yes
67	an unexpected occurrence	splenic artery coil embolization	No	no
68	residual mild pancreatic tail ductal dilatation	splenic vein patent	No	no
69	post-radiation change	stable low density infiltration	No	yes
70	mitral valve repair	status post tricuspid	No	no
71	mri findings	subacromial impingement secondary	No	no

72	occasional honeycombing	subpleural fibrosis	No	no
73	glenoid attachment	superior glenohumeral ligaments	No	no
74	a para	suprascapular ganglion	No	no
75	caliber of aneurysmal	the aortobifemoral endovascular stent	No	no
76	rectus	the deep fascia	Yes	
77	fat density focus	the gastric antral lumen	No	no
78	mesenteric and omental infiltration	the greater sac	No	no
79	shaped vertebral bodies	the humeral heads and h	No	no
80	hepatofugal flow	the intrahepatic portal	No	no
81	post procedural	the left hemithorax	No	no
82	predominantly air-filled	the left-sided ducts	No	no
83	fluid distention	the peroneal longus tendon sheath	No	no
84	advance narrowing	the radial carpal interface	No	no
85	benign thyroid disease	the thyroid glandis	No	yes
86	irregular sigmoid colon	the tract caudally	No	no
87	oblong soft tissue nodular	the upper anterior mediastinum	No	no
88	an associated soft tissue mass	the upper sternal sclerotic	No	no
89	malignancy elsewhere	thoracic nodal metastases	No	no
90	the common iliac vein compatible	trace fluid	No	no
91	cirrhotic liver morphology	trace upper abdominal ascites	No	no
92	recent anoxic injury	transependymal flow	No	no
93	a biliary obstruction	transhepatic biliary catheter	No	yes
94	a cyst with placement	ultrasound guided aspiration	No	no
95	parietal and right frontal lobes	underlying small vessel	No	no
96	intrinsic bone lesions	left elbow	No	no

		radiographs		
97	a subtle occult fracture	unremarkable portable pelvis radiograph	No	no
98	high attenuation material	ureteral calculi or neoplasm	No	no
99	colon likely secondary	wall edema	No	no
100	marked loss of gray	white matter differentiation	No	no

BIBLIOGRAPHY

1. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform. 2006 Sep;7(3):256-74.

2. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2008 Jan;36(Database issue):D344-50.

3. Gruber TR. A translation approach to potable ontology specifications. Knowledge Acquisition. 1993;5:199-220.

4. Smith B, Kusnierczyk W, Schober D, Ceusters W, Towards a reference terminology for ontology research and development in the biomedical domain. The Second International Workshop on Formal Biomedical Knowledge Representation: "Biomedical Ontology in Action" (KR-MED 2006) 2006.

5. Sowa JF. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Pacific Grove, CA: Brooks Cole Publishing Co.; 1999.

6. Cimino JJ. In defense of the Desiderata. J Biomed Inform. 2006 Jun;39(3):299-306.

7. Smith B. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. J Biomed Inform. 2006 Jun;39(3):288-98.

8. de Keizer NF, Abu-Hanna A. Understanding terminological systems II: terminology and typology. Methods Inf Med. 2000;39:22-9.

9. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonl JHM. Understanding terminological systems I: terminology and typology. Methods Inf Med. 2000;39:16-21.

10. Cornet R, De Keizer NF, Abu-Hanna A. A framework for characterizing terminological systems. Methods Inf Med. 2006;45:253-66.

11. William WC, Sunita S. Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. Proceedings of the 10th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining; Seattle, WA, USA: ACM. 2004.

12. Navigli R, Velardi P. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2005;27(2):1075-86.

13. Poesio M, Vieira R, Teufel S, Resolving bridging references in unrestricted text. Proceedings of the ACL Workshop on Operational Factors in Robust Anaphora Resolution 1997. pp 1-6.

14. Soon WM, Ng HT, Lim DCY. A machine learning approach to coreference resolution of noun phrases. Comput Linguist. 2001;27(4):521-44.

15. Ng V, Cardie C. Improving machine learning approaches to coreference resolution. Proceedings of the 40th Annual Meeting of the ACL; Philadelphia, Pennsylvania: ACL. 2001.

