Knowledge Acquisition Framework from Unstructured Biomedical Knowledge Sources

Demeke Asres Ayele

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

KNOWLEDGE ACQUISITION FRAMEWORK FROM UNSTRUCTURED BIOMEDICAL KNOWLEDGE SOURCES

DEMEKE ASRES AYELE

A THESIS SUBMITTED TO IT DOCTORAL PROGRAM
ADDIS ABABA UNIVERSITY

PRESENTED IN FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN
INFORMATION TECHNOLOGY
(LANGUAGE TECHNOLOGY)

Jury:

Prof. Luciano Serafini	External Examiner
Dr. Solonron Tefera	Internal Examiner
Dr. Jean-Pierre Chevallet	Supervisor
Dr. Getnet Mitikie	Co-Supervisor
Dr. Milion Mcshesha	Co-Supervisor
Dr. Dida Midekso	Chairman

ADDIS ABABA, ETHIOPIA
9 August 2016

Knowledge Acquisition Framework from Unstructured Biomedical Knowledge Sources

DEMEKE ASRES AYELE

Addis Ababa University
Oct 2016

Addis Ababa University

School of Graduate Studies

This is to certify that the thesis prepared by Demeke Asres Ayele, entitled *Knowledge Acquisition Framework from Biomedical Knowledge Sources* and submitted in fulfillment of the requirement for the Degree of Doctor of Philosophy in Information Technology (Language Technology) complies with the regulations of the university and meets the accepted standards with respect to originality and quality.

Signed By Examining Committee:

External Examiner: _____Signature: _____ Date: _____

Internal Examiner: _____Signature: _____ Date: _____

Principal Advisor: _____Signature: _____ Date: _____

Co-Advisor: _____Signature: _____ Date: _____

Co-Advisor: _____Signature: _____ Date: _____

_____

Chair of Department or Graduate Program Coordinator

Abstract

In biomedicine, the explosion of textual knowledge sources has introduced formidable challenges for knowledge-aware information systems. Traditional knowledge acquisition methods have been proved costly, resource intensive and time consuming. Automation of large scale knowledge acquisition systems requires narrowing down the semantic gap between biomedical texts and structured representations. In this context, this study proposes a knowledge acquisition framework from biomedical texts. This contributes towards reducing efforts, time and cost incurred to minimaize ontology acquisition bottlenecks.

The proposed framework approximates, models, structures and ontologizes implicit knowledge buried in biomedical texts. In the framework, the semantic disambiguator approximates biomedical artefacts from biomedical texts. The conceptual disambiguator models and structures the biomedical knowledge abstracted from the domain texts. Ontologization presents an explicit interpretation of biomedical artefacts and conceptualizations. The components of the framework are instantiated with scientific and clinical text documents and produced about four million concepts and seven million associations. This set of artefacts is structured into the lower ontological knowledge structure where the upper ontology structure is reused from existing ones. The conceptual structure is represented with graph formalism. The formal interpretation is based on OWL DL language primitives and constructs, which generates a set of OWL DL axioms. The set of OWL DL axioms is referred as the OWL ontology (K_o).

The extent of approximation and quality of structural design are evaluated using criteria-based methods. A set of metrics is used to measure each criterion and showed encouraging results. Correctness measurements for concept entity are 70% for accuracy, 82% for completeness, 68% for conciseness and 100% for consistency. Quality measurement showed complex ontology structure with metrics values of 986,448 for vocabulary size, 18.73 for connectivity density, 145,246 for tree impurity and 226, 698 for graph entropy. The ontology schema potential metrics values are also 0.80 for relationship richness, 3 for attribute richness and 13,253 for inheritance richness. Ontology clarity showed an average readability, which is 3 attributes on average. The proposed framework has limitations to address the acquisition of individuals and entity attributes, losing cardinality information in the acquisition of the ontological knowledge. These lead to limitations on the formal interpretation of biomedical semantics, which in turn lead to deploy only existential restriction based interpretations. Thus, a way forward has been recommended to enhance semantic disambiguation and ontologization of the proposed framework so that they enable to accommodate the acquisition of cardinality and attribute information.

Keywords: Semantic Disambiguation, Conceptual Disambiguation, Ontologization, Knowledge Acquisition Framework, Biomedical Knowledge Source, Ontological Knowledge

Dedication

To my parents

Acknowledgements

First of all I would like to thank the almighty God, who helped me to succeed in this dissertation work with His power and love. He was with me through the ups and downs and supported me to achieve this success, and I gave the glory to Him. The Bible says “But we have this treasures in earthen vessels, that the Excellency of the power may be of God, and not off us” (II CORINTHIANS 4: 7). Nothing is impossible for God! I also thank Addis Ababa University, IT Doctoral Program, for all financial, academic and technical support in pursuing my study. I will also never forget the MRIM team of LIG lab, Grenoble, in supporting me to develop good technical experiences through my supervisor, Jean-Pierre Chevallet.

My gratitude also goes to my supervisors. I really thank my principal supervisor, Jean-Pierre Chevallet (Assoc. Prof.), for his supervision, dedicated help and advice through out this dissertation work. He has been a true inspiration throughout my research period. He thought me how to think critically and confidently in the research I was doing. He was encouraged me to publish and work hard without losing patience. My gratitude also goes to my co-supervisors, Dr Million and Prof. Getnet. My deep gratitude goes to Dr Million for his unlimited professional guidance and patience, from the planning of the research to its write up. I also appreciate Dr Million for his dedicated help on the academic and administrative matters throughout the research work. My thankfulness also goes to Professor Getnet for his brotherly advice and concern towards the biomedical understandings of each part of the research work. I also like to appreciate Prof. Getnet on his concern on the over all research accomplishment and finalization of the dissertation.

My deepest gratitude also goes to the IT Doctoral Program community. It was very helping, positive, cooperative, which has given me moral, strength and strong commitment towards successful completion of my dissertation. It is experienced in establishing research environments, particularly encouraging effortful students towards successful completion of their studies.

Table of Contents

List of Acronyms

CDSS	Clinical Decision Support Systems
CLEF	Clinical E-Science Framework
DAG	Direct Acyclic Graph
DSS	Decision Support System
EHR	Electronic Health Record
EMR	Electronic Medical Records
EPR	Electronic Patient Record
GEHR	Good Electronics Health Record
GENIA	Genome Information Acquisition
GLIF	GuideLine Interchange Format
HIS	Health Information Systems
HL7	Health Level 7
ICT	Information and Communication Technology
IE	Information Extraction
ILP	Inductive Logic Programming
KR	Knowledge Representation
LVG	Lexical Variant Generator
MLT	Medical Language Texts
MT	Metathesaraus
NLP	Natural Language Processing
NLU	Natural Language Understanding
NLM	National Library of Medicine
SNOMED CT	Systematic Nomenclature of Medical/Clinical Terminology
SN	Semantic Network
SP	SPECIALIST Lexicon
TM	Text Mining
UMLS	Unified Medical Language System
VS	Vital Signs

List of Figures

List of Tables

Chapter One Introduction

1.1 Background

With escalating advancement of technology-enabled innovations in biomedical and healthcare industries, Information and Communication Technology (ICT) has been playing crucial roles since many decades ago [1]. In healthcare, for instance, ICT is found to enhance quality, accessibility and cost of healthcare services [2]. Quality can be enhanced by reducing medical errors and usable resource mobilizations. Health data and service accessibility can also be enhanced by creating collaborative environments among healthcare providers, patients and other stakeholders to improve healthcare service deliveries. These, in turn, reduce costs incurred for healthcare services and enable to develop healthy societies. Consequently, ICT is being adopted in most developed nations', Europe and USA, health programs to use its daunting roles [1], [2]. However, in developing nations, especially in Sub-Saharan African, ICT is under utilized to support their resource-constrained healthcare programs and practices [3], [4].

In the developed nations, the clinical practices are progressively embracing innovative ICTs to enable better cures and means for early detection of diseases despite existence of suspicious in processes speed [2]. For example, many cutting edge European ehealth projects have led to significant technological achievements, covering a wide range of health services, which have nevertheless failed to reach sufficient deployment in a real practice [1], [2]. The adoption rate of technologies for better management practices and administrative needs is reportedly slower than other ehealth applications [1], [2]. Consequently, the developed nations are practicing the adoption of advanced ICTs to enable intelligent healthcare practices and services for patients' safety and virtual physiological human [2], [3].

Efficient adoption of technology-enabled healthcare practices, however, requires advanced methods, such as modeling and simulation, at large scale [1]. These reveal the potential of the methods and the recognition they enjoy in supporting medical decision making [2]. The methods have significant application history in numerous healthcare decision support system practices, where their objectives are of managerial or policy nature [3], [4], [5]. These Decision Support Systems (DSS) have shown moderate potential in solving managerial problems, which correlate

with the developed nations' health program sets forth as highest priorities. In the priority set, health data quality is the primary burden [4], [5]. Hence, to meet the current demands arising from adoption of advanced technologies and intelligent healthcare practices at large scale, new tools, methods and business models are being emerging [4], [5].

For example, Clinical Decision Support Systems (CDSSs) are emerging to support healthcare practitioners and providers for their decision making towards improved clinical practices and real-time information access facilities [6]. In the recent increase of attention to prevent medical errors, such as computer-based order entry systems, CDSS has been proposed as a key element to improve patient safety [7]. The SAPPHIRE project[1] is also a CDSS-based system to support healthcare service delivery at home-environment. The healthcare environment attempts to narrow the distance between healthcare institutions' IT infrastructures and the patients' home infrastructures. The environment has a communication protocol based on data privacy requirements, semantically enriched patient data, their integration to hospital information system and Electronic Patient Record (EPR). These adoptions of technologies led to the generation of huge amount of data and information in diagnosis, testing, monitoring, health treatment and management of patients, billing of healthcare services and management of healthcare resources [3]. Numerous healthcare guidelines, policy and scientific literatures are also generated in the adoption. This data and information might also be stored at heterogeneous and distributed Health Information Systems (HIS) with different formats, which are mainly proprietary [4].

The generated data and information are required to be accessible for biomedical stakeholders in general and healthcare practitioners and patients in particular as required by the treatment path of the patients, in a uniform and transparent way anywhere and anytime [3], [4]. For example, healthcare providers may require sharing of data and information, such as clinical notes, observations, laboratory tests, imaging reports, treatments, therapies, drugs administered, allergies and letters, x-rays and bills. These data are heterogeneous in their terminologies, schema, syntax, semantics, data types, formats and constraints. This data heterogeneity leads to significant data interoperability, sharing, accessibility and integrity challenges, which results healthcare systems and biomedical research to be characterized with increased cost, high error rate and knowledge mismanagement [5]. Furthermore, the unstructured patients' health

[1] http://www.srdc.metu.edu.tr/webpage/projects/

information, scientific texts, clinical guidelines, narrative documents and policy documents might not be understood easily by automated systems for advanced use of technologies, for instance CDSSs. The core components of these technologies are knowledge bases for storing and maintaining domain knowledge and inference engines for retrieving relevant knowledge and inferring new ones from the knowledge bases [6], [7], [8]. In this study, biomedical applications that use advanced technologies are referred as knowledge-intensive systems.

In these systems, domain knowledge is required to be represented explicitly and formally for efficient processing, human and machine readability, accurate specification, portability and reusability among healthcare and biomedical institutions and stakeholders. This enables to enhance the integrity, interoperability, intelligibility and precise information access requirements of intelligent systems and services, both in healthcare and biomedical environments [7], [8]. Conventionally, guidelines and terminologies are knowledge representation formalisms in which they present ambiguous and imprecise semantics [9] - [12]. Guidelines are developed based on consensus and evidence in medical research and practice, and enable decision makings in diagnosis and treatment procedures. The electronic version is represented as a standard format, the GuideLine Interchange Format (GLIF). Where as, clinical terminologies allow healthcare professionals to use widely agreed sets of terms and concepts for communicating clinical information among healthcare professionals and institutions around the world for the purposes of diagnosis, prognosis and treatment of diseases [13]. They facilitate identifying and accessing information pertaining to healthcare practices and hence improve the provision of healthcare services.

Guidelines and terminologies, however, provide very ambiguous and vague representations of knowledge in addition to their limited expressiveness and reasoning services [11], [13]. But, the emergent information systems require semantically rich representation formalisms for their precise information access, intelligibility, interoperability, information sharing and collaborative decision makings, for example ontologies [11]. The SAPPHIRE project[1] uses ontologies rather than terminologies to represent vital signs and patient records. Ontological formalism is found to be very promising for semantic analysis and representation of Electronic Health Records (EHRs), healthcare services, unstructured biomedical data and vital signs that are proprietary

[1] http://www.srdc.metu.edu.tr/webpage/projects/

[11]. Thus, ontologies are becoming the crucial component of knowledge-aware biomedical applications and services [14], [15].

There are several information systems, which have used ontologies to support the integrity, interoperability, intelligibility and precise information access of biomedical applications [15] - [19]. For example, Snae and Brueckner [15] developed ontology-based personnel health counseling system based on personal health records. In this work, ontologies are used for matching personal health data with medical treatments, which enables to maintain data transmission between patients and a system. In mobile health platforms, the shared features are extracted, registered and manipulated using ontology-based representations [16]. It can also enable to provide ways of presenting reusable and adaptive healthcare services [16], [17]. In ubiquitous computing environment, ontologies provide context awareness for personalized healthcare services to users at anytime and anywhere [18]. Dang and Hedayati [19] are also developed a personalized healthcare application that retrieve the necessary information about patients care, insurance policies and drug prescriptions with the help of ontology-based knowledge representation system. The ontology allows users and physicians to manage and create context sensitive medical workflows without the intervention of IT people. Nardon and Moura [17] stated the methodology of sharing knowledge based on ontologies by describing how to integrate heterogeneous information for complex queries in real environmental settings.

Despite the crucial contribution of ontologies to enhance intelligibility, integrity, interoperability and precise information access, the design and construction of them and their adoption to a specific application context have been provided less attention yet [17]. Existing ontologies are handcrafted with domain experts and knowledge engineers, and thus, development of large scale ontologies has become resource-intensive and expensive [17]. Quality, scalability, expressivity, flexibility and exhaustivity of these ontologies are low and error-prone, and even there are very little ontologies, such as gene ontology and BioTop, which are formally and explicitly expressed. Consequently, adapting ontologies to a specific application context has become an emergent research interest, which requires well developed methods, techniques and tools in the field. Furthermore, large scale noiseless ontology design and learning, which could be tailored to specific application contexts, has also become challenging due to their scalability, integrity, independency and rigor in addition to the knowledge acquisition bottleneck.

1.2 Statement of the Problem

Knowledge acquisition is motivated with large volume and rate of accumulation of unstructured knowledge sources [3]. In a domain world, unstructured knowledge sources are the highest proportion (accounts for 80%) as compared with other sources, semi-structured and structured [20]. As human knowledge is expressed using natural languages, the accumulation of this knowledge is also the highest rate. This is evident with the proportion of knowledge sources on the Web, in corporates, organizations and institutions [8], [18]. Hence, ontology acquisition can be more prominent and richer if the source of knowledge is unstructured. Knowledge acquisition is also motivated with the representation formalisms and its richness to express domain semantics [13], [17]. Ontologies are semantically-rich representation formalisms, which enable inferencing and tractablity [19]. They are also the core component of semantic web technologies and enable interoperability, integrity, intelligibility and precise information access for domain applications and services. Semantic networks and frame nets are intractable and less expressive representation formalisms. For example, they are failed to express negations, disjunctions and non-taxonomic relations. Conceptual Graphs (CG) and FOPL are highly expressive, but intractable, which do not support inferencing and reasoning [17], [19].

This research is, therefore, motivated to leverage knowledge rich sources, such as unstructured knowledge, and highly expressive and tractable representation formalisms in knowledge acquisition. Particularly, this research is motivated with problems related to medical language text complexities and ambiguities, ontology extraction and structuring based on granularities, and limitations of data-driven methods, such as noise filtering and consensus reaching. These problems are the knowledge acquisition bottlenecks, which have made ontology learning resource-intensive, time consuming and expensive [3], [4]. Existing ontologies are hand-built by experts and knowledge engineers and are highly expensive and time-consuming in addition to error-prone, inflexible, small scale and impracticable [9]. Although existing ontology learning methods and frameworks can support ontology acquisition, involvement of experts and ontology engineers are not minimized to an acceptable level, even unimaginable for large scale ontology learning. Furthermore, they are less domain binded (uses TFIDF for domain relevance), dependent to other ontology tools, lacks rigor and integrity, and they are also noisy. The frameworks are used data-driven methods, and thus, acquisition of shared and noise-free domain

conceptualization is impracticable. Data-driven-based frameworks are also shallow and failed to represent semantic phenomenon such as disjunctions, negations and quantifications [20], [21].

Generally, resource-intensiveness, expensiveness, time-consumption, rigor, scalability, independent and shared ontology acquisitions from biomedical texts are the major challenges in the existing ontology learning methods and frameworks. These challenges are aroused from the inherent complexity and ambiguity problems of natural language texts. The ambiguities and complexities lead to multiple interpretations and views of the biomedical knowledge, which result different conceptualizations of a domain. Multiple interpretations may introduce different understandings for different peoples, which results different knowledge modeling and structuring problems [4], [9]. Consequently, knowledge abstraction with less ambiguity and efficient structuring of the conceptualization has become challenging in the field of knowledge acquisition and representation.

NLP techniques aim to acquire, understand and comprehend texts, which are successful in meeting medical language problems as far as syntax is concerned [22], [25], [27]. But, it has to go a long way in areas of semantics and pragmatics [30], [31]. In semantics, unresolved issues are finding the meaning of a word or a word sense, recognizing quantifiers and its scopes, recognizing concepts and individuals, recognizing associations between concepts or individuals, co-reference resolutions, relation of modifiers to nouns, identifying meaning of tenses to temporal objects and identifying semantic cues across sentences in different paragraphs in a discourse. In pragmatics, a simple declarative sentence stating facts is not only a statement of fact but also serves as some communication functions [30], [31]. The function may be to inform, to mislead about a fact or speaker's belief about a fact, to draw attention, to remind previously mentioned event or object related to fact. All these problems hinder disambiguation and interpretation of semantic phenomenon from biomedical texts, which are highly problematic and less attention have been provided yet.

Representation formalisms must be precise and unambiguous, and enable to capture the intuitive structure of natural language sentences and discourses [36], [37]. Context-independent meaning of sentences can be represented using their logical forms, which encode the possible word senses and identify semantic relationships between words and phrases. This abstract set of semantic

relationships between verbs (relational phrases) and its noun phrases (argument phrases) can be used to capture these relationships [36]. But, the key problem is to consider what combinations of the individual word meanings can combine to create coherent sentence meanings at the sentence and discourse contexts [37]. Furthermore, the representation of the structure of a sentence and its logical form, and to map this into expressions in the representation formalism has been provided less attention yet [37]. As above-mentioned, this may require integration of the different representation formalisms to bring closer semantic structures between biomedical texts and ontological theories, which are hardly possible yet. Thus, for generic, rigor and integrated knowledge acquisition, these problems may require to answer questions how consistent is knowledge acquisition across different formalisms and their interpretations of contextual knowledge?

Empirical NLP has provided significant number of sound techniques, but quite opposed to learning ontological theories [23]. Ontologies are logical theories and declarative by their nature whereas empirical methods are concerned with analytical models that explain data. Despite these methods are not declarative, there are cases that can learn logical theories from data using Inductive Logic Programming (ILP) [24]. However, theories learned from data through ILP differ crucially from shared ontological theories [23]. Ontological theories reflect a shared understanding of domain of interest, which can be developed as a consequence of reflection and consensuses within a certain community and thus representing a commitment to a specific conceptualization [17]. However, in logical theories derived inductively from data, it is not clear how far they can be seen as expressing a shared conceptualization. Thus, one of the concept-driven methods might support an interpretive-based knowledge acquisition, which might also enable to answer to questions what method can enable better consensus reaching than data-driven methods in knowledge acquisition from biomedical texts?

The knowledge acquisition and representation community has also provided less attention to integrate linguistics and ontology learning to knowledge engineering methodologies, and knowledge representation to the way knowledge is expressed in natural language texts [34]. These distinctions have been neglected largely, which might be useful in bridging the gap between semantic structures in natural language texts and ontological theories. While there are works on integrating machine learning to traditional knowledge acquisition and engineering

methodologies such as CommonKADS [34], integration of ontology learning with more recent ontology engineering methodologies such as On-To-Knowledge [35], DILIGENT [36] or METHONTOLOGY [37] haven't been provided adequate attention. The lack of such integration results difficulties to develop integrated, rigorous and scalable ontology acquisition frameworks from unstructured sources. These problems may require an integrated answer to questions in addition to consistency issues across formalisms. For example, how to integrate concept-driven methods to ontology learning and engineering? How to integrate medical language texts, domain conceptualization, domain modeling and structuring formalisms, and formal interpretations?

Generally, to alleviate these problems a little further, an answer is required to a generic question "*how to make closer semantic structures between biomedical language texts and ontological theories*"? This is further tailored to the following research questions:

- *To what extent can biomedical artefacts and their associations be approximated from their unstructured sources, such as biomedical texts*?
- *How to conceptualize, model, structure and interpret biomedical knowledge artefacts*?
- *How consistent and integrated is the knowledge acquisition and representation across formalisms*?

Thus, to address these questions, the following general objective with a specific question is addressed by the corresponding specific objectives.

1.3 Objective of the Study

1.3.1 General Objective

The general objective of this research is to design and develop a rigorous framework for the acquisition and representation of ontological knowledge from unstructured biomedical knowledge sources, the biomedical texts.

1.3.2 Specific Objectives

In order to achieve the general objective, the following specific objectives are formulated:

- *To disambiguate and acquire biomedical artefacts and their associations from biomedical texts.* This approximates biomedical knowledge from the biomedical texts.

- *To model, structure and interpret biomedical knowledge and its conceptualization.* This enables to structure and represent the biomedical knowledge.

- *To disambiguate agreed upon ontological elements: biomedical concepts, relations, their associations and axioms.* The biomedical artefacts and their associations are common understandings among experts, engineers and users in biomedicine, which results a shared conceptualization of the domain knowledge.

- *To keep consistencies across the different formalisms during ontological knowledge acquisition.* The integration of NLP techniques, ontology engineering methodologies and knowledge representation formalisms enabled for, somewhat, consistent acquisitions of ontological elements and axioms.

- *To evaluate the proposed framework for assessing its correctness and quality.*

1.4 Scope and Limitation

The scope of the study is to design and develop a knowledge acquisition framework from unstructured biomedical knowledge sources. Specifically, the knowledge acquisition considers biomedical texts, such as scientific documents (e.g. literatures, books, journals and reports), clinical texts (e.g. clinical notes, radiologic reports, diagnosis results and prescriptions), guidelines, policy and standard documents. The knowledge sources can be any unstructured type and format, such as text, pdf, html, XML, tagged corpora, as far as they can be converted or chunked into phrases, clauses and sentences. This research, however, is not designed to acquire knowledge from other unstructured sources such as images, audio and video. Thus, the proposed framework disambiguates and interprets the hidden knowledge from biomedical texts. In the process of disambiguation, a set of biomedical artefacts, entities and their associations, are generated for abstracting and conceptualizing the biomedical domain.

Disambiguation of situation-specific scenarios is limited to knowledge-based interpretations, where text scenarios instantiate interpretations and the knowledge-base suggest the

interpretation. However, disambiguation based on data-driven or hybrid techniques is beyond the scope of this research. For unambiguous representation and inferencing support, domain conceptualization and its structuring are explicitly interpreted using formal language primitives and constructs. Consequently, the ontology structure is limited to a direct graph based structuring. Particularly, a Direct Acyclic Graph (DAG) is used to structure and represent the ontology structure. The use of other formalisms such as conceptual graphs, frame nets and semantic networks are beyond the scope of this research. Formal interpretation of the conceptualization is also limited to the use of OWL DL language primitives and constructs. But, the use of other formal logics and OWL Lite or OWL Full primitives and constructs are beyond the scope of this research as well.

The knowledge base is limited to biomedical domain where it provides necessary and sufficient biomedical semantics. Furthermore, disambiguation, conceptualization and structuring are limited to the possible semantics suggested by the knowledge base. Consensus reaching is also determined based on the knowledge base semantics, which are already agreed upon biomedical knowledge. That is, consensus of the framework is achieved based on whether the knowledge base is developed collaboratively by different experts and users in the domain or not. This enables to design and develop knowledge acquisition framework, which constructs a shared conceptualization and then shared ontology structure automatically.

Evaluation of the proposed framework is limited to measure its graph-centric representation. Evaluating the functional and usability dimensions is beyond the scope of this research. The structural dimension is measured based on criteria, which describe the ontology structure properties. The use of other evaluation approaches, such as layers-based, gold standard, application and data-driven based, are beyond the scope of this research. The structural evaluation is also limited to measure the extent of approximating biomedical artefacts and the quality of its structural design. Degree of approximation is evaluated by measuring its correctness; where as quality of the framework is evaluated by measuring the efficiency and effectiveness of the ontology structure design. Furthermore, only eight criteria are applied to measure correctness and quality of the proposed framework. Where, four of them are used to measure correctness and others are used to measure quality

While scope and delimitation are as above-stated, the collection of textual knowledge sources for sub-fields of biomedicine is hardly possible. Consequently, collecting clinical narratives such as prescriptions, diagnosis and radiologic reports, books, health standard and policy documents become impossible due to its huge budget and longer time requirement to convert into its electronic version from their fragmented paper based version. Only scientific textual documents and clinical texts (e.g. clinical notes/reports) are used as a knowledge sources. Thus, only biomedical concepts and their associations from scientific and clinical texts are included to the set of biomedical knowledge artefacts. However, as far as any biomedical texts can be segmented or chunked into valid sentences, phrase or clauses, the framework can be applied directly.

The proposed framework is also limited to disambiguate biomedical concepts, roles and their associations. Disambiguation and interpretation of entity attributes and individual entities are not considered in this research implementation, which is left as work on progress. But, it is easily adapted to disambiguate and interpret entity attributes and individuals if the knowledge base supports them. Disambiguation of named entities requires further knowledge base from what we have in the biomedical domain. However, it may require semantically annotated corpora, which is very expensive and effort intensive. As a result, attributional and assertional knowledge in the framework are empty or nullified.

In formalizing the conceptualization, a restricted interpretation is used for unambiguous representations of the set of axioms, which based on either universal quantifier (∀) or existential quantifier (∃) or number restrictions. The technicality of universal quantifier (∀) and number restriction is left as work on progress as they are determined based on the number of individuals involved in the restriction. The evaluation of the framework is also limited to measure values of the set of metrics for each criterion and their analysis results. However, the evaluation doesn't consider the correlation of set of metrics values to subject judged values. This is because judging each metrics values by subjects is labor intensive, time consuming and expensive. Furthermore, checking the well-formedness and syntactic consistency of the set of OWL DL axioms, OWL DL ontology, is left as work on progress. It is technically laborious, expensive and time consuming to check its syntactic consistency and well-formedness by opening, for example, in protégé environment. Although the proposed framework has visible differences (related to scalability, rigor, expert and engineer involvement) with existing frameworks (e.g. TextToOnto,

Text2Onto and CRCTOL), it doesn't attempted comparative evaluations. This is due to large effort, time and budget requirements of the existing frameworks to bring them into an experimentation.

1.5 Significance

There are multi-fold applications of the research concerning to the development of ontological knowledge. Firstly, the research enables to develop biomedical ontology, which can be deployed in knowledge-intensive applications to enhance their intelligibility, integrity, interoperability and preciseness information access. Ontologies have semantically rich taxonomical knowledge, which provides effective and efficient reasoning services. This enables inferencing services while utilized with applications of the domain, and thus, enhances intelligibility. The ontological knowledge also serves as an integration schema of multiple applications' data sources so that it enables the sharing of various data sources among the applications, where it enhances integrity. The ontology also provides shared understanding of the biomedical knowledge among multiple applications, and thus, enhances interoperability. This way, therefore, the proposed ontology acquisition framework has multiple significances for knowledge-intensive applications in biomedicine.

Consequently, the first significance of the proposed framework is to enable precise information access. For precise information access, ontology is required to support the interpretation. The second significance is to integrate the different biomedical information systems in such a way that they able to run uniformly and precisely. The framework enables to use a common knowledge for applications of the domain of interest. And the different applications are able to interoperate each other using the ontological knowledge as a common knowledge base, domain understanding. Lastly, ontologies are the core components of the semantic web technologies, and thus, the have significant contributions to build semantic web applications in addition to semantic markup.

As ontological knowledge is well structured and rich in its semantic representation of a particular domain, the reasoning services it can provide is plentiful. This enables to enhance the intelligibility of the applications that use the ontology for its interpretation and presentation.

Intelligibility is, therefore, one of the significances of the ontological knowledge. The research output could also benefit several communities that are also relevant parties in evaluating the proposed framework. That is, intelligent applications and services could be benefited from the results of the research to access knowledge. Researchers could also be the beneficiaries of the study and the techniques and algorithms developed in the research. Physicians, patients and health institutions are also indirect beneficiaries of the study by having the services provided by the health care that interact actively with them.

1.6 Contribution

The study contributes to the field of knowledge acquisition and representation and its communities as a whole. It has also technological, social and economic contributions. Contribution to the field is the scientific contributions where as technological contribution is to advance technologies inline with the research, knowledge acquisition and representation. Furthermore, this research has also shown the potential of language technologies for knowledge acquisition from unstructured sources, for example biomedical texts. Economic and social contribution is related to resource intensive (e.g. labor, time and cost) nature of the field of knowledge acquisition and representation, and the benefits obtained as a result of deploying the framework for intelligent information systems.

In relation to scientific contribution, the study contributes to open up multiple research dimensions of knowledge acquisition and representation. For example, one dimension is discourse segmentation and disambiguation. In this dimension, each phrase should be segmented and disambiguated semantically. Thus, the study opens many insights in the direction of discourse disambiguation and interpretation, conceptual knowledge structuring and its formal interpretation. It also inspires researches that enable to unlock problems hindering the deployment of intelligent information systems in practical application scenarios. It enables to support integribility, intelligibility and interoperability between interacting information systems. The study also enables to initiate research and investigate for precise fact retrieval and access, semantic search and healthcare applications.

The study contributes to the scientific communities at different dimensions. Firstly, the knowledge acquisition framework enables the scientific communities to construct biomedical ontology at reduced time, labor and cost. Secondly, it initiates research interests in the area. Thirdly, it pushes forward solving the problem of knowledge acquisition from unstructured sources. Technological contribution is the maturity and practicality of utilizing intelligent biomedical applications and services. In this context, the study contributes to support biomedical technologies, health information systems and applications in the domain. Technology maturity can brought social benefits by enabling to utilize health information systems. These social benefits, in turn, can bring economic benefits, for example by delivering health care services everywhere and any time. This reduces cost incurred for health services at institutions and time to get services.

1.7 Methodology

The aim of this research is to design and develop ontology acquisition framework from biomedical texts. In biomedicine, scientific sentences are different from sentences about a diagnosis results or sentences in health standard documents, in which at least they differ pragmatically. Furthermore, the linguistics structure in scientific documents is more formal than clinical texts. This indicates that there is no uniformity not only in the content but also in complexity of the linguistics structures. Thus, determining the population and sampling technique of the study are found to be crucial.

The study employed experimental research and considers all biomedical documents, potentially converted to texts, as population of the study. Technically, the population of the study is approximated to the set of all biomedical artefacts found in the knowledge base. A cluster-based sampling technique is chosen for its easiness and the text collection is in scientific and clinical sub-domains of biomedicine. The choice of cluster-based sampling technique may lead to accept the resulting skewed knowledge of the scientific and clinical sub-domains. But, the proposed framework is independent from the semantic content of textual sentences, but in the linguistic elements and their structures to makeup sentences. The framework is also independent from sentences in any sub-domain of biomedicine as far as they are valid biomedical sentences. That

is, the proposed framework can directly be applied to other sub-domains of biomedicine to complement the less representativeness of the cluster-based sampling technique.

Recently, knowledge acquisition has become more of interpretive and contextual than data-driven [41]. Information is becoming less and less atomic pieces of data and is becoming more and more semantic concept, which carries an interpretation and exists in a context. Thus, interpretive-based methods might have better quality and correctness than data-driven methods for declarative knowledge acquisition [41]. Interpretations of situation-specific scenarios in the bioMed text collection are suggested by the knowledge base. Consequently, knowledge-based method is used for semantic disambiguation and structuring of the implicit knowledge in the biomedical texts. Furthermore, the study applies different tools and techniques and follows several steps to achieve its objectives. The study also uses a criteria-based evaluation method to measure the correctness and quality of the proposed framework.

1.7.1 Research Design

After formulation of the research problems and questions, an experiment is designed for gathering textual knowledge sources and determines the population size and sampling technique. The research design enables to analyze the knowledge sources and abstract the set of biomedical artefacts, conceptualize, structure and interpret them. It also applies a criteria-based evaluation method to measure the structural properties of the graph-centric representation (G_o).

In the research design, therefore, all biomedical sentences are considered as the population of the study. Each sentence is clustered into either scientific or clinical texts. Thus, textual sentences in each cluster are a population in that cluster. The less domain representativeness nature of cluster based sampling is complemented by the framework, as it is dependent only on the knowledge base and knowledge sources in addition to linguistics structures of each sentence. This means that the framework can be applied to any categories as far as the knowledge base is generic to biomedicine and each knowledge source is valid biomedical sentences.

Biomedical artefacts are abstracted and conceptualized based on graph representation formalism. The set of knowledge artefacts is abstracted by instantiating the semantic disambiguation model, where as ontology structuring is instantiated by integrating the upper ontology structure with the

lower ontology structure. The upper knowledge structure is reused from existing ones and the lower knowledge structure is acquired from the set of knowledge artefacts. The conceptual knowledge is interpreted into its formal representation based on OWL DL language primitives and constructs. Finally, the proposed framework is evaluated by measuring the extent of approximating the biomedical knowledge and quality of its structural design. In the following sections, the research design for each component of the framework is stated.

1.7.2 Background Knowledge

In this research, a biomedical knowledge base, which suggests the interpretation of situation-specific scenarios in biomedical texts, is referred as background knowledge. It is also a knowledge known before hand about the domain. In this respect, we choose UMLS as background knowledge for interpreting situation-specific scenarios from bioMed text collection. This is because, UMLS has been developed by the National Library of Medicine (NLM) and is an integration of more than 150 biomedical vocabulary sources into its Metathesaraus in which it consists of more than twelve million concepts and their associations [49]. It has three semantically correlated knowledge layers that represent biomedicine at different level of semantic granularity: the Semantic Network (SN); the Metathesaraus (MT) and its Specialist Lexicon (SL).

The UMLS semantic network represents the high level conceptual abstraction of biomedicine with broader semantic classes and relations, named as semantic types and relationships. The Metathesaraus represents the fine-grained concepts and synonymous terms as well as relationships among concepts. The lexicon represents lexical knowledge sources, which consist of morphological and syntactic attributes of each term in the Metathesaraus. It creates linkage between Metathesaraus concepts and span of texts in the bioMed text collection. Consequently, as it is integrated with UMLS tools (semRep), the UMLSAB2012 version of UMLS is used to suggest interpretation of situation-specific scenarios.

1.7.3 Knowledge Source

The bioMed text collection is prepared as a knowledge source of the proposed ontology acquisition framework. The bioMed text collection is composed of 55,536 scientific and clinical

textual documents. The scientific knowledge sources are literatures, articles, journals and books. The clinical texts are clinical notes and reports. BioMed text collection is prepared as a combination of textual documents from pubmed, ClinicalTrial, Genia and CLEF text collections. From these sources, we collected about 65,736 textual document collections, but we used only 55,536 text documents where the rest has introduced preprocessing difficulties.

Generally, any textual knowledge sources are collected from their various repositories to be used as a knowledge source as far as they can be chunked into phrases, clauses and sentences. Although there is several daunting unstructured knowledge sources in biomedicine in which they represent biomedical phenomenon or situation, this study instantiates the knowledge acquisition framework with scientific and clinical textual knowledge source as they can be easily available for research purposes.

1.7.4 Modeling Language

Conceptual knowledge is expressed using graph formalism, particularly direct graph. Thus, conceptual ontology is represented using a directed acyclic graph, which completely eliminates cyclic redundancies in its representation and guarantees inferencing or tractability. Furthermore, the proposed framework expressed ontologies explicitly and formally using mathematical semantics, which has unambiguous interpretation and support inferencing. Formal representation of ontologies requires formal modeling language, for example Description Logic (DL) due to its tractability. Thus, the OWL DL primitives and constructs are used for unambiguous interpretation of the conceptualization (C_o).

1.7.5 Preprocessing

The bioMed text collection is cleaned to eliminate unwanted part of the text, such as non-ASCII characters and punctuation. This enhances the performance and accuracy of the semantic disambiguation model. Thus, to do the preprocessing, a technique, which handles two tasks, is applied. Firstly, it converts all the different knowledge source types and formats into one common representation format (e.g XML format).