16. Friedman C, Borlawsky T, Shagina L, Xing HR, Lussier YA. Bio-Ontology and text: bridging the modeling gap. Bioinformatics. 2006 July 26, 2006;22(19):2421-9.

17. Liang T, Lin Y-H, Anaphora resolution for biomedical literature by exploiting multiple resources. Second International Joint Conference on Natural Language Processing 2005.

18. Gomez F. An algorithm for aspects of semantic interpretation using an enhanced WordNet. Second meeting of the North American Chapter of the ACL on Language Technologies Pittsburgh, Pennsylvania: ACL. 2001.

19. Gomez-Perez A, Manzano-Macho D. An overview of method and tools for ontology learning from texts. The Knowledge Engineering Review. 2005;19(3):187-212.

20. Girju R, Moldovan DI, Knowledge acquisition for question answering. Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference 2001: AAAI Press.

21. National Institutes of Health. Research Portfolio Online Reporting Tools (RePORT). 2010 [cited 2010 May 25]; Available from: http://projectreporter.nih.gov/project_info_history.cfm?aid=7941562&icde=2611544.

22. BioInform. Stanford's Mark Musen on the New National Center for Biomedical Ontology. 2005 [cited 2010 May 25]; Available from: http://www.genomeweb.com/informatics/stanford-s-mark-musen-new-national-center-biomedical-ontology.

23. Du L. DUMC gets $1.25M for ontology. The Chronicle [serial on the Internet]. 2009 [cited 2010 May 25]: Available from: http://dukechronicle.com/node/148034.

24. National Science Foundation. The Hymenoptera Ontology: Part of a Transformation in Systematic and Genome Science. 2009 [cited 2010 May 25]; Available from: http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0850223.

25. United States National Library of Medicine. FAQs: SNOMED CT® in the UMLS®. 2003 [updated 2008-12-15; cited 2010 May 25]; Available from: http://www.nlm.nih.gov/research/umls/Snomed/snomed_faq.html.

26. United States National Library of Medicine. SNOMED Clinical Terms® To Be Added To UMLS® Metathesaurus®. 2003 [updated 2006-05-24; cited 2010 May 25]; Available from: http://www.nlm.nih.gov/research/umls/Snomed/snomed_announcement.html.

27. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W170-3.

28. Tudorache T, Noy NF, Tu SW, Musen MA. Supporting collaborative ontology development in Protege. Seventh International Semantic Web Conference; Karlsruhe, Germany. 2008.

29. Campbell KE, Cohn SP, Chute CG, Shortliffe EH, Rennels G. Scalable methodologies for distributed development of logic-based convergent medical terminology. Methods Inf Med. 1998 Nov;37(4-5):426-39.

30. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251-5.

31. Payne PRO, Mendonça EA, Johnson SB, Starren JB. Conceptual knowledge acquisition in biomedicine: A methodological review. Journal of Biomedical Informatics. 2007;40(5):582-602.

32. Bruce GB, David CW, editors. Readings in knowledge acquisition and learning: automating the construction and improvement of expert systems: Morgan Kaufmann Publishers Inc.; 1993.

33. Buitelaar P, Cimiano P, Magnini B. Ontology learning from text: method, evaluation and applications. Breuker J, Dieng R, Guarino N, Mantaras RLd, Mizoguchi R, Musen M, editors. Amsterdam, Berlin, Oxford, Tokyo, Washington DC: IOS Press; 2005.

34. Navigli R, Velardi P, Gangemi A. Ontology learning and its application to automated terminology translation. IEEE Intelligent Systems. 2003;18(1):22-31.

35. Fensel D, Studer R. Knowledge acquisition, modeling and management. Fensel D, Studer R, editors: Springer 2008.

36. Shadbolt N, O'hara K, Schreiber G. Advances in knowledge acquisition: Springer; 2008.

37. Hoffmann A. Advances in knowledge acquisition and management: Pacific Rim knowledge acquisition workshop. Guilin, China: Springer 2006.

38. Hearst MA, Automatic acquisition of hyponyms from large text corpora. Proceedings of ACL 1992.

39. Harris ZS. Mathematical structures of language. New York, NY, USA: Krieger Pub Co; 1968.

40. Firth JR. Papers in linguistics. London: Oxford University Press; 1934 -1957.

41. Gamallo P, Agustini A, Lopes G, Selection restrictions acquisition from corpora. Proceedings EPIA, Springer 2001.

42. Grefenstette G. Explorations in automatic thesaurus discovery. Boston, MA: Kluwer Academic Publisher; 1994.

43. Church KW, Hanks P, Word association norms, mutual information, and lexicography. Proceedings of 27th Annual Meeting of the ACL 1989. pp 76-83.

44. Smadja F. Retrieving Collocations from Text: Xtract. Comput Linguist. 1993;19(1):143-77.

45. Caraballo S, Charniak E, Determining the specificity of nouns from text. Proceedings of SIGDAT 1999.

46. Lin D, Automatic retrieval and clustering of similar words. Proceedings of COLING 1998.

47. Hindle D, Noun classification from predicate-argument structures. Proceedings of 28th ACL 1990. pp 268-75.

48. Reinberger M-L, Spyns P, Unsupervised text mining for the learning of DOGMA-inspired ontologies. Proceedings of ECAI and EKAW 2004.

49. Agirre E, Ansa O, Hovy E, Martínez D, Enriching very large ontologies using the WWW. Proceedings of the Ontology Learning Workshop, ECAI, Berlin, Germany 2000.

50. Gulla JA, Brasethvik T, Kvarv GS. Association rules and cosine similarities in ontology relationship learning. Enterprise Information Systems: Springer Berlin Heidelberg; 2009. p. 201-12.

51. Cherfi H, Toussaint Y, How far association rules and statistical indices help structure terminology? Proceedings of the15th ECAI: Workshop on Machine Learning and Natural Language Processing for Ontology Engineering, Lyon, France 2002.

52. Bodenreider O, Aubry M, Burgun A. Non-lexical approaches to identifying associative relations in the gene ontology. Pac Symp Biocomput. 2005:91-102.

53. Collier N, Nobata C, Tsujii J, Extracting the names of genes and gene products with a Hidden Markov Model. Proceedings of COLING, Sarrebruck 2000.

54. Morgan A, Hirschman L, Yeh A, Colosimo M, Gene name extraction using FlyBase resources. Proceedings of the ACL workshop on Natural language processing in biomedicine 2003. pp 1-8.

55. Shen D, Zhang J, Zhou G, Su J, Tan CL, Effective adaptation of Hidden Markov model-based named entity recognizer for biomedical domain. Proceedings of the ACL Workshop on Natural Language Processing in Biomedicine 2003. pp 49-56.

56. Kazamay Ji, Makinoz T, Ohta Y, Tsujiiy Ji, Tuning support vector machines for biomedical named entity recognition. Proceedings of the ACL workshop on Natural Language Processing in Biomedicine 2003. pp 1-8.

57. Yamamoto K, Kudo T, Konagaya A, Matsumoto Y, Protein name tagging for biomedical annotation in text. Proceedings of the ACL workshop on Natural Language Processing in Biomedicine 2003. pp 65-72.

58. Alfonseca E, Manandhar S, Extending a lexical ontology by a combination of distributional semantics signatures. Proceedings of EKAW 2002. pp 1-7.

59. Alfonseca E, Manandhar S, An unsupervised method for general named entity recognition and automated concept discovery. Proceedings of the 1st International Conference on General WordNet 2002.

60. Hasting PM. Automatic acquisition of word meaning from context [Ph.D Thesis]: Univ. of Michigan; 1994.

61. Hahn U, Schnattinger K, Towards text knowledge engineering. Proceedings of AAAI98, IAAI98 1998. pp 524-31.

62. Basili R, Pazienza MT, Velardi P. An empirical symbolic approach to natural language processing. Journal of Artificial Intelligence. 1996;85:59-99.

63. Hamon T, Nazarenko A. Detection of synonymy links between terms: experiment and results. In: Bourigault D, Jacquemin C, L'Homme M-C, editors. Recent Advances in Computational Terminology2001. p. 185-208.