Secondly, the technique eliminates unnecessary (noise) characters, figures, tables and percentages of each textual sentence. After these two tasks, each textual knowledge source is segmented into sentences, phrases or clauses. Finally, each textual knowledge sources is presented as a set of sentences separated by a period followed by a space and a capitalization, making ready for situation specific scenario, concept and their relationship, disambiguation.

1.7.6 Semantic Disambiguation

In semantic disambiguation, situation-specific scenarios in text fragments are disambiguated with MetaMap and semRep. MetaMap disambiguates biomedical concepts, where as semRep disambiguates semantic propositions, associations between pair of concepts. While disambiguating concepts, syntactic parser is used to segment phrases, both argument and relational, in each biomedical sentence [316]. The parser, therefore, disambiguates the syntactic structure of sentences in the bioMed text collection, which we referred as surface semantics disambiguation.

MetaMap uses a lexical matching algorithm to compute the matching of argument phrases with strings referring concepts in the background knowledge. After matching computation and selection of candidate terms, the corresponding concepts are recognized as concepts referred by the span of texts in bioMed text collection. Matching of a term with multiple concepts is also disambiguated based on the concepts' supper-class categorization and the terms context. SemRep uses indicator rules, the UMLS and the set of concepts from MetaMap output. Thus, it disambiguates ontology predicates that correspond to semantic indicators in the bioMed text collection. It also disambiguates appropriate semantic arguments (e.g. concepts) for each semantic predicate. SemRep is also enhanced to include ontological predicates, which are not belonging to the UMLS semantic network associations. To accomplish this, a pair of concepts is searched for their possible associations in the background knowledge and if it exists, the pair will be considered as a predicted proposition.

1.7.7 Conceptual Disambiguation

The conceptual knowledge structure is instantiated by integrating the upper ontology structure (G_u) and the lower ontology structure (G_L). The upper ontology structure is reused from

existing sub-domain and semantic categories, the UMLS semantic groups and semantic types. The lower ontology structure is learned from the set of knowledge artefacts (K_p). Thus, the ontology structure (G_o) and its conceptualization (C_o) is modeled at two knowledge levels: the upper and lower knowledge levels. The upper knowledge level is modeled by sub-domain categories and semantic categories, where the semantic categories are sub-partitions of sub-domain categories.

The lower knowledge level is modeled by the fine-grained concepts and their associations disambiguated from bioMed text collection. During disambiguation, each fine-grained concept is categorized by their semantic categories. This created overlap between the upper and lower knowledge layers, where the set of concept overlaps are used to integrate the two ontology structures. Finally, the ontology structure is represented as a direct acyclic graph.

1.7.8 Ontologization

The ontology structure (G_o) is represented using graph formalism. Graph formalism is very ambiguous and provides poor reasoning support and intractable. The conceptualization (C_o) is, therefore, explicitly interpreted and represented using formal language semantics, OWL DL. OWL DL has semantically-rich primitives and constructs that enable to express the interpretation of domain semantics. It is also computationally tractable and supports reasoning services. Thus, unambiguous interpretation of biomedical ontology is represented as a set of OWL DL axioms.

Consequently, the biomedical ontology (K_o) is formulated as a set of logically integrated biomedical axiom types, which is formulated as six tuples, $K_o = (\Psi, \psi, H_c, H_r, \Phi, A)$. Instantiating this formulation provides four tuples, $K_o = (\Psi, H_c, H_r, \Phi)$. This is because of that assertional and attribution axioms are nullified as they are not disambiguated and interpreted. But, two additional axioms are also introduced, equivalent axioms (ω) and disjoint axioms (ϖ). Thus, the instantiated formulation becomes $K_o = (\Psi, H_c, H_r, \Phi, \omega, \varpi)$, where $\psi = \{\ \}$ and $A = \{\ \}$.

1.7.9 Evaluation Approach

A criteria-based evaluation method is employed to measure the correctness and quality of the proposed framework. Each criterion describes a set of ontology structure properties. Each structural property is also measured using metrics, whose value is determined using a formulation that characterizes the proposed framework or otology structure. Thus, a set of metrics is used to measure each criterion, which can be measured by analyzing the set of its metrics values, which are used to evaluate each criterion, and thus, the results of evaluation, in turn, used to obtain insights about either correctness or quality of the proposed framework. For each criterion, the results of analyzing metrics values and the value for each criterion, including insights of correctness or quality of the framework, are used to perform comparative analysis and discussion.

1.7.2 Research Method

Information processing has brought a paradigm shift towards concept-driven to address the demand shifts from data-driven to semantic-driven services [39]. That is, the basic unit of information is becoming less and less atomic pieces of data and is becoming more and more a semantic entity, which carries interpretations and exists in contexts [41]. Concept-driven methods are based on semantic interpretations and provide natural understanding (cognitive understandings) to domain semantics. It is also one of semantic-driven methods, which support interpretations and provide cognitive understandings. Knowledge-driven methods support shared understandings of domain semantics. Thus, a concept-driven method is used to suggest the interpretations of situation-specific scenarios in the biomedical text. The method is also used to suggest the interpretation of the ontology structure (G_o) to a direct acyclic graph formalism.

Consequently, situation-specific scenarios are interpreted suggested by a set of scenarios in the background knowledge. This means that situation-specific scenarios in the biomedical text suggest which scenarios in the background knowledge are relevant, and scenarios in the background knowledge suggest ways of interpretating and disambiguating the biomedical text. Similarly, scenarios in the knowledge abstraction suggest which scenario contexts in the

background knowledge are relevant, and the scenarios in the background knowledge suggest ways of interpreting and structuring the set of knowledge abstractions.

1.8 Organization of the Dissertation

The rest of the study is organized into six chapters. Chapter two provides detail reviews of the state of the arts in the field of knowledge acquisition, representation and discovery. It also reviews recent studies related to knowledge discovery and acquisition frameworks, which lead to connections with the proposed framework. Chapter three reviews more surveys related to the proposed knowledge acquisition framework. Chapter four provides detail investigations of the design and modeling of the proposed knowledge acquisition framework in a general way. The proposed knowledge acquisition framework and its components are well explained in the same chapter.

While chapter five instantiates the proposed framework to concrete instances such as the UMLS knowledge base, bioMed text collection and MetaMap and semRep programs, chapter six states the evaluation approaches and evaluates the proposed framework. The framework is evaluated using eight criteria, each of which corresponds to a set of metrics. Thus, each criterion is evaluated by measuring a set of structural metrics. Chapter seven concludes the study and recommends further research investigations for more enrichment and maturity of the knowledge acquisition framework.

Chapter Two State of the Art

Ontologies have become core components of Semantic Web (SW) where they enable software agents to communicate each other to resolve conflicts [61], [62]. Gruber [63] defined ontology as a *formal, explicit specification of a shared conceptualization*, where one of the reasons why business and knowledge management communities had a growing interest [64] - [67]. As the volume of biomedical texts is extremely increasing, the manual techniques to acquire ontological knowledge have become impractical. Plus, unstructured knowledge is rapidly changing and difficult to keep up-to-date ontologies. These revolutionalize to large text collections where ontology construction has access to large quantities of unstructured data from different sources [68], [69], such as manually collected or crawled from websites. The challenge is, however, to automate ways of acquiring ontology from the unstructured sources and bring to adequate approximations and representations of the knowledge available for a human reader of the texts [68], [69], [72], [73].

As ontologies represents human knowledge about a domain and human beings successfully communicate their knowledge through medium of texts, it is possible to learn ontologies from large quantities of unstructured texts [70], [71]. Many investigations have been concentrated on the knowledge acquisition challenges, working within Natural Language Processing (NLP), Knowledge Discovery (KD) and Knowledge Representation (KR) techniques. However, the intention has been to consider a range of computational and representational techniques to push the boundaries forward and gain deeper understanding of the relationship between natural language texts and structured knowledge, which however remained resource-intensive and expensive [71], [74], [75]. This chapter, therefore, grants deeper understandings of the state of the art of knowledge bases, knowledge sources, ontology acquisition and representation, and ontology evaluation approaches, methods and techniques.

2.1 Knowledge Bases

Knowledge bases are domain representations in the internal structure of computer programs. The extent to approximate and represent domain knowledge has brought opportunities for intelligent applications to function successfully and interact with the environment humanly. But, although

there is a long history to abstract and represent a domain, most of them are handcrafted and have limited scope and coverage [69]. Traditionally, abstractions and representations are to build knowledge bases from the information provided with the analysis of interviews or dialogue between experts and their customer by means of protocol analysis, which is time consuming, effort intensive and expensive [72].

With the emergence of NLP, Text Mining (TM) and Knowledge Representation (KR) technologies, however, the development of knowledge bases is supported with software tools, environments and techniques [70], [71]. Particularly, these technologies support the development of knowledge bases with the analysis of unstructured (e.g texts) knowledge sources. These enable to capture the hidden knowledge from the unstructured sources, abstract and represent the specified domain. These, in turn, have contributed to reduce time, effort and expensiveness of developing knowledge bases [70], [71]. These technologies may require knowledge bases for acquisition and representation of domain knowledge [69]. In textual sources, for example, meaningful sentences are composed of meaningful words, and any system that processes natural language texts must have information about the words and their meanings. This information can be provided with dictionaries [76]. While dictionaries are developed for ease of human readers but not for machines, WordNet provides effective combination of lexicography and computational information. Thus, wordnet is a repository of linguistics and domain knowledge, which is related to thesauri, taxonomies and ontologies [52].

Ontologies are knowledge bases about entities, abstract and concrete, their associations and attributes, which make up the domain [63], [77]. Thesaurus is a wordbook where words and terms are organized under headings and sub-headings. Taxonomy is a conceptually organized reality [78], [79]. The distinctions among these knowledge structures are a matter of background than substantive differences in approach and objectives. Thus, a brief survey of the three types of knowledge bases and their distinctions are presented in the following sections.

2.1.1 Thesaurus

A thesaurus is a set of words, which mean the same but spell differently. That is, thesaurus is a set of words, which share meaning, sense or are otherwise associated in the minds of people [78]. A dictionary explains meanings of words but in thesaurus the converse is true. An idea is given

to search a word or words by which that idea may be mostly expressed successfully [80]. Thus, words and phrases of a language are classed according to their signification [81]. For example, Roget [78] provided taxonomy of languages and it is entirely inaccessible in most electronic versions. Taxonomy of a thesaurus is important that it contains a great deal of latent information concerning to interrelation between different areas of vocabularies.

As categorization of words has become a useful tool to handle word sense disambiguation, thesaurus is recognized in natural language processing [79]. For example, Sparck Jones [80] attempts to use thesaurus for information retrieval. In the work, the emphasis is on sub-levels, which consist of close synonyms, which are substitutable in a given context. The thesaurus rows are taken as a set of features and clustering techniques are used to see if they could be classified in accordance with their membership of thesaurus heads. One of the key problems was that clusters did not have labels and it became difficult in the research [77]. For example, the Roget's thesaurus was used for word sense disambiguation by a number of researchers. For this reason, the researchers attempted to adapt a thesaurus to a corpus and then use it to perform word sense disambiguation [82].

In Information Retrieval (IR), thesauri are also used for query expansion [83]. Qiu and Frei [84] and Crouch and Yang [85] presented ways to develop thesaurus automatically from text for query expansion and shows improved retrieval performance. Although Voorhees [86] showed that WordNet is not useful in improving text retrieval, Kekäläinen [87] and Sormunen et al. [88] argued that domain specific thesauri can significantly improve text retrieval. Recently, Clough and Stevenson [89] also showed that using EuroWordNet for cross-language information retrieval is effective.

2.1.2 Taxonomies

Taxonomies are a step towards formal specification of knowledge structure in which a specific relationship is usually implied between a parent node and the children [63], [90]. In taxonomies, a child has set-subset relation with parents. Taxonomies have greater depth than a thesaurus, with more levels between leaf and root nodes. In addition, taxonomies can be conceived as a representation of a domain [63], [90]. For example, libraries have long used sophisticated taxonomies in order to classify books and other media, for instance, the Library of Congress

Classification System[1] and the Dewey Decimal System[2]. These systems make unstated claim that each topic in the taxonomy is more general than its individual children, parent includes children. The problem with library classification schemes is that they are not trying to classify objects that appear to be limited in the number of dimensions relevant to classification but rather they classify documents in which by their nature discuss a number of topics [63].

Library classification systems facilitate access to objects, which cover many topics or perceived from many dimensions [91]. The problem where to classify an item is also apparent with one of the most widely used taxonomies of the Internet era, Yahoo![3]. Yahoo! provides access to documents, which are web pages. Yahoo! has been incredibly influential to determine the design of internet and intranet portals of the web content, layout and underlying structure. However, it is an incredibly unwieldy object, which attempts all things to all men [83]. The significance is, however, in the commercial world, there is far greater demand of taxonomies[4] and less for ontologies where many academias are trying to build and promote [92].

2.1.3 WordNet

It is an alphabetical structuring of lexical information and put words together that is spelled alike and scatters words with similar meanings randomly through the list [76]. Wordnet is a proposal for effective combination of lexicographic information and modern high-speed computation, where its aim is to create dictionary and thesaurus [52]. Another objective of wordnet is to support automatic text analysis and artificial intelligence [52]. It is useful to determine connections between synonym sets and tracing morphological connections between words. In WordNet, the taxonomy is structured not only by a *synonym-of* relation, but verbs and nouns are hierarchically organized via hypernym/hyponym relation. Furthermore, EuroWordNet is a multilingual database of WordNets for European languages [93]. Each WordNet is structured in the same way as Princeton WordNet in terms of synsets with basic semantic relations between them.

[1] http://lcweb.loc.gov/catdir/cpso/
[2] http://www.oclc.org/dewey/
[3] http://www.yahoo.com
[4] http://www.dmoz.org/

Thus, WordNets are lexical databases for NLP research and applications. While the current versions have broad domain coverage, they manifest many defects that reflect the lack of domain expertise on the part of lexicographers and it is not built for domain-specific applications. In biomedicine, for example, the research community has been aware of these defects [94], [95]. With a goal of eliminating noises associated with applications of WordNet and similar resources in biomedical domain, Medical WordNet (MWN) is developed as a free-standing lexical database designed specifically for medical NLP [325].

In Wordnets, the common semantic relations are synonymous, antonymy, meronymy, hyponymy/hypernymy and morphological relations [52]. Synonymy is the basic relation as wordnet uses sets of synonyms to represent word senses. It is a symmetric relation between word forms. Antonymy is also a symmetric relation between word forms, especially important in organizing meanings of adjectives and adverbs. Hyponymy and its inverse, hypernymy, are transitive relations between synsets. Because, there is only one hypernym, the relation organizes meanings of nouns to hierarchical structure. Meronymy and its inverse, holonymy, are complex relations.

2.1.4 Ontologies

Ontologies have vibrating in many disciplines ranging from Semantic Web to Knowledge Management (KM) [96], [97]. It also shares and enables to reuse domain knowledge among peoples in a discipline. It interweaves formal semantics understandable to a computer with real world semantics understandable to human [96]. They are also built for knowledge representation [97]. For example, the IEEE Standard Upper Ontology working group defines ontology similar to dictionary with detailed structure that enables computer programs to process contents [98]. Thus, ontology consists of a set of concepts, axioms and relationships that describe a domain of interest [98]. Although there is a consensus of the need for ontologies, each of the following requirements raises its own set of problems. Firstly, although ontology is believed to be shared to a group of agents, in the real world it is extremely difficult to reach consensus in any domain. Secondly, there are disagreements about what the appropriate ontology terms are. That is, there are disagreements to what terms mean and how to relate to external world. Thirdly, there are also disagreements in how to represent these terms.

Human beings are notoriously slow at reaching consensus and ontology represents a commitment to a conceptualization or model of the real world. For explicit ontology, every possible logical relation between concepts is either explicitly specified or logically derivable. Thus, all ontologies to date are partial descriptions of a domain and they are neither complete nor fully explicit. Since ontologies are ideally formal, they must not have any of the vague and metaphorical properties of natural language, which gives it much of its expressive power and practicality. There are different kinds of ontologies, for example top-level and foundational ontologies, which attempt to provide general categories of a domain. These efforts are the result of a combination of philosophic, linguistic and logical motivations and are generally top-down impositions of a specific perspective upon the world [77]. Further examples can be obtained in IEE Standard Upper Ontology [98], Generalized Upper Model (GUM) [99], Beale et al [100] and OpenCyc [101]. The basic expectation is that top-level ontologies enable interoperability between different information systems and automated reasoning [77].

Domain ontologies are built to cover a specific domain, and many of them have built for specific applications and a great deal are available over the web [97]. Navigli et al [97] distinguish between core and specific domain ontologies. Core domain ontology consists of the general characteristic of a domain and they can be found in general resources like WordNet, where specific ontology is specialized to a domain like Gene Ontology (GO). This is missing in most application areas and it is the main challenge to automate ontology learning. The problem is acute because many domains of human knowledge are changing so rapidly that it is impossible to rely on small scaled hand-built ontologies [102], [103]. Generally, ontologies are semantically-rich representation of domain knowledge than others such as thesaurus, taxonomies and wordnets. According to Gruber's [63] definition, they are formal, explicit and shared representations of a domain knowledge. Thus, in this research, a formal ontological formalism is chosen for explicit and unambiguous representation of knowledge acquired from biomedical texts.

2.2 Knowledge Sources

Although unstructured knowledge sources are found to be rich in explicit, implicit and tacit knowledge and their distinctions, transforming them for applying NLP techniques are found to

be difficult and challenging [72]. In the context of texts and NLP, there are a number of implicit and explicit knowledge, which can be acquired using NLP methods and techniques that may, for example, find associations among terms [69]. Knowledge in the text is either explicit or implicit where the implicit are implied or inferred from the explicit. Thus, the explicit or implicit knowledge is in the text but it is difficult to transform into usable formalism for knowledge engineers to deal with them using automatic means [104].

Traditional knowledge sources (e.g. domain experts) have limitations to cover the domain of interest, error-prone, small scale and inconsistent [72], [69]. Knowledge acquired from experts is intentional and inefficient for large scale information processing. The reason is that experts may perform automatically and the knowledge required may not be retrieved and come into mind. Or when experts are unable to provide actual knowledge, they may say nothing or else give reasonable textbook explanations. Another reason is the time where the acquisition process takes. For example, interviews and dialogues take longer time, and even, subsequent transcriptions and analysis may take many hours or days to complete [104]. The third is the high value placed upon experts' time; gaining access to experts may be difficult. Furthermore, inadequate grasp of the domain, knowledge abstraction and representation techniques, misinterpretation, lack of rapport with the expert or misunderstanding the aims of current knowledge acquisition processes enable to undermine the knowledge sources [104]. This way, traditional knowledge sources, domain experts, are inefficient for ontological knowledge acquisition and even highly expensive, time consuming and resource intensive for knowledge acquisition techniques [69] ,[72], [104].

However, with the emergence of NLP and TM technologies, unintentionally presented NL texts, such as books, guidelines, policy and standard documents, are becoming tremendous sources of domain knowledge [70], [71]. These knowledge sources provide richer domain knowledge than the traditional sources, but challenging to acquire the knowledge in them as they are unintentionally written natural language texts. In these sources, the explicit and implicit knowledge are represented naturally in the texts written for different purposes and contexts but not intended to knowledge acquisition systems. However, the sophistication of these sources has meant that researchers must turn to more expedient techniques, which promise a decrease in the amount of time spent and efforts consumed and less expensive on knowledge acquisition and an increase in the quality of knowledge acquired [70], [71].

2.2.1 Knowledge Types

In textual sources, explicit and implicit knowledges are not easily acquired and transformed to useful forms [104]. A great deal of explicit and implicit knowledge can be acquired using NLP techniques, which finds, for example, associations among terms. Although, acquiring term associations are a major step for ontology engineering, it does not provide propositional knowledge and enable inferences as expected in ontological theories [106], [107]. Term associations, explicit knowledge in the text, don't inform the nature of ontological relationships between two terms [106]. Thus, explicit knowledge is expressed with verbal statements of a language [105], [106]. In this context, while explicit knowledge is expressed and communicated linguistically, implicit knowledge is inferred or implied from the explicit knowledge [108].

That means once a proposition is classified as explicit, there are undrawn consequences of the proposition, the implicit knowledge [109]. However, explicit and implicit knowledge have number of distinctions, which are specifically related to personal-level storage and processing of information. Dennett [110] introduced a distinction by stating '*Let us have it that for information to be represented implicitly, we shall mean that it is implied logically by something that is stored explicitly*.' In this expression, the undrawn consequences of the explicit proposition are implicitly represented. In fact, relative to a notion of explicit storage, it is possible to define a notion of implicit representation, differing over the inferential resources that used to draw consequences from the information stored explicitly [109], [110].

Although there is no one correct account of explicit storage, the use of languages provides personal-level reflections of explicit knowledge [110]. Since problems of accessing explicitly stored information may result failures to use information, it has to be accessed before its uses [27], [110]. Thus, explicitly known information are either explicitly stored or implicitly represented. But, explicit storage, even within a human subject, is not sufficient for explicit knowledge. Information may be explicitly stored within a human subject but not available to the subject for verbal report, not accessible to the subject's consciousness and can't be conceptualized by the subject [111]. For example, tacit knowledge of rules involve explicit storage of this kind where an internal representation encodes rules in a structured format and a

process of searching and accessing the representation is required before tacit knowledge of rule can contribute causally to any transitions between input and output representations [112], [113].

Generally, the emphasis is on explicit and implicit knowledge resides in textual knowledge sources, which are found very challenging to acquire them accurately with the current potential of NLP and IE technologies. For example, in a text "*During sexual reproduction, a male sperm joins with a female egg. This is called fertilization*", the antecedent of *this* can't be resolved because it is not clear that there is one. Anaphora resolution is dependent on a great deal of world knowledge, which can not be brought easily in the process of ontological knowledge acquisition. However, explicit and implicit knowledge can be understood as domain knowledge, which can be disambiguated and interpreted from domain texts using advanced NLP techniques. This knowledge can, therefore, be easily structured and encoded into an ontological formalism.

2.2.2 Ontological Nature

The notion of explicitness has very precise meaning that ontological relationship between two terms is expressed with lexico-syntactic patterns of the type identified by Hearst [27]. In this context, it might be found extremely rare for background knowledge to be explicitly expressed in any text [122], [123]. Particularly, it is true for scientific texts as they are attempts to change community accepted ontologies. However, this is not true for texts in textbooks, manuals or glossaries where by their nature don't assume domain specific background knowledge, but they assume general, possibly top-level ontologies. One can also find specifications of ontologies at the borders of a domain, where it might be in time or intellectual space [123]. Thus, one can expect when a concept or rather its corresponding lexicalization is first introduced, there may be statements defining or explicating the idea. When a concept is borrowed from one discipline into another, the term is likely to be defined [124]. The important points to be considered here is that the foundation for efforts for automatically building ontologies from texts is the assumption that there are texts, which specify in a coherent manner the ontological relations one is interested in, and these textual specifications can be read and processed by a machine. However, a number of efforts at automating the ontology acquisition have encountered substantial difficulties due to data sparseness as well [124], [125].

No matter how large a corpus is, the major part of ontologies may not be specified because it is assumed to be part of the background knowledge that the reader brings to the text. This assumption is applied to ordinary texts without didactic intentions, such as textbooks or manuals [123], [124]. This is because; they probably have explicit definitions of terms and explanations if they are intentionally written to instruct [125]. The potential problems related to textbooks and manuals are: firstly, they tend to reflect knowledge at a particular point in time and fall behind the changes occurring in a field; Secondly, explicit knowledge is very sparse because the writers tend to assume sufficient if a definition is provided once; and finally, there are significant areas of endeavor where there are no textbooks [126], [127]. It is not possible to prove empirically data sparseness problem because one could always imagine a larger collection of texts in a specific domain, somewhere in the collection one might also find the missing text expressing knowledge one is seeking to identify. However, experience has shown that a certain number of textual contexts are required for ontological knowledge explicitly available. Finally, at a certain level, background knowledge is presupposed by specialized texts but it is important to keep in mind when designing ontology acquisition systems and understanding their failings [124], [125].

With these understandings of knowledge presented in texts, there are a number of potential sources of ontological knowledge where all of which present certain challenges [143], [144]. Encyclopaedias are the ideal source of ontological knowledge that may include defining or explanatory texts. Google Glossary (GG) is also a source of ontological knowledge. It is a new experimental service in Google Labs, which provides definition texts for each term one enters. Another significant ontological knowledge source is the Internet [143], [144].

2.3 Knowledge Acquisition Methods

Knowledge acquisition is a method for acquiring information and its structures from different knowledge sources, for example biomedical texts [69]. There are different methods for knowledge acquisition from free texts, such as conventional methods, pattern-based methods, memory-based methods, knowledge-based methods and machine learning-based methods.

2.3.1 Conventional Methods

Domain experts are the core sources of knowledge and interviews are a natural way to access domain knowledge from these experts [71], [72], [73]. For example, TEIRESIAS [128], MOLE [129], AQUINAS [73] and Blythe and Ramachandran [74] are technology supported interview-based knowledge acquisition systems. In these systems, the interview is integrated with knowledge-bases and prompts the expert to propose knowledge add-ons for remedy of erroneous answers [73]. The strength of the method is that knowledge enter into the system with experts exclusive of human intermediary, minimizing the engineering effort, and the interview system can make use of existing knowledge-bases and knowledge of appropriate problem-solving methods to provide a context of its questions [72], [73].

In these systems, a dialogue is produced through instantiating question patterns to internal names of domain concepts, which restricts the number of responses available to experts [73], [74]. The method requires the basic structure of knowledge-bases and underlying problem-solving techniques so that any change to knowledge may readily contained [74]. To succeed with these methods, a set of criterion are required to be meet whose details can be found in [75]. The methods are also required the availability of experts in the construction period [75]. Ripple Down Rules (RDR) is another technique, which enables experts to articulate knowledge [130]. RDR is based on realizations to justify conclusions and the justification may change based on contexts [130] [131].

Another traditional knowledge acquisition method is protocol analysis [72], which is a tool to support engineers, for instance the Acquist hypertext-based tool is used to support data analysis such as separating texts into chunks [131], [132]. The text fragments are concept labels and the concepts are structured into hierarchies where experts define the link between concepts [133]. KRITON is another tool, which attempts to support a protocol analysis [130]. Thus, text analysis adapts NLP techniques to support Knowledge Acquisition (KA) systems [134].

2.3.2 Pattern-Based Methods

Assuming that language understanding involves identifying linguistics patterns, another method for knowledge acquisition is to match input texts against pre-determined notion of knowledge, a

set of patterns [146]. In this method, patterns are at a coarser granularity and explicit incorporation assumes the nature of natural language texts, the representation, content of domain knowledge and the target knowledge to be acquired [147]. For example, PETRARCA attempts to obtain knowledge of term descriptions from a corpus of texts [146], [147]. Then, the system derives semantic interpretations of unidentified terms in the unstructured texts. For this, the system uses elements of general knowledge, syntax to semantics rules [135]. For instance, patterns that match words to concepts associating them are used to generate possible interpretations [148].

The Wit system, for instance, acquires knowledge of concepts from texts using a parsing technique that recognizes a small number of syntactical phenomena [146], [147]. Hull and Gomez [149] also presented a system for acquiring knowledge from bio-graphical entries in encyclopaedia. Entries are parsed and passed to a semantic interpreter, which instantiates verbal concepts with appropriate subjects and objects. Additional knowledge is provided as a set of verbs, which are interesting along with indication of verbal concepts and a general knowledge. This enables to disambiguate phrases and complete instantiations [148], [149].

2.3.3 Memory-Based Methods

A memory-based method is also another technique for knowledge acquisition [150]. According to Lebowitz [151], memory-based method is used to acquire knowledge of concepts by integrating bottom-up processing of texts and top-down reuse of domain concepts stored in memory. Understanding a text, which describes a new concept, becomes recognizing which concepts in memory are similar to it and the way it differs from these. In this technique, the goal is to acquire conceptual descriptions of physical structures of disk drives [151]. Systems are also developed frame-based models to express drives and their components along with a set of relations, which describe physical associations of components. Thus, texts are manipulated and labeled to identify memory pointers and terms, usually noun phrases and semantic indicators [150], [151].

Goel et al [152] also stated a method for knowledge acquisition with more sophisticated notions of how devices could be stored in memory. The objective of the system is to acquire the

structural-behavioral-functional models of devices. In this system, ontology of concepts is provided to describe models, together with existing memory of devices. Inputs to the system are descriptions of new devices [152]. Texts are manipulated conceptually using domain knowledge to identify the potential cues of memory of models of devices, which are comparable to new device. The texts are also parsed and translated to a conceptual interpretation, where the interpretation is used to identify functional and structural differences between the new device and the devices whose model have been retrieved. The distinctions are, therefore, utilized to choose a suitable plan and enable retrieved models to be changed and describe the new devices that are indexed and stored in memories [151], [152].

2.3.4 Machine Learning Methods

Data-driven methods are recognizable in knowledge acquisition from natural language texts [153]. Pattern and memory based acquisition methods use data-driven methods in their knowledge acquisition processes. Machine learning is also a data-driven technique, which uses existing machine learning algorithms to acquire knowledge from free texts. Since learning algorithms are developed to capture and represent knowledge, machine-learning-based methods are attractive to be adapted. For example, a machine learning technique is used to learn non-taxonomic relationships among concepts [45]. In the method, texts are tokenized and matched with a lexicon and then subjected to lexico-syntactic analysis. The resulting texts are analyzed to form pairs of concepts, which are linguistically linked in free texts.

The pair of concepts is subjected to associative learning algorithms, which aim to recognize co-occurring concepts. Furthermore, a technique of learning ontology of domain concepts is proposed by Faure and Nédellec [154]. In the technique, concepts are used to generalize sub-categorization frame descriptions. More machine learning methods can also be found in literatures [155]-[159]. Furthermore, there are interactive-based knowledge acquisition methods, which consider the interaction of systems with users or experts in the course of knowledge acquisition. More survey of such methods can be found in [160], [161], [162]. Integration is also another method for knowledge acquisition in which more surveys can be obtained in [163], [164], [165].

2.3.5 Knowledge-Based Methods

Many knowledge acquisition methods use knowledge bases to suggest disambiguation and interpretation of knowledge items from textual sources [135], [136]. The use of knowledge base is suitable to integrate semantics to a system and present a link between natural language texts and semantics in knowledge bases. Several knowledge bases are available recently where their use as background knowledge has the potential of reducing resources needed to develop knowledge acquisition systems [78], [79], [98]. In this context, many knowledge-based methods have been introduced to acquire knowledge items from textual sources towards extending background knowledge or to create new ones [137], [138]. In these techniques, texts are parsed to add potential domain concepts into the background knowledge [139]. Knowledge-based methods have, therefore, been used to disambiguate and interpret syntactic and semantic knowledge that exist in textual documents. Syntactic knowledge bases are used to disambiguate linguistics structure of textual documents, for example Lexicons and WordNets [52]. Semantic knowledge bases, such as semantic graphs, thesauri and ontology, are used to disambiguate and interpret semantics and pragmatics of textual knowledge sources [140]. These methods have been successful in relation to quality and accurate disambiguation and interpretation of situation-specific scenarios in text fragments [140].

Consequently, researchers have been investigating the use of knowledge-based methods for practical applications such as medical knowledge acquisition [141] and medical literature indexing [142]. These methods are outperforming in quality and accuracy of semantic disambiguation and interpretation from unstructured textual sources (e.g. biomedical texts) [139] [140], [141], [142]. Rules are formulated by disambiguating terms, constants and context keywords from textual sources. Then, a recognizer is used to organize rules as tuples of the generated knowledge base schema. Although exploiting these techniques practically remained difficult due to resources required to annotate many texts and the need for templates that stipulate information types, recently several knowledge bases are emerging that enable to reduce these problems. In biomedicine, there have been several knowledge bases that have been utilized in disambiguating and interpreting implicit knowledge hidden in biomedical textual sources. The Unified Medical Language System (UMLS) is one of the largest Knowledge Base (KB) where

most information acquisition systems rely on it. AQUA [140], PROTEOUS-BIO [145] and semRep [142] are three acquisition systems based on UMLS as a knowledge base [71].

Generally, knowledge acquisition methods can be classified into two, data-driven and concept-driven. The above-stated methods, except knowledge-based one, are data-driven techniques. They are used either frequencies or patterns to generate analytical models. Although these methods are well developed, their interpretations are resource-intensive, where the analysis results are empirical models, which don't have declarative nature. On the other hand, ontological theories are declarative in nature. In acquiring ontologies, data-driven methods are failed to accommodate, at least, the shared characteristics of ontologies. Knowledge-based methods are, however, interpretive or cognitive that are concept-driven and enables to interpret semantic knowledge reside in textual sources. In this research, therefore, a knowledge-based method is used to suggest each instance of textual scenarios in the knowledge-base.

2.4 Natural Language Processing

Recently, Natural Language processing (NLP) has gained an escalating complexity [169], [215]. Generic engines are emerging to convey semantic representations of sentences or generate sentences from their representations [216]. NLP is also enabled to build targeted systems of specific purposes, for example, finding index terms in free texts and the ability to judge what level of syntax analysis is appropriate. Consequently, NLP techniques are becoming crucial for creating user-friendly decision-support systems, particularly in areas of knowledge discovery and acquisition. However, these systems must have substantial knowledge about the structure of a language, including what the words are; how to combine the words into sentences; what the words mean; and how these word meanings contribute to sentence meanings [216]. They also require techniques of encoding and using knowledge in a way that can produce appropriate behavior. Plus, situational knowledge plays crucial roles in determining how a system interprets a particular sentence [269].

In biomedicine, researchers are developing and using NLP techniques, which are varying along several dimensions, but the complexity of natural language dictates that semantic interpretation is focused in scope, for instance, applications that are designed to interpret clinical texts such as

discharge summaries [168], [169]. These works are knowledge-based and the specificity of the domain suggests the type and amount of knowledge used [167]. Existing knowledge sources, such as UMLS [168] or GALEN ontology [169], are used but commonly locally developed knowledge bases are used. Furthermore, system restrictions are imposed based on syntactic structures. For instance, NLP techniques process noun phrases or phrases covered by semantic grammars. Different linguistic grammars might be used, including semantic grammars, definite clause and dependency grammars and bottom-up parsers [137], [166].

For example, MedLEE developed semantic models derived from Linguistic String Project (LSP) and is guided by a semantic grammar that consists of patterns of semantic classes [170], [171]. These classes are defined in a semantic lexicon and Friedman et al [170] discussed the use of UMLS in constructing this lexicon. MedLEE has also been evaluated for several clinical applications [172], [173]. The AQUA system was developed to interpret natural language queries issued by users for information retrieval system [140]. The parser used standard definite clause grammars enhanced by an operator grammar, with a support of a semantic lexicon compiled from the UMLS Metathesaraus and Semantic Network. The final semantic representation is in the form of conceptual graphs [140].

The RECIT system concentrates on processing noun phrases and is composed of a proximity processor, a typology of concepts, a dictionary with syntactic and semantic information, a set of conceptual relationships and a set of canonical concepts [174]. The semantic information relies on the model developed by the GALEN project [175]. Rosario et al [177] describe an approach to semantic interpretation of noun phrases and nominal compounds based on semantic information contained in a large lexical hierarchy, the National Library of Medicine's Medical Subject Headings (MeSH) [177]. Part of the challenge addressed by their research is to determine the possible semantic relations that can obtain among the components of a nominal construction. SymText uses probabilistic Bayesian networks to represent semantic types and relations [176]. Syntactic knowledge comes from augmented transition networks and the system depends on a set of reports to train the network for a specific medical domain. SymText has been evaluated for clinical applications [178], [179], [180]. In a recent upgrade to SymText, MPLUS, Bayesian networks are represented in an object-oriented format and a bottom-up chart parser

provides syntactic analysis. In addition, MPLUS uses an abstract semantic language to link Bayesian network types to each other in a predication format [181].

The MENELAS system is a multilingual text understanding system built to extract information from patient discharge summaries [182]. The domain knowledge resides in ontologies, and linguistic relations are projected to the reference model using morpho-syntactic analysis. Thus, the output is an annotated parse tree, which is subject to a semantic analyzer that heuristically selects the best representation using semantic lexicon and rules. MENELAS is also evaluated for coding a subset of discharge summaries whose details is provided by Zweigenbaum et al [183]. Hahn et al [184] also developed a natural language processor, MEDSYNDIKATE, to automatically acquire knowledge from medical reports. Grammatical knowledge comes from a lexicon and a fully specified dependency grammar. Conceptual knowledge comes from a locally developed ontology that consists of a set of axioms for concept roles with corresponding type restrictions for role fillers. In addition to sentence level analysis, MEDSYNDIKATE uses a centering algorithm to resolve anaphoric expressions at the discourse level [185], [186].

In addition to the above-mentioned NLP systems, there are also tools and techniques for semantic information acquisition from biomedical language texts. They enable to identify and extract entities such as individuals, concepts, roles and their attributes and values from the biomedical texts. For example, MetaMap and semRep are biomedical text processing tools developed by National Library of Medicine (NLM) [187], [188].