64. Downey D, Etzioni O, Soderland S, Weld DS, Learning text patterns for Web information extraction and assessment. Proceedings of the AAAI workshop on Adaptive Text Extraction and Mining 2004.

65. Moldovan DI, Girju R. An Interactive tool for the rapid development of knowledge bases. International Journal on Artificial Intelligence Tools. 1999.

66. Grefenstette G, Automatic thesaurus generation from raw text using knowledge-poor techniques. Ninth Annual Conference of the UW Centre for the New OED and text Research - Making Sense of Words 1993.

67. Geffet M, Dagan I, The distributional inclusion hypotheses and lexical entailment. Proceedings of the 43rd Annual Meeting of the ACL 2005. pp 107-14.

68. Faatz A, Steinmetz R. Ontology enrichment with texts from the WWW. Semantic Web Mining Workshop, Helsinki, Finland. 2002.

69. Bikel D, Miller S, Schwartz L, Wesichedel R, Nymble: a high-performance learning name-finder. Proceedings of the Fifth Conference on Applied Natural Language Processing 1997. pp 194-201.

70. Chanlekha H, Collier N. A methodology to enhance spatial understanding of disease outbreak events reported in news articles. International Journal of Medical Informatics. 2010;79(4):284-96.

71. Caraballo S, Automatic construction of a hypernym-labeled noun hierarchy from text. Proceedings of the 37th Annual Meeting of the ACL 1999.

72. Cederberg S, Widdows D, Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. Proceedings of CoNLL 2003. pp 111-8.

73. Fiszman M, Rindflesch TC, Kilicoglu H, Integrating a hypernymic proposition interpreter into a semantic processor for biomedical texts. Proceedings of AMIA Annual Symp 2003. pp 239-43.

74. Snow R, Jurafsky D, Ng AY. Learning syntactic patterns for automatic hypernym discovery. Advances in Neural Information Processing Systems 2004.

75. Riloff E, Automatically generating extraction patterns from untagged text. Proceedings of the Thirteenth National Conference on Artificial Intelligence 1996.

76. Velardi P, Navigli R, Cucchiarelli A, Neri F, Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. Proceedings of ECAI and EKAW 2004.

77. Cimiano P, Pivk A, Schmidt-Thieme L, Stabb S, Learning taxonomic relations from heterogenerous sources of evidence. Proceeding of EKAW 2004.

78. Rinaldi F, Yuste E, Schneider G, Hess M, Roussel D, Exploiting technical terminology for knowledge management. Proceedings of ECAI and EKAW 2004.

79. Morin E, Jacquemin C. Automatic acquisition and expansion of hypernym links. Computer and the Humanities. 2004;38(4):343-62.

80. Bodenreider O, Rindflesch TC, Burgun A, Unsupervised, corpus-based method for extending a biomedical terminology. Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain 2002. pp 53-60.

81. Ryu P-M, Choi K-S, Measuring the specificity of terms for automatic hierarchy construction. Proceedings of the ACL-SIGLX Workshop on Deep Lexical Acquisition June, 2005. Ann Arbor, Michigan.

82. Witschel HF, Using decision trees and text mining techniques for extending taxonomies. Proceedings of Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at ICML 2005.

83. Berland M, Charniak E, Finding parts in very large corpora. Proceedings of the 37th Annual Meetings of the ACL 1999. pp 57-64.

84. Sundblad H, Automatic acquisition of hyponyms and meronyms from question corpora. Proceedings of the 15th ECAI; Lyon, France 2002.

85. Girju R, Badulescu A, Moldovan D, Learning semantic constraints for the automatic discovery of part-whole relations. Proceedings of HLT 2003.

86. Nenadć G, Spasić I, Ananiadou S. Automatic discovery of term similarities using pattern mining. COLING on COMPUTERM: 2nd International Workshop on Computational Terminology ACL. 2002. p. 1-7.

87. Kavalec M, Svatek V. A study on automated relation labeling in ontology learning. In: Buitelaar P, Cimiano P, Magnini B, editors. Ontology Learning from Text: Method, Evaluation and Applications. Amsterdam, Berlin, Oxford, Tokyo, Washington DC: IOS Press; 2005. p. 44-58.

88. Blaschke C, Valencia A. Automatic ontology construction from the literature. Genome Inform. 2002;13:201-13.

89. Turney PD, Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the Twelfth European Conference on Machine Learning 2001.

90. Harris ZS. A grammar of English on mathematical principles. New York: Wiley; 1982.

91. Harris ZS. A theory of language and information: a mathematical approach. Oxford: Clarendon Press; 1991.

92. Friedman C, Alderson PO, Austin J, Cimino JJ, Johnson SB. General natural language text processor for clinical radiology. JAMIA. 1994;1(2):161-74.

93. Sager N, Lyman M, Buchnall C, Nhan NT, Tick LJ. Natural language processing and representation of clinical data. JAMIA. 1994;1(2):142-60.

94. GENIA. GENIA. Available from: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi.

95. Lafferty J, McCallum A, Pereira F, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of ICML 2001. pp 282-9.

96. Riloff E, Shepherd J, A corpus-based approach for building semantic lexicons. Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP) 1997.

97. Roark B, Charniak E, Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. Proceedings of ACL 1998. pp 1110-6.

98. Widdows D, Dorow B, A graph model for unsupervised lexical acquisition. 19th International Conference on Computational Linguistics 2002. pp 1093-9.

99. Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990;41(6):391-407.

100. Baeza-Yates R, Ribiero-Neto B. Modern information retrieval. Boston, MA: Addison Wesley/ACM Press; 1999.

101. Rindflesch TC, Rajan J, Hunter L, Extracting molecular binding relationships from biomedical text. Proceedings of the 6th Applied Natural Language Processing Conference, ACL 2000. pp 188-95.

102. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L, EDGAR: extraction of drugs, genes and relations from the biomedical literature. Proceedings of PSB 2000. pp 514-25.

103. Snow R, Jurafsky D, Ng AY. Learning syntactic patterns for automatic hypernym discovery. Advances in Neural Information Processing Systems. 2004.

104. Riloff E, Automatically Constructing a Dictionary for Information Extraction Tasks. Proceedings of the Eleventh National Conference on Artificial Intelligence 1993. pp 811-6: AAAI Press.

105. Thelen M, Rilloff E, A Bootstrapping method for learning Semantic lexicons using Extraction Patterns Contexts. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2002.

106. Markó K, Schulz S, Hahn U. Automatic lexeme acquisition for a multilingual medical subword thesaurus. Int J Med Inform. 2007 2007/3//;76(2-3):184-9.

107. Cover TM, Thomas JA. Elements of information theory. New York, NY: John Wiley and Sons Inc.; 1991.

108. Witschel H. Terminologie-extraktion – M¨oglichkeiten der Kombination statistischer und musterbasierter Verfahren. W¨urzburg: Ergon Verlag. 2004.

109. Kodratoff M. Comparimg machine learning and knowledge discovery in databases: An application to knowledge discovery in text. ECCAI summer course. 1999.

110. Agrawal R, Imielinski T, Swami A, Mining association rules between sets of items in large databases. Proceedings of the SIGMOD international conference on management of data 1993. pp 207-16. Washington, D.C., United States: ACM.

111. Borgelt C, Efficient implementations of Apriori and Eclat. Proceedings of CEUR Workshop 2003. Aachen, Germany.

112. Grefenstette G, Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis. Proceedings of the 30st annual meeting of the Association for Computational Linguistics 1992.

113. Wilbur WJ, Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Computers in Biology and Medicine. 1996;26(3):209-22.

114. Faure D, Nedellec C, Rouveirol C. Acquisition of Semantic Knowledge using Machine Learning Method: The System ASIUM. Technical Report #ICS-TR-88-16, Laboratoire de Recherche en Informatique, University Paris Sud. 1998.

115. Peat HJ, Willet P. The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of the American Society for Information Science. 1991;42(2):378-83.

116. Diday E. Introduction a L'analyse des Donnees Symboliques. INRIA No 1074. 1989.

117. Faure D, Nedellec C, ASIUM: Learning subcategorization frames and restrictions of selection. ECML Workshop on text mining 1998.

118. Yamaguchi T, Acquiring conceptual relationships from domain-specific texts. Proceedings of IJCAI Workshop on Ontology Learning (OL) 2001. Seattle, USA.