In MetaMap [58], an input text undergoes lexico-syntactic analysis consisting of: tokenization, sentence boundary determination and acronym/abbreviation identification; part-of-speech tagging; lexical lookup of input words in the SPECIALIST lexicon; and a final syntactic analysis consisting of a shallow parse in which phrases and their lexical heads are identified by the SPECIALIST minimal commitment parser. Then, each phrase is analyzed first for variant generation, where variants of all phrase words are determined; this is followed by candidate identification, where matching of phrase texts are computed and evaluated as to how well they match the input texts; the third step is mapping construction, where candidates found in the previous step are combined and evaluated to produce a final result that best matches the phrase

texts; and, optionally, word-sense disambiguation (WSD), where mappings involving concepts that are semantically consistent with surrounding text are favored.

SemRep is an NLP system designed to disambiguate semantic propositions from biomedical text using underspecified syntactic analysis and structured domain knowledge from the UMLS [187] [188], [189], [190]. After input and tokenization, text is submitted to underspecified parser that relies on syntactic information in the SPECIALIST Lexicon. Part-of-speech ambiguities are resolved with the MedPost Tagger [191]. The interpretation of semantic propositions depends on the underspecified analysis enriched with domain knowledge and is driven by syntactic phenomena that indicate semantic predicates, including verbs, prepositions, nominalizations and the head-modifier relation in simple noun phrases. Rules are used to map syntactic indicators to predicates in the UMLS. For example, there is a rule that links the nominalization *treatment* with the predicate *TREATS*. Domain restrictions are enforced by a meta-rule stipulating that all semantic propositions identified by SemRep must be sanctioned by a predication in the UMLS Semantic Network (SN) [142].

This rule ensures that syntactic arguments associated with *treatment* must have been mapped to UMLS concepts with semantic types that match one of the permissible argument configurations for TREATS, such as *Pharmacologic Substance* and *Disease or Syndrome* [142]. Further syntactic constraints on argument identification are controlled by statements expressed in a dependency grammar [142]. For example, the rules for nominalizations state that one possible argument configuration of an *object* is marked by with the preposition ***of*** occurring to the right of the nominalization and that one possible location for the subject is anywhere to the left of the noun phrase containing the nominalization [142]. Generally, NLP technologies are increasingly supporting knowledge discovery and acquisition systems. In this research, MetaMap and semRep are used to recognize domain entities and their associations from biomedical texts.

2.5 Information Extraction

In the last decades, rapid proliferation of textual information has been available in a myriad of repositories on the Internet and intranets. That is, large proportion of information has been transmitted through free-text documents and is hard to search them [192]-[203]. Consequently, a

growing need of techniques to analyze free-texts and discover valuable information from them led to the emergence of Information Extraction (IE) technologies [204]. The task of IE is to identify a predefined set of entities in a specific domain, where a domain consists of a corpus of texts along with clearly specified information need [192], [202]. Thus, IE is a process of extracting structured information, such as identifying small-scale structures like noun phrases, denoting a person or group of persons, geographical references and numeral expressions, and finding semantic relations between them. However, domain specific knowledge is required for correct aggregation of partially extracted information into a structured representation [204].

IE is a non-trivial task due to the complexity and ambiguity of natural language structures. For example, there are different ways of expressing the same fact distributed across multiple sentences of documents or repositories, and the relevant information might also be implicit and difficult to differentiate [206]. IE is narrower than full text understanding, which computes the possible interpretations and grammatical relations in natural language texts whose realization is still impossible from the technical point of view [205]. Thus, the use of less sophisticated linguistic analysis techniques might be advantageous since they might be sufficient for extracting and aggregating relevant pieces of information. Particularly, the recent advances of NLP with robust, efficient and high-coverage shallow text processing techniques, opposed to deep linguistic analysis, have contributed to the wide spread deployment of IE techniques in real-world applications [205], [206].

The goal of IE is, however, to extract prominent facts about pre-specified types of events, entities or relationships and build more meaningful, rich representations of their semantic content [198]. IE systems are, therefore, used to populate databases or knowledge bases to provide structured input for mining more complex patterns in text collections. Recently, IE provides spectacular advances in converting raw textual information to structured data and they are increasingly being deployed in commercial applications. Consequently, IE constitutes to machine translation, question answering, text summarization and opinion mining [206] - [210].

IE systems are also support to develop new ontology or populate existing ontology with entities and their associations extracted from texts [211] - [214]. For example, Hearst [215] and Cimiano et al [216] proposed techniques to learn and populate ontologies leveraging IE technologies. A

common use of patterns involve searching of phrases, which explicitly show the existence of 'is-a' or 'part-of' relations between two lexical units. Unfortunately, these phrases do not appear often in standard texts, and thus, systems limited with this kind of extraction suffer from low recall. Thus, data-driven ontology learning and populations are supported with supervised and unsupervised learning techniques. Supervised learning outperforms unsupervised one but require manual annotation of training data sets, which is expensive [217], [218].

There are different software platforms, libraries and web services, which are used for different IE tasks. GATE[1] (General Architecture for Text Engineering) is a free software platform for natural language processing. It provides a component framework for creating processing resources, which operate on documents and corpora. ANNIE[2] (A Nearly New Information Extraction) system contains components for common NLP tasks, such as tokenization, sentence splitting, POS tagging, named entity recognition and a simple Coreferences resolution. GATE has also been used as a base for IE systems, such as GATE-SVM [219].

There are also publicly available web services dealing with IE tasks. Calais[3] aims at extracting named entities, relations and events in the domain of news articles. Several tag recommending plugins for web content management systems are based on Calais, for example Tagaroo[4]. Another tag recommending service is provided by Zemanta[5], which can suggest annotations in the form of links to popular sites, such as Wikipedia or Amazon. In biomedical domain, the UMLS platform and associated tools, such as MetaMap[6] and semRep[7], enabled biomedical information extraction.

2.6 Representation Formalisms

Knowledge representation is a core technology in Artificial Intelligence (AI), where AI emphasizes on storing, manipulating and computation of information using computer programs to achieve human intelligence [220]. Knowledge must be represented in machine understandable

1 www.gate.ac.uk
[2] https://gate.ac.uk/sale/tao/splitch6.html
3 http://www.opencalais.com
[4] http://tagaroo.opencalais.com
[5] http://www.zemanta.com
[6] http://metamap.nlm.nih.gov/
[7] http://semrep.nlm.nih.gov/

formalisms for these operations and automated reasoning services [220]. Techniques of automated reasoning allow computer programs to infer new knowledge from existing representations. Thus, knowledge representations appear to be in different formalisms where the most prominent are semantic networks, frame networks, logics and ontologies [220].

2.6.1 Semantic Networks

Semantic networks stem from existential graphs, which express logical sentences as graphical node and link diagrams [222]. Latter, similar notations have been introduced as conceptual graphs differing slightly in syntax and semantics [77]. Despite the differences, semantic graph formalisms concentrate on expressing taxonomic structure of concepts and the relations between them [77]. Semantic network is a graph whose nodes represent concepts and whose arcs represent relations between the concepts in which they provide structural representation of statements about the domain of interest. In biomedicine, for example, *bacteria* and *infections* are typical concepts, while the relations between them are *caused*, *caused_by*, *affected* and *affected_by*. Thus, semantic networks provide a means to abstract knowledge from natural language formalism. That is, it enables to capture knowledge in texts in a form suitable to computational reasons [223], [224].

Concepts represent meanings of noun phrases while relations represent meanings of verb phrases, nominalizations, prepositions or comparatives. A semantic network fragment *CAUSES* (*bacteria*, *infections*) is read as *bacteria causes infections*, expressed as a binary relation between the two concepts, *bacteria* and *infections*. The concepts and their associations are generic and stand for anything relevant in the domain of interest. However, some particular relations for standard knowledge representation and reasoning involve particular instances (objects) instead of concepts. For example, considering a particular bacterium (*bact#1*) and a particular infection (*infect#2*). The latter represents concrete individuals in a domain of interest, the former serves as the classes to group these individuals which have certain properties in common. A particular relation that links individuals to their classes is that of instantiation, denoted by *instance_of*, and thus, *bact#1* is an *instance_of* a concept bacterium [223], [224].

The lower part of semantic network is concerned with knowledge of individuals reflecting about a particular situation of bacterium, *bact#1*, which causes a particular infection, *infect#1*. The

upper part is concerned with knowledge of general concepts, reflecting various possible situations. The most prominent relation in semantic network is subsumption, denoted by *kindOf* relation, which relates two general concepts and expresses generalization-specialization [222]. In the above text, *parasiticBacteria* is a special kind of *bacteria* and it is subsumed by *bacteria*. Subsumption is associated with the notion of inheritance in that a specialized concept inherits all the properties from its most general parent concepts.

In general, semantic networks distinguish between concepts, denoted by generic nodes, and individuals, denoted by individual nodes, and between subclass/supperClass edges and property edges. Using subclass/supperClass links, concepts are organized in a subsumption hierarchy. Using property edges, properties are associated with concepts, to individuals belonging to concepts whose properties are associated with. The two kinds of edges interact with each other in such a way that a property is inherited along subclass/supperClass edges if not modified in a more specific class. However, concrete individuals and data values are not represented well in addition to negations, quantifiers and disjunctions.

2.6.2 The Frame Networks

A frame system has been introduced as alternative to semantic network formalism [221], [223]. Frame systems use data structures to represent knowledge concerning situations and objects, which include defaults and multiple perspectives. Thus, a frame system is a structured representation of semantic networks where concepts and attributes are described as frames. The aim of a frame system is to collect relevant knowledge about a situation in one object instead of distributing across different axioms [221], [224]. That is, a situation is represented in a frame where it contains slots to represent properties of the situation. Therefore, frames provide a structured representation of objects or class of objects [221]. But, frame systems have also limitations similar to semantic networks, which include vague and ambiguous representation, logical inadequacies, unable to represent negations, disjunctions and quantifications.

For example, one frame may represent an automobile and another frame a class of automobile. In a frame language, constructs and primitives are available for organizing frames that represent classes into taxonomies. The constructs allow a knowledge base designer to describe each class as a specialization of generic classes. For instance, an automobile is described as vehicles plus a

set of properties that distinguish automobiles from other vehicles. Reasoning in frames comes in two shapes: the first is using a *partial matching*, more specific frames are embedded into more general ones, thus giving meaning to a new situation or classifying an object as a *kindOf* relationship; the second is searching for slot fillers to collect more information concerning a specific situation. Varieties of expert systems are based on a frame-based formalism and are further enhanced with rules, triggers and daemons [221], [224].

The advantage of frame-based languages, over semantic networks, is that they capture knowledge in a way experts think. This provides a structural representation of useful relations, support a concise definition by specialization technique, and this is easy to use [77], [220], [221] [223], [224]. A special purpose deductive algorithm can be developed, which exploits the structural characteristics of frames to rapidly perform a set of inferences required in knowledge-intensive applications. In this context, frame languages are powerful as the taxonomic associations enable shared descriptive information among multiple frames through inheritance and the internal structure of frames enable to maintain semantic integrity constraints automatically [225], [226]. One of the basic tenets of knowledge system technologies is that domain knowledge can effectively be utilized by a system and easily understood by its users if it is represented in declarative rather than procedural formalism. Frame systems, therefore, provide direct facilities to declaratively describe how the knowledge stored in frames is used [226].

2.6.3 Logic Formalism

Knowledge representations with semantic networks and frame nets are vague, ambiguous and has logical inadequacies which are problematic for computational and inferencing services. They also don't support quantifications, negations and disjunctions in their representation. In the conceptual fragment *CAUSES (Bacteria, Infection)*, for example, it is not clear whether every *bacteria* causes *infection* or some of them. Thus, semantic graphs and frame nets are evolved to logic-based formalisms for expressivity, unambiguous interpretations and computational efficiency [221], [225].

First Order Predicate Logic (FOPL) enables to describe a domain of interest as a set of objects and construct formulas around these objects formed by predicates, functions, variables and logical connectives [223]. Natural language statements are expressed using logical sentences of

objects of a domain of interest with appropriate choice of predicate's and function's symbols. In formulas, concepts and relations are mapped to unary and binary predicates respectively. Whereas, Description Logic (DLs) are computationally tractable subset of FOPL with a typical Tarskian model-theoretic semantics but restricted to unary and binary predicates to capture the notion of concepts and relations [226]. DLs enable to represent knowledge of application domain in a structured and formally understood manner [226].

DLs are motivated in that the important notion of a domain is described by concept descriptions and binary predicates, supported with the use of concept and role constructors and primitives. DLs differ with the semantic network and frames nets as they are equipped with formal semantics and its tractability. That is, DLs are decidable fragment of FOPL and expressive enough that they become major knowledge representation formalism, particularly in the emerging semantic web [226], [227], [228]. Description logic theory consists of statements about concepts, individuals and their relationships [226]. Individuals correspond to constants in first order logic and concepts correspond to unary predicates. The DL concepts can be either named or anonymous. Named concepts consist of a name, say mammal, which can also be mapped into a unary predicate in FOL. Composite or anonymous concepts are formed from named concepts with the use of DL concept constructors, similar to the formation of complex formulas out of atomic formulas in FOL [226]. For example, if C and D are DL concepts, then $C \cap D$, $C \cup D$, and $\neg C$ are DL composite or anonymous concepts.

Exceptionally, DLs provide two special classes, namely $\bot$ and $\top$. They are defined by means of equivalences $\bot \equiv C \cap \neg C$ and $\top \equiv C \cup \neg C$, where C is some arbitrary concept. The concept $\bot$ is the NULL concept, a concept under which everything falls or SINK concept. The concept $\top$ is the TOP concept, a concept that subsumes every concept in the knowledge. Thus, the TOP and NULL concepts are the highest and the lowest in DLs knowledge base respectively. Furthermore, DLs allow a restricted use of quantifiers through role restrictions. Roles are named entities, which are binary predicates in FOL. For example, in DLs given a role, r, and a concept C, the composite concept $\forall r.C$ or $\exists r.C$ can be formed. In addition, DLs are equipped with terminological and assertional axioms. Terminological axioms (denoted as Tbox) are used to introduce names for complex descriptions. Assertional axioms (denoted as Abox) are used to state properties of individuals. A set of assertions in an Abox and named individuals that occur in

the Abox assertions are Abox individuals. The Tbox consists of a set of statements of the form C $\subseteq$ D, C $\supseteq$ D or C$\equiv$ D, where C and D are named or composite concepts. Any Tbox can be translated to first order logic, and thus, inherits a logical consequence relation from it [226].

DLs allow to state individuals as instances of concepts. For example, C (a) states that an individual *a* belongs to a concept *C*. Similarly, a statement r (a, b), where r is a role means that the individual *a* and *b* stand in relation to r. Thus, the DL Abox consists of a set of statements of the form C (a) or r (a, b), where C is a named or anonymous concept, r is a role and *a* and *b* are individuals. The DL knowledge base (κ), therefore, is the *Tbox* and *Abox*. In κ, a concept *C* is subsumed by *D* iff all instances of *C* are necessarily instances of *D* [226], [227], [228]. An individual *i* is an instance of a concept *C* iff *i* is always interpreted as an element of *C* ($i \in C$) [226], [227], [228].

Consequently, DLs are ideal candidates as ontology languages and thus, the W3C proposed Web Ontology Language (OWL), evolved from DLs, as ontology language [226], [227], [228]. OWL has a syntax based on RDF Schema, but it basis for its design with the expressive DL, *SHIQ*, and developers have tried to find a good compromise between expressiveness and complexity of reasoning [226], [228]. Although reasoning in *SHIQ* is decidable, it has a rather high worst case complexity. However, highly optimized reasoner such as RACER [229], Pellet [231] and Fact++ [232] behave quite well in practice. Though *SHIQ* has different features, it has been argued with the DL and ontology communities that these features play a central role when describing properties of aggregate objects and when building ontologies [226], [228]. The actual use of DL provides these features as the underline logical formalism of the web ontology language (OWL), which substantiates the claim [227]. Thus, OWL is an extension of RDF Schema in the sense that it uses the RDF meaning of classes and properties (rdfs: Class and rdfs: subClassOf) and its adds language primitives to support richer expressiveness. Simple extension of RDF schema may clash with the trade of between expressive power and efficient reasoning. RDF schema has powerful modeling primitives (e.g. rdfs: Class, rdfs: Property), which are very expressive and may lead to uncontrollable computational properties [228].

OWL comes with different flavors with a compromise between expressivity and tractability. Usually, ontology developers take into considerations which sublanguage best suits their needs.

For example, the choice between OWL Lite and OWL DL depends on the extent to which users require more expressive constructs. The choice between OWL DL and OWL Full is mainly depends on the extent to which users require the meta-modeling facilities of RDF schema, such as defining classes of classes or attaching properties to classes. When using OWL full as compared to OWL DL, reasoning support is less predictable as complete OWL full implementations is impossible. But, OWL DL language-based reasoning is complete and guaranteed in its decidability. Although, it doesn't have compatibility in meta-modeling of RDF(s)'s primitives, it can apply indirect approaches and techniques to supplement its limitations [226], [227], [228].

Generally, semantic networks and frame nets are emphasized on taxonomic structures and thus, they have limited logical adequacies (increased vagueness), have high heuristic inadequacies (lack of preciseness) and difficult to incorporate negations, disjunctions and non-taxonomic knowledge. FOL is highly expressive but computationally intractable and hence, it doesn't support reasoning services. Thus, as ontologies are semantically-rich concept-based formalisms, and their structuring and representation is based on computationally tractable language formalisms, the DL based formalism is always recommended [228]. In this research, ontologies are, therefore, chosen as formalisms to represent biomedical knowledge. We used the decedent of DL formalism, the OWL DL language primitives and constructs for explicit interpretation of the ontological axioms and assertions.

2.7 Ontology Acquisition Evaluation

According to Gruber [63], ontology is defined as explicit and formal specification of a shared conceptualization, where the conceptualization refers to an abstraction of a domain of interest. Domain abstraction has been increasingly used in information access, data integration and the biggest of which is in the semantic web applications. The apparent increase of using ontologies has lead to an increase in the existence of ontologies, which have heightened the need for evaluating ontologies [236], [237]. Thus, ontology evaluation has become an emerging field, which introduced a number of frameworks and methodologies [240]. The importance of evaluating ontologies is also evident with roles they play in the semantic web and ontology-enabled applications [237], [238], [239]. Hence, ontologies are the centerpiece of knowledge

descriptions in the semantic web allowing for the definition of a shared knowledge, which is acted upon by agents performing on behalf of humans [38], [235]. Ontologies, therefore, attracted lots of interests from academia and industries leading to the proliferation of several ontologies [238], [239]. However, it presents a challenge in deciding how good ontologies are and hence the field of measuring how far ontologies approximate real domain world, correctness, and how good are the qualities, structural and representational efficiency and effectiveness, of ontologies [240], [241].

In this context, ontology evaluation defined as a progression of deciding and quantifying correctness and quality of Ontology Learning (OL) with criteria-based hierarchical assessment techniques [241], [323], [324]. However, a definition of ontology evaluation is provided by Gómez-Pérez et al [236], which later echoed by Vrandecic et al [36]. Accordingly, ontology evaluation is defined with two interesting contexts, verification and validation, in which they also offer a way to categorize current ontology evaluation endeavors. Another definition of ontology evaluation is introduced by Brank et al [41] as a layer-based approach. Ontology is a complex structure and it could be better to evaluate each layer separately than targeting the entire ontology at once. The last definition is related to either comparison against gold standard, uses in applications, judged by human experts or using domain corpus [238], [239]. The first and third approaches may be used at one or more layers of the second approach as required and their optimality [243] - [246].

In evaluating ontology learning frameworks, semiotic-based dimensions, such as structural, functional and usability, are assessed [39], [246]. The structural dimension assesses the graphical structure and formal semantics of ontologies. The functional dimension assesses the intended use of ontologies and their components, the ontology's functions in a context. Usability assesses the level of ontology annotations, its usability, and addresses the communication aspects of ontologies. The three dimensions are analogous to semiotic assessments in linguistics, which embraces syntax, semantics and pragmatics [39], [246]. Thus, to assess ontology learning using its semiotic dimensions, the semiotic characteristics are identified and analyzed. The functional and usability dimensions are resource intensive and costly as they require domain expert and ontology engineer involvement. They also require longer time for evaluations and even in some cases they may not be practical and easy. For good understanding of ontology evaluation, we

introduced different ontology evaluation methods, focusing on the evaluation of structural dimensions using criteria-based methods.

2.7.1 Use-Based Evaluation

Another aspect of evaluating ontologies is to measure their effectiveness in the context of applications. Less attention has been provided to consider application environment and measure different ontologies to assess which one is appropriate in the context. In machine-readability vision of Semantic Web (SW), ontologies are enabling technologies of interoperability. It may be entirely inappropriate for humans to read and assess ontologies, but only technological effects of them are being judged [264]. For example, Velardi and Navigli [265] have done relevant work in OntoLearn system. The task was very challenging, which required bilingually aligned corpora or usage of the Web [114], [266]. Navigli et al used OntoLearn to extract technical terminologies and associate each component element of complex terms with appropriate synsets in EuroWordNet and then select the correct synset from other languages. However, although the results were reputable, the system does not actually use standalone ontology in order to produce its output. The ontology is generated automatically as part of the process and thus the terminology translation task is not well-suited to compare ontologies or successive versions of ontologies [114], [244].

Velardi and Navigli [265] extended their approach to evaluate OntoLearn as ontology learning system, which presented a technique to generate natural language definitions. The ontology learning approach is entirely centered on linking and extending WordNet; and presenting to ontology engineers with a combination of WordNet synset numbers is not very useful, thus, automated generation of glosses provides human readable access for conceptual choices of the ontology generation component. Although the evaluation results are reported as acceptable, a major challenge is to what extent the approach can capture the subtlety of terminological changes as technology and language usage changes. A comparatively low acceptability statistics indicate that this approach is challenging even if it is interesting and provoking. Porzel and Malaka [267] was also undertaken a relevant application-based evaluation methodology. In the methodology, the scenario is to identify the correct speech recognition hypotheses in dialogue systems, where correct hypotheses were identified by hand and act as a gold standard. The system is about

ranking of speech recognition hypotheses where the rank of each hypothesis is determined by a coherence score derived using ontologies [267], [268].

Generally, the task is a speech recognition problem, where evaluation of the task output is interpretations of sentences as compared with a gold standard. Although use-based evaluation methods are elegant for evaluating the effectiveness and efficiency of ontologies through their uses, they have also several limitations: ontology is good or bad when used in a particular context for a particular task, thus, generalization is very difficult with this observation only; the ontology may be a small component of the application and its effect on the outcome may be relatively small and indirect; comparison among different ontologies is possible if they can all be plugged into the same application scenario; the method is very expensive in the sense that it requires specifically tailored or generic ontologies fitting to each application scenarios.

2.7.2 Data-Driven Based Evaluation

Data-driven based ontology evaluation attempts to measure the equivalence between ontologies and a domain of knowledge, domain corpus [244]. The basic notion of the technique is to measure the extent of fitting between ontologies and domain knowledge [267]. Thus, this technique is referred as data-driven ontology evaluation, which essentially identifies a means to compare ontologies with the actual knowledge of a domain. For example, Brewster et al argued that a corpus of texts is an effective source of information to construct large proportion of ontologies [270]. In the technique, a set of domain terms are extracted from the corpus, and a set of concept and relation names are extracted from ontologies and finally count the number of term overlaps between the ontologies and the corpus.

In this technique, ontologies are penalized for terms present in the corpus but not in them, and terms absent in the corpus but present in the ontologies. Another view of this technique is to use a vector representation of the terms in the corpus and ontologies and compute the semantic distance between the two [267], [270]. Brewster et al [268] used a set of domain-specific terms extracted from a corpus of documents, for example using latent semantic analysis to compute the amount of term overlaps between the domain knowledge and ontologies, for example names of concepts and relations. These set of overlapped terms enabled to measure the fit between ontologies and the corpus. Furthermore, precision and recall measures are used to have insights

about the proportion of the overlap. For more wide-ranging and complicated ontologies, which integrate lots of factual information, such as Cyc ontology, text collections are used as a source of facts in the external world.

In this context, the evaluation measure is a proportion of these facts, which are derivable from information in the ontologies. Thus, the approach proposed the idea of developing congruence based measurement between a corpus and a specific domain knowledge, which are used to evaluate ontology learning. The method has limitations, for example evaluating the suitability of organizational structures of ontologies is difficult because basically there are always a large number of different ways to structure the understanding of a domain. In the contexts of data-driven methods, the only thing which can be determined is a range of possible ways of organizing vocabularies, terms or concepts. Another limitation may be data sparsity and mismatch between the domain corpus and the ontologies, which directly affect the amount of overlapped terms between ontologies and the domain knowledge. Furthermore, the domain corpus requires annotations, which are very expensive and time-consuming.

2.7.3 Gold Standard-Based Evaluation

The gold standard-based evaluation of ontology learning measures the semantic distance between learned ontologies and priori developed benchmark ontologies. The learning system is better when the learned ontologies scored highest similarity with the benchmark. In this context, a number of gold standards were developed and a number of techniques to measure the extent to which the learned ontologies corresponds to the gold standard are also developed [271]. For example, Maedche and Staab [273] proposed a method for comparing ontologies at two layers. In the lexical layer, lexical labels are compared using Levenshtein edit distance and derived an averaged string match. The second is the conceptual layer where the notion of semantic cotopy is used as a means to express the semantics of a concept. Semantic cotopy is a set of super- and sub-concepts and is used to calculate the taxonomic overlap between two ontologies. They also defined a relation overlap measure where a relation is defined lexically and conceptually as a pair (C_1, D_1). Where, C_1 is a concept that the relation belongs to and its range restriction D_1. Experimental results of their measures show disagreements between experts constructing ontologies irrespective of the amount of material being predefined. This raises the issue of how

effective the technique is to compare ontologies as lexicalization of concepts and relations varies, while the semantics that Maedche and Staab [274] proposed for concepts and relations are unnecessary and insufficient to establish semantic similarity.

Brank et al [271] extend Maedche and Staab's technique and provided a general account of including instances as part of the evaluation. In contrast to Maedche and Staab, they weren't relied on natural language descriptors of concepts, thus eliminating the string edit distance method. That is, there was no assumption regarding the representation of instances, but only distinguishing of one instance from another. Brank et al extended the Rand index to OntoRand index to measure the distance between different clusters containing any two instances and then comparing the distance across two ontologies [275]. They also proposed that by using different functions to measure the distance between placements of different instances, it provides a family of measures comparing different ontologies. Generally, the gold standard-based evaluation can be used to measure the precision and recall of learned ontologies compared to the gold standard. Evaluation results can also be replicated and are comparable if the same corpus, learning algorithm and gold standard are used. Besides, it is only the first run of evaluation, which is expensive for creating the gold standard. Successive runs of evaluations are fully automatic.

Although gold standard-based methods seem simple and easy, there are basic problems in using them. The first problem is that the technique is somewhat arbitrary. It may be comparable to asking a student to know a textbook for an exam, while the real objective is to find out if the student knows a topic in the textbook. Thus, a gold standard method may have questionable validity, which depends on whether one is evaluating the closeness of ontologies to another or whether one is evaluating ontology is a reasonable representation of a domain, in which they have significant distinction. The main reason gold standards are used is that they are easy and not too complicated. But, excessive dependence on this technique is indicative of immaturity of ontology evaluation in general. Another basic problem with a gold standard is how to measure the difference and similarity between any two ontologies. Generally, similarity-based techniques between learned and gold standard ontologies remained open problems [271], [273].

2.7.4 Layer-Based Evaluation

As ontology acquisition is a complex structure, separating into different learning layers make it more manageable and practical. That is, each layer may be developed separately by different authors [276]. Thus, a layer-layer based evaluation focuses to measure the different layers of ontology learning in a more practical way than to measure the entire ontology [276]. Layer-based evaluation is effective for automatic means than other evaluations techniques [235]. The other reason is that when automatic learning techniques are used to build ontologies, the techniques may be different for the different layers and separately evaluating them is very effective [277]. In this method, each layer may also be defined with different authors, but the definitions may generally tend to be similar and require evaluations at each of the following layers [235], [238] [278], [279]:

Evaluation at *Lexical, vocabulary or data layer* focuses on domain artefacts and their labels. These are concepts, instances and facts in the ontology learning and vocabularies used to label domain artefacts. Evaluation tends to engage contrasts between different sources of data in the problem domain such as domain specific corpus, and techniques such as cosine based similarity measures. Evaluation at *Hierarchies or Taxonomies layer* emphasizes on the ontology hierarchies. Taxonomies are the back-bone of ontologies defined as a set of structures linked by *is-a* relation, subsumption structures. Subsumption *(is-a)* structures are important and they are the focus of specific evaluation efforts, for example to measure inheritance or inferencing capabilities of ontologies. In this context, the measures may be errors related to circulatory definitions, redundancies and wrong partitioning. Evaluation at *other semantic relations* also focuses on non-hierarchical structures, which refers non-taxonomic structures. These structures may also be the focus of evaluation efforts, which measures connectivity densities of ontology learning systems.

Evaluation at a context focuses to measure the usability of ontologies at a particular application or use contexts. In this context, ontology may be seen as part of a larger collection of ontologies, which either reference or being referenced by other ontologies [278], [280], [281]. Another form of context is the application where ontologies are being used. Thus, it may be practical to evaluate ontologies within the context of an application and observe how the results are affected

in using the ontologies. It may also be focused on evaluation from a particular point of view of individual user or organization such as company, which may use the ontologies [282], [283]. Evaluation at the *Syntactic Layer* may have a particular interest for ontologies developed manually. The constructed ontologies are described with formal language and must match the syntactic requirements of that language. The presence of natural language documentations and avoiding loops between definitions may also be considered at the syntactic layer [235].

Evaluation at the *structure, architecture, design* layer focuses on higher-level design decisions that are used during the development of ontologies, which is primarily of interest for manually constructed ontologies. Assuming that a kind of design principles has been agreed upon prior to constructing ontologies, evaluation means checking to what extent the resulting ontology matches those criteria. Structural concerns involve the organization of ontologies and their suitability for further development, such as addition of new concepts, modification of old ones [235], [281]. For some applications, it is also important that the formal definitions and statements of ontologies are accompanied by appropriate natural language documentation, which must be meaningful, coherent, up-to-date and consistent with the formal definitions and sufficiently detailed. Evaluation of these qualities must usually be done largely or even entirely manually by experts and domain engineers.

In general, at each layer of ontology acquisition, one or combination of the above-mentioned evaluation techniques can be used depending on the suitability of the methods. Thus, the layer-based evaluation is easily integrable with any evaluation methods at a particular layer of ontology learning. Moreover, evaluating the structural dimension of ontologies requires lesser resources and cheaper than other methods. The structural dimension evaluation is also easier to be measured automatically based on a set of criterion describing the ontology graph properties and behaviors.

2.7.5 Structure-Based Evaluation

Ontology evaluation can be pursued based on principles of its construction [235], [236]. This is considered as a formative evaluation since the emphasis is on continuous evaluation throughout the ontology development life-cycle where a number of criteria are used as assessments [241] [323], [324]. For example, c*onsistency* is one criterion, which determines whether inconsistent

conclusions are derivable from ontologies or not. It may arise within the definition of concepts or individuals, for instance formal and informal definitions may not coincide. Inconsistencies may also arise from errors in the taxonomy, such as circularity errors and subclass partitions with common classes [235], [236]. Another criterion is c*ompleteness*, which determines whether all knowledge in ontologies is explicitly stated or inferred. But, Gómez-Pérez investigated that completeness can not be proven but incompleteness can be [235], [236]. C*onciseness* is also a criterion, which determines the relevance of knowledge to a domain of interest. Ontology is concise if it does not store unnecessary knowledge and no redundancies in the definition of terms. However, it is extremely difficult to determine what may be unnecessary in ontologies, just as subjective as determining what is needed [238], [239].

Adaptability is concerned with ease of which new definitions and concepts are added to ontologies without re-organizing the ontology structures. If ontology reflects a model or theory about a domain then the ease with which new concepts or phenomena are added is a good indicator of soundness of the model [246]. Accuracy [241] and complexity [323] are also other criteria to measure the structural dimension of ontologies. Accuracy, completeness and consistency enable to determine the extent of approximating the real world by the domain model. Adaptability, clarity, ontology schema potential and complexity enables to determine the quality of ontology structure design [324], [326].

In this context, structural evaluation of ontologies focuses on the graph-centric representation of an ontology learning system and its logical adequacies. In this representation, nodes and arcs are structural elements whereas depth, breadth, modularity and connectivity are topological properties. Consistency, complexity, satisfability and subsumption are logical adequacies and characteristics of ontology structures. Thus, the set of ontology elements, S_o, is a constitute of ontologies that makeup the graph-like structure. The structural elements (e.g. nodes, arcs, and root and leaf nodes) are part of the ontology graph structure (G_o). Ontology graph properties enable to determine the characteristics of ontology graph structure (G_o), for example average depth and breadth of the ontology structure and its complexity, cohesiveness and coupling. Complexity and cohesiveness, in turn, enable to measure the design quality of an ontology acquisition system and its artefact, the ontology.

Thus, quality and correctness of ontology learning systems can be assessed by measuring the characteristics of the ontology graph structure (G_o) with minimal resource requirement and less expensively. And the ontology structure properties and characteristics can easily be adapted to criteria-based evaluation methods with lesser resource requirement, computational complexity and cheaply. With this reason, we adapted criteria-based evaluation techniques in the structural dimension of the proposed ontology acquisition framework.

2.7.6 Criteria-Based Evaluation

Criteria-based evaluation is a technique of evaluating ontologies focusing on measuring correctness and quality of ontology acquisition system's structural dimension [321]. For example, criteria-based hierarchical evaluation is comprised of four evaluation-layers where a set of criterion is defined for each layer. At first layer, the ontology acquisition system and its artefacts are measured. At the second layer, the correctness or quality or both of the ontology acquisition system is measured. At the third layer, each criterion is measured and lastly, metrics of each criterion are measured. Thus, the criteria-based evaluation considers multiple dimensions of a knowledge acquisition system and its artefacts. The dimensions are measures used to assess a set of metrics for each criterion. Each of the dimensions, therefore, expresses at least one aspect of a knowledge acquisition system and then its artefact, the ontological knowledge [323] [324].

In this research, multiple criteria are used to measure different perspectives of the structural dimensions of the proposed framework [321], [323], [324]. The adoption of criteria-based methods is due to resource-intensiveness of other methods and their requirement to involve large number of experts for annotation and subjective evaluation. Plus, it is easily automatable or objectively measureable than others. Criteria and metrics selections are not exhaustive but to demonstrate correctness and quality of the knowledge acquisition framework at lesser resource utilization, computational efficiency and complexity. In this context, each criterion measures an aspect of the structural dimension, which characterizes the ontology acquisition framework. Thus, measuring each criterion based on their metrics evaluates either quality or correctness of the proposed framework. Generally, the following criteria and their metrics are chosen to evaluate either correctness or quality of the proposed framework [321], [323], [324].

Correctness Criteria: these criteria are introduced to measure the extent of approximating biomedical artefacts and their associations from biomedical texts. Correctness of biomedical knowledge abstraction is evaluated using four criteria, such as accuracy, completeness, conciseness and consistency. Accuracy is measured in terms of precision and recall, where as completeness and conciseness is measured in terms of domain coverage and relevance. Consistency is measured in terms of cyclic redundancy and inconsistent partitioning.

Quality Criteria: these criteria are introduced to measure the quality of proposed framework and its artefact. The criteria are focused to evaluate the efficiency of ontology structure design of the proposed framework. Four major criteria are identified for evaluation, such as complexity, adaptability, clarity and potential of the ontology schema. Complexity is assessed using Vocabulary Size (SOV), Edge-Node Ratio (ENR), Tree-ImPurity (TIP) and Graph Entropy (EOG). Adaptability is measured in terms of coupling and cohesiveness, where coupling is measured using Number of External Concepts (NEC) and Roles (NER) involved in the acquisition and cohesiveness is measured using Number of Root (NoR) nodes, Number of Leaf (NoL) nodes and Average Depth of Inheritance Tree of Leaf Node (ADIT-LN). Clarity of the proposed framework is measured in terms of natural language descriptions, such as *owl: labels* and *owl: comments*, involved in the acquisition to easy its understanding by domain experts and users. The last one, potential of ontology schema is measured in terms of schema metrics, such as Relationship Richness (RR), Attribute Richness (AR) and Inheritance Richness (IR) of the ontology graph structure.

2.8 Related Works

Manual acquisition of domain ontologies requires much time, resource-intensive, incomplete and inaccurate, error-prone, biased towards their developers, inflexible and specific to the purpose that motivated their development [48], [70]. Consequently, in order to minimize the expenditure of ontology learning practices, ontology acquisition environments, systems and techniques have been developed [273], [281], [284], [287], [288]. Several ontology learning frameworks and methods have also been developed to support ontology engineers to develop domain ontologies [234], [300], [301], [302], [303], [327]. For example, the ASIUM system acquires taxonomic relations and subcategorisation frames of verbs using syntactic inputs. It clusters noun phrases

based on syntactically related verbs [234]. The ASIUM system extends its lexicon, set of domain concepts and concept hierarchies using cooperative machine learning techniques [234]. OntoLT is also another ontology learning framework, which makes linguistics analysis at protégé interface [303]. OntoLearn is another framework, which focuses on word sense disambiguation. The framework proposed an algorithm based on the structure of general ontologies. The algorithm contains an explanation component, which generates domain relevant concept definitions [265]. An example of ontology learning environment is the work of Mo'K, which enables to build concept hierarchies from text collections using unsupervised machine learning technique [327]. The environment emphasizes on mass clustering techniques and supports ontology engineers to practice with different constraints of ontology learning [327].

All of these systems hardly support user modeling and non-taxonomic relations extraction. They are also used TFIDF to determine domain relevance, which is inappropriate to discriminate domain concepts, even for large collections and big document size. Particularly, TFIDF may provide erroneous results in the case of polysemous terms. That is, the term may have highest frequency for other domains than the domain of interest. Furthermore, these systems require expert and knowledge engineer involvement at larger extent. Thus, the development of accurate and large scale ontologies in a reasonable time and resources are hardly expected [327]. Although many frameworks and methods have been developed to support ontology learning, the involvement of domain experts and ontology engineers are not significantly reduced. Furthermore, the issue of scalability, rigorousity, shared semantics and independency hasn't been raised in the previous works [300], [302], [313]. These have remained ontology learning resource-intensive and time consuming yet.

2.8.1 Ontology Acquisition Methods and Limitations

As mentioned previously, manual development of ontologies are expensive, tedious, error-prone, biased towards their developers, inflexible and time consuming [208]. To enhance these limitations, (semi-) automatic techniques are investigated to learn ontologies from domain texts. Automation of ontology learning not only minimizes expenditures, but also enables to build ontologies that better suite for applications [201]. Consequently, several frameworks and systems have been proposed to develop ontologies, which can be viewed in two perspectives [194]-[205].

The first is to develop ontology building tools and support knowledge engineers and experts, for example protégés and OntoEdit [287], [288]. The second is to support ontology learning from various unstructured sources (semi-) automatically [289], [290].

Although there are many tools to support ontology building (e.g. protégé and OntoEdit), we emphasized on ontology learning techniques from unstructured sources, focusing on data-driven and semantic-driven techniques. In this context, ontology learning is understood as a set of methods and techniques used to develop ontologies in a (semi-) automatic manner from free texts [291], [292]. Hence, learning ontologies from free texts depends on natural language processors, which is the lexico-syntactic analysis components of ontology learning systems. Data-driven ontology learning uses techniques from machine learning, text mining, natural language processing and probabilities to generate analytical models. For example, Cimiano and Volker [302] used surface text processing with statistical analysis. Recently, semantic-driven methods and techniques are used to develop ontologies, focusing context based searching of complex relationships among concepts and intelligent applications. The success of these methods and techniques depends on the availability of semantically-rich background knowledge [71], [298], [299]. In the semantic-driven techniques, lexico-syntactic analysis is used to chunk valid linguistics units. Thus, the use of NLP is common to data-driven and semantic-driven ontology learning methods and techniques.

2.8.1.1 Data-Driven Methods

The data processing community provides large number of sound techniques for data-driven learning, which are quite opposed to the idea of learning ontologies that are logical theories and declarative by nature [42]. It is concerned with analytical models that explain data. In one of data-driven approach, supervised learning, these models serve for classifying training examples, where as in unsupervised learning, they discover patterns in the data, referred as clusters [43], [44]. These techniques are used for ontology learning, for example Maedche and Staab [45] used association rules to discover relations between concepts and Cimiano et al [46] used clustering techniques to group and hierarchically arrange words. However, these models are generally not declarative in the sense of logical theories.

Although theories learned from data through Inductive Logic Programming (ILP) differ crucially from ontologies, there are methods and techniques that attempt to learn logical theories from data [47]. But logical theories derived inductively from data is not clear how far they can express a shared conceptualization. This is because; ontological theories reflect a shared understanding of a domain of interest, which represent a commitment to a specific domain conceptualization. Furthermore, the large volume of data generated analytically requires knowledge engineers and experts for its interpretation, which is highly expensive and resource-intensive. Generally, the future machine learning methods require a systematic analyzes of inductively derived models, classifications and associations to support ontology engineers to formulate their conceptualization in the form of ontological theories. This assumes that ontology learning is an interactive and cooperative process between ontology engineers and a system [48]. In this context, involvement of domain experts and ontology engineers can be supported with data-driven methods and techniques but still they have significant roles in the ontology learning process.

2.8.1.2 Semantic-Driven Methods

Information processing has brought shifts from data processing towards concept processing to address demand shifts from data-driven to semantic-driven applications [39]. It means that the basic unit of information processing is becoming less and less atomic pieces of data and is becoming more and more semantic entities, which easy interpretations and exists in contexts. This changes information processing from syntactic computing to semantic computing paradigm [41]. The semantic-driven processing emerged from artificial intelligence attempts to model cognition, which is interpretive-based information processing as compared to analytical models [41]. Semantic-driven methods depend on the availability of semantically-rich background knowledge, which suggests situation-specific scenario interpretation in the free texts. We considered such semantic-driven interpretation methods and techniques as one of the knowledge-based methods, a concept-driven technique.

With the availability of background knowledge, the concept-driven technique enables to instantiate already consensus reached semantics to support ontological theories and develop a shared domain conceptualization. Thus, knowledge-based methods support ontological theories

than the data-driven counterparts. In this context, we applied a concept-driven processing technique to learn ontologies from biomedical texts. In this method, situation-specific scenario is interpreted suggested by a set of instances in the background knowledge. That is, the situation specific scenario in a text suggest a knowledge base scenario, and knowledge base instances suggest ways of interpretating and disambiguating text fragments from biomedical texts. In this way, a knowledge-based method enables to learn ontologies, which have common understandings to a group of peoples or experts. Plus, it minimizes expenditures of ontology learning from unstructured sources than data-driven methods and techniques.

Thus, knowledge-based methods and techniques are outperforming in accuracy and quality of knowledge acquisition from textual sources [71], [142]. There are many investigations, which used knowledge-based methods for a range of applications, including medical knowledge acquisition, medical literature indexing and searching, automatic coding of clinical texts and processing molecular biology information [141], [142]. In these systems, after extraction of terms, constants and keywords as rules, a recognizer is used to organize them as tuples of the generated knowledge-base schema. Although exploiting these methods remains difficult because of time and effort needed to manually annotate large texts and the need for templates that stipulate information types to extract, several knowledge-bases are emerging that reduce these problems [49].

In biomedicine, for example, there are many knowledge-bases, which are utilized to disambiguate implicit knowledge buried in the biomedical texts. UMLS is one of the largest biomedical knowledge-base where most knowledge acquisition systems rely on it. It is an integration of more than 150 biomedical vocabulary sources, including SNOMED CT and MeSH [49]. Furthermore, UMLS has been developed, maintained and used since 1986 by different domain experts and users internationally in addition to those in the National Library of Medicine (NLM). Thus, the semantics in the UMLS knowledge base is already consensus reached and shared among these experts and users. AQUA [140], PROTEOUS-BIO [141], MetaMap [58] [187] and semRep [142] are knowledge acquisition systems based on the UMLS knowledge-base. MetaMap is a program that disambiguates concepts from biomedical texts [58], [187]. On the other hand, semRep disambiguates semantic propositions from biomedical texts using underspecified syntactic analysis and structured domain knowledge, the UMLS [142].

2.8.1.3 Comparisons

A comparison is made between data-driven and knowledge-based methods for domain ontology learning based on identified parameters. The parameters are scalability, rigor, consensus reaching, domain binding, domain adaptive and support ontological theories. Table 3.1 illustrates the parameters supported by the methods. In the table, tick-mark (✓) illustrates the method supports the parameter and the cross (x) shows that the parameter is less supported by the method.

Table 2.1 – Parameters for Comparison

Methods	Rigor	Domain binding	Consensus reaching	Domain adaptive	Ontological theories
Data-driven methods	×	×	×	✓	×
Knowledge-based methods	✓	✓	✓	based on KB avaliability	✓

The data-driven methods are less rigorous in relation to scalability and integrity of ontology learning. This is due to the need to involve ontology engineers and experts in the acquisition process. For example, the TextToOnto framework [300] can extract about 5,000 terms and thus hardly possible to structure these concepts using ontology engineers. Data-driven methods also use TFIDF techniques to determine terms' domain relevance, which is less efficient for discriminating terms with multiple meanings, particularly across domains [300], [302]. Although data-driven methods are highly domain adaptive, it doesn't support consensus reaching and ontological theories. In contrast, knowledge-based methods support all parameters illustrated in the table. Consequently, with all these reasons, we adapted a knowledge-based method in the proposed ontology acquisition framework from biomedical texts.

2.8.2 Ontology Acquisition Frameworks and Limitations

Ontology engineering has emerged as a science and thus, ontology learning tools, methods and techniques have become necessary to easy the difficulties to model knowledge relevant to a domain of interest [36], [37]. As above-stated, data-driven methods and techniques have supported ontology engineering and enabled ontology learning, which have indeed the potential to reduce cost, time and effort of developing and maintaining ontologies [300], [302]. This is the

reason why a plethora of ontology learning frameworks have been developed in the last few decades and integrated with standard ontology engineering tools and environments for actual ontologization and interpretation of domain artefacts and their interactions [300], [302], [303], [313]. For instance, TextToOnto [300] is integrated with KAON ontology engineering environment [301]. OntoLT [303] is integrated with protégé and Text2Onto [302] is integrated with NeOn Toolkit[1], a networked ontology learning environment. These frameworks are developed to support ontology engineers in acquiring domain artefacts, concepts, relations and their associations, from free domain texts. Ontology learning technologies are far less developed as compared with data mining and text mining techniques [300], [302], [303], [313]. Consequently, these frameworks underlie with baseline architecture whose components are ontology management, coordination, resources processing and algorithm library [300]. The components enable ontology engineers to accomplish different tasks of ontology modeling and development such as adding entities to the ontology model, coordinating the components, processing the resources and accessing the algorithm libraries. Interactions of these components are shown in Figure 3.1 as proposed by Maedche and Volz [300].

The text processing component analyzes texts suggested by a lexical database and domain lexicon. But, the database is general purpose knowledge rather than representing specific knowledge of biomedicine [325]. The learning and discovering component extracts lexical relationships, example *hypernym* and hyponym, rather than domain knowledge relationships, example concept and role taxonomies, in addition to instances of each concepts and roles. Thus, ontology structuring is made manually supported by ontology tools, for example OntoEdit, which requires longer time, more efforts and expensive.

[1] neon-project.org

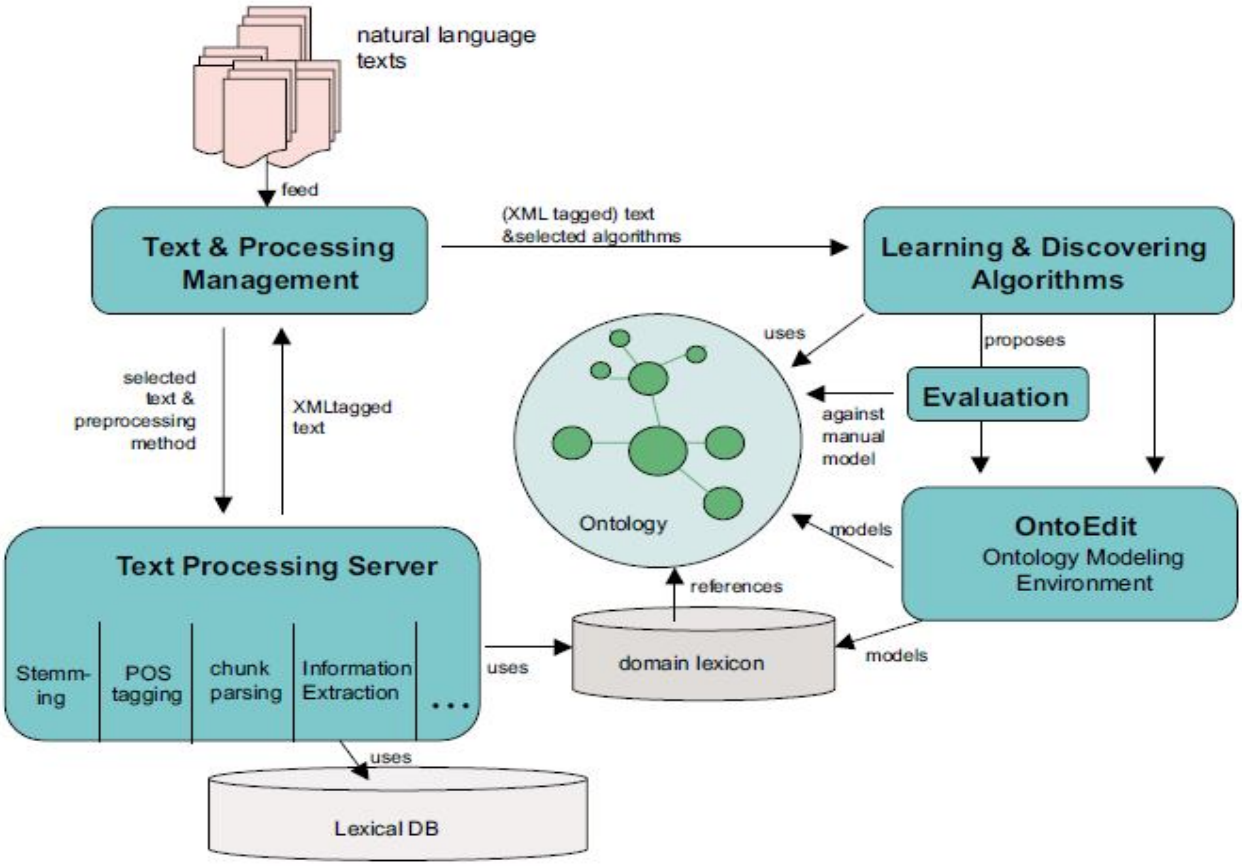

Figure 2.1 – Architecture of Ontology Learning Framework [300]

The ontology learning framework focuses on a specific ontology engineering task and applies shallow NLP tools for their document analysis. For example, OntoLT [303] is a plug-in for protégé ontology editor and focuses on end-users and linguistics analysis. Basically, it uses the internal structure of noun phrases to develop ontologies from domain texts. TextToOnto [300] develops a variety of algorithms for different ontology learning tasks. In particular, it developed different term relevance measures, different algorithms to support taxonomy relation extraction and techniques for learning relations between concepts. Generally, the focus of TextToOnto is on the algorithmic stamina, which consequences the combination of different algorithms and their interactions. But, in TextToOnto end-users were abandoned.

The successor of TextToOnto, Text2Onto [302], focuses on handling the challenge of user interactions and the dynamic change of text documents by introducing Probabilistic Ontology Model (POM) as a container for the different algorithms as well as adding probabilities to the

learned structures to facilitate the interaction with the user. Plus, OntoLearn [265] emphasizes on the problem of sense disambiguation with respect to a lexical database. The CRCTOL system [313] is focused on the NLP tools, particularly on text parsing. CRCTOL utilizes full parsing techniques in document analysis to enhance the performance of concept recognition. The shared characteristics of these frameworks are the use of natural language processing tools to derive features from domain texts and learn ontological structures. Commonly, the NLP tools for analyzing text documents are morphological analyzers, part-of-speech (POS) taggers and chunk parsers. Then, these tools employed text-mining techniques to build ontologies from text document collections.

However, these frameworks provided less attention on the issue of integrity, independency, rigor, domain binding or relevance and consensus reaching to support the development of ontologies from free texts. Thus, ontology acquisition remained requiring much time, resource and expensive. As above-stated, TextToOnto focuses on the algorithmic backbone, Text2Onto focuses on the dynamicity and user interactions and CRCTOL concentrates on the NLP components. Thus, the frameworks lack an integrated acquisition, which seriously impacts on the rigor of the frameworks. Furthermore, the frameworks used TFIDF or DFIDF for domain relevance measure and WordNet to support ontology structuring by ontology engineers, which also supported by ontology tools. For example, OntoLT uses protégé and TextToOnto uses OntoEdit for structuring and interpreting domain artefacts and their associations. In these frameworks, generally, fully automatic acquisition of ontology remains at a distant future, overall acquisition is considered to be semi-automatic with human intervention supported by ontology tools [304].

The frameworks may also have generality and adaptability but the issue of senses disambiguation, explicit interpretation and scalability are highly ignored. They haven’t used sense disambiguation techniques. But, CRCTOL [313] used variant of LESK algorithm, which uses window based context overlap. Lesk’s approach is very sensitive to the exact wording of definitions, so the absence of certain word can radically change the results. The algorithm determines overlaps only among glosses of senses being considered. This is a limitation in that dictionary glosses tend to be fairly short and do not provide sufficient vocabulary to relate fine-grained sense distinctions.

The frameworks are also designed for small scale, in less than five thousand concepts and their associations, as the learning requires expert and engineer involvement [300], [302], [303], [313]. Furthermore, the frameworks used data-driven methods, particular machine learning techniques, which are mainly based on weighting schemes, frequency counts and term collocations. On the other hand, ontological theories are highly interpretive and declarative. Thus, reaching consensus is almost unachievable in data-driven approach. TextToOnto [300], Text2Onto [302] and CRCTOL [313] are used WordNet for suggesting their ontology structure. But, WordNet doesn't have well defined taxonomic structure and have very limited domain specific coverage, particularly in biomedical domain [325].

Generally, existing ontology learning frameworks can be characterized with different dimensions. These may include scalability, integrity, consensus reaching, domain relevance, rigor, ontology structuring, formal interpretation, use of ontology tools and knowledge bases, and finally weather the framework is ontology learning or not. These dimensions are used to clearly demonstrate the weakness and strengths of existing ontology learning frameworks, particularly OntoLT, TextToOnto, Text2Onto and CRCTOL. In the last column of the table, the Pikes framework (one of linked-data extraction framework) [329] is indicated to show that PIKES is not ontological knowledge extraction framework. As illustrated in the table, scalability is less than five thousand concepts and their associations. The ontology developed using this set of artefacts is very small scale and insufficient for practical application usage. Each framework is also specialized only in one component of a framework. Thus, there is a lack of integrated ontology learning, which, in turn, consequences lack of rigorous acquisition. While the ontology learning frameworks are domain adaptable, consensus reaching is very challenging and never achieved yet.

Ontology structuring and interpretation are also manual supported with ontology engineering tools. It is resource intensive in terms of time, effort and cost. Domain relevance is determined using TFIDF and DFIDF techniques. These techniques have their own weakness even if the document collection is large and each document size is long. Using WordNet for ontology structuring and lexical interpretation may make the resulting ontology more general rather than specific to a given domain. Wordnet is a generic lexical knowledge structure, which has no clear taxonomic hierarchies. Plus, it is comprised of very limited domain specific structures and

lexical information. These necessitated highly scaled, integrated, rigor and consensus reached framework with rich semantics knowledge for better ontology learning and engineering. In table 3.2, the cross (x) illustrates the absences of supporting the ontology dimensions or characteristics by the frameworks.

In conclusion, it is understood that knowledge acquisition is still facing a big challenge to model ontologies relevant to a domain due to the knowledge acquisition bottlenecks. This requires the development of generic, rigorous and independent ontology acquisition frameworks, which minimize the involvement of ontology engineers and domain experts. To contribute towards alleviating these problems and narrow down ontology acquisition bottlenecks, we proposed to design and develop a rigorous, scalable and independent ontology learning framework from biomedical texts. In addition to rigorousity and scalability, the proposed framework minimizes the involvements of ontology engineers and domain experts towards automating ontology acquisition, and thus, reduces time, effort and cost incurred in conventional ontology learning frameworks.

Table 2.2 – Characteristics of Ontology Learning Frameworks

		OntoLT	TextToOnto	Text2Onto	CRCTOL	PIKES
1	Scalability	<5,000	<5,000	<5,000	<5,000	>5000
2	Integrity	×	×	×	×	×
3	Consensus reaching	×	×	×	×	×
4	Independency	×	×	×	×	∨
5	Domain relevance	TFIDF	TFIDF	TFIDF	KFIDF	not used
6	Rigor	×	×	×	×	∨
7	Ontology structuring	manual	manual	manual	manual	semi-auto
8	Interpretation	Protégé	ontoEdit	ontoEdit	ontoEdit	not used
9	Onto-enviroment	Protégé	ontoEdit	ontoEdit	ontoEdit	not used
10	KB	×	wordnet	wordnet	wordnet	wordnet
11	Is Ontology learning Framework	∨	∨	∨	∨	×

Chapter Three Knowledge Acquisition Framework

Knowledge acquisition models and frameworks are crucial technologies to support large scale ontology engineering at minimal resource, cost and time requirement [300], [313]. Consequently, researchers have investigated several methods to support ontology engineering, and thus, derived meaningful domain artefacts and their associations. These set of artefacts and their associations support to model ontologies and represent a domain [300], [313]. In this chapter, an ontology acquisition framework is proposed and designed. The proposed Knowledge Acquisition Framework (KAF) disambiguates implicit biomedical knowledge (concepts, relations and their associations) from biomedical texts. The set of concepts, relations and their associations are referred as biomedical knowledge artefacts and denoted by K_p. After domain abstraction, the proposed framework conceptualizes, models, and structures biomedical knowledge disambiguated from unstructured knowledge sources (K_s). The conceptualization (C_o) and its structure (G_o) are interpreted with ontology language, OWL DL, primitives and constructs for unambiguous interpretations (denoted as K_o). This is referred as the formal ontology model, which is a set of OWL DL axioms.

The set of knowledge artefacts (K_p) is disambiguated from biomedical texts using lexico-syntactic and semantic analysis techniques. The lexico-syntactic analysis determines meaningful linguistics units of biomedical sentences leveraging Part Of Speech (POS) taggers and syntactic parsers, referred as phrase segmentation or chunking. Phrase segmentation chunks each textual sentence to a set of text spans, which are either argument or relational phrases. Argument phrases (e.g. NPs) are disambiguated to biomedical concepts or individuals, whereas relational phrases (e.g. VPs and nominalizations) are disambiguated to associations between argument phrases, semantic predicates. Furthermore, lexico-syntactic analysis predicts the association of syntactic arguments and relations, referred as syntactic-associations or syntactic-relations.

The semantic analysis recognizes biomedical artefacts referred by meaningful linguistics units, argument and relational phrases. It also predicts the associations of entities, argument and relational phrases, which are performed at three steps: Firstly, argument phrases are mapped to biomedical concepts/individuals using lexical matching techniques supported by a background

knowledge [58]; Secondly, relational phrases (semantic indicators) are mapped to semantic/ontological predicates using indicator rules; Finally, association of biomedical concepts/individuals and relationships, semantic propositions, are predicted using a semantic disambiguator supported with a set of rules (indicator rules) and the background knowledge [142]. Thus, the semantic analyzer used a set of rules and background knowledge to disambiguate semantic associations of concepts with their relationships.

The biomedical domain is modeled for structuring a set of biomedical artefacts (K_p) and for developing the domain conceptualization (C_o). The model structures biomedical domain horizontally and vertically. Horizontally, the domain is partitioned into M disjoint sub-domain categories, where each of them can further partitioned into disjoint semantic categories and then more fine-grained concepts in a disjoint manner. Vertically, the domain is structured to *N* knowledge levels in a taxonomic manner. The taxonomic structure, linked by *ISA* relation, is modeled to satisfy the partial order relation properties. Thus, the biomedical knowledge is modeled to *N* knowledge levels hierarchically, where each level is partitioned disjointly to *M* domain and semantic categories and fine-grained concepts. In chapter 5, the biomedical domain model is instantiated at two knowledge layers: upper and lower knowledge. The upper knowledge layer represents upper ontology structure and its conceptualization [314], [315]. It is comprised of two knowledge levels: top and bottom. While the top knowledge level is modeled with sub-domain categories, such as *Anatomy* and *chemicalsAndDrugs*, the bottom knowledge level is modeled with sub-partitions of sub-domain categories, such as *diseaseOrSyndrome* and *pathologicFunctions*.

The lower knowledge layer is modeled with a set of biomedical knowledge artefacts (K_p) abstracted from biomedical texts. A set of knowledge artefacts is comprised of very fine-grained biomedical concepts subsumed by semantic categories (coarse-grained concepts) in the upper ontology structure. Integration of the two ontology structures is, therefore, performed using concept-overlap technique. While biomedical concepts are disambiguated, their semantic classes are also predicted and assigned to them. This created semantic associations between concepts in the lower ontology structure and the bottom knowledge level of upper knowledge layer. This association, referred as concept-overlap, is used to merge the two ontology structures.

After the conceptual knowledge is modeled and structured, a formal ontology model axiomatizes with a logical formalism. Consequently, unambiguous representation of biomedical knowledge (K_o) is formulated as a combination of six logically designed ontological axioms. A primitive axiom (Ψ) is a set of atomic and defined concept and role axioms. A primitive attribute axiom (ψ) is a set of concept, individual and role attribute axioms. A concept taxonomy axiom (H_c) is a set of concept subsumption relation axioms. A role taxonomy axiom (H_r) is the set of role subsumption relation axioms. A non-taxonomic relation axiom (Φ) is the set of non-taxonomic relation axioms. The assertional axiom (A) is a set of individuals, their associations and instantiations. Thus, the biomedical ontological knowledge (K_o) is formulated as six tuples:

$$K_o \equiv (\Psi, \psi, H_c, H_r, \Phi, A) \tag{3.1}$$

Figure 3.1 depicts the general architecture of the proposed ontology acquisition framework interacting components. In the architecture, the components are denoted with shapes of drawings. Cylinders denote repositories and rectangles denote processes where as document collection denotes a set of domain texts. Arrows denote component interactions.

The framework has three major components: semantic disambiguation, conceptual disambiguation and ontologization. Semantic disambiguation enables to acquire a set of biomedical artefacts (K_p) from domain knowledge sources (K_s) suggested by the background knowledge (K_b).Conceptual disambiguation enables to model and constructs domain ontology structure (G_o).Ontologization enables to interpret the conceptual ontology into a set of OWL Dl axioms (K_o), producing OWL DL ontology (K_o) .

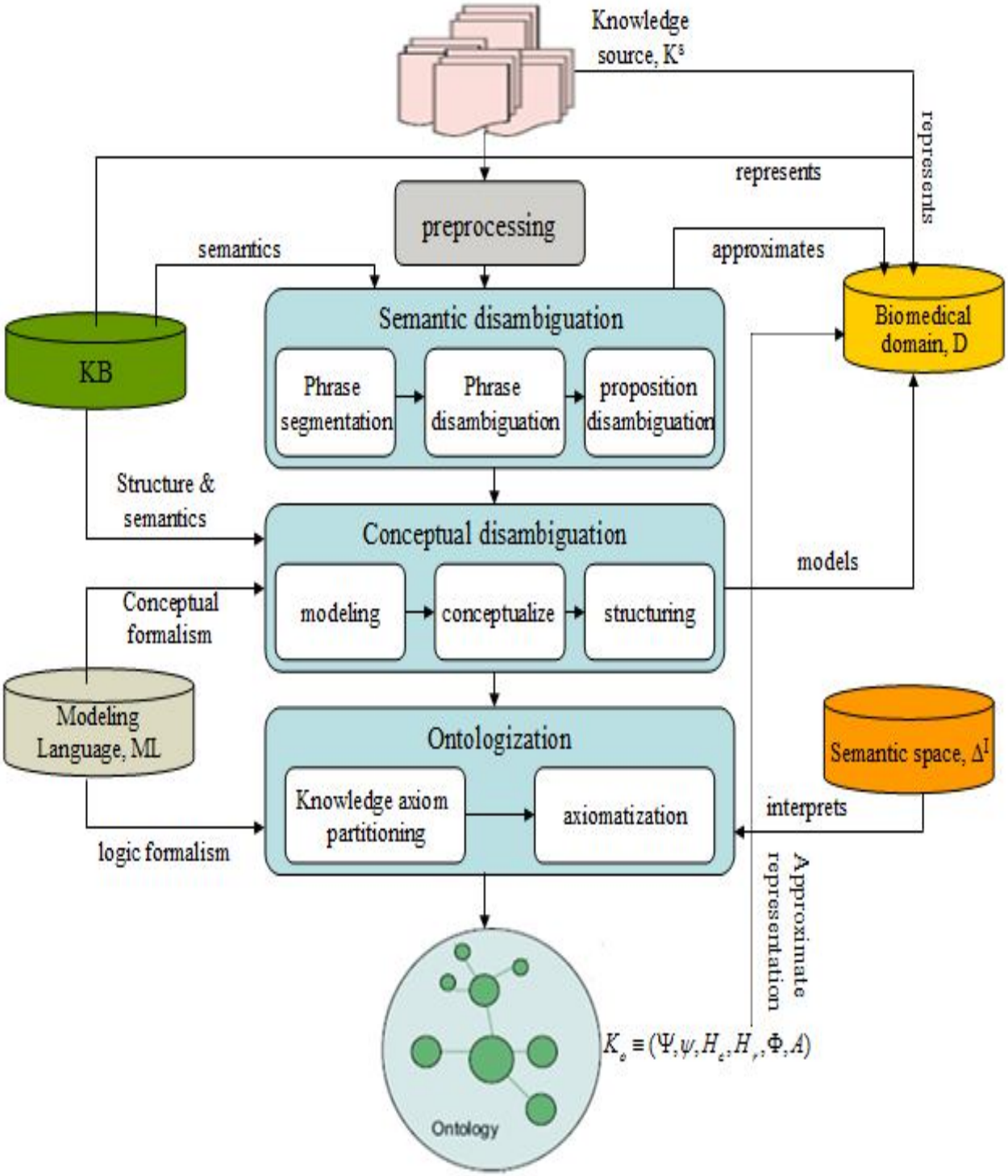

Figure 3.1 – Knowledge Acquisition Framework Architecture

3.1 Preprocessing

Human knowledge about a biomedical domain world, D_w, the knowledge of people about the biomedical domain, is described with medical language texts. This knowledge is expressed in a way that human can comprehend and understand it. In the expression of medical language syntax and semantics, biomedical artefacts are expressed either explicitly or implicitly. This knowledge is crucial for knowledge-enable applications and services, and thus, must be expressed in a way that automated systems able to understand it. Thus, for accurate and quality disambiguation and interpretation of the biomedical artefacts, alphabets (e.g. diacritic marks), numerals, punctuations and span of texts, are not necessary and meaningless for knowledge-aware systems. Furthermore, interpreting these terms are very difficult and complex for the state of the art techniques. Thus, to prepare the biomedical knowledge source for accurate, quality and computationally efficient acquisition and representation of the knowledge artefacts, the biomedical texts must be cleaned from these texts. In addition, most knowledge sources might have meta-information that describe them as a whole and other visual and tabular effects in them (e.g. figures, tables and graphs) to clarify its contents. As the focus of the knowledge acquisition and representation systems is on the content of the sources rather than meta-knowledge, the meta-information is also eliminated and cleaned.

Unstructured biomedical knowledge source: the unstructured knowledge source is a collection of biomedical documents with different formats, types and categories in the domain. Particularly, the unstructured knowledge source is the biomedical textual document collection, which can be normalized into phrases, clauses or sentences containing biomedical artefacts and their interactions either implicitly or explicitly stated in the text. More formally, for domain knowledge source (S_{d_i}) for each document (d_i), the domain knowledge source collection (S) is the set of biomedical documents, which comprised of biomedical artefacts and their interactions (S_{d_i}), is formulated as:

$$S \equiv \{S_{d_1}, S_{d_2}, ..., S_{d_n}\}, \text{ where } , S_{d_i} = \{c_{i1}, c_{i2}, ... c_{ik}, r_{i1}, r_{i2}, ... r_{in}\} \quad (3.2)$$

Where c_i denotes a set of biomedical concepts and r_i denotes a set of conceptual relations in the document collection. This way, *m* number of biomedical documents are collected and prepared as the knowledge sources of the proposed framework.

Biomedical knowledge source format: to ease document analogue, every biomedical knowledge sources are converted to one and common standard format. Originally, every biomedical knowledge sources may have its own format, such as pdf, .doc and xls, which are required to be converted into XML format. For its persistency and standard representation, XML based format is used as a common representation format. Thus, every knowledge source, for instance a textual document (d_i), are converted to XML format. It means that the whole text is tagged with XML and stored with XML extension.

Meta-information: is additional information about biomedical knowledge sources. For example, information related to size, location, format, type, category, belongingness and authors of the knowledge source, is meta-information and talks about the knowledge source or document. This information is not describing domain knowledge facts, and thus, they are not mentioning domain artefacts or entities and their interactions. Consequently, they are cleaned from the knowledge source collection.

Titles and sub-Titles: they are title texts of each knowledge source, textual document, and texts in sub-titles. These texts are kept delimiting with full stops followed by a space and a capital letter. Similarly, in abstracts and summary texts, titles are dropped and the texts are kept merging with the body of the documents.

Visual information: most knowledge sources are elaborated using diagrams, graphs, figures and charts. However, these don't express domain artefacts to be abstracted with knowledge acquisition frameworks. However, captions may provide important domain artefacts. Thus, all visual information such as figures, graphs and diagrams are removed except their captions.

Tabular Information: biomedical knowledge sources may be elaborated using tabular information, such as numerals, using tables and other columnar presentations. Though this information describes domain artefacts, their acquisition is very problematic. In this research,

therefore, tabular information is supposed to be eliminated from each biomedical knowledge sources.

Acronyms or Abbreviations: all acronyms or abbreviations, except standard and Author defined ones, are changed into their long form. If the symbols in the acronyms are separated using full stop or space, the full stop or space is eliminated and the symbols are capitalized.

Punctuations: all punctuations, except comma, colon, semi-colon and full stop, are eliminated in each knowledge sources. Exclamation marks, question marks and full stops that are considered at the end of a sentence are replaced with full stop followed by a space and a capital letter. Where as, exclamation marks, question marks and full stops that are assumed to appear in other locations (if any) are eliminated.

Numbers and Amounts: these are quantities or measurement values in the knowledge sources. Numbers and amounts are eliminated during knowledge acquisition if they are not referring to amounts, such as volumes and distances, however, proportions such as percentages and fractions are replaced with the corresponding textual representations.

Generally, preprocessing knowledge sources is contextual, which can be preprocessed depending on the contexts of the next process to be applied on the knowledge sources. In this research, the objective is to remove potentially all natural language expressions that have no equivalent scenario specific semantic representations in the background knowledge. In this context, the above-mentioned expressions are eliminated even if perfectly accurate preprocessing is an ideal task.

Biomedical Text Collection (BTC): after preprocessing of the biomedical knowledge source collected from the different source repositories, bioMed text collection is built as a set of cleaned textual biomedical knowledge source. Every sentence is delimited with a full stop and a space followed by a capital letter. Thus, the bioMed text collection, **bioMed**, is formulated as a set of sentences (s_i) filtered out after cleaning the original knowledge source collection.

$$bioMed \equiv \{ s_1, s_2, ..., s_n \} \tag{3.3}$$

Where, *n* is the number of sentences in the bioMed text collection.

3.2 Semantic Disambiguation

Meaning disambiguation and interpretation are modeled as *situation-specific scenario representations* of domain texts, where elements of the representation denote concepts and their associations. In this context, the extent of understanding meaning is reflected by the ability and quality of responding questions about scenarios the text describes. For example, in a text:

(1) The reason behind this study was to compare the efficacy and safety of intra-articular triamcinolone hexacetonide and triamcinolone acetonide in children with oligoarticular juvenile idiopathic arthritis.

Situation-specific scenario representation presents that the *disorder* described in the text is *juvenile idiopathic arthritis* or that there are two drugs, *triamcinolone hexacetonide* and *triamcinolone acetonide*. Moreover, it presents that *disorder* is *childhood arthritis*; *oligoarticular juvenile idiopathic arthritis* is a disease; *intra articular triamcinolone hexacetonide* and *triamcinolone acetonide are* drugs. Thus, semantic analysis demonstrates how detail is situation-specific scenario representations of a text-fragment, which is the potential capability of a semantic processor to understand it. However, most scenario representations are required to come from strong prior expectations about the way the domain might be and meaning processing involves matching, combining and instantiating these prior expectations with knowledge stated in the text fragments. For example, in text (1) above, surface (phrasal) analysis addresses the identification of *intra articular triamcinolone hexacetonide* and *oligoarticular juvenile idiopathic arthritis*, isolating the relevant strings. But, these phrases alone do not show that the first is a *drug* and the second is a *disease*. Moreover, it does not provide the information that *childhood arthritis* is another name for *disorder*.

Semantic processing is leveraged to enhance phrase analysis and incorporate semantic information with situation-specific scenario representations of biomedical texts. For example, phrases in text (1) can be mapped to concepts in the prior expectation (background knowledge), *intra articular triamcinolone hexacetonide* mapped to *Triamcinolone Hexacetonide*, and *oligoarticular juvenile idiopathic arthritis* mapped to *Chronic Childhood Arthritis*. From the information in the background knowledge, therefore, it is possible to determine that *Triamcinolone Hexacetonide* is a *drug* and *Chronic Childhood Arthritis* is a *disease*. Thus, disambiguation of

biomedical concepts provides enriched meaning representation of textual scenarios. Additional level of processing, however, combines these concepts into their associations that explicitly represent their interactions. Associations of concepts are referred as semantic predications or semantic propositions and they are made up of arguments and syntactic indicators. For example, disambiguation of a text in (1) enables to predict a semantic proposition, *triamcinolone hexacetonide* **TREATS** chronic *childhood arthritis*.