119. Hearst M, Schütze H, Customizing a lexicon to better suit a computational task. Proceedings of the ACL SIGLEX Workshop on Acquisition of Lexical Knowledge from Text 1993. Columbus, OH.

120. Shamsfard M, Barforoush AA. Learning ontologies from natural language texts. International Journal of Human-Computer Studies. 2004;60:17- 63.

121. Etzioni O, Cafarella M, Downey D, Kok S, Popescu A-M, Shaked T, et al., Web Scale information extraction in know it all (preliminary results). Proceedings of the 13th International World Wide Web Conference 2004. pp 100-11. New York, USA.

122. Hahn U, Romacker M, Schulz S. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. Pac Symp Biocomput. 2002:338-49.

123. Hahn U, Romacker M, Schulz S. Discourse structures in medical reports-watch out! The generation of referentially coherent and valid text knowledge bases in the MEDSYNDIKATE System. Int J Med Inform. 1999;53(1):1-28.

124. Strube M, Hahn U. Functional centering: grounding referential coherence in information structure. Comput Linguist. 1999;25(3):309-44.

125. Śarić J, Jensen LJ, Ouzounova R, Rojas I, Bork P. Extraction of regulatory gene/protein networks from Medline. Bioinformatics. 2006 March 15, 2006;22(6):645-50.

126. Maedche A, Volz R, The ontology extraction & maintenance framework Text-To-Onto. Proceedings of the ICDM Workshop on the Integration of Data Mining and knowledge management, San Jose, CA, USA November 31, 2001.

127. Srikant R, Agrawal R. Mining generalized association rules. Future Generation Computer Systems. 1997;13:161-80.

128. Kietz JU, Maedche A, Volz R, A method for semi-automatic ontology acquisition from a corporate intranet. Proceedings of Workshop Ontologies and Text 2000.

129. Maedche A, Staab S, Discovering conceptual relations from text. Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, 2000. pp 321-5.

130. Maedche A, Staab S, Semi-automatic engineering of ontologies from text. Proceeding of the 12th Internal Conference on Software and Knowledge Engineering Chicago, USA 2000.

131. Maedche A, Staab S, Mining ontologies from text. Proceedings of EKAW, Springer Lecture Notes in Artificial Intelligence (LNAI-1937) 2000.

132. Cimiano P, Völker J. Text2Onto - A framework for ontology learning and data-driven change discovery. In: Montoyo A, Mućoz R, Métais E, editors. NLDB; Alicante, Spain: Springer, Heidelberg 2005. p. 227–38.

133. Mima H, Ananiadou S, Nenadić G, Tsujii J, A methodology for terminology-based knowledge acquisition and integration. Proceedings of COLING 2000. pp 667-73. Sarrebruck.

134. Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries. 2000;3:115-30.

135. Ushioda A, Hierarchical clustering of words. Proceedings of COLING 1996. Sarrebruck.

136. Craven M, DiPasquo D, Freitag D, McCallum A, Mitchell T, Nigam K, et al. Learning to construct knowledge bases from the world wide web. Artificial Intelligence. 2000;118:69-113.

137. Ogren PV, Cohen KB, Acquaah-Mensah GK, Eberlein J, Hunter L. The compositional structure of Gene Ontology terms. Pac Symp Biocomput. 2004:214-25.

138. Ogren PV, Cohen KB, Hunter L. Implications of compositionality in the gene ontology for its curation and usage. Pac Symp Biocomput. 2005:174-85.

139. Spackman KA, Campbell KE, Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. Proceedings of AMIA Annual Symp 1998. pp 740-4.

140. Pakhamov S, Coden A, Pakhomov S, Ando R, Duffy P, Chute C. Domain-specific language models and lexicons for tagging. Journal of Biomedical Informatics. 2005;38:422-30.

141. Stetson PD, Johnson SB, Scotch M, Hripcsak G, The sublanguage of cross-coverage. Proceedings of AMIA Annual Symp 2002. pp 742–6.

142. Taira RK, Soderland SG, Jakobovits RM. Automatic structuring of radiology free-text reports. Radiographics. 2001;21(1):237-45.