In general, semantic disambiguation relies on the recognition of domain artefacts (concepts, roles and individuals) and prediction of their associations (semantic predicates) asserted among these entities in the knowledge source. To achieve this, the semantic disambiguator analyzes each biomedical sentence at two levels: lexico-syntactic and semantics. Lexico-syntactic analysis enables to chunk argument and relational phrases within each sentence of the biomedical text. The analysis generates a set of argument and relational phrases in addition to a set of associations between the argument and relational phrases. See section 4.2.1 below for detailed discussions of phrase segmentation and syntactic associations. Semantic analysis is performed at two stages: phrase disambiguation and proposition disambiguation. Phrase disambiguation maps argument and relational phrases to domain entities (concepts and individuals) and their associations, where as proposition disambiguation predicts semantic associations between concepts and their semantic relationships. That is, proposition disambiguation predicts semantic associations corresponding to syntactic associations.

Figure 3.2 illustrates the architecture of semantic disambiguation and its interacting components. Cylinders denote knowledge repositories and curved rectangles denote processes, where as document collections denote biomedical knowledge sources. Arrows denote component interactions. The semantic disambiguation enables to acquire biomedical artefacts (concepts, individuals, and relationships) from free texts and predicts their semantic associations. It is comprised of three components: phrase segmentation, phrase disambiguation and semantic association disambiguation. Phrase segmentation chunks argument (e.g. NPs) and relational phrases (e.g. VPs, PPs, etc) for each biomedical sentence, and then, produce syntactic associations (a set of syntactic triplets). Phrase disambiguation produces semantic mappings for each argument and relational phrases. While argument phrases are mapped into domain individuals or concepts, relational phrases are mapped into semantic (ontological) predicates.

Semantic association disambiguation enables to predict semantic relations between a pair of concepts or individuals, which results a set of biomedical semantic predictions or triples.

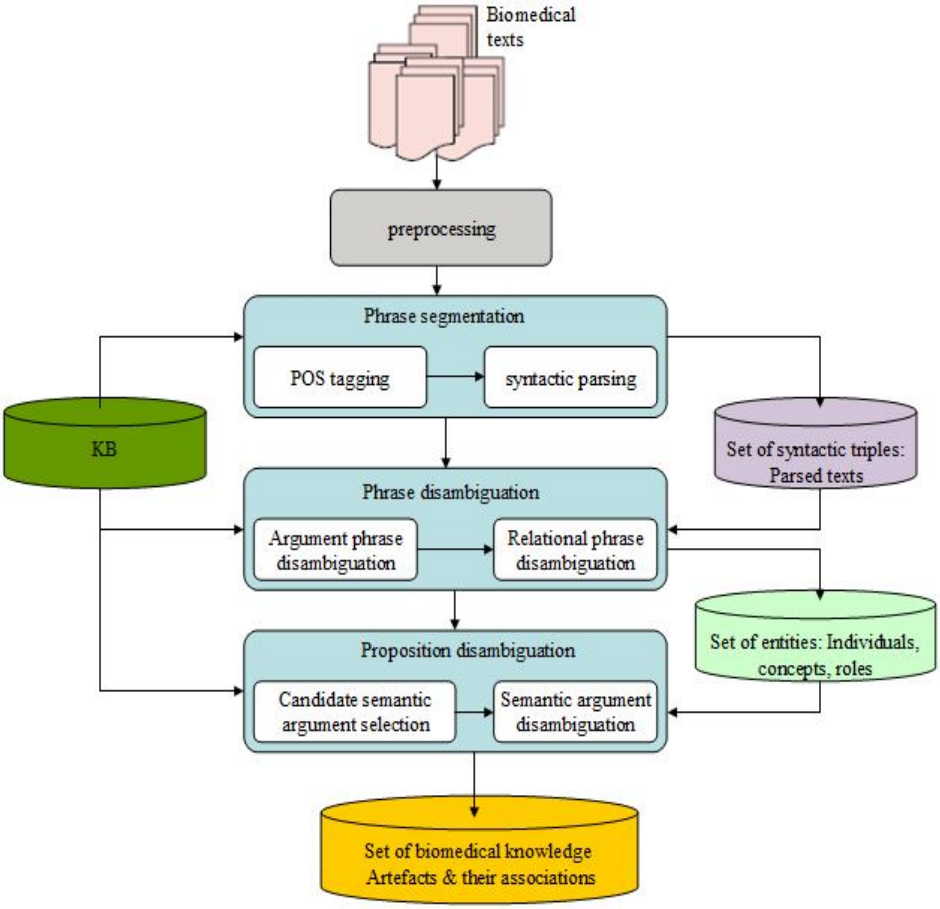

Figure 3.2 – Semantic Disambiguation Architecture

3.2.1 Phrase Segmentation

Phrase segmentation analyzes the lexico-syntactic aspects of biomedical texts, focusing on syntactic disambiguation of biomedical sentences. This is referred as surface semantics analysis as it disambiguates the surface meanings of text fragments. An example of surface semantic analysis is the generation of knowledge fragments in (2) from a textual sentence in (1).

(2) [(“”, “Study”, “efficacy and safety”), (“Intra-articular triamcinolone hexacetonide”, “in”, “children with disorder”), (“Triamcinolone acetonide”, “in”, “children with disorder”), (“children”, “with”, “oligoarticular juvenile idiopathic arthritis”), (“childhood arthritis” “synonymous”,

"disorder"),("intra articular triamcinolone hexacetonide", "is", "drug"), ("oligoarticular juvenile idiopathic arthritis", "is"," disease")]

Generally, surface semantic analysis presents each text fragment based on the following syntactic formulations:

Syntactic cues: [<subject>, or <object>, or <syntacticrelation>]
Syntactic triplets: [<Subject, Syntacticrelation, Object>]

Where, the *subject* and *object* are syntactic arguments and *syntacticrelation* (e.g. *in* and *with* in the text fragment) is semantic indicators or relational phrases in the text fragment. Arguments (e.g. *subjects* and *objects*) are referred by noun phrases, where as *syntacticrelation* (semantic indicators) are referred by verbs, nominalizations, prepositions (e.g. *in* or *with*) or comparatives in the textual knowledge sources.

Thus, each biomedical sentence is analyzed and its syntactic knowledge fragments are disambiguated and chunked in the parse tree. Syntactic disambiguation is performed by looking for *subject-semanticCues-object* patterns rooted at the main verb or in a relative clause or in a preposition in the parse tree. For example, given the textual sentence in (1), one can have the knowledge fragments in (2). In order to parse each sentence, lexical entries including multiword forms (e.g. Doppler Echocardiography in (3)) is assigned a part-of-speech label and lexical ambiguities are assigned more than one part-of-speech label. For example, the term *used* has labels "verb" and "adj" in the background knowledge, where as the term *left* has "adj," "adv," "noun," and "verb". Thus, lexical ambiguities are resolved by enabling part-of-speech taggers to consider the syntactic contexts of each sentence in the biomedical text. For example, considering the text fragment in (3), part-of-speech labeled output is provided in (4).

(3) Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes

(4) NP(Doppler echocardiography) modal(can) aux(be) verb(used) adv(to) verb(diagnose) adj(left) adj(anterior) adj(descending) NP(artery) NP(stenosis) prep(in) NP(patients) prep(with) NP(type) num(2) NP(diabetes)

Part Of Speech (POS) tagging determines the syntactic functions of words in the biomedical texts. It labels words with a particular part of speech tag based on its definition and context, relationship with adjacent and related words in a phrase or sentence. For example, for input text in (3), the POS tagged output text is as in (4). The use of a specific tagger is left to the implementation of the phrase segmentation component in which this research has used the MedPost Tagger [317] to annotate the bioMed text collection.

Syntactic analysis is based on part-of-speech labels of the input text fragments and the background knowledge, for example the SPECIALIST lexicon. One technique of syntactic analysis may be to identify simple noun phrases, i.e. noun phrases where the head is the right most elements and has no right modification. In this technique, heads are identified and terms to the left of the head are labeled as modifiers, for example the text in [3] is analyzed as in [5]. Segmentation is based on barrier words, which serve as boundaries between phrases. An example of barrier words are modals (e.g. can in (4)), auxiliaries (e.g. be in (4)), verbs (e.g. used and diagnose in (4)) and prepositions (in and with in (4)). Using these barrier words, for example, part-of-speech labeled texts in (4) is syntactically analyzed (or chunked) as in (5).

(5) [head('Doppler echocardiography')], [modal(can)], [aux(be)], [verb(used)], [adv(to)], [verb(diagnose)], [mod(left), mod(anterior), mod(descending), mod(artery), head(stenosis)], [prep(in), head(patients)], [prep(with), head('type 2 diabetes')]]

Thus, during the process of chunking, a parser uses barrier words to close the current phrase and open the next one. Any phrase containing a noun constitutes a noun phrase. The rightmost noun is labeled as *head* and terms to the left of the head, other than determiners and prepositions, are labeled as *modifiers (mod)*. These noun phrases comprise of vocabularies where the concepts they refer to are computed using semantic mapping to match each noun phrase to concepts in the background knowledge. The semantic mapper examines all the term combinations in the noun phrase and then determines the best match, taking into account term variations (e.g. inflectional and derivational) and allowing for partial and multiple matching.

The syntactic phenomena, such as *verbs*, *prepositions*, *nominalizations* and the *head-modifier* relations, are semantic indicators and they are mapped to semantic predicates in the background knowledge. For example, the semantic indicators in (5) are the verb *diagnoses*, the prepositions

in and *with* and the *modifier-head* structure in the noun phrase whose head is *stenosis*. Semantic indicators are syntactic predicates that anchor the disambiguation and interpretation of syntactic structures as semantic propositions. Consequently, disambiguation of syntactic associations asserted in the input text depends on the syntactic information contained in the parse structure, which are the argument and syntactic relation identification. Argument identification is constrained by a grammar that establishes a syntactic relation between semantic indicators and heads of noun phrases serving as arguments, i.e. arguments syntactic position relative to the semantic indicator. Grammar rules can be stated in general terms for each class of indicators, such as verbs, prepositions and nominalizations. For example, the argument identification rules for verbs stipulate that subjects must occur to the left of the verb and objects to the right.

Although the proposed framework is generic and a particular grammar rule adoption is left to the implementation of the framework, this research adopted logic-based grammar formalism for shallow syntactic analysis as implemented in minimum commitment specialist parser [328]. Thus, a grammar either for shallow, dependency or deep syntactic analysis can easily be adopted into the framework with complexity and effectiveness trade off. The syntactic constraint imposed by the grammar serves as a necessary condition for disambiguating and interpretating syntactic indicators and its arguments as syntactic associations. In (5), for example, the grammar rules applied to the verb *diagnose* must limit the subject of the verb to the noun phrase *Doppler echocardiography*. The object, however, can be any of the three noun phrases to the right of *diagnose*: *left anterior descending artery stenosis*, *patients*, or *type 2 diabetes*. Such syntactic ambiguities are resolved by the syntactic analyzer. In this research, however, semantic constraints are applied in determining which of the three is the right object of the verb *diagnose* in (5).

Generally, syntactic structures of text fragments are expressed into syntactic associations that form *subject-syntacticrelation-object* triplets. The subjects and objects are noun phrases and syntacticrelations are semantic indicators such as verbs, prepositions or nominalizations. These structures enable computation of lexical matching between terms in the syntactic associations and terms referring to entities in the background knowledge. Figure 3.3 depicts the architecture of phrase segmentation and interaction of its components. In the figure, cylinders are repositories

and curved rectangles are processes. The document representations are unstructured knowledge sources and arrows are component interactions.

The phrase segmentation architecture has three major interacting components: lexical analysis, syntactic analysis and syntactic disambiguation. The lexical analyzer determines the functional categories of each word in a sentence, referred as POS tagging. Syntactic analyzer determines the syntactic structures of each sentence by identifying the argument (subject and object arguments) and relational phrases (syntactic indicators). Syntactic disambiguation enables to predict the syntactic associations among subject and object arguments, and syntactic indicators, referred as syntactic associations or triplets.

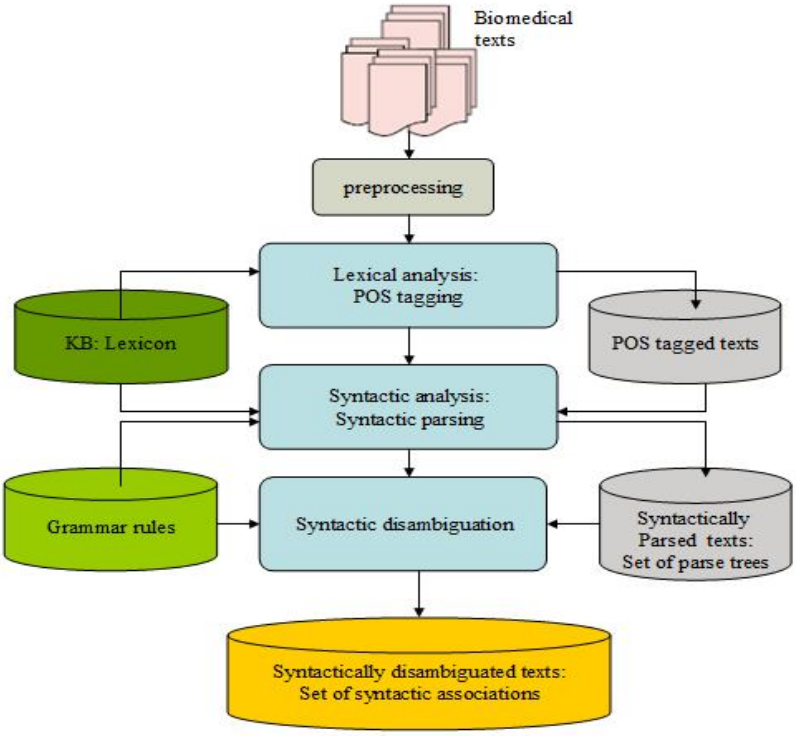

Figure 3.3 – Phrase Segmentation

3.2.2 Phrase Semantic Disambiguation

Situation-specific scenario representations of biomedical texts require disambiguating explicitly stated contents and the implied semantics. Disambiguation of these semantics requires to instantiate constitutes of the syntactic structures, referred as semantic interpretations. Phrase semantics disambiguation models the interpretation of constitutes of syntactic structures,

syntactic arguments and syntactic indicators, to participating entities in the background knowledge. This enables to disambiguate situation-specific scenarios where its elements are denoted with entities (concepts and individuals) and their associations (relations) in the biomedical texts.

Thus, the disambiguation model involves computation of lexical matches between terms that refer to domain entities in the background knowledge and terms mentioned in the biomedical text, for example Noun Phrases (NPs) and Verb Phrases (VPs), prepositions and nominalizations. The model looks for modifier-head relations in the syntactic structures and computes lexical matches with terms referring to domain entities in the background knowledge. The model also looks for semantic indicators, which relates two noun phrases (argument phrases) from the syntactic structure, and computes semantic matches with conceptual relations in the background knowledge. In a scenario representation, therefore, as many as possible *subject* and *object* terms and *syntacticrelations* from the syntactic structure match with *subject* and *object entities* and *semanticrelations* respectively in the background knowledge.

More formally, let t_s and t_o are syntactic subject and object of a syntactic structure of a text fragment respectively. And let S_t and S_o are set of synonymous terms referring to semantic subjects and objects of scenarios in the background knowledge respectively. Then, lexical matching computes the similarity of t_s with $S_t = \{t_1, t_2, ..., t_n\}$ and t_o with $S_o = \{t_1, t_2, ..., t_n\}$. Terms with the highest similarity values are ranked as a matching term and concepts referred by these terms in the background knowledge are taken as an interpretation of the syntactic subject t_s or object t_o. Syntactic relations are interpreted based on their syntactic and semantic functions in the syntactic structure. In this research, however, semantic indicators are interpreted using rules, which map the syntactic indicators to their corresponding semantic predicates in the background knowledge. Each mapping rule associates a set of semantic indicators with a semantic predicate defined in the background knowledge. This set of rules is referred as indicator rules hereafter.

Thus, indicator rules are required to disambiguate and interpret biomedical texts, for example the indicator rules for texts in (5) are as provided in (6). The syntactic phenomena, such as part-of-

speech or syntactic structure, occurs to the left of the rules (or colon) and semantic predicates occur to the right of the rules (or colon).

```
(6)  (Diagnose, verb): DIAGNOSES
     ([Left anterior descending artery, stenosis], Mod-head):
     LOCATION_OF
     (In, preposition)  : OCCURS_IN
     (With, preposition) : CO-OCCURS_WITH
```

The matching function works with syntactic information in the syntactic structures and semantic information in the background knowledge. That is, the phrase semantics disambiguator uses lexical information (e.g. NPs and VPs) and terms that refer to entities in the background knowledge (e.g. a set of synonymous strings referring to each entity) for lexical matching computation. Each entity in the background knowledge has a set of associated terms/phrases expressing it and a term in a text fragment matches the entity if that term is a member of the associated terms for that entity or one of its specializations or generalizations. In the matching computation, best matching is determined using a scoring function that assess the degree of match, looking for scenarios with the maximum number of matching text fragments, and in the case of a tie preferring a scenario with the maximum number of entities potentially matching terms in the text fragment. When the scoring function computes irresolvable tie values, it is responsible to handle disambiguation of word senses and semantic relations. For example, in the case of text fragments extracted from (1), for the best match with *anatomical disorder* scenario, the *disorder* in the text fragment is taken to mean the anatomical *disorder* (≈ word sense or context) as opposed to *visual disorder* or *vocal* disorder.

Phrase semantics disambiguation is carried out at two levels: the matching of syntactic arguments (e.g. NPs) to semantic arguments (e.g. concepts), and the matching of syntactic indicators (e.g. prep. *with*) to semantic predicates (e.g. CO-OCCURS_WITH). Consequently, disambiguation of syntactic arguments is comprised of five major components in which the detailed description can be found in [58]: lexical variant generation, candidate identification, candidate evaluation, mapping construction and Word Sense Disambiguation (WSD). Variant generation determines the variant generators of noun phrases and candidate identification looks for and determines candidate terms in the background knowledge. Candidate evaluation

computes the matching strength of each candidate terms. Mapping construction computes the combined matching strength and selects the best matches. The word sense disambiguator determines the semantics of noun phrases based on the surrounding texts. Particularly, the word sense disambiguator resolves if the input text maps to two or more candidate terms of the background knowledge.

In variant generation, variants are essentially consists of one or more noun phrase terms together with all of its spelling variants, abbreviations/acronyms, synonyms, inflectional and derivational variants and meaningful combination of these. The validity of the generators and the variants are determined based on their existence in the background knowledge. That is, they must be either meaningful single words or multi-word terms that exist in the background knowledge. Candidate identification is based on rules: every term, referring to entities in the background knowledge, contains one or more terms in the text fragment (e.g. NPs). The assumption is that all terms referring to entities in the background knowledge are potential information of the matching function. Thus, all terms referring to domain entities in the background knowledge is collected as a member of the candidate set. Candidate evaluation computes matching scores for each candidate term and determines the matching strength of the terms using evaluation metrics and then orders the candidates by the mapping strength. The evaluation metrics are centrality, variation, coverage and cohesiveness [187]. Centrality measures the involvement of head words in the candidate term. Variation measures the average variation distance of a candidate term to their variants, for example, spelling variation, inflectional and derivational variants, synonymous, acronym/abbreviation or combination of them. Coverage determines the amount of matching words between candidate terms and the noun phrases, where as cohesiveness measures word links or positions in each matching terms.

Mapping construction combines the candidates involved with disjoint parts of terms in the text fragment and then re-computes the matching strength based on the combined candidates. After computation, it selects candidates that have the highest score to form a set of best matching terms of the original terms in the text fragment. In case of ambiguity, a tie of matching, the mapping function applies word sense disambiguation based on syntactic and semantic information in the chunked text and background knowledge. For example, in a phrase “ocular complication” (7), terms scoring 694 and 861 respectively are in a tie.

```
(7)      861   Complications (Complication) [Pathologic Function]
          861   complications (Complication Aspects) [Pathologic Function]
          694   Ocular (Eye) [Body Part, Organ, or Organ Component]
          694   Ocular (Vision) [Organism Function]
          694   Ocular (Ocular (qualifier)) [Spatial Concept]
```

The word sense disambiguator chooses one of the matching terms proposed by the mapping function. Generally, interpretation of syntactic arguments is determined by computing lexical matches and imposing syntactic and semantic constraints. The syntactic constraint enables to determine noun phrases from the syntactic structures and semantic constraints enable to resolve ambiguities in multiple mappings. The mapping function computes the similarity of syntactic arguments (e.g. NPs) to terms referring to semantic arguments (e.g. concepts), where similarity is computed by the scoring function during candidate evaluation.

In disambiguating syntactic indicators, such as verbs, prepositions, nominalizations and comparatives, all semantic indicators are chunked and analyzed to group them according to their semantic correspondence with semantic predicates in the background knowledge [142]. The syntactic contexts of semantic indicators are determined using their syntactic argument expressions, and semantic contexts of the semantic predicates are determined using their semantic argument expressions. After chunking of semantic indicators, they are classified under each semantic predicates extracted from background knowledge. Then, semantics correspondence rules (indicator rules) are built to develop the matching of semantic indicators to semantic predicates in the background knowledge.

Generally, relational phrase disambiguation is a rule construction function that chunks semantic indicators, retrieves semantic predicates and constructs matching rules between syntactic indicators and semantic predicates. The assignment of semantic indicators to semantic predicates is performed with a semantic classifier, which understands the syntactic and semantic functions of them. In this research, interpretation of syntactic indicators and construction of indicator rules are performed manually for semantic predicates in the background knowledge. For example, in a text fragment in (8), there is a rule in (9) that links the nominalization *treatment* (semantic indicator) with the semantic relation TREATS (semantic predicate).

(8) *Treatment of fracture with surgery*

*(9) (*treatment, *nominalization) :* (TREATS, *semantic predicate)*

Thus, syntactic relation disambiguation is comprised of four components: syntactic relation chunking, context identification (e.g. *NPs*, *VPs*, *nominalizations*, *etc*.), semantic predicate extraction, classification and rule construction. Syntactic relation chunking segments each sentence of the knowledge source and identifies syntactic indicators (e.g. VPs, prepositions, nominalizations, comparatives, etc) that link argument phrases (NPs). Context identification determines whether the syntactic indicator is verb, nominalization, preposition, and so on. The semantic predicate extraction enables to acquire semantic predicates either explicitly or implicitly (inferred ones) stated in the background knowledge. While semantic classification assigns each semantic indicator to one or more semantic predicates as its category, rule construction builds mapping rules based on the information on categories of semantic indicators, for example building indicator rules as in (9).

Figure 3.4 illustrates architecture of phrase semantics disambiguation, where two components are interacting: argument and relational phrase disambiguation. The architecture gets input from phrase segmentation and produces a set of domain artefacts: individuals, concepts and ontological predicates. Argument phrase disambiguator generates a set of biomedical individuals and concepts, whereas, the relational phrase disambiguator produces a set of indicator rules, which associate semantic indicators with semantic predicates.

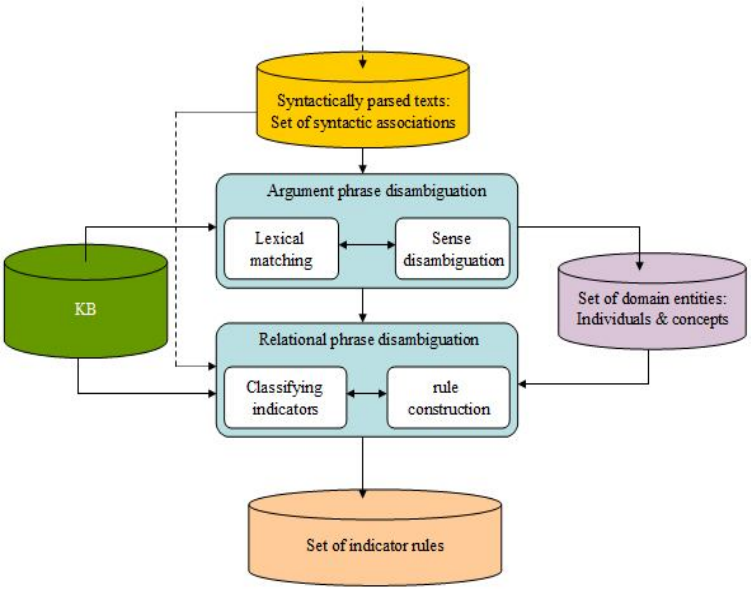

Figure 3.4 – Phrase Semantics Disambiguation

3.2.3 Proposition Disambiguation

After domain entity disambiguation, referring expressions such as *fluoroquinolones* (10) are augmented with concepts, individuals and their semantic classes. Relational expressions such as *treatment* in (10) are also augmented with semantic predicates and their relation types for predicting semantic propositions. Disambiguation of semantic associations depends on these analyses supported with the background knowledge and driven by syntactic cues or semantic indicators, such as verbs, prepositions, nominalizations and head-modifier relations. Semantic restrictions are enforced by meta-rules stipulating that *identified semantic associations must be sanctioned by predications in the background knowledge*. For example, the rule may ensure that syntactic arguments associated with *treatment*, in the analysis of (10), must be mapped to concepts in the background knowledge with one of its parent concept match with permissible argument configurations for TREATS. In (10), the semantic classes *Pharmacologic Substance* and *Disease or Syndrome* fulfill these requirements.

(10) New fluoroquinolones such as ofloxacin are beneficial in the treatment of chronic obstructive airways disease exacerbation requiring mechanical ventilation

Where as, syntactic constraints for argument identification is controlled by statements expressed in the grammar rules. For example, rules for nominalizations may state that one possible argument configuration is that, for the *object* marked by the preposition *of* occurring to the right of the nominalization, one possible location for the *subject* is anywhere to the left of the noun phrase containing the nominalization. For semantic proposition interpretation of *treatment* in (10), for example, choosing the noun phrase *ofloxacin*, which maps to a concept with semantic class Pharmacologic Substance, as subject and *chronic obstructive airways disease exacerbation*, which maps to a concept with semantic class Disease or Syndrome, as object allow both constraints (syntactic and semantic) to be satisfied. Thus, the final interpretation is the semantic proposition in (11), where concepts from the background knowledge are arguments of the predicate.

(11) Ofloxacin TREATS *Chronic obstructive airways disease exacerbated*

Figure 3.5 illustrates the architecture of semantic proposition disambiguation. In the architecture, two components are interacting: candidate argument selection and argument sense disambiguation. Candidate argument selection produces a set of candidate biomedical entities (e.g. individuals and concepts), where as sense disambiguation interprets and produces the correct arguments of each semantic predicate in a sentence. Lastly, the proposition disambiguation produces a set of semantic associations, semantic predictions, as interpretations of each syntactic association.

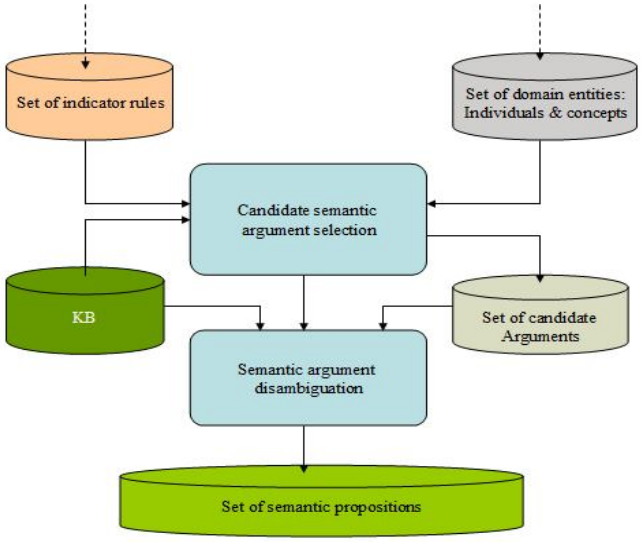

Figure 3.5 – Semantic Proposition Disambiguation

For selecting the candidate arguments and disambiguate them, the following steps are followed.

- For each sentence, *s*, identify and collect all syntactic arguments, *args*, of a semantic indicator, s_i within *s*;
- For each sentence, *s*, identify and collect semantic mappings, *semArgs*, of the syntactic arguments of semantic indicator, s_i, within *s*;
- Identify and collect the semantic mapping, sem_s_i, for the semantic indicator, s_i, in *s*;
- Identify or disambiguate a set of correct semantic arguments for the semantic indicator, sem_s_i;
- Formulate the semantic triplets in the form of:

$$< subj\_semArgs, sem\_s_i, obj\_semArgs >$$

Generally, semantic proposition interpretation is achieved using syntactic and semantic analyses. The syntactic analysis produces the syntactic triplets, {*syntactic_subj*, *syntactic_rel*, *syntactic_obj*} where *syntactic_subj* and *syntactic_obj* are noun phrases, which are syntactic arguments where as the *syntactic_rel* is a semantic indicator, which link the two arguments. As a result of semantic interpretation, a set of knowledge artefacts (K_p) (*concepts, roles and their associations*) are generated to represent the domain world (D_w) as it is described or represented in the knowledge sources, bioMed text collection. The set of knowledge artefacts is then modeled to construct the conceptual ontology structure (G_o) and its conceptualization (C_o).

3.3 Conceptual Disambiguation

As abovementioned, disambiguation of biomedical artefacts depends on scenarios in the background knowledge, which had been stated either explicitly or implicitly about the way the domain world would be [46], [316]. It has also stated that disambiguation involves matching, combining and instantiating situation-specific scenarios in text fragments. Thus, understanding is modeling the interactions of situation-specific scenarios with the background knowledge. Thus, conceptual disambiguation models and structures the knowledge resides in the biomedical texts, where their interpretations are suggested by scenarios in the background knowledge.

Knowledge modeling partitions the biomedical domain to M independent and disjoint sub-domain categories, where each of them further partitioned to independent and disjoint sub-categories. The disjoint partitioning continues to N knowledge levels, where the narrowest categories are comprised of very similar class of biomedical individuals. While domain partitioning models the biomedical knowledge with M disjoint sub-domain categories and their descendent categories, generalization-specialization models the domain taxonomically with N knowledge levels satisfying partial order relation properties (reflexivity, transitivity and antisymmetricity). Thus, biomedical knowledge is modeled hierarchically in a taxonomic manner and horizontally in a disjoint manner. More formally, let m_i is the number of partitions at each

knowledge level l_i, where i runs from top $(i=1)$ to bottom $(i=N)$. Thus, the biomedical knowledge (K_{bio}) can be modeled as:

$$K_{bio} = \bigcup_{i=1}^{N}(l_i, (\bigcup_{j=1}^{m_i} m_{ij})), (l_{i+1}, m_j) \subseteq (l_i, m_j), 1 \le i \le N, 1 \le j \le m_i \quad (3.4)$$

Where, N is the depth of the knowledge hierarchies and m_i is the number of partitions at each knowledge level l_i.

Consequently, domain conceptualization (C_i) is a set of knowledge hierarchies and categories in the knowledge model, K_{bio}, which make up a particular knowledge abstraction of biomedical domain. More formally, a knowledge model (K_{bio}) can be abstracted with several conceptualizations, where each conceptualization (C_i) is implied by the knowledge model (K_{bio}), i.e.

$$K_{bio} = \bigcup_{i=1}^{m}(C_i),\ K_{bio} \succ C_i, or, C_i \subseteq K_{bio} \quad (3.5)$$

Conceptual knowledge structure (G_o) is modeled based on a conceptual modeling language formalism. For example, the digraph-based language can be used to instantiate the structuring of the biomedical knowledge conceptualization (C_o). Particularly, a Direct Acyclic Graph (DAG) is used to enhance intelligibility by supporting inferencing and tractability. Thus, structural design is based on knowledge granularities and disjoint categorizations of a domain. Where, hierarchical knowledge is structured based on subsumption relationship between granularities. Conceptually, the subsumption relation is an 'isa' link, denoted with $\subseteq$, and thus, the hierarch is a taxonomic structure. That is, 'ISA' link is a partial order relation satisfying transitivity, reflexivity and anti-symmetric properties.

Partial Order Relation Definition: a binary relation ***R*** on a set ***A*** is a partial order if and only if it is: (i) reflexive, $A_i R A_i$; (ii) Anti-symmetric, $A_i \subset A_j$, where A_j is predecessor of A_i; and (iii)

Transitive if $A_i \subseteq A_j, A_j \subseteq A_k, then, A_i \subseteq A_k$. Where, the ordered pair **<A, R>** is a POSET (*Partially Ordered Set*) when ***R*** is a partial order.

In this research, the partial order relation (R) is the subsumption relation ($\subseteq$) and the partial order set (A) is the set of biomedical concepts (C). Thus, the ordered pair is **<C,** $\subseteq$**>** and $\subseteq$ is a partial order. More formally, let C_1, C_2, C_3 be biomedical concepts along the same hierarchy, and $C_1 \subseteq C_2 \subseteq C_3$, then the following are always true based on the partial order relation definition.

$$\begin{aligned} &C_1 \subseteq C_3, transitivity \\ &C_1 \subseteq C_1, C_2 \subseteq C_2, C_3 \subseteq C_3, reflexitivity \\ &C_1 \subset C_3, C_2 \subseteq C_3, C_1 \subseteq C_2, anti-symmetricity \end{aligned} \tag{3.6}$$

Horizontally, biomedical knowledge is partitioned into disjoint and independent categories at each knowledge granularities or levels. Disjointness refers to absence of overlaps between semantic categories. Independency is related to functioning and further partitioning capability of each category. This means that each category functions independently and its partitioning is independent of other categories. An independent biomedical category, such as Anatomy, is comprised of all anatomy individuals only and can further be partitioned into its sub-categories independently. More formally, let *bioMed* denotes the biomedical domain at the highest granularity, let C_1, C_2, C_3 are sub-partitions of *bioMed* at lower granularity and C_{11}, C_{12}, C_{21}, C_{22} are sub-partitions of C_1 and C_2 respectively, then the following are always true.

$$\begin{aligned} &C_1 \subseteq bioMed\ , C_2 \subseteq bioMed\ , C_3 \subseteq bioMed \\ &bioMed\ \equiv C_1 \cup C_2 \cup C_3 \\ &C_1 \cap C_2 \cap C_3 \equiv \phi \\ &C_1 \equiv C_{11} \cup C_{12}, C_2 \equiv C_{21} \cup C_{22} \\ &C_{11} \cap C_{12} \equiv \phi, C_{21} \cap C_{22} \equiv \phi \end{aligned} \tag{3.7}$$

Figure 4.6 illustrates biomedical knowledge modeling based on disjoint partitioning and generalization-specialization or subsumption relationships. In this case, the broadest concept is the biomedical domain and the finest concept is a class of similar individuals.

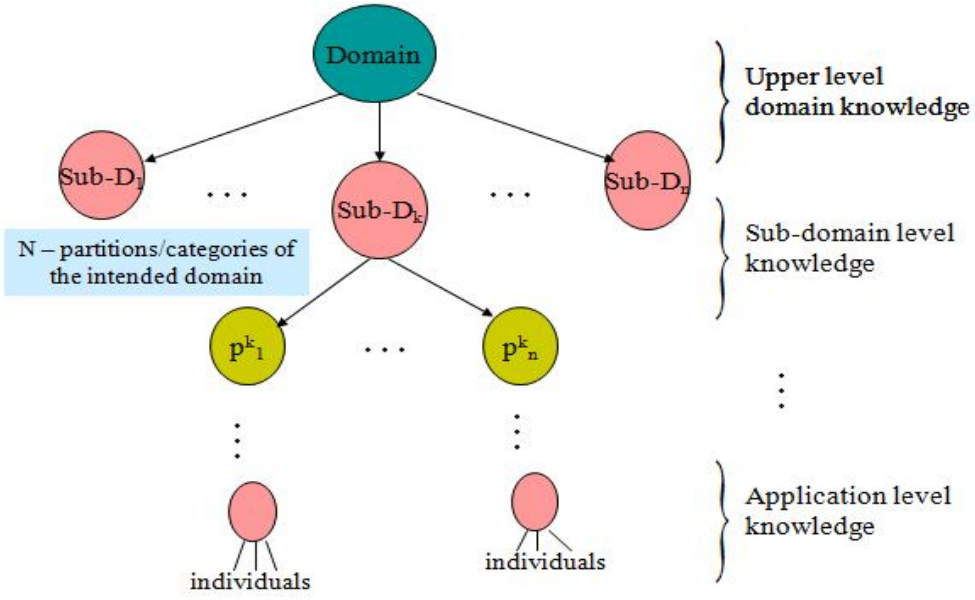

Figure 3.6 – Domain Knowledge Modeling

3.3.1 Conceptualization

A conceptualization (C_o) of a domain knowledge is independent of any modeling language formalism and of particular state of affairs, because, the same domain conceptualization (C_o) can be expressed using multiple modeling languages. For example, a conceptualization comprised with words patient, treatment, medical insurance, physician and medical unit refer to exactly the same ontological entity, namely the natural kind denoted by these terms. The conceptualization (C_o) is neutral to the actual existence of a patient p, in contrast it stipulates that if there is a situation where p exists as a patient, then p also exists as a human in the situation.

However, the conceptualization must be captured in some concrete structure to reason about biomedical characteristics. The representation must characterize it as an intensional structure that encompasses all state of affairs and independent of a particular language vocabulary. Thus, a conceptualization of biomedical knowledge structure (G_o) is intensionally modeled independent

of the state of affairs in the intended world (w) based on a set of domain artefacts and a set of n-ary relations (R) of the conceptualization (C_o). In general, a conceptualization of the biomedical knowledge is characterized as an intensional structure encompassing all state of affairs and independent of a particular modeling language. Hence, conceptualization is supplemented with the following definitions.

Definition (*Conceptualization*): a conceptualization is an intensional structure defined over (W, D, R), where W is a (non-empty) set of possible domain worlds and D is a set of individuals in the universe of interpretations and R is the set of n-ary relations that are considered in C_D. The elements $\rho \in R$ are intensional (or conceptual) relations with signatures, $\rho^n : W \rightarrow \rho(D^n)$. Each n-ary relation is a function from W to n-tuples of individuals in the domain.

For instance, one can have ρ accounting for the meaning of the natural kind *apple*. In this case, the meaning of *apple* is captured by the intensional function ρ, which refers to all instances of apples in every possible domain world. Thus, the definition considers all state of affairs in all possible domain worlds. The proposed framework, however, considers state of affairs in a world, which we name it as Intended World Structure (G_w). The intended world structure is defined based on the conceptualization (C_w), a possible world (w), and a set of domain individuals (D_w).

Definition (*Intended World Structure (G_w)*): for every world $w \in W$, according to C_w, one can have an intended world structure G_w as a structure $< D_w, R_{wc} >$ such that $R_{wc} = \{\rho(w) \mid \rho \in R\}$.

According to the definition, one can say that every intended world structure G_{wc} is the characterization of some state of affairs in a world w deemed admissible by the conceptualization C_w. That is, C_w defines all the admissible state of affairs in that domain, which could be represented by the set $G_c = \{G_{wc} \mid w \in W\}$.

However, considering a language L with a vocabulary T that contains terms to represent every entity in C_w, one can associate the intensional structure (G_w), conceptualization C_w, to the language vocabulary T as the interpretation of the intensional structure.

Definition (*Intensional Interpretation*): given a modeling language *L* with vocabulary *T* and a conceptualization C_w, intensional interpretation is given by $< C_w, \sigma >$, where: $C_w \equiv < W, D, R >$ is the domain conceptualization and $\sigma : T \rightarrow D \cup R$ is the intensional interpretation function, which assigns elements of *D* to constant symbols in T, and elements of *R* to predicate symbols in *T*.

This definition can be generalized to the conceptualization of all state of affairs exist in all possible domain worlds (W) and the set of individuals in the domain of interpretation (D) as well as to the set of all relations in the domain of interpretations (R). With these formulations and models of particular conceptualization, world knowledge K_W is formulated into four tuples for intensional structuring and interpretation:

$$K_W \equiv (W, D, R, T) \tag{3.8}$$

Where, *W* is the possible domain worlds, *D* is the set of domain individuals, *R* is the set of relations *R*, and *T* is the set of vocabularies of a modeling language *L*.

3.3.2 Structural Model

A structural model is a representation of a conceptualization (C_i) based on a particular conceptual language formalism. It is also a structural instantiation of a conceptualization using a specific conceptual modeling language, for example Direct Acyclic Graph (DAG). Thus, structural modeling is a concretization of a conceptualization using a specific conceptual language syntax and semantics to construct a knowledge structure (G_o). In this research, the conceptualization (C_o) and its conceptual structure (G_o) is instantiated using the biomedical artefacts and their n-ary relations, such as concepts and their associations. While nodes of the structure are biomedical concepts and individuals, arcs of the structure are semantic links, relationships between concepts or individuals. In the structure, however, the broader semantic categories, which are supposed to appear at the highest level of the conceptual structure, don't exist or have very limited existence in the knowledge sources.

The proposed framework disambiguates the fine-grained biomedical entities and their

associations in addition to few coarse-grained entities. Largely, the set of biomedical artefacts represent the fine-grained knowledge granularities at the lower conceptual structure (G_L).The upper conceptual structure representation comes from sub-domain categories, coarse-grained concepts and their associations, which builds the upper knowledge structure (G_u). Thus, the upper knowledge structure (G_u) is reused from existing upper-level ontologies [53] [314] or built with the help of domain experts and ontology engineers. It is, then, integrated with the lower knowledge structure (G_L) acquired from the biomedical texts.

The upper knowledge structure (G_u) is conceptualized to upper knowledge conceptualization (C_u) and the lower knowledge structure is conceptualized to lower knowledge conceptualization (C_l) of the biomedical knowledge conceptualization (C_o), $C_o \equiv C_L \cup_{\subseteq} C_U$. Thus, the conceptualization C_o is the union of the two conceptualizations, C_u and C_l, over subsumption operator, $\subseteq$, using the overlapping concepts. More formally, the integration is defined based on subsumption relation on the overlapping concepts of the two knowledge structures. The subsumption relation is defined at the point of overlapping, i.e.

$$C_o \equiv \begin{cases} C_u \cup C_l, if, C_l \subseteq C_u, and, C_u \cap C_l \neq \phi \\ non-\text{int}\, egrable, if, C_u \cap C_l = \phi \end{cases} \tag{3.9}$$

For completeness, the structural model combines two relational models: the semantic overlap and *InstanceOf* relation models. The overlap relation model computes the subsumption relation between the two conceptualizations, C_l and C_u. The overlap computation is based on semantic classes that exist in C_l and C_u, their intersections. For example, in a text (10), the concepts *Ofloxacin* and *Chronic obstructive airways disease exacerbated* with their semantic class *Pharmacologic Substance* and *Disease or Syndrome* respectively exist in C_l structure, but *Pharmacologic Substance* and *Disease or Syndrome* are also exist in the upper knowledge structure creating an overlap between the two knowledge conceptualizations. Thus, the overlap relational model creates a subsumption relation between the lower and upper ontology schema structures. It is also possible to consider the lower ontology structure (G_l) as an instance of the

upper ontology schema structure (G_u), i.e. elements in C_l can be considered as instances of elements in C_u. This is because G_l is an intensional structure and it is the instance of the upper knowledge schema structure.

The *InstanceOf* relational builds an elemental relationship between each individual and its semantic class or concept in the lower ontology structure (G_l). The model classifies each individual $i \in I$ in one and only one semantic class in C_l. To keep the consistency of the classifier, the model is constrained by the disjointness axioms at each semantic class. For example, in (10), if Ox#1 and Ox#2 are individuals belonging to *Ofloxacin*, Ox#1 and Ox#2 should not belong to other semantic classes in the conceptualization (C_o). To keep the uniqueness of the semantic class (*Ofloxacin)* for Ox#1 and Ox#2, the disjointness axiom is set between sibling semantic classes in the instantiation of the conceptualization (C_o).

Generally, a set of domain knowledge artefacts (K_p) can be articulated into a knowledge conceptualization (C_o) in terms of the possible domain worlds (W), domain of individuals (D), and a set of semantic relations (R). This set is articulated to conceptual structures supported with background knowledge for its interpretation. The structural articulation of the biomedical artefacts is represented with concrete specifications based on a conceptual modeling language constructs and primitives, leading to intensional interpretations of the structure. Lastly, the conceptualization (C_o) is extensionally interpreted using logical language primitives and constructs.

3.4 Ontologization

The biomedical conceptualization articulates knowledge artefacts (K_p) and models to a conceptual structure (G_o). However, conceptual knowledge has ambiguous interpretations, poor reasoning support and less computationally tractable [221]. For this reasons, the biomedical knowledge conceptual structure (G_o) are interpreted extensionally for their unambiguous representations and better reasoning support. In this context, ontologization presents the way how the conceptual structure is interpreted and instantiated to a set of individuals in the domain

of discourse, the intended world (w). Thus, ontologization considers the following constraints to check satisfability and redundant instantiation of individuals within multiple concepts.

Constraint 4.**1**: Every individual in the conceptualization (C_o) must be an instance of a conceptual type, semantic class, that represents a semantic category in the conceptual structure (G_o). This definition enables ontologization to interpret each semantic class to a set of individuals in the domain of interpretation (Δ^I).

Constraint 4.**2**: An individual in the conceptualization (C_o) must instantiate exactly in one conceptual type or semantic class. This definition enables ontologization to partition each individual distinctly in their semantic class.

Ontologization also considers two instance levels in the conceptualization (C_o) or conceptual structure (G_o). While the first instantiates semantic classes in the intensional structure to a set of individuals in the domain of interpretation, the second instantiates the schema structure to semantic classes in the intensional structure. Consequently, a conceptualization (C_o) is instantiated with a set of individuals in the domain of discourse (Δ^I).

Generally, semantic categories and semantic classes in the conceptual structure (G_o) and its conceptualization (C_o) is instantiated either directly or indirectly to a set of individuals in the domain of interpretation (Δ^I). The semantic classes are instantiated directly to a set of individuals in the domain of discourse, where as semantic categories are instantiated indirectly to their individuals through their instances, i.e. semantic classes. Consequently, semantic categories and semantic classes are interpreted to their extensional representation. Ontologization, therefore, maps semantic classes and categories in a similar fashion to their set of individuals in the intended domain world, *w*.

3.4.1 Interpretation Model

The interpretation model adopts set theoretic model to interpret every concept in the conceptualization (C_o) to a set of individuals in the intended world *w*, where $w \in W$, the

possible domain worlds. The extension of each biomedical concept or category is world invariant in cases where a static possible domain world *w* is considered. This is also generalized to all possible domain worlds, *W*. In this research, the interpretation model (***M***) considers the extensions of biomedical concepts or categories in a static biomedical domain world. That is, interpretation of biomedical concepts and categories are world invariant.

Definition (*Concept Extension function*): Let *W* be a non-empty set of possible domain worlds and let $w \in W$ be a specific world. The extension function χ maps a concept c to its interpretation, $c_w^{\ I} \subseteq \Delta_w^{\ I}$, which is the set of individuals in a specific world w. Formally, the extension function $\chi_w(c)$ maps a concept *c* to set of instances in a world *w*. Consequently, the extension function $\chi(c)$ provides a mapping of a concept *c* to a set of instances that exist in the possible domain worlds, *W*.

The interpretation model *M* can, therefore, be generalized as:

$$M \cong \chi(c) \cong \bigcup_{w \in W} \chi_w(c),\ \chi_w(c) = c_w^{\ I} \subseteq \Delta_w^{\ I} \tag{3.10}$$

Where, *c* is a concept in the conceptualization C_o, $\chi_w(c)$ is the interpretation of a concept *c* in a world *w*, and $\chi(c)$ is the interpretation of a concept c in the possible domain worlds, *W*. The interpretation model (M) can also be supplemented with definitions related to specialization relations and rigidity of a concept interpretation.

Definition (*Subsumption Relation*): Let c_1 and c_2 are two concepts in the conceptualization (C_o) such that c_1 is a specialization of c_2 in the conceptual structure (G_o). Then, $\forall w \in W$, the model interprets as:

$$\chi_w(c_1) \subseteq \chi_w(c_2) \tag{3.11}$$

The subsumption relation in the conceptual structure (G_o) is, therefore, interpreted as a subset relation in the semantic interpretation model, M.

Definition (*Rigid Concept*): A concept c is rigid or modally constant if and only if for

any $w_1, w_2 \in W$.

$$\chi_{w_1}(c) \equiv \chi_{w_2}(c) \tag{3.12}$$

Combining the previous definitions, for any rigid biomedical concept c, the following is valid.

$$\chi_W(c) \equiv \chi(c) \equiv \chi_w(c), \forall w \in W \tag{3.13}$$

Where, a rigid concept is one that applies to its instances necessarily, i.e., in every possible domain world.

Definition (*Non-Rigid Concept*): A concept *c* is non-rigid if and only if for $w \in W$, there is an individual x such that $x \in \chi_w(c)$ and there is a $w' \in W$ such that $x \notin \chi_{w'}(c)$.

In the conceptualization (C_o), each concept node is associated with other concept nodes with their semantic links, semantic relations. The interpretation model maps these semantic relations to associations of set of individuals. The following definitions formulate how the model interprets these semantic relations.

Definition (*Role Extension Function*): let *W* be a non-empty set of possible domain worlds and let $w \in W$ be a specific world. The extension function χ_w maps each semantic role *r* to its interpretation domain Δ^I, i.e. the set roles in the intended world *w*. More formally, the extension function $\chi_w(r)$ maps a role r to a set of role instances in a world *w*. Consequently, the extension function $\chi(r) = \chi_W(r)$ provides a mapping to the set of role instances that exist in the possible domain worlds, *W*.

Let the domain of a relation *r* in a world *w* be defined as $Dom_w(r) = \{x \mid (x, y) \in \chi_w(r)\}$ where x and y are individuals, then for $\forall w \in W$, $Dom_W(r) = (x \mid (x, y) \in \chi_W(r)\}$. Similarly, the range of a relation r in a world w be defined as $Range_w(r) = \{y \mid (x, y) \in \chi_w(r)\}$ where x and y are individuals, then for $\forall w \in W$, $Range_W(r) = \{y \mid (x, y) \in \chi_W(r)\}$. Generally, $\chi(r) = \chi_W(r)$ can be formulated as:

$$\chi(r) \equiv \bigcup_{w \in W} \chi_w(r), where, \chi_w(r) \subseteq dom_w(R) \times Ran_w(R) \quad (3.14)$$

Thus, the interpretation model is generalized as $M \equiv (w, S, I)$, where $w \in W$ is the intended world, $S \equiv (D, R)$, where D is the domain of individuals and R is the extensional relations, i.e. relations in the domain of universe. And τ is a mapping function defined as $\tau : T_w \rightarrow D_w \cup R_w$, where T_w is the vocabulary of language L in a world w. The mapping function, τ, interprets each individual in D of w to constant symbols in T of w, and element r in R of w to predicate symbols in T of w.

In general, the interpretation function maps the conceptualization (C_o) to elements (individuals) in the domain of discourse (Δ^I). Let Δ^I be a non-empty domain of interpretation and χ is the interpretation function, then χ assigns:

- An element $i^I \in \Delta^I$ for each individual $i \in D$,
- A subset $c^I \in \Delta^I$ for each atomic concept, $c \in C_o$,
- A relation $r^I \subseteq \Delta^I x \Delta^I$ for each role, $r \in R$.

Combining all, the interpretation function can be generalized to the following formulation:

$$\chi : T \rightarrow I \cup C \cup R, where, I^I, C^I, R^I \in \Delta^I \quad (3.15)$$

Figure 4.7 depicts the extensional interpretation model of biomedical individuals, concepts and their associations. It enables further understanding of how the conceptual knowledge is mapped to a set of individuals in the domain of discourse (Δ^I). The figure demonstrated the interpretation function using two individuals (Ayele and Senait), three concepts (Lawyer, Doctor, & Vehicle), and two roles (hasChild and owns). The diagram illustrated the mapping of individuals (e.g. Senait), a set of individuals (e.g. Doctor), and the binary relation (e.g. owns) in the domain of interpretation (Δ^I). The circle on the right is the domain of interpretation and the individuals, concepts and the roles are conceptual entities. The interpretation function τ map the semantics of the conceptual entities (Senait, Ayele, Lawyer, Doctor, hasChild, owns) to the domain of

interpretation (Δ^I) based on the extension function (χ).

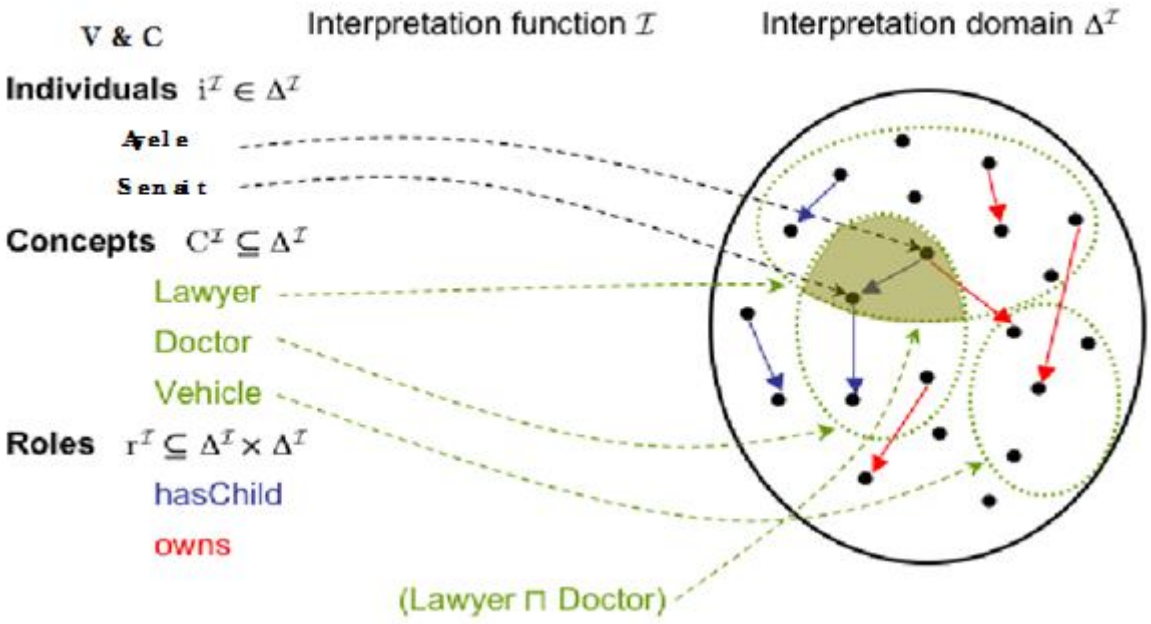

Figure 3.7 – Extensional Interpretation Model

The interpretation function, therefore, maps the conceptualization (C_o) to its semantics in the domain of discourse (Δ^I). Δ^I is a non-empty domain of interpretation and τ is the interpretation function, which assigns an element $i_i^I \in \Delta^I$ for each individual $i_i \in D$; a subset $c_i^I \subseteq \Delta^I$ for each atomic concept, $c_i \in C$; a relation $r_i^I \subseteq \Delta^I \times \Delta^I$ for each role $r_i \in R$. That is, the interpretations function $\tau : T \rightarrow D \cup C \cup R, where, D^I, C^I, R^I \subseteq \Delta^I$.

3.4.2 Construction Approach

Interpretation of the conceptualization (C_o) or its conceptual structure (G_o), represents the ontological knowledge (K_o) axioms. In the proposed framework, it is designed as eight step procedure, namely:

- Inventory of the biomedical artefacts, set of concepts (C), set of roles (R) and set of individuals (I). The inventory classifies the biomedical artefacts into individuals, concepts, roles and their attributes from the set of knowledge artefacts (K_p).

- Define the non-primitive biomedical concepts and roles based on the primitive ones. The non-primitive concepts are referred as unnamed concepts and roles. These concepts are defined based on the primitive concepts and roles using union, intersection and quantifier operators.

- *Assertional Axioms* (A): it is a knowledge comprised of individuals and their associations in the biomedical domain. It also includes the link knowledge of every individual to their semantic classes. That is, the instantiation of intensional knowledge to its extensional counterpart.

- *Primitive Attribute Axioms* (ψ): this is the axiomatic interpretation of primitive entities' attributes such as concepts, individuals and roles.

- *Concept Taxonomy Axioms* (H_c): the hierarchical knowledge in the schema and intensional structure level of the conceptual structure (G_o). It also defines the properties of hierarchical relations ('ISA',$\subseteq$) to fulfill the partial order relation characteristics or not.

- *Role Taxonomy Axioms* (H_r): the hierarchical knowledge in the schema and intensional structure level of the conceptual structure (G_o). It also defines the properties of hierarchical relations ('ISA',$\subseteq$) to fulfill the partial order relation characteristics or not. That is, the transitivity, reflexivity and anti-symmetricity properties.

- *Non-taxonomic Axioms* (Φ): it is non-hierarchical knowledge in the schema and intensional structure of the conceptual structure (G_o). It also defines the inheritability properties of each non-taxonomic relation in the conceptual structure (G_o).

3.4.3 Domain Ontology Model

The domain ontology model partitions the biomedical knowledge and its conceptual structure (G_o) into six logically consistent axioms. These are the primitive axioms (Ψ), the primitive attribute axioms (ψ), the concept taxonomy axioms (H_c), the role taxonomy axioms (H_r), the non-taxonomic axioms (Φ) and the assertional axioms (A). An ontological knowledge is

modeled in a logically consistent integration of these knowledge axioms. Thus, the formal domain ontology (K_o) model is expressed as a set of knowledge axioms, which is formulated as six tuples:

$$K_o \equiv (\Psi, \psi, H_c, H_r, \Phi, A) \tag{3.16}$$

The model interprets each tuple by mapping to the domain of interpretation (Δ^I). The logical integration model is defined as a mapping function (χ), which interprets each tuple in K_o based on the domain of discourse (Δ^I), where:

- For conceptualization (C_o), its interpretation is ${C_o}^I \in \Delta^I{}_{C_0}, where, \Delta^I{}_{C_0}$ is the possible conceptualizations in the domain of discourse (Δ^I).
- For primitive axioms (Ψ), its interpretation is $\Psi^I \in \Delta^I{}_{\Psi}, where, \Delta^I{}_{\Psi}$ is the possible primitive attribute axioms in the domain of discourse (Δ^I).
- For primitive attribute axioms (ψ), its interpretation is $\psi^I \in \Delta^I{}_{\psi}, where, \Delta^I{}_{\psi}$ is the possible primitive attribute axioms in the domain of discourse (Δ^I).
- For concept taxonomy axioms (H_c), its interpretation is ${H_c}^I \in {\Delta_c}^I, where, {\Delta_c}^I$ is the possible concept taxonomy axioms in the domain of discourse (Δ^I).
- For role taxonomy axioms (H_r), its interpretation is ${H_r}^I \in {\Delta_r}^I, where, {\Delta_r}^I$ is the possible role taxonomy axioms in the domain of discourse (Δ^I).
- For role taxonomy axioms (Φ), its interpretation is $\Phi^I \in {\Delta_\Phi}^I, where, {\Delta_\Phi}^I$ is the possible role taxonomy axioms in the domain of discourse (Δ^I).
- For assertional axioms (A), its interpretation is $A^I \in \Delta^I{}_A, where, {\Delta_A}^I$ is the possible assertional axioms in the domain of discourse (Δ^I).

The integration model, therefore, maps each tuple in the ontological knowledge formulation to

its semantics in the intended world *w*. The mapping function λ is defined formally as a projection of the domain of interpretation onto the cross product of the six tuples in K_o :

$$\lambda : \Delta^I \rightarrow \Psi^I \times \psi^I \times H^I{}_c \times H^I{}_r \times \Phi^I \times A^I \tag{3.17}$$

3.5 Semantic Space

The unstructured biomedical knowledge source (W_T) is a heterogeneous source of knowledge in the domain. The proposed framework considers the knowledge source a comprised of a set of individuals *I*, which is a subset of all individuals in the possible domain world, *W, i.e.* $W_T \subseteq W$. This implies that $I \subseteq I_W$. The set of all individuals, in the possible domain world *W*, is the interpretation or semantic space in the specified domain. In this research, semantic space is also referred as the domain of interpretation. Formally, the semantic space is denoted using delta symbol raised with capital *I* (Δ^I).

$$\begin{aligned} &\Delta \equiv \{i_1, i_2, ..., \ i_{W_N}\}, \ i.e. \\ &\Delta \equiv \bigcup_{i=0}^{i=W_N} \{i_i\} \end{aligned} \tag{3.18}$$

Where, W_N is the number of individuals in the possible domain world (*W)*. The set of individuals in the semantic space can be structured and conceptualized in several ways based on the purpose of conceptualization and the intended applications. That is, there are several possible conceptualizations of the set of individuals in the semantic space (Δ) of the domain. A specific unstructured biomedical knowledge source can provide one possible conceptualization of the domain. That is, each semantic class is a set of related individuals in the semantic space (Δ^I). In turn, these classes of individuals and their relationships must correspond to semantic concepts and their interactions in the conceptualization (C_o), i.e.

$$C_o^I \subseteq \Delta^I \tag{3.19}$$

Chapter Four Instantiating the Proposed Framework

Once the proposed framework designed, it is necessary to instantiate the framework based on domain text collection and knowledge structure as domain knowledge source and background knowledge respectively. Consequently, in this chapter a knowledge source is instantiated with *bioMed text collection*, which is prepared as a combination of texts in the PubMed, Genia, CLEF, and ClinicalTrial corpus collections. The background knowledge is instantiated with a specific version of the UMLS knowledge, UMLS 2012AB.

MetaMap and enhanced semRep tools are used to instantiate the disambiguation of biomedical concepts and their associations in the *bioMed text collection* respectively. MetaMap enables to disambiguate biomedical concepts from biomedical sentences. The MetaMap word sense disambiguator is used to resolve multiple mappings of a term in a text to background knowledge. SemRep enables to disambiguate biomedical concept associations, semantic predictions. However, the coverage of interpreting concept associations is limited to semantic predicates defined in the UMLS semantic network. In this research, coverage of disambiguating semantic associations between concepts is enhanced with an algorithm that generates a matrix of concepts and look for their associations in the background knowledge. The quality and accuracy of the semantic associations are validated with the existence of a pair of concepts in a sentence and their semantic association in the background knowledge. The algorithm prunes unnecessary concept pairs, for example self and non-existent associations.

The ontology structure (G_o) is instantiated with a Direct Acyclic Graph (DAG) formalisms at two knowledge levels, upper and lower ontology structures. While the upper ontology structure is reused from existing ones, the lower ontology structure is acquired from the bioMed text collection. A concept-overlap technique is applied to integrate the two ontology structures. Plus, pruning and alignment are applied to construct 135 ontologies rooted at each of 135 semantic categories. Finally, the ontology structure (G_o) is interpreted with OWL DL primitives and constructs, producing a set of OWL DL axioms (K_o). A restriction-based interpretation is used to interpret each biomedical concept associations acquired from the bioMed text collection. In

this study, existential restriction-based interpretation is used after an experimental analysis of determining the trends of biomedical concept associations.

4.1 BioMed Text Collection

In this study, the bioMed text collection is prepared as an instance of the biomedical knowledge source. It is composed of 55,536 textual documents, where each document contains an average of 30 sentences. It is used to experiment the instantiation of the proposed framework. Each textual document is tagged with an XML version 1.0 and encoded with UTF-8. Within the documents, the additional information is the documents' meta-data such as their file name, the URL where they are located in and the full text tag that delimits the body text of the document. The average size of each textual document is 30.51 KB, and thus, the total size of the collection is about 2.0 GB. Most of the documents are related to scientific literatures and clinical notes. The contents of each document are tagged with an XML, including sources, figures and tables. These enabled to easily identify the different contents of the textual document during the cleaning phase of the text collection. The bioMed text collection contains information of non-specific nature and could potentially be judged as a good biomedical knowledge source.

4.2 Cleaning bioMed Text Collection

In bioMed text collection, each text document contains different irrelevant text parts, such as figures, tables, and bad characters like diacritics. These spans of texts and characters are cleaned for accurate processing of sentences in the bioMed text collection. Every sentence is delimited with full stop followed by a space and capital letter for easy recognition by the phrase and proposition disambiguation. The text in appendix A.1 illustrates a sample biomedical text document before cleaning. In the text, there are several irrelevant text parts and bad characters such as "*a 6 – 24 *", citations like "*[1 , 2 , 12 , 15 , 16 , 19 , 23 – 27]"* and figures like "*Fig. & #160"*.

In cleaning and delimiting each sentence, the content of the text documents are considered to have different sections. Scientific documents, for example, have abstracts, captions for images, graphs and tables in which all of them are annotated with XML tags. The XML tags enabled the cleaner to detect Meta-information such as captions from each text document. Sentence

delimitation is based on regular expression that detects full stop followed by space and a capital letter. The cleaner examined each sentence in the textual documents for bad characters (diacritics) and irrelevant texts in it.

The cleaner accomplished its task in three steps: recognition and removal of XML tags in the content of the bioMed text collection, sentence delimitation, and elimination of bad characters and unwanted texts in the textual sentences. Appendix A.2 illustrates a sample document snapshot after cleaning and delimitation of sentences.

BioMed Text Collection Cleaner: it is a module implemented for cleaning a bioMed text collection. It refers different algorithms, which eliminates the different irrelevant part of the bioMed text collection. Let M_{uc} denote the uncleaned bioMed text collection and M_c denote the cleaned bioMed text collection, the following pseudo-algorithm is developed using these as input and output.

Pseudo-algorithm **-** bioMed text collection cleaner, which refers different algorithms and produces cleaned bioMed text collection.

```
//input:  uncleaned bioMed text collection, M_uc
//output: cleaned bioMed text collection, M_c
For each document d in M_uc
  Remove XML tags (d)
  Detect sentence boundaries (d)
  Normalize sentences (d)//remove irrelevant characters
End
```

XML Tag Identification and Cleaning: algorithm is developed for identifying and removing these parts of the texts. The algorithm reads each document, identifies XML tags and removes the section or the tag.

Pseudo-algorithm - XML tag cleaner, identifies and eliminates XML tags/sections in each text document, d.

```
//ID is a section identification number attached in each document
//M_c is the bioMed text collection
```

```
//input is text documents with XML tags and sections
//output is text documents with XML tag removed & merged
For each document, d , in M_c
    For each XML tag ID
      If ID is document ID, ignore the XML tag,
      If ID is abstract ID, clean the XML tag,
      If ID is body text ID, clean the XML tag,
      Else if image, figure, graph, table captions
         Eliminate the section itself
    End
    Eliminate spaces between lines
    Merge body text with the abstract
End
```

Sentence Boundary Detection: the boundary of each sentence is identified and delimited with a sentence delimiter. Boundary detector scans left-to-right and line-by-line each biomedical document (d), with a pattern of full stop, space and an uppercase letter ($regExp_1$), or full stop, upper case letter ($regExp_2$). Once the boundary detector recognized either of these two patterns in a line of text, it puts a new pattern that concatenated a period, a single space and a single upper case letter as a sentence delimiter. The pseudo-algorithm for accomplishing this task is presented as follows:

Pseudo-algorithm **–** sentence boundary detector delimits sentence boundaries in each text document (d).

```
// regExp_1 = [.\sA-Z], regExp_2 = [.A-Z] and regExp_3 = [.\sA-Z]
// M_c is the bioMed text collection
//input is XML tag removed biomedical documents
//output is sentence delimited biomedical documents
For each document, d, in M_c
    For each line of text l in d
      Scan regExp_2 and regExp_1
      If regExp_2 or regExp_1 is found,
               Replace with new regular expression regExp_3
```

```
      Else
        Replace full stop with single space
    End
End
```

Sentence Normalization: the cleaner truncate or drop out words with number or percentage as part of its characters. However, it keeps abbreviations as it appears in the knowledge sources. RegExp is applied to detect numbers, measurements and percentages in a word of texts. If the regExp is detected in a word of texts, the word in a text is truncated as a whole.

Pseudo-algorithm – sentence normalizer cleans bad characters and amounts.

```
//M_c is the bioMed text collection
//input is XML tag removed and sentence delimited biomedical documents
//output is a collection of cleaned biomedical documents
For each knowledge source, d, in M_c
  For each sentence S in each d
    Scan bad characters and amounts
    If regExp is detected,
        Truncate it
    End if
 End for
End for
```

At the end of cleaning the bioMed text collection, each sentence is delimited with a full stop followed by a space and capital letter. Bad characters (e.g. diacritics) and unwanted part of texts in each sentence of the biomedical documents are also cleaned. The resulting text collection is, therefore, used as input for semantic disambiguation, which involves concept disambiguation and semantic proposition prediction.

4.3 Semantic Disambiguation

The semantic disambiguation model enables to acquire biomedical concepts and their associations. In this section, an attempt is made to instantiate this model and acquires biomedical concepts and their associations. MetaMap and enhanced-semRep programs are used for concept

and semantic proposition disambiguation respectively. The MetaMap program is used to recognize and extract biomedical concepts from each sentence of the bioMed text collection. The MetaMap WSD server is used to resolve multiple mapping of a term. Concept associations are predicted using an enhanced-semRep program. For example, in a text fragment "*Bacteria cause Infections*", MetaMap disambiguated the concepts *Bacteria* and *Infections*, where as the role *CAUSE* and the semantic proposition {*Bacteria, CAUSE, Infection*} is disambiguated using the semRep program. Thus, *bacteria* and *infection* are argument phrases (NPs) where as *cause* is a relational phrase (e.g. VPs). Similarly, *Bacteria* and *Infection* are argument concepts and *CAUSE* is a semantic predicate.

4.3.1 Concept Disambiguation

In recognizing and disambiguating biomedical concepts, lexical matching between terms in the bioMed text collection and background knowledge are computed. The lexical matching algorithm used an exhaustive technique to generate term variants, identify potential candidates, compare the candidates and compute the mapping for accurate disambiguation of the concepts. The MetaMap WSD server is configured to resolve multiple mapping of a term based on the semantic contexts of the matching concepts. For example, in an input text "*ocular complications of myasthenia gravis*", MetaMap disambiguated candidate concepts as depicted in Figure 4.1 below.

```
Processing 00000000.tx.1: Ocular complications of myasthenia gravis.

Phrase: Ocular complications of myasthenia gravis.

Meta Mapping (807):

   677   Myasthenia Gravis, Ocular [Disease or Syndrome]
   760   Complications (Complication) [Pathologic Function]

Meta Mapping (807):

   677   Myasthenia Gravis, Ocular [Disease or Syndrome]
   760   complications (Complication Aspects) [Pathologic Function]
```

Figure 4.1 – Multiple Mapping

In the figure, two phrases, *myasthenia gravis* and *ocular complications*, are chunked and disambiguated. The phrase *myasthenia gravis* is interpreted as a concept *Myasthenia Gravis (Ocular)* whose class is *Disease or Syndrome* and the phrase *complications* is interpreted as a concept *Complication* or *Complication Aspects* where their class is *Pathologic Function*. In case of a ties, for example *Complications* and *Complication Aspects*, MetaMap used its WSD server to choose the more correct one, in this case *complications*, based on their parent classes' contexts. Consequently, more than 4 million concepts are recognized and disambiguated from the bioMed text collection. This set of concepts is, therefore, used as input for semantic proposition prediction. Thus, the set of biomedical concepts (C_m) recognized and disambiguated from the bioMed text collection can, generally, be formulated as two tuples:

$$C_m = (T, C) \tag{4.1}$$

Where, T is a set of terms referring to biomedical concepts C, and C is a set of biomedical concepts disambiguated from the bioMed text collection.

```
Processing 00000000.tx.1: Scapular winging is a rare debilitating condition that leads to
limited functional activity of the upper extremity.

Phrase: Scapular winging
Meta Mapping (1000):
  1000   Scapular winging [Finding]

Phrase: a rare debilitating condition
Meta Mapping (913):
   913   rare condition (Rare Diseases) [Disease or Syndrome]

Phrase: leads to limited functional activity of the upper extremity.
Meta Mapping (679):
   742   LEAD (Cardiac Lead Procedure) [Therapeutic or Preventive Procedure]
   742   Limited (Limited (extensiveness)) [Functional Concept]
   742   Functional (Function) [Functional Concept]
   742   *Activity (*Activity (kind of quantity)) [Idea or Concept]
   764   Upper Extremity [Body Location or Region]
Meta Mapping (679):
   742   LEAD (Cardiac Lead Procedure) [Therapeutic or Preventive Procedure]
   742   Limited (Limited (extensiveness)) [Functional Concept]
   742   Functional (Function) [Functional Concept]
   742   Activity (Active) [Functional Concept]
   764   Upper Extremity [Body Location or Region]
Meta Mapping (679):
   742   LEAD (Cardiac Lead Procedure) [Therapeutic or Preventive Procedure]
   742   Limited (Limited (extensiveness)) [Functional Concept]
   742   Functional (Function) [Functional Concept]
   742   Activity (Activities) [Activity]
   764   Upper Extremity [Body Location or Region]
```

Figure 4.2 – Biomedical Concept Disambiguation

Figure 4.2 illustrates a snapshot of concept disambiguation output. More outputs of concept disambiguation can be obtained in appendix A.3 before post-processing.

4.3.2 Semantic Proposition Disambiguation

Biomedical concepts are interacting in the bioMed text collection. This set of concept interactions is referred as semantic associations and show their relationships as described in the text collection. These associations are recognized and instantiated to generate a set of semantic propositions. Semantic propositions are a set of semantic triples in which each of them forms a graphical structure. Such graphical structures are formulated as triplets such as {*subject_argument, predicate, object_argument}*, where the argument set is biomedical concepts and the predicate set is ontological relations, which links the subject and the object argument sets.