143. Schadow G, McDonald CJ, Extracting structured information from free text pathology reports. Proceedings of AMIA Annual Symp 2003. pp 584-8.

144. Embarek M, Ferret O, Learning patterns for building resources about semantic relations in the medical domain. Proceedings of the 6th International Language Resources and Evaluation 2008. pp 2006-12. Marrakech, Morocco.

145. Riloff E, Automatically generating extraction patterns from untagged text. Proceedings of the 13th National Conference on Artificial Intelligence 1996. pp 1044-9. Portland, OR.

146. Guarino N, Welty CA. An overview of OntoClean. Handbook on Ontology. 2004:151-72.

147. Faatz A, Steinmetz R, An evaluation framework for ontology enrichment. Proceedings of ECAI and EKAW 2004.

148. National Cancer Institute Thesaurus (NCIT) 2010; Available from: http://ncit.nci.nih.gov/.

149. Mejino JLV, Rubin DL, Brinkley JF, FMA-RadLex: an application ontology of radiological anatomy derived from the Foundational Model of Anatomy reference ontology. Proceedings of the Annual Symposium of American Medical Informatics Association 2008. pp 465. Washington, DC.

150. Hearst MA, Automatic acquisition of hyponyms from large text corpora. Proceedings of the 12th Conference on Computational Linguistics 1992. pp 539-45. Nantes, France.

151. Berland M, Charniak E, Finding parts in very large corpora. Proceedings of the 37th Conference on Computational Linguistics 1999. pp 57-64. College Park, MD.

152. Liu K, Chapman W, Hwa R, Crowley RS. Heuristic Sample Selection to Minimize Reference Standard Training Set for a Part-Of-Speech Tagger. J Am Med Inform Assoc. 2007 September 1, 2007;14(5):641-50.

153. GATE. June, 2010; Available from: http://gate.ac.uk/.

154. Zou Q, Chu W. IndexFinder: A Knowledge-based Method for Indexing Clinical Texts. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.841.

155. Caraballo S, Automatic construction of a hypernym-labeled noun hierarchy from text. Proceedings of the 37th Conference on Computational Linguistics 1999. pp 120-6. College Park, MD.

156. Cederberg S, Widdows D, Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. Proceedings of the 7th Conference on Natural Language Learning 2003. pp 111-8. Edmonton, Canada.

157. Downey D, Etzioni O, Soderland S, Weld DS, Learning text patterns for Web information extraction and assessment. Proceedings of the American Association for Artificial Intelligence Workshop on Adaptive Text Extraction and Mining 2004. pp 50-5. San Jose, CA.

158. Xu R, Morgan A, Das AK, Garber A, Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon. Proceedings of the Workshop on BioNLP 2009. pp 63-70. Boulder, Colorado.

159. Pantel P, Ravich D, Hovy E, Towards terascale knowledge acquisition. Proceedings of Conference on Computational Linguistics 2004. pp 771-7. Barcelona, Spain.

160. Snow R, Jurafsky D, Ng AY, editors. Learning syntactic patterns for automatic hypernym discovery. Cambridge, MA: MIT Press; 2005.

161. Bodenreider O, Rindflesch TC, Burgun A, Unsupervised, corpus-based method for extending a biomedical terminology. Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain 2002. pp 53-60.

162. Liu K, Chapman WW, Savova G, Chute CG, Sioutos N, Crowley RS. Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. Methods Inf Med. 2010;49(6).

163. Lin D. Dependency-based evaluation of MINIPAR. Workshop on the Evaluation of Parsing Systems; Granada, Spain. 1998.

164. Xu F, Kurz D, Piskorski J, Schmeier S, An domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. Proceedings of LREC, The 3rd International Conference on Language Resources and Evaluation 2002. Las Palmas, Canary Islands, Spain.

165. Pantel P, Ravich D, Hovy E, Towards terascale knowledge acquisition. Conference on Computational Linguistics 2004. pp 771-7. Barcelona, Spain.

166. Brewster C, Alani H, Dasmahapatra S, Wilks Y. Data Driven Ontology Evaluation. In Proceedings of International Conference on Language Resources and Evaluation. 2004.