A semantic processing program is used to predict semantic associations between biomedical concepts. For example, the SemRep program is used to instantiate concept associations as they exist in the bioMed text collection and thus, semRep generated more than 5 million biomedical concept associations from bioMed text collection. But, semRep considers only 54 semantic predicates from UMLS semantic network, which results limited coverage in predicting concept associations. For example, an XML formatted output of the semRep program is illustrated for a text "*ocular complications of myasthenia gravis"* as follows.

```
<? Xml version="1.0" encoding="UTF-8"?>
<SemRepAnnotation>
 <Document id="D00000000" text="ocular complications of myasthenia gravis." >
 <Utterance id="D00000000.tx.1" section="tx" number="1" text="ocular
complications of myasthenia gravis.">
  <Entity id="D00000000.E1" cui="C0015392" name="Eye" semtypes="bpoc"
text="ocular" score="888" begin="0" end="6" />
  <Entity id="D00000000.E2" cui="C0009566" name="Complication"
semtypes="patf" text="complications" score="888" begin="7" end="20" />
  <Entity id="D00000000.E3" cui="C0026896" name="Myasthenia Gravis"
semtypes="dsyn" text="myasthenia gravis" score="1000" begin="24" end="41" />
  <Predication id="D00000000.P1">
   <Subject maxDist="0" dist="0" entityID="D00000000.E1" relSemType="bpoc" />
```

```
   <Predicate type="LOCATION_OF" indicatorType="MOD_HEAD" begin="0" end="20"
/>
   <Object maxDist="0" dist="0" entityID="D00000000.E2" relSemType="patf" />
  </Predication>
 </Utterance>
</Document>
</SemRepAnnotation>
```

From the result, there are three disambiguated concepts, where their Concept Unique Identifications (CUIs) are *C0015392*, *C0009566* and *C0026896.* The semRep program, however, predicted only one association, between the concepts *C0015392* and *C0009566*, {*C0015392*, *LOCATION_OF*, *C0009566*}. The possibility of other associations, for example between *C0015392* and C0026896, or *C0009566* and *C0026896,* are not predicted. To complement this limitation, an algorithm to enhance *semRep program coverage* is developed, which predicts other concept associations by generating a matrix of concepts and predicting their relationships in the background knowledge, the UMLS. Furthermore, the algorithm considers concept associations with 682 semantic predicates (see appendix A.5 for more semantic predicates), which reside in the lower level of the background knowledge, including the 54 semRep semantic predicates.

***Pseudo-Algorithm* –** for predicting overlooked semantic links by semRep program and looks for their associations in the background knowledge if exists. The inputs are a set of concepts disambiguated by semRep program (C_{sr}), a set of sentences (s_{sr}) where the concepts appear, and the background knowledge ($K_{bk} \equiv UMLS$).

//let M_{sr} = bioMed text collection processed by semRep program

//let s_{sr} = set of sentences processed by semRep program in M_{sr}

//let K_{kb} = the background knowledge

For each sentence in M_{sr},

  Collect a concept set, c_s, in C_{sr}

  Generate a concept-pairing matrix

  Prune unnecessary concept-pairings, e.g. self-associations

  Identify semRep predictions, P_{sr}

```
Retrieve possible associations from K_bk
Compute argument matching, i.e subject and object of the association
If the arguments matched with concept-pairings and valid predicate:
      Set as valid associations
Else
      Set as invalid associations
End
```

The algorithm produces more than two million semantic associations where the arguments (concepts) are from biomedical sentences and semantic links (the predicates are from the background knowledge). Figure 4.3 illustrates the output of enhanced semRep program coverage algorithm. In the figure, all semantic propositions are not included in semRep program predications. Consequently, it scales up the set of semantic propositions by the semRep program. The semantic processing program, therefore, generated a set of semantic graphs (or a set of conceptual graphs) at a larger scale and complements to some extent the coverage of semRep program. However, enhancing semRep coverage algorithm recognizes concept associations within sentences but not across sentences, paragraphs or discourses and exists in the UMLS knowledge.

{C0000163,has_component,C0363709}
{C0000163,has_component,C0363806}
{C0000163,has_component,C0363807}
{C0000163,has_component,C0482554}
{C0000163,has_component,C0482555}
{C0000163,has_component,C0482556}
{C0000163,has_component,C0482557}
{C0000163,has_component,C0482558}
{C0000163,has_component,C0482559}
{C0000163,has_component,C0482560}
{C0000163,has_component,C0799464}
{C0000163,has_component,C0799465}
{C0000163,has_component,C0943554}
{C0000163,has_component,C0943651}
{C0000163,has_component,C1978839}
{C0000163,has_component,C2713054}
{C0000163,has_component,C2735891}
{C0000163,has_component,C2735892}
{C0000163,has_component,C0363798}
{C0000163,has_component,C0363799}
{C0000163,has_component,C0363800}
{C0000163,has_component,C0484618}
{C0000163,has_component,C0797744}
{C0000163,has_component,C0803832}
{C0000163,has_component,C0941912}
{C0000163,has_component,C0942044}
{C0000163,has_component,C0942045}
{C0000163,has_component,C1316681}
{C0000163,has_component,C1316682}
{C0000163,has_component,C1526473}
{C0000152,permuted_term_of,C0000152}
{C0000163,permuted_term_of,C0000163}
{C0000107,has_permuted_term,C0000107}
{C0000107,has_permuted_term,C0000107}
{C0000107,has_sort_version,C0000107}
{C0000107,has_entry_version,C0000107}
{C0000107,has_sort_version,C0000107}
{C0000119,has_permuted_term,C0000119}
{C0000119,permuted_term_of,C0000119}
{C0000119,sort_version_of,C0000119}
{C0000119,has_sort_version,C0000119}
{C0000132,has_permuted_term,C0000132}
{C0000132,has_permuted_term,C0000132}
{C0000132,permuted_term_of,C0000132}
{C0000132,permuted_term_of,C0000132}
{C0000137,has_permuted_term,C0000137}
{C0000137,permuted_term_of,C0000137}
{C0000139,has_permuted_term,C0000139}
{C0000139,permuted_term_of,C0000139}
{C0000139,permuted_term_of,C0000139}
{C0000139,sort_version_of,C0000139}
{C0000139,entry_version_of,C0000139}
{C0000139,permuted_term_of,C0000139}
{C0000139,inverse_isa,C0003216}
{C0000139,has_permuted_term,C0000139}
{C0000139,has_sort_version,C0000139}
{C0000139,has_entry_version,C0000139}
{C0000139,has_permuted_term,C0000139}
{C0000151,has_permuted_term,C0000151}
{C0000151,has_permuted_term,C0000151}
{C0000152,sort_version_of,C0000152}
{C0000152,exhibited_by,C1151265}
{C0000163,has_permuted_term,C0000163}
{C0000163,sort_version_of,C0000163}
{C0000163,has_component,C1631262}
{C0000163,has_component,C1715074}
{C0000163,measures,C0363800}
{C0000163,measures,C0484618}
{C0000163,measures,C0797744}
{C0000163,measures,C0803832}
{C0000163,measures,C0941912}
{C0000163,measures,C0942044}
{C0000163,measures,C0942045}
{C0000163,measures,C1316681}
{C0000163,measures,C1978080}
{C0000163,measures,C1978081}
{C0000163,measures,C1978082}
{C0000163,measures,C1978083}
{C0000163,measures,C2735893}
{C0000163,measures,C2735894}
{C0000163,measures,C2735895}
{C0000163,measures,C2735896}
{C0000163,measures,C0363709}
{C0000163,measures,C0363806}
{C0000163,measures,C0363807}
{C0000163,measures,C0482554}
{C0000163,measures,C0482555}
{C0000163,measures,C0482556}
{C0000163,measures,C0482557}
{C0000163,measures,C0482558}
{C0000163,measures,C0482559}
{C0000163,measures,C0482560}
{C0000163,measures,C0799464}
{C0000163,measures,C0799465}
{C0000163,measures,C0943554}
{C0000163,measures,C0943651}

Figure 4.3 – Snapshot of Predicted Semantic Propositions

Generally, the semantic processing program, in a text "*ocular complications of myasthenia gravis*", produced concept associations *{C0015392, LOCATION_OF, C0009566}*. The semantic indicator 'of' is instantiated to an ontological predicate, *LOCATION_OF*, suggested by a mapping rule '*of → LOCATION_OF'*. The prediction also instantiated syntactic arguments of a preposition 'of' as *eye (myasthenia gravis, ocular)* and *complications*, which are interpreted into concept arguments *Eye (C0015392)* and *Complication (C0009566)* respectively. Consequently, the two concept association interpretation is *C0015392 (Eye) LOCATION_OF C0009566 (Complications)*, which states that *eye* is a location of *complications*. Similarly, the semantic classes of the two concepts are *Body Part, Organ, or Organ Component (bpoc)* and *Pathologic Functions (patf)* respectively. The general interpretation of each text fragment is in the form that "*Eye* is a type of *Body part, Organ, Organ Component*, which is the location of *Complications* whose type is *Pathologic Function*".

From this, one can see that every semantic prediction is a semantic Graph (SG). Examples of SG are: Eye (*C015392*) *isa bpoc*; *Complication* (*C0009566*) *isa patf*; and *Eye LOCATION_OF Complication*. From the principle of inheritance, it also implies that *bpoc LOCATION_OF patf*. Consequently, more than seven million concept associations are disambiguated from the bioMed text collection. Figure 4.4 illustrates a snapshot of concept associations generated by the semantic processing (enhanced semRep coverage) program where numerical codes represent concepts, capitalized letters represent relationships between concepts and underscores are separators. More output of the semantic processing program is illustrated in appendix A.4 and A.6.

Finally, the acquired knowledge artefacts (K_p) is generalized to a formulation of four tuple biomedical knowledge:

$$K_p \equiv (T, C, R, C_R) \tag{4.2}$$

Where T is the set of terms referring to concepts and roles in the bioMed text collection; C is the set of atomic concepts interpreted from the text collection; R is the set of atomic roles interpreted from the same text collection in the indicator rules; and C_R is the set of binary associations between the concepts in C and the roles in R.

C0029365_TREATS_C0241473	C1096593_CAUSES_C0917801	C0332835_PART_OF_C1305735
C1293130_TREATS_C1096593	C0184661_TREATS_C0030193	C0522224_PROCESS_OF_C0237401
C0161479_PROCESS_OF_C0030705	C0011307_TREATS_C0442726	C0037004_LOCATION_OF_C0030193
C1281575_LOCATION_OF_C0006086	C0087111_METHOD_OF_C0543467	C0221198_PROCESS_OF_C0030705
C1269567_LOCATION_OF_C0522224	C0037004_LOCATION_OF_C1457887	C0037004_LOCATION_OF_C0234238
C0037004_LOCATION_OF_C0031037	C0087111_TREATS_C0522224	C0748691_PROCESS_OF_C0030705
C0459914_TREATS_C0175677	C0037047_LOCATION_OF_C1285497	C0026845_PART_OF_C0036277
C0949766_TREATS_C0522224	C0817096_LOCATION_OF_C1306645	C0434255_ISA_C0043251
C1304649_PART_OF_C1269567	C0543467_TREATS_C0748691	C0518031_PROCESS_OF_C0030705
C0027530_LOCATION_OF_C1306645	C0580841_PROCESS_OF_C0030705	C0019552_LOCATION_OF_C0018563
C0949766_ISA_C0087111	C0021153_ISA_C0543467	C0021153_ISA_C0087111
C0038293_LOCATION_OF_C0018670	C0175677_COEXISTS_WITH_C0012691	C0026845_LOCATION_OF_C0009917
C1533685_TREATS_C0234233	C0196542_TREATS_C0030705	C0079896_TREATS_C1457887
C0043189_PART_OF_C0507206	C0037004_LOCATION_OF_C0029365	C0030193_PROCESS_OF_C0030705
C0043189_PART_OF_C0507205	C0034542_CAUSES_C0152180	C0234230_PROCESS_OF_C0030705
C1280976_LOCATION_OF_C1285497	C0037004_LOCATION_OF_C0271548	C1457887_PROCESS_OF_C0037047
C0949766_TREATS_C0004093	C0522224_PROCESS_OF_C0030705	C0037011_PROCESS_OF_C0237401
C0029423_PART_OF_C1281575	C0021153_PRECEDES_C0021153	C0543467_TREATS_C0030193
C1280976_LOCATION_OF_C0522224	C0037763_CAUSES_C0030193	C0231484_PROCESS_OF_C0030705
C0027530_LOCATION_OF_C0225006	C0037949_PART_OF_C1281575	C1280066_LOCATION_OF_C1306645
C0196878_TREATS_C1457887	C0021400_ISA_C0042769	C0043189_LOCATION_OF_C0085639
C0817096_LOCATION_OF_C0221198	C0205076_LOCATION_OF_C0181620	C0043251_CAUSES_C1265748
C0020164_LOCATION_OF_C0018670	C1140618_LOCATION_OF_C0580846	C0043189_PART_OF_C1281575
C0037004_LOCATION_OF_C1306645	C0043189_PART_OF_C0030705	C0043189_LOCATION_OF_C0522224
C0543467_TREATS_C0030705	C1280230_LOCATION_OF_C0522224	C0556664_TREATS_C0175677
C0185473_METHOD_OF_C0185470	C0000905_LOCATION_OF_C0175677	C1304649_PART_OF_C1280976
C0240953_PROCESS_OF_C1524106	C0543467_ADMINISTERED_TO_C0030705	C1442903_PROCESS_OF_C0335081
C0196878_ISA_C0543467	C0522224_ISA_C1457887	C0023685_PART_OF_C0016068
C0087111_PRECEDES_C0021153	C0000905_LOCATION_OF_C0221198	C1306645_CAUSES_C0238656
C0434255_COEXISTS_WITH_C0332667	C0434255_CAUSES_C1265748	C0024485_DIAGNOSES_C0399342
C1279046_LOCATION_OF_C0005558	C0026845_PART_OF_C1281575	C0410112_PROCESS_OF_C0335081
C1293130_TREATS_C0175677	C1281575_LOCATION_OF_C1285497	C1442903_COEXISTS_WITH_C0021368
C0205166_ISA_C1444754	C0037004_LOCATION_OF_C0018670	C0399342_COEXISTS_WITH_C1442903
C1281575_PART_OF_C1281575	C1281575_LOCATION_OF_C0935623	C1442903_PROCESS_OF_C0030705
C1281575_LOCATION_OF_C0181620	C1281575_LOCATION_OF_C0522224	C0003086_LOCATION_OF_C0222032

Figure 4.4 – Snapshot of Semantic Propositions

4.3.3 Contribution and Challenges

In ontology learning, the basic domain elements are acquired from free texts. Existing ontology learning frameworks used simple term extraction techniques based on TFIDF for domain relevance determination. In this study, biomedical concepts and their associations are determined using semantic analysis techniques, which better recognizes domain concepts as compared with TFIDF techniques. Furthermore, concept associations are disambiguated suggested by a prior biomedical knowledge, which provide multiple instances of text fragments from free domain

texts. Existing concept association predictor considers only fifty four semantic predicts that are in the upper ontology layer. This study, however, considers about 682 semantic predicates, including the existing fifty four. This enhances the predicted concept association by two million (from five million to 7 million semantic predictions).

Even though these enhancements, semantic disambiguation is limited to acquire biomedical concepts and their associations from free biomedical texts. Acquiring biomedical individuals and attribute information is left as way forwards in the future. Furthermore, involving large number of semantic predicates could improve the number of semantic predictions acquired from free texts. That is, at least considering the 736 semantic predicates found in the background knowledge might bring significant improvements.

4.4 Conceptual Structure

The set of disambiguated biomedical artefacts (i.e. concepts and roles) and semantic graphs (i.e. semantic predictions) are structured into a lower conceptual ontology structure (G_l). The upper conceptual ontology structure (G_u) is reused from the UMLS semantic groups and semantic classes [314] [315]. The conceptual ontology (G_o) is instantiated with direct acyclic graph formalism. The lower ontology structure is built as an integration of a set of semantic predictions acquired from the bioMed text collection, in which 135 conceptual ontologies, rooted at each of the 135 semantic categories (i.e. semantic classes), are constructed. In the process, semantic predictions are aligned based on their hierarchical knowledge granularity. In addition, redundancies and cycles are pruned to minimize inconsistencies and improve inferencing.

These ontologies are comprised of fine-grained concepts and their associations. In each of the 135 ontologies, the root concept is overlapped with a leaf concept in the upper ontology structure (G_u). In the upper ontology structure (G_u), each semantic class (e.g. semantic type) is subsumed by only a semantic group and every semantic group represents sub-domain categories of the biomedical domain, the ***bioMed*** concept. Consequently, the upper ontology structure (G_u) is rooted at the concept *bioMed*, which is partitioned disjointly into sub-domain categories (i.e.

semantic groups). Each sub-domain categories are further partitioned disjointly into semantic categories, a set of leaf nodes or concepts.

Integration of the two ontology structures (upper and lower) is built by computing concept-overlaps between the two structures. For example, in a text '*ocular complications of myasthenia gravis*', *Eye* is a fine-grained concept whose type (root concept) is *Body Part, Organ or Organ Components* and *Complications* is also a fine-grained concept whose type (root concept) is *Pathologic Function*. But, *Body Part, Organ or Organ Components* and *Pathologic Function* are semantic categories appearing as leaf concepts of the upper ontology structure, creating overlaps with the root concepts of the lower ontologies. In this way, the two ontology structures are integrated consistently and generated the large conceptual ontology (G_o) as a direct acyclic graph.

4.4.1 Upper Ontology Structure

The UMLS semantic groups, semantic types and their associations are re-used to construct the upper ontology structure. Semantic groups partitioned the biomedical domain into 15 sub-domain categories. The categories are partitioned disjointly with the help of domain experts and they form biomedical sub-domain categories where any semantic classes are belong to only in one of them. The 15 sub-domain categories are further partitioned disjointly into 135 coarse-grained semantic classes (semantic types) [314]. Each semantic class is subsumed with a sub-domain category. For example, *Finding* and *Pathologic Function* are coarse-grained concepts, which are subsumed by the *Disorder* sub-domain category. Consequently, taxonomically, the upper ontology is structured using subsumption relation, the 'ISA' link. That is, all sub-domain categories are subsumed by the broadest biomedical concept, the ***bioMed***. The bioMed concept is the root of the upper ontology structure (G_u). The leaf concepts of the upper ontology are the semantic classes or coarse-grained concepts, which overlaps with root concepts of the lower ontology structure (G_l).

To construct the upper ontology structure, a technique that uses the bioMed concept, sub-domain categories and coarse-grained concepts and their semantic associations, is developed. In the technique, the bioMed concept is considered as a set of all biomedical individuals and it is the

top concept ($\top$) in the ontology's lattice structure (G_o). The bioMed concept ($\top$) is then further partitioned disjointly into domain concepts (categories), which are subsumed by the bioMed concept ($\top$). These categories are associated with the broadest biomedical relations such as ISA, ASSOCIATED_WITH, CONCEPTUALLY_RELATED_TO, PHYSICALLY_RELATED_TO, SPATIALLY_RELATED_TO and TEMPORARLY_RELATED_TO. These relations are structured hierarchically where ASSOCIATED_WITH is the top ($\top$) biomedical relation and all others are subsumed by it.

An algorithm is, therefore, developed to structure the bioMed concept, sub-domain concepts and their associations. The algorithm creates the top concept first and then it creates the sub-partitions, their associations and the structures of associations. The inputs for the algorithm are the bioMed concept, C_T, sub-domain concepts, C_D, the top relation, R_T, and its partitions (sub-domain relations, R_D). The output of the algorithm is the structure of the highest (except the leaf nodes) knowledge level of the upper ontology structure (G_u). Figure 5.5 depicts the output of the algorithm, which partitions into 15 sub-domain categories. In the figure, items in the first column, for example chemicalsAndDrugs, Devices, Disorders and Objects are sub-domain categories. These sub-domain categories are semantically related (ISA link) with the root concept, bioMed.

```
ActivitiesAndBehaviors        ISA   bioMed
Anatomy                       ISA   bioMed
ChemicalsAndDrugs             ISA   bioMed
ConceptsAndIdeas              ISA   bioMed
Devices                       ISA   bioMed
Disorders                     ISA   bioMed
GenesAndMolecularSequences    ISA   bioMed
GeographicAreas               ISA   bioMed
LivingBeings                  ISA   bioMed
Objects                       ISA   bioMed
Occupations                   ISA   bioMed
Organizations                 ISA   bioMed
Phenomena                     ISA   bioMed
Physiology                    ISA   bioMed
Procedures                    ISA   bioMed
```

Figure 4.5 – Sub-Domain Categories

Pseudo-algorithm: the highest knowledge structure, except leaf nodes, of upper ontology (G_u).

```
//let C_T be top concept, C_D sub-partitions of top concept
//let R_T be top relation, R_D sub-partitions of top relation
//let C_i be a sub-concept partitions, R_i a sub-relation partition
Create C_T ;//creates bioMed concept (T)
Create R_T ;//creates the ASSOCIATED_WITH relation
For each C_i , R_i
    Set C_i as disjoint sub-partition of C_T , C_i isa C_T ;
    Set R_i as disjoint sub-partition of R_T , R_i isa R_T ;
End
```

Every sub-domain concepts are further partitioned into coarse-grained concepts (semantic classes), which are also subsumed by sub-domain concepts. Technically, the coarse-grained concepts are the leaf nodes of the upper ontology structure, G_u. Thus, an algorithm is developed to structure the coarse-grained concepts as leaf nodes of the upper ontology structure, G_u. The algorithm iterates over the 15 sub-domain partitions and looks for coarse-grained concepts subsumed by sub-domain categories. A total of 135 coarse-grained concepts are structured into the leaf nodes of the upper ontology structure, G_u. These nodes overlap with root concepts in the lower ontology, G_l. In each algorithm, siblings of each node are set as disjoint to ensure absence of redundant individuals in the different categories or concepts.

Pseudo-algorithm: structuring the leaf nodes of the upper ontology structure (G_u).

```
//considering all notations defined in the above algorithm and
//let c_i denote each coarse-grained concepts
//let r_i denote each coarse-grained relations
For each c_i , r_i
      Set c_i as disjoint sub-partition of C_i , c_i isa C_i ;
      Set r_i as disjoint sub-partition of R_i , r_i isa R_i ;
End
```

The output of this algorithm is the upper ontology structure (G_u) at three knowledge levels, where the bioMed concept is the top knowledge, sub-domain categories are the middle knowledge and the coarse-grained concepts are the lower knowledge levels. Each sub-domain categories are further partitioned into disjoint coarse-grained concepts. For example, the *Disorders* sub-domain category is partitioned into 12 disjoint coarse-grained concepts as illustrated in Figure 4.6.

Finding	isa	Disorder
InjuryOrPoisoning	isa	Disorder
PathologicFunction	isa	Disorder
ExperimentalModelOfDisease	isa	Disorder
DiseaseOrSyndrome	isa	Disorder
SignOrSymptom	isa	Disorder
AnatomicalAbnormality	isa	Disorder
NeoplasticProcess	isa	Disorder
MentalOrBehavioralDysfunction	isa	Disorder
CellOrMolecularDysfunction	isa	Disorder
AcquiredAbnormality	isa	Disorder
CongenitalAbnormality	isa	Disorder

Figure 4.6 – Partitioning of Disorder Category

Moreover, each sub-domain categories are associated with each other for certain reasons. For example, **living_beings** is *associated_with* **disorders** or **chemicals** and **drugs** may also be *associated_with* **Disorders**. Thus, semantic associations are represented using five sub-relations under the top relation, *ASSOCIATED_WITH*. The five relations are obtained with disjoint partitioning of the top relation, *ASSOCIATED_WITH*, in a similar fashion of sub-domain categories.

Thus, the upper relationship structure is represented using the five relation partitions as illustrated in Figure 4.7. All other relations in the ontology structure are sub-specializations of the five relations.

Conceptually_related_to	isa	Associated_With
Spatially_realted_to	isa	Associated_With
Physically_related_to	isa	Associated_With
Functionally_related_to	isa	Associated_With
Temporally_related_to	isa	Associated_With

Figure 4.7 –Upper Relation Structure

For example, relationships related to spatial are structured taxonomically under the *SPATIALLY_RELATED_TO* relation, which are, for example, *ADJACENT_TO, LOCATION_OF, SURROUNDS and TRAVERSES*. Another example is the relation *CARRIES_OUT*, which is a specialization of the relation *FUNCTIONALLY_RELATED_TO* and the relation *CONCEPTALLY_PART_OF* is a specialization of *CONCEPTUALLY_RELATED_TO*. Figure 4.8 depicts the partitioning of the relation *CONCEPTUALLY_RELATED_TO* into its sub-relations. In a similar fashion, 54 relations are partitioned to represent the upper ontology structure (G_u).

Generally, the upper ontology structure is represented using a set of sub-domain and coarse-grained concepts associated with these relationships. Formally, it can be formulated as:

$$G_{sc} \subseteq bioMed, G_m \subseteq G_{sc}, G_m \cup G_{sc} \subseteq G_u \tag{4.3}$$

Where, G_u is the upper ontology structure, G_{sc} is sub-domain categories and G_m is coarse-grained concept structure.

issue_in	isa	conceptually_related_to
measurement_of	isa	conceptually_related_to
measures	isa	conceptually_related_to
method_of	isa	conceptually_related_to
property_of	isa	conceptually_related_to
analyzes	isa	conceptually_related_to
assesses_effect_of	isa	conceptually_related_to
conceptual_part_of	isa	conceptually_related_to
degree_of	isa	conceptually_related_to
derivative_of	isa	conceptually_related_to
developmental_form_of	isa	conceptually_related_to
diagnoses	isa	conceptually_related_to
evaluation_of	isa	conceptually_related_to
issue_in	isa	conceptually_related_to
measurement_of	isa	conceptually_related_to

Figure 4.8 – Partitioning of Relations

Receptor(T192)	isa	Chemicals&Drugs(CHEM)	isa	bioMed
Steroid(T110)	isa	Chemicals&Drugs(CHEM)	isa	bioMed
Vitamin(T127)	isa	Chemicals&Drugs(CHEM)	isa	bioMed
Classification(T185)	isa	Concepts&Ideas(CONC)	isa	bioMed
Conceptual Entity(T077)	isa	Concepts&Ideas(CONC)	isa	bioMed
Functional Concept(T169)	isa	Concepts&Ideas(CONC)	isa	bioMed
Group Attribute(T102)	isa	Concepts&Ideas(CONC)	isa	bioMed
Idea or Concept(T078)	isa	Concepts&Ideas(CONC)	isa	bioMed
Intellectual Product(T170)	isa	Concepts&Ideas(CONC)	isa	bioMed
Language(T171)	isa	Concepts&Ideas(CONC)	isa	bioMed
Qualitative Concept(T080)	isa	Concepts&Ideas(CONC)	isa	bioMed
Quantitative Concept(T081)	isa	Concepts&Ideas(CONC)	isa	bioMed
Regulation or Law(T089)	isa	Concepts&Ideas(CONC)	isa	bioMed
Spatial Concept(T082)	isa	Concepts&Ideas(CONC)	isa	bioMed
Temporal Concept(T079)	isa	Concepts&Ideas(CONC)	isa	bioMed
Drug Delivery Device(T203)	isa	Devices(DEVI)	isa	bioMed
Medical Device(T074)	isa	Devices(DEVI)	isa	bioMed
Research Device(T075)	isa	Devices(DEVI)	isa	bioMed
Acquired Abnormality(T020)	isa	Disorders(DISO)	isa	bioMed
Anatomical Abnormality(T190)	isa	Disorders(DISO)	isa	bioMed
CellorMolecularDysfunction(T049)	isa	Disorders(DISO)	isa	bioMed
CongenitalAbnormality(T019)	isa	Disorders(DISO)	isa	bioMed
DiseaseorSyndrome(T047)	isa	Disorders(DISO)	isa	bioMed
ExperimentalModelofDisease(T050)	isa	Disorders(DISO)	isa	bioMed
Finding(T033)	isa	Disorders(DISO)	isa	bioMed
InjuryorPoisoning(T037)	isa	Disorders(DISO)	isa	bioMed
MentalorBehavioralDysfunction(T048)	isa	Disorders(DISO)	isa	bioMed
Neoplastic Process(T191)	isa	Disorders(DISO)	isa	bioMed
Pathologic Function(T046)	isa	Disorders(DISO)	isa	bioMed
Sign or Symptom(T184)	isa	Disorders(DISO)	isa	bioMed
Amino Acid Sequence(T087)	isa	Genes&MolecularSequences(GENE)	isa	bioMed

Figure 4.9 – Snapshot of Upper Ontology Structure

4.4.2 Lower Ontology Structure

The lower ontology structure (G_l) is represented with the integration of 135 ontologies where each of them rooted at each of the 135 coarse-grained biomedical concepts, which are also the leaf nodes of the upper ontology structure. The set of root nodes in the lower ontology and the set of leaf nodes in the upper ontology are overlappings and enabled to align the two ontologies. In the lower ontology structure, each of the 135 ontologies is constructed and rooted at each of the 135 concepts (nodes), and redundancies and cycles are pruned. In the construction, argument concepts are aligned to their immediate parent concepts, which are already aligned to the root concept either directly or through their ancestors. In the process of pruning, semantic predictions

that can be inferred from taxonomies are eliminated. Finally, each of the 135 ontologies is structured into a direct acyclic graph and merged with the upper ontology structure using concept-overlap technique.

In this way, the lower ontology is structured using more than 4 million concepts, 682 relationships and their associations, semantic predictions. The semantic predictions are structured with either taxonomically (e.g. 'ISA') or non-taxonomically (e.g. CAUSE and CAUSED_BY). The taxonomic structure eliminates the redundant subsumption relations implied from transitivity. For example, for the semantic prediction {*C0196878, ISA, C0543467*}, {*C0543467, ISA, C0087111*} and {*C0196878, ISA, C0087111},* the third semantic prediction is redundant, which can be inferred from the first two. Consequently, these types of semantic predictions are eliminated and referred as pruning.

The non-taxonomic structure is simply represented with non-taxonomic associations of each concept for sorting and correct structuring. For example, the concept, *C0196878*, in the previous taxonomy, has associations with a set of concepts {*C0196878, TREATS, C1457887; C0196878, USES, C0037494; and, C0196878, TREATS, C0030705*}. The non-taxonomic structuring sorts them for correct representation as a set of semantic predictions {(*C0196878, TREATS, C1457887), (C0196878, USES, C0037494), (C0196878, TREATS, C0030705)}.*

4.4.2.1 Aligning Semantic Predictions

Semantic prediction alignment enables to structure each of the 135 ontologies in the form of direct acyclic graph. In the structuring, arguments of semantic predictions (i.e. concepts) are structured into hierarchies where broader concepts subsume narrower concepts and linked with an *isa* relationship. Consequently, arguments of hierarchical semantic predictions (i.e. linked with *isa* relationship) are aligned based on concepts' knowledge levels, broader-narrower relationships. For example, section 4.3.2 predicts that the concept *ganciclovir* (*C0017066*) is subsumed by its type *nucleic acid* (*nnon*) and by the concept *drugs (C0013227*). It further predicts the concept *drugs (C0013227)* is subsumed by its type *Pharmacologic Substance* (phsu). This is structured as in the following, where dashed-arrows indicate subsumption relation, which requires alignment and eliminating the redundancies to generate a direct acyclic graph structure.

```
C0017066 --> nnon, C0017066 --> C0013227, C0013227 --> phsu
```

Aligning the above predictions, therefore, generated the following hierarchical structures. These structures belong with two root nodes, *nnon* and *phsu,* due to the two parents of a concept *drugs* (*C0013227*).

```
C0017066 --> C0013227--> nnon, C0017066 --> C0013227 --> phsu
```

In order to build these structures, a technique is developed to align each hierarchy and construct taxonomies of the ontologies, forming direct acyclic graphs. The algorithm first constructs concepts, followed by doing it with the next broader concepts. In a similar fashion, this structuring continuous to the next broader sub-concepts up until the leaf concepts are reached, the narrowest concepts in the ontology structure. For implementing this technique, a generalized breadth first search algorithm is implemented after re-structuring of a set of semantic predictions disambiguated.

Pseudo-algorithm – Aligning Hierarchical Semantic Predictions to build direct acyclic graph

```
//Input =a set of semantic predictions, k_p , linked with isa
//Output=hierarchical structure of ontologies, g_i
//Let M=adjacent matrix of concepts, M_ij =elements of M
//let sp_i = each semantic prediction in k_p
//let N=number of concepts involved in k_p
For each sp_i in k_p
    Get argument concepts, M_ij , of sp_i
End
Construct N x N matrix, M, of concepts
For each row concept, c_i , in M
      Get recursively the ancestors of c_i
      Identify the broadest concept sets, c_b
End
For each column concept, c_i
      Get recursively the descendents of c_i
      Identify the lowest concept sets, c_l
End
For each broadest concepts, c_bi , in c_b
      Set c_bi isa c_ri , where c_ri is a root concept
End
```

Consequently, the algorithm selects a set of concepts linked with *isa* relationship and represents with an adjacent matrix where the columns are parent concepts and rows are child concepts. The matrix represents the presence of subsumption between row and column concepts by a one and by a zero otherwise. Thus, a generalized breadth first algorithm iterates recursively into the parent and child of a row concept (c_i). The recursive iterations continuous up until the parents or Childs are exhausted. The top parent concepts are considered as broadest concept and the lowest Childs are considered as leaf concepts in the ontology structure. Finally, the set of top (broadest) concepts are linked to a root concept (c_r). A snapshot of part of the output hierarchies is illustrated in Figure 4.10.

C0013227, ISA, phsu	C0070166, ISA, C0013227
C0087111, ISA, topp	C0003280, ISA, C0013227
C0012634, ISA, dsyn	C0027415, ISA, C0013227
C0009450, ISA, dsyn	C0286079, ISA, C0013227
C0012634, ISA, dsyn	C0017066, ISA, C0087111
C1274040, ISA, ftcn	C0034991, ISA, C0087111
C0017066, ISA, C0013227	C0087071, ISA, C0087111
C0699142, ISA, C0013227	C0013216, ISA, C0087111
C0700899, ISA, C0013227	C0701307, ISA, C0017066
C0036557, ISA, C0013227	C0029118, ISA, C0012634
C0002771, ISA, C0013227	C0150055, ISA, C0012634
C0003211, ISA, C0013227	C0003864, ISA, C0012634
C0043031, ISA, C0013227	C0149756, ISA, C0012634

Figure 4.10 – Snapshot of Aligned Hierarchies

4.4.2.2 Pruning Redundancies

In disambiguating semantic predictions in sec 4.3.2 above, several of them are either explicitly or implicitly repeated or redundancies. For example, in the following structural listings that are acquired from sec 4.3.2 above, a semantic prediction {*C0017066*, *isa*, *nnon*} repeats itself explicitly three times. Furthermore, in a set of semantic predictions {(*C0439228*, *isa*, *tmco*), (*C0439228*, *isa*, *C0040223*), (*C0040223*, *isa*, *tmco*)}, the semantic prediction (*C0439228*, *isa*, *tmco*) is redundant as it is defined implicitly or transitively via the other two semantic predictions, (*C0439228, isa, C0040223*) and (*C0040223*, *isa*, *tmco*). In a similar fashion, there are several such redundancies in the set of semantic predictions (K_p) generated in sec 4.3.2 above.

```
(Lst. 4.1)
      C0017066, ISA, nnon           C0029118, ISA, dsyn
      C0017066, ISA, C0013227       C0029118, ISA, C0012634
      C0013227, ISA, phsu           C0012634, ISA, dsyn
      C0439228, ISA, tmco           C0206178, ISA, dsyn
      C0439228, ISA, C0040223       C0206178, ISA, C0009450
      C0040223, ISA, tmco           C0009450, ISA, dsyn
      C0286079, ISA, nnon           C0206178, ISA, dsyn
      C0286079, ISA, C0013227       C0206178, ISA, C0012634
      C0013227, ISA, phsu           C0012634, ISA, dsyn
      C0017066, ISA, nnon           C0206178, ISA, dsyn
      C0017066, ISA, C0087111       C0206178, ISA, C1274040
      C0087111, ISA, topp           C1274040, ISA, ftcn
      C0205447, ISA, qnco           C0701307, ISA, nnon
      C0205447, ISA, C0449851       C0701307, ISA, C0017066
      C0449851, ISA, ftcn           C0017066, ISA, nnon
```

In this study, therefore, an algorithm is developed that eliminates either explicitly or implicitly (transitively) defined redundancies in the set of semantic predictions (K_p).

Pseudo-algorithm – Pruning hierarchical semantic predictions and generate direct acyclic graph

```
//let G_l is a set of hierarchies in all the 135 ontologies
//Input =a set of hierarchies, G_l , of the ontology structures,
//Output=direct acyclic graphs of ontology structures, g_DAG
//let N=number of hierarchies involved in each ontologies, g_oi
For each ontology, g_o , in G_l
    For each hierarchy, g_oi , in g_o
          Scan redundant subsumptions, H_rp
          Eliminate the redundancies, H_rp
    End
    Set consistent hierarchies, g_DAG , forming a DAG structure

End
```

Firstly, the algorithm scans explicit redundancies and eliminates them iteratively. In the above listings (Lst. 4.1), for example, the algorithm scans from top to bottom until all the three repetitions of a semantic prediction (*C0017066, isa, nnon*) is found. Secondly, the algorithm scans each hierarchy iteratively for implicit or transitive redundancies up until the hierarchies are exhausted. In above listings, the semantic prediction (*C0439228, ISA, tmco*) is a repetition in the hierarchies defined as follows where dashed arrows indicate subsumption relations.

```
C0439228 --> tmco, C0439228 --> C0040223, C0040223 --> tmco
```

The algorithm, therefore, scans each subsumption relation and eliminates the implicitly repeated hierarchies to generate a consistent hierarchy, where a snapshot is shown in Figure 4.11. Consequently, the semantic prediction, *C0439228 --> tmco,* is eliminated for consistency reasons. In the hierarchy, the semantic class, `tmco,` has become a root concept.

```
C0439228 --> C0040223 --> tmco
```

```
C0013227, isa, nnon
C0013227, isa, phsu
C0017066, isa, C0013227
C0040223, isa, tmco
C0439228, isa, C0040223
C0286079, isa, C0013227
C0087111, isa, topp
C0017066, isa, C0087111
C0449851, isa, ftcn
C0449851, isa, qmco
C0205447, isa, C0449851
C0012634, isa, dsyn
C0029118, isa, C0012634
C0009450, isa, dsyn
C0206178, isa, C0009450
C0206178, isa, C0012634
C1274040, isa, ftcn
C0206178, isa, C1274040
C0701307, isa, C0017066
```

Figure 4.11 – Snapshot of Redundancy Eliminated Hierarchies

4.4.3 Non-Hierarchical Associations

The set of non-hierarchical semantic predictions is unordered collection of semantic triples, where several of them appear with unnecessary repetitions, redundancies. An algorithm is developed to eliminate such redundancies and to order semantic triples based on the left argument, c_l. For example, in the following output listings (Lst.4.2) generated in sec. 4.3.2, the semantic association (`C0030705, PROCESS_OF, humn`) is repeated six times as well as unsorted with the left concept (`C0030705`). Consequently, in this study, an attempt is made to implement an algorithm that eliminates such redundancies and changes the unordered set into ordered set of semantic triples.

(Lst. 4.2)

```
C0001175, PROCESS_OF, dsyn
C0001175, PROCESS_OF, C0030705
C0030705, PROCESS_OF, humn
C0009450, PROCESS_OF, dsyn
C0009450, PROCESS_OF, C1265292
C1265292, PROCESS_OF, bact
C0030705, ADMINISTERED_TO, humn
C0013227, ADMINISTERED_TO, phsu
C0013227, ADMINISTERED_TO, C0030705
C0030705, ADMINISTERED_TO, humn
C0043474, ADMINISTERED_TO, phsu
C0043474, ADMINISTERED_TO, C0030705
```

```
C0009450, PROCESS_OF, dsyn          C0030705, ADMINISTERED_TO, humn
C0009450, PROCESS_OF, C0030705      C0043474, ADMINISTERED_TO, phsu
C0030705, PROCESS_OF, humn          C0043474, ADMINISTERED_TO, C0030705
C0043474, ADMINISTERED_TO, phsu     C0030705, ADMINISTERED_TO, humn
C0043474, ADMINISTERED_TO, C0030705 C0043474, ADMINISTERED_TO, phsu
C0030705, ADMINISTERED_TO, humn     C0043474, ADMINISTERED_TO, C0030705
C0043474, ADMINISTERED_TO, phsu     C0030705, ADMINISTERED_TO, humn
```

Firstly, the algorithm eliminates redundant semantic predictions in the unordered set of semantic triples, and then sorts out the set of semantic triples from broader to narrower tuples by the set of left concepts (c_i). The inputs to the algorithm is a set of non-hierarchical associations (K_{nha}) from the set of semantic predictions (K_p) disambiguated in sec. 4.3.2.

Pseudo-algorithm – non-redundant ordering of non-hierarchical associations

```
//input=set of non-hierarchical associations, K_nha in K_p
//output=set of ordered set of semantic associations
//N=number of non-hierarchical associations
//let c_i=each left concept in K_nha
//let c_j=each right concepts in K_nha
//k=number of triples whose left argument is c_i
//r=semantic relationships
//sp_i=each semantic prediction
For each sp_i in K_nha
      arg_r=get right argument of sp_i
      arg_l=get left argument of sp_i
      prd_lr=get semantic predicate
      For each sp_j in K_nha
            Compare (arg_l, prd_lr, arg_r )and sp_j
            If comparison=true
                  Eliminate sp_j
            End
     End
End
For each c_i, r_i in K_nha //ordering the tuples
      Search k number of c_j, where (c_i, r_ij, c_j) is true
      Set (c_i, r_ij, c_j) as an ordered tuple
End
```

The output of the algorithm is a non-redundant and ordered set of semantic triples. Figure 4.12 illustrates non-redundant and ordered set of semantic predictions.

C0087111,TREATS,topp	C0021440,METHOD_OF,C1276413	C0001175,TREATS(INFER),dsyn	C0242656,TREATS,patf
C0087111,TREATS,C0206178	C0302339,ASSOCIATED_WITH,hops	C0001175,PROCESS_OF,dsyn	C0242656,PROCESS_OF,patf
C0206178,TREATS,dsyn	C0302339,ASSOCIATED_WITH,C0206178	C0001175,PROCESS_OF,C0030705	C0242656,PROCESS_OF,C0030705
C0206178,PROCESS_OF,dsyn	C1275992,LOCATION_OF,medd	C0001175,PROCESS_OF,dsyn	C0001675,TREATS,aggp
C0206178,PROCESS_OF,C0030705	C1280202,LOCATION_OF,bpoc	C0001175,PROCESS_OF,C0030705	C0001675,PROCESS_OF,humn
C0206178,TREATS,dsyn	C1280202,LOCATION_OF,C1275992	C0001175,PROCESS_OF,dsyn	C0556895,USES,topp
C0206178,ASSOCIATED_WITH,dsyn	C1280202,LOCATION_OF,bpoc	C0001175,PROCESS_OF,C0030705	C0556895,USES,C0043474
C0206178,PROCESS_OF,dsyn	C1280202,LOCATION_OF,C0206178	C0221198,LOCATION_OF,fndg	C0556895,USES,topp
C0206178,PROCESS_OF,C0030705	C0019693,TREATS,dsyn	C0006104,LOCATION_OF,bpoc	C0556895,USES,C0043474
C0206178,PROCESS_OF,dsyn	C0019693,TREATS,dsyn	C0006104,LOCATION_OF,C0221198	C0556895,USES,topp
C0206178,PROCESS_OF,C0030705	C0019693,TREATS,dsyn	C1276413,METHOD_OF,topp	C0556895,USES,C1579410
C0206178,LOCATION_OF,dsyn	C0019693,TREATS,dsyn	C1276413,TREATS,topp	C0556895,USES,topp
C0206178,TREATS,dsyn	C0019693,TREATS,dsyn	C1276413,TREATS,C0030705	C0556895,USES,C1579410
C0206178,TREATS,dsyn	C0001175,PROCESS_OF,dsyn	C1276413,TREATS(INFER),topp	C0556895,TREATS,topp
C0206178,TREATS(SPEC),dsyn	C0001175,PROCESS_OF,C0030705	C1276413,TREATS(INFER),C0001175	C0556895,TREATS,C0001675
C0206178,AFFECTS,dsyn	C0001175,PROCESS_OF,dsyn	C1276413,TREATS(INFER),topp	C0009450,PROCESS_OF,dsyn
C0206178,AFFECTS,C0030705	C0001175,PROCESS_OF,C0030705	C1276413,TREATS(INFER),C0206178	C0009450,PROCESS_OF,C1265292
C0206178,PROCESS_OF,dsyn	C0001175,PROCESS_OF,dsyn	C0012132,TREATS,phsu	C0009450,PROCESS_OF,dsyn
C0206178,PROCESS_OF,C0030705	C0001175,PROCESS_OF,C0030705	C0012132,TREATS,C0242656	C0009450,PROCESS_OF,C0030705
C0206178,TREATS(INFER),dsyn	C0001175,PROCESS_OF,dsyn	C0012132,TREATS,phsu	C0009450,PROCESS_OF,dsyn
C0206178,PROCESS_OF(SPEC),dsyn	C0001175,PROCESS_OF,C0030705	C0012132,TREATS,C0019693	C0009450,PROCESS_OF,C0030705
C0206178,PROCESS_OF(SPEC),C0030705	C0001175,PROCESS_OF,dsyn	C0012133,compared_with,nnon	C1265292,PROCESS_OF,bact
C0206178,TREATS,dsyn	C0001175,PROCESS_OF,C0030705	C0012133,TREATS,phsu	C0042291,TREATS,phsu
C0206178,PROCESS_OF,dsyn	C0001175,PROCESS_OF,dsyn	C0012133,TREATS,C0019693	C0042291,TREATS,C0030705
C0206178,PROCESS_OF,C0030705	C0001175,PROCESS_OF,C0030705	C0012133,TREATS,phsu	C0042291,INHIBITS,lipd
C0009671,PROCESS_OF,inbe	C0001175,PROCESS_OF,dsyn	C1517925,compared_with,topp	
C0009671,PROCESS_OF,C0030705	C0001175,PROCESS_OF,C0030705	C1517925,compared_with,C0040808	
C0456909,TREATS,dsyn	C0001175,PROCESS_OF,dsyn	C0242656,OCCURS_IN,patf	
C0021440,METHOD_OF,topp	C0001175,PROCESS_OF,C0030705	C0242656,OCCURS_IN,C0030705	

Figure 4.12 – Snapshot of Ordered Semantic Triples

To eliminate redundancies, the algorithm considers tuple (semantic predictions) uniqueness. That is, the algorithm represents each tuple uniquely. The algorithm compares the arguments and the predicate separately, and if it founds similar tuple arguments and predicate, drops it out. In this way, the algorithm removes redundancies accurately and sorts the output.

4.4.4 Integrating the Ontologies

In this section, the 135 ontologies in the lower ontology structure (G_l) are integrated to the upper ontology structure (G_u). Each of these ontologies are rooted at a semantic class (a coarse-grained concept), which is also a leaf node (concept) of the upper ontology structure (G_u). Thus, the set of leaf concepts in upper ontology and the set of root concepts in the lower ontologies are overlappings or the same concepts. This enables to easily align or merge the two ontologies,

upper and lower. For example, the semantic classes of *Eye* and *Complications* are *Body Part, Organ or Organ Components* (*bpoc*) and *Pathologic Functions* (*paf*) respectively. These coarse-grained concepts are root nodes of lower ontology structures containing the fine-grained concepts *Eye* and *Complications*. They, *bpoc and paf*, are also the leaf concepts of the upper ontology structure. Consequently, they are overlaps of the two ontologies, which enable their integration.

The integration is further elaborated by considering the interpretation of a text *"ocular complications of myasthenia gravis"*. In the text, three semantic predictions, *{(eye, isa, bpoc), (complication, isa, patf), (eye, LOCATION_OF, complication)}* are disambiguated and interpreted. These hierarchies are rooted at *bpoc (Body part, Organ or Organ component*) and *patf* (*Pathologic Functions*). Consequently, bpoc and patf are overlaps of the two ontologies, where patf and bpoc are sub-categories of Disorder (DISO) and Anatomy (ANAT) sub-domain categories respectively. The relation LOCATION_OF, which links the concepts *eye* and *complication*, is inherited from the higher relation SPATIALLY_RELATED_TO. This relation is created as an association between *Anatomy* and *Disorder* sub-domain categories. In general, each hierarchical semantic prediction is interpreted and instantiated to integrate with the upper ontology structure using concept-overlap technique.

Figure 4.13 depicts an example of the integration of the two ontology structures based on the fine-grained concepts e*ye* and c*omplications* and their parent concepts and ancestors. Thus, *Body part, Organ or Organ components (bpoc)* and *Pathologic Functions (patf)* are coarse-grain concepts, which create an overlap between the two ontologies. This overlap enabled to build a semantic link (isa, subsumption relation) between the two ontology structures. Consequently, each hierarchy is integrated to the upper ontology structure in a way the lower ontology structure is structured to its immediate upper ontology structure concepts. That is, each of the broadest fine-grained concepts in the lower ontology is a specialization of the narrowest coarse-grained concepts in the upper ontology, leaf nodes.

A technique is developed to compute the overlaps and integrate the two ontology structures. The technique implements an algorithm for computing the overlaps based on similarities (exact matching) of root concepts in the lower ontologies and leaf concepts in the upper ontology

structure. The algorithm inputs a set of leaf concepts (135 in number) in the upper ontology (G_u) and a set of root concepts in the lower ontology structure (G_l) (135 in number). The algorithm outputs the integrated (merged) ontology structure (G_o) in the form of a directed acyclic graph.

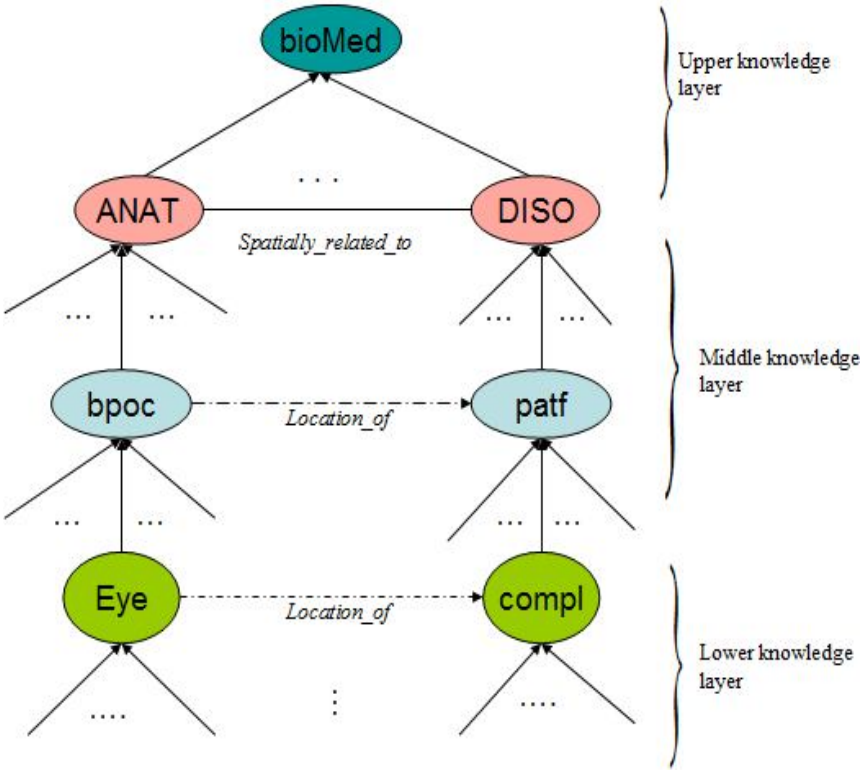

Figure 4.13 – Ontology Structure Integration

A snapshot of part of the ontology structure, descendent from disorder (DISO) sub-domain category and Disease and Syndrome (dsyn), is illustrated in Figure 5.14. Thus, the snapshot illustrates part of the lower ontology structure descendent from disorder sub-domain category.

Pseudo-algorithm – concept-overlap based ontology integration

```
//let G_u = upper ontology structure
//let G_l = lower ontology structures
//let g_ln = each leaf concept in G_u
//let g_rn = each root concept in G_l
//G_o = the integrated ontology structure
For each leaf node, g_ln, in G_u
  Scan if g_ln =root concept in G_l
  If g_ln is a root concept
```

Set g_{rn} =leaf concept of G_u

Set g_{rn} =root concept of G_l

End

Set $G_o = G_u \cup G_l$ //the whole ontology structure

DISO:ISA:BIOMED	C0524910:ISA:C0009450	C0150055:ISA:C0012634	C0026769:ISA:C0012634	C0041296:ISA:C0029118
patf:ISA:DISO	C0276340:ISA:C0009450	C0003864:ISA:C0012634	C0027765:ISA:C0012634	C0343888:ISA:C0029118
sosy:ISA:DISO	C0206526:ISA:C0009450	C0149756:ISA:C0012634	C0030807:ISA:C0012634	C0010414:ISA:C0029118
dsyn:ISA:DISO	C0275959:ISA:C0009450	C0006142:ISA:C0012634	C0030809:ISA:C0012634	C0085436:ISA:C0029118
fndg:ISA:DISO	C1175175:ISA:C0009450	C0003864:ISA:C0012634	C0025202:ISA:C0012634	C0085436:ISA:C0029118
C0009450:ISA:dsyn	C1175175:ISA:C0009450	C0036341:ISA:C0012634	C0263361:ISA:C0012634	C0004032:ISA:C0029118
C0206178:ISA:dsyn	C1175175:ISA:C0009450	C0337672:ISA:C0012634	C0015397:ISA:C0012634	C0041296:ISA:C0029118
C0029118:ISA:dsyn	C0041296:ISA:C0009450	C0002895:ISA:C0012634	C0751366:ISA:C0012634	C0085436:ISA:C0029118
C0012634:ISA:dsyn	C0021400:ISA:C0009450	C0036341:ISA:C0012634	C0205853:ISA:C0012634	C0010414:ISA:C0029118
C0032302:ISA:C0009450	C0339901:ISA:C0009450	C0005758:ISA:C0012634	C0006664:ISA:C0012634	C0041296:ISA:C0029118
C0263387:ISA:C0009450	C0032300:ISA:C0009450	C0162835:ISA:C0012634	C0041296:ISA:C0012634	C0041296:ISA:C0029118
C0026946:ISA:C0009450	C0032285:ISA:C0009450	C0263583:ISA:C0012634	C0162839:ISA:C0012634	C0006849:ISA:C0029118
C0016513:ISA:C0009450	C0021400:ISA:C0009450	C0520594:ISA:C0012634	C1367970:ISA:C0012634	C0041296:ISA:C0029118
C0006309:ISA:C0009450	C0021400:ISA:C0009450	C0006142:ISA:C0012634	C0343101:ISA:C0012634	C0041296:ISA:C0029118
C0016514:ISA:C0009450	C0042769:ISA:C0009450	C0235347:ISA:C0012634	C0037274:ISA:C0012634	C0041296:ISA:C0029118
C0042214:ISA:C0009450	C0035243:ISA:C0009450	C0030807:ISA:C0012634	C0162839:ISA:C0012634	C0041296:ISA:C0029118
C0021400:ISA:C0009450	C0032285:ISA:C0009450	C0005758:ISA:C0012634	C0005758:ISA:C0012634	C0041296:ISA:C0029118
C0021311:ISA:C0009450	C0021400:ISA:C0009450	C0027627:ISA:C0012634	C0877055:ISA:C0012634	C0041296:ISA:C0029118
C0021400:ISA:C0009450	C0006267:ISA:C0009450	C0038165:ISA:C0012634	C0011603:ISA:C0012634	C0041296:ISA:C0029118
C0021400:ISA:C0009450	C0032302:ISA:C0009450	C0005758:ISA:C0012634	C0162835:ISA:C0012634	C0041296:ISA:C0029118
C0035235:ISA:C0009450	C0740766:ISA:C0009450	C0021100:ISA:C0012634	C0021832:ISA:C0012634	C0038826:ISA:C0029118
C0041296:ISA:C0009450	C0275518:ISA:C0009450	C0263313:ISA:C0012634	C0011991:ISA:C0012634	C0026918:ISA:C0029118
C0032285:ISA:C0009450	C0021400:ISA:C0009450	C0702166:ISA:C0012634	C0043528:ISA:C0012634	C0006849:ISA:C0029118
C0021311:ISA:C0009450	C0021400:ISA:C0009450	C0001144:ISA:C0012634	C0036114:ISA:C0012634	C0021311:ISA:C0029118
C0009443:ISA:C0009450	C0021400:ISA:C0009450	C0019618:ISA:C0012634	C0036457:ISA:C0012634	C0041296:ISA:C0029118
C0021400:ISA:C0009450	C0036457:ISA:C0009450	C0011847:ISA:C0012634	C0026848:ISA:C0012634	C0041318:ISA:C0029118
C0021400:ISA:C0009450	C0016627:ISA:C0009450	C0030809:ISA:C0012634	C0524851:ISA:C0012634	C0085315:ISA:C0029118

Figure 4.14 – Part of the Ontology Structure Hierarchies

In general, the biomedical ontology structure (G_o) is formulated as semantic integration of the two ontology structures, G_u and G_l, as:

$$G_o \equiv G_u \cup G_l, \quad where, G_l \subseteq G_u \subseteq G_o \subseteq bioMed \tag{4.4}$$

Conceptual representation of biomedical semantics is highly ambiguous and less machine understandable. For example, a semantic proposition {*bacteria, CAUSES, infections*} produces the following ambiguous interpretations as it may mean one or more of them at the same time.

Bacteria $\cap$ $\forall$CAUSE.Infections, i.e. {x| $\forall$x $\in$ B**acteria**^{I}, and (x, y) $\in$ CAUSE^I implies y$\in$ Infections^I}

$Bacteria \cap \exists CAUSE.Infections$, i.e. $\{x| \exists y \in Infections^I$, and $(x, y) \in CAUSE^I$ and $y \in Infections^I\}$

$Bacteria \cap \leq nCAUSE.Infections$, i.e. $\{x| \#\{y| (x, y) \in CAUSE^I$ and $y \in Infections^I\} \leq n\}$

$Bacteria \cap \geq nCAUSE.Infections$, i.e. $\{x| \#\{y| (x, y) \in CAUSE^I$ and $y \in Infections^I\} \geq n\}$

Where interpretations of the concepts, *Bacteria* and *Infections*, and that of the relation, CAUSE, are:

For Bacteria, $Bacteria^I \subseteq \Delta^I \equiv \{x \mid x \in Bacteria^I$ and $x \in \Delta^I\}$

For Infection, $Infection^I \subseteq \Delta^I \equiv \{x \mid x \in Infection^I$ and $x \in \Delta^I\}$

For CAUSE, $CAUSE^I \subseteq \Delta^I \times \Delta^I \equiv \{(x, y) \mid x, y \in \Delta^I$ and $x^I\ CAUSE^I\ y^I\}$

Unambiguous interpretation must consider one of them according to the discourse context. For the machine, however, this requires further information (inputs), such as cardinalities, to decide on the correct semantic context. Consequently, instantiations of biomedical concept associations are opted to use one of the restriction-based interpretations, existential restriction. Other restrictions are stricter in which their interpretations can be, at least partially, inclusive by the existential restriction-based interpretation.

In this research, the choice of existential restriction-based interpretation is based on experimental analysis where interpretation of most semantics associations laid in this class of restriction. Consequently, interpretation of biomedical artefacts (i.e. concepts, roles and their associations) results a set of knowledge axioms, referred as set of ontology axioms.

4.4.5 Contributions and Challenges

Construction of biomedical ontology structures (conceptual ontology structure (G_o)) had been a challenge. Consequently, works in the state of the art were supported with ontology engineers and ontology tools (e.g. OntoEdit) to structure ontologies. In this study, conceptual structure of the biomedical ontology is constructed independently (with out support of ontology engineers and ontology tools) and automatically. In learning the ontology structure, each component is highly integrated to other components of the framework, which results consistent ontology

acquisition process. Furthermore, construction of the ontology structure is independent from a number of biomedical concepts and their associations; and thus, the proposed framework is scalable. These have significant impacts to reduce/ minimaize resources (e.g. time, effort and cost) in ontology learning from free texts.

In conceptual structuring, the upper ontology structure is reused from existing top-level ontologies, the UMLS semantic network. The chose of the UMLS semantic network is due to its easy integration with the set of ontologies acquired from free biomedical texts. However, more semantically rich (multi-knowledge level upper ontology), which better covers the biomedical domain seems sounding and left as way forwards in the future. This can be obtained with comparative studies among the existing top-level ontologies or creating new ones.

4.5 Ontologization

The formal ontology is a set of OWL DL axioms and denoted by K_o. Axiomatization enforces restrictions for interpreting biomedical artefacts (the set of concepts and roles) and their associations. Restriction-based interpretation enabled unambiguous representations on the expense of losing certain semantic information and human understandability, i.e. more of mathematical semantics than natural one. The OWL DL primitives and constructs support two classes of restrictions: quantification (universal ($\forall$) and existential ($\exists$)) and number restrictions ($= n, \geq n, and \leq n$). Consequently, the OWL DL existential restriction primitives are used to instantiate interpretation of biomedical artefacts, and thus, a set of OWL DL axioms (K_o) is generated.

4.5.1 Experimental Analysis

In this section, an experiment is conducted to observe the trends of interpretations of biomedical concept associations. In the experiment, a randomly selected five hundred (500) semantic predictions are considered and described into human readable representations. Two human subjects, where the first is an expert of microbiology and the second is an expert of medicine, are trained about restriction-based interpretations, particularly related to the three restriction-based interpretations: universal ($\forall$), existential ($\exists$) and number restrictions (either =n, $\leq$n or $\geq$n).

Interpretational approximations are also made between existential and number restrictions. Number restrictions are considered as a special case of existential restrictions where the number of individuals (n) in the right concept related to an individual in the left is known. The subjects are also trained to consider these assumptions and they could judge each semantic association as either existential or universal restriction-based interpretations. The subjects write either ∃ for existential-based interpretation, ∩ for number restrictions, or ∀ for universal-based interpretations.

After the training, each subject is presented for five hundred human readable semantic associations and requested to judge the type of interpretations each of the associations experienced. The subjects are also supported by the researcher in case of disagreements and further clarifications of the interpretations. In case of disagreements of the two subjects, the subjects are discussed with the researcher and three of them decide what to do on the disagreed semantic associations. The decision is either to agree in one of them or reject (25 triples) the disagreed semantic triple (semantic prediction). Table 4.1 illustrates part of the semantic associations judged by the two subjects. In the table, universal restriction-based interpretations (denoted by∀), number restriction-based interpretations (denoted by∩) or existential restriction-based interpretations (denoted by ∃) are indicated without the human readable descriptions.

In the judgmental evaluation, 25 (5%) semantic triples are rejected due to disagreements between the subjects, 55 (11%) semantic associations are judged to have universal restriction-based interpretations, 105 (21%) are judged to have number restriction-based interpretations. Others, 315 or 63%, are judged to have existential restriction-based interpretations. According to the assumption that number restriction-based interpretation is approximated by existential restriction-based interpretation, 420 semantic associations, which account 84%, experienced existential restriction-based interpretations. Consequently, we experienced an approximated restriction-based interpretation of biomedical concept associations. Note that hierarchical (*isa*) and *part_of* relations are existential restriction-based interpretations in which the above proportion includes these relationships as well.

Table 4.1 – Subject Judged Semantic Associations

Semantic Triples	Rest_ Type	Semantic Triples	Rest_T ype	Semantic Triples	Rest_T ype
C0029365_TREATS_C0241473	∃	C1096593_CAUSES_C0917801	∃	C0332835_PART_OF_C1305735	∃
C1293130_TREATS_C1096593	∃	C0184661_TREATS_C0030193	∃	C0522224_PROCESS_OF_C0237401	∩
C0161479_PROCESS_OF_C0030705	∩	C0011307_TREATS_C0442726	∃	C0037004_LOCATION_OF_C0030193	∃
C1281575_LOCATION_OF_C0006086	∃	C0087111_METHOD_OF_C0543467	∀	C0221198_PROCESS_OF_C0030705	∩
C1269567_LOCATION_OF_C0522224	∃	C0037004_LOCATION_OF_C1457887	∃	C0037004_LOCATION_OF_C0234238	∃
C0037004_LOCATION_OF_C0031037	∃	C0087111_TREATS_C0522224	∃	C0748691_PROCESS_OF_C0030705	∩
C0459914_TREATS_C0175677	∃	C0037047_LOCATION_OF_C1285497	∃	C0026845_PART_OF_C0036277	∃
C0949766_TREATS_C0522224	∃	C0817096_LOCATION_OF_C1306645	∃	C0434255_ISA_C0043251	∃
C1304649_PART_OF_C1269567	∃	C0543467_TREATS_C0748691	∃	C0518031_PROCESS_OF_C0030705	∃
C0027530_LOCATION_OF_C1306645	∃	C0580841_PROCESS_OF_C0030705	∩	C0019552_LOCATION_OF_C0018563	∃
C0949766_ISA_C0087111	∃	C0021153_ISA_C0543467	∃	C0021153_ISA_C0087111	∃
C0038293_LOCATION_OF_C0018670	∃	C0175677_COEXISTS_WITH_C0012691	∀	C0026845_LOCATION_OF_C0009917	∃
C1533685_TREATS_C0234233	∃	C0196542_TREATS_C0030705	∃	C0079896_TREATS_C1457887	∃
C0043189_PART_OF_C0507206	∃	C0037004_LOCATION_OF_C0029365	∃	C0030193_PROCESS_OF_C0030705	∃
C0043189_PART_OF_C0507205	∃	C0034542_CAUSES_C0152180	∃	C0234230_PROCESS_OF_C0030705	∩
C1280976_LOCATION_OF_C1285497	∃	C0037004_LOCATION_OF_C0271548	∃	C1457887_PROCESS_OF_C0037047	∩
C0949766_TREATS_C0004093	∃	C0522224_PROCESS_OF_C0030705	∩	C0037011_PROCESS_OF_C0237401	∩
C0029423_PART_OF_C1281575	∃	C0021153_PRECEDES_C0021153	∀	C0543467_TREATS_C0030193	∃
C1280976_LOCATION_OF_C0522224	∃	C0037763_CAUSES_C0030193	∃	C0231484_PROCESS_OF_C0030705	∩
C0027530_LOCATION_OF_C0225006	∃	C0037949_PART_OF_C1281575	∃	C1280066_LOCATION_OF_C1306645	∃
C0196878_TREATS_C1457887	∃	C0021400_ISA_C0042769	∃	C0043189_LOCATION_OF_C0085639	∃
C0817096_LOCATION_OF_C0221198	∃	C0205076_LOCATION_OF_C0181620	∃	C0043251_CAUSES_C1265748	∃
C0020164_LOCATION_OF_C0018670	∃	C1140618_LOCATION_OF_C0580846	∃	C0043189_PART_OF_C1281575	∃
C0037004_LOCATION_OF_C1306645	∃	C0043189_PART_OF_C0030705	∃	C0043189_LOCATION_OF_C0522224	∃
C0543467_TREATS_C0030705	∃	C1280230_LOCATION_OF_C0522224	∃	C0556664_TREATS_C0175677	∃
C0185473_METHOD_OF_C0185470	∀	C0000905_LOCATION_OF_C0175677	∃	C1304649_PART_OF_C1280976	∃
C0240953_PROCESS_OF_C1524106	∀	C0543467_ADMINISTERED_TO_C0030705	∀	C1442903_PROCESS_OF_C0335081	∩
C0196878_ISA_C0543467	∃	C0522224_ISA_C1457887	∃	C0023685_PART_OF_C0016068	∃
C0087111_PRECEDES_C0021153	∀	C0000905_LOCATION_OF_C0221198	∃	C1306645_CAUSES_C0238656	∃
C0434255_COEXISTS_WITH_C0332667	∀	C0434255_CAUSES_C1265748	∃	C0024485_DIAGNOSES_C0399342	∃
C1279046_LOCATION_OF_C0005558	∃	C0026845_PART_OF_C1281575	∃	C0410112_PROCESS_OF_C0335081	∩
C1293130_TREATS_C0175677	∃	C1281575_LOCATION_OF_C1285497	∃	C1442903_COEXISTS_WITH_C0021368	∀
C0205166_ISA_C1444754	∃	C0037004_LOCATION_OF_C0018670	∃	C0399342_COEXISTS_WITH_C1442903	∀
C1281575_PART_OF_C1281575	∃	C1281575_LOCATION_OF_C0935623	∃	C1442903_PROCESS_OF_C0030705	∩
C1281575_LOCATION_OF_C0181620	∃	C1281575_LOCATION_OF_C0522224	∃	C0003086_LOCATION_OF_C0222032	∃

4.5.2 Knowledge Axioms

Axiomatization of a set of concepts, roles and their associations requires their inventories before interpreting them to their respective axioms. Consequently, the semantic disambiguation is instantiated with 4 million concepts and 682 roles. About 7 million concept associations (semantic propositions) are in the text collection. Having these counts of biomedical concepts, roles and their associations, knowledge axioms are interpreted and instantiated as in the following procedure:

Firstly, **primitive** (Ψ) concepts and roles are considered and identified. Primitive concepts and roles are a set of concept or role axioms comprised with two set of axioms, atomic and defined. Atomic axioms are interpretation of atomic concepts or roles, where as defined axioms is

interpretation of anonymous concepts or roles defined based on atomic concepts or roles respectively. Consequently, the primitive axioms (Ψ) is set to be a logical combination of four primitive knowledge axioms. More formally, Ψ is formulated as:

$$\Psi \cong (\beta, \delta, \alpha, \rho) \tag{4.5}$$

Where, β is a set of atomic concept axioms, δ is a set of atomic role axioms, α is set of defined concept axioms and ρ is set of defined role axioms.

Atomic concept axioms (β) is an interpretation of a set of biomedical concepts using OWL DL primitives and constructs. Each concept axiom interpretation includes three attributes: its identification, label and description. For example, the concept ***bioMed*** and its two sub-domain categories (*Anatomy* and *Disorder*) are interpreted and represented as:

```
<owl: Class RDF: about=”#bioMed”>
      <rdfs: label xml: Lang=”en”>Biomedicine</rdfs: label>
      <rdfs: comment xml: Lang=“en“>The highest biomedical concept, which
          consists of all biomedical individuals</rdfs: comment>
</owl: Class>
<owl: Class RDF: about=”#ANAT”>
      <rdfs: label xml: Lang=”en”>Anatomy</rdfs: label>
      <rdfs: comment xml: Lang=“en“>All anatomical organs and structures
          </rdfs: comment>
</owl: Class>
<owl: Class RDF: about=”#DISO”>
      <rdfs: label xml: Lang=”en”>Disorder</rdfs: label>
      <rdfs: comment xml: Lang=“en“>All type of body disorders of animals
          </rdfs: comment>
</owl: Class>
```

Atomic role axioms (δ) is an interpretation of a set of biomedical roles using OWL primitives and constructs. Each role interpretation includes its attributes: identification, label and description. For example, a role *associated_with* and the only subsumption relation are interpreted and represented as:

```
<owl: ObjectProperty RDF: ID="Asso_with">
```

```
      <rdfs:  label xml: Lang=”en”>ASSOCIATED_WITH</rdfs: label>
      <rdfs:  comment  xml:  Lang=”en”>  Semantic  Association  of  two
          concepts</rdfs: comment>
</owl: ObjectProperty>
<owl: ObjectProperty RDF: ID="ISA">
      <rdfs: label xml: Lang=”en”>A Type of Relation</rdfs: label>
     <rdfs:  comment  xml:  Lang=”en”>A  core  relationship  for  taxonomic
            structure</rdfs: label>
      <rdf: type rdf: resource="&owl; TransitiveProperty"/>
      <rdf: type rdf: resource="&owl; SymmetricProperty"/>
      <rdf: type rdf: resource="&owl; ReflexiveProperty"/>
</owl: ObjectProperty>
```

To generalize concept and role axiomatization, all atomic concepts and roles are interpreted and represented to a set of atomic concept axioms (β) and atomic role axioms (δ) in a similar fashion. In order to accomplish these, an algorithm is developed to produce the set of atomic concept and role axioms. The algorithm uses the set of biomedical concept (C_a) and role (R_a) inventories as its inputs and produces a set of primitive concept axioms (β) and primitive role axioms (δ), where part of the outputs are illustrated in the above listings.

Pseudo-algorithm – to **g**enerate a set of atomic concept and role axioms

```
//let the inputs are the set of atomic concepts (C_a) and roles (R_a)
//let the outputs are the set of atomic concept and role axioms, β and δ
For each atomic entity e in C_a or R_a
      If e equals to atomic concept c
         //generate atomic concept axioms
         <owl: Class RDF: about=”#concept_id”>
            <rdfs: label xml: Lang=”en”>concept_name</rdfs: label>
            <rdfs: comment xml: Lang=“en“>concept_description</rdfs: comment>
         </owl: Class>
      End
      Else if e equals to atomic role r
            //generate atomic role axioms
            <owl: ObjectProperty RDF: ID="role_id">
                <rdfs: label xml: Lang=”en”>role_name</rdfs: label>
```

```
            <rdfs: comment xml: Lang=”en”>role_description</rdfs: comment>
        </owl: ObjectProperty>
    End
End
```

Defined concept axioms (α) is an interpretation and representation of a set of defined concepts. Defined concepts are produced when binary association of two concepts is interpreted. For example, interpreting a causative association of *bacteria* and *infection* defined a new class of bacteria, called parasitic bacteria, which causes infection, i.e. a set of infections bacteria. This defines new concepts, which have no associations before. This class of bacteria is interpreted and represented as:

```
    <owl: Class RDF: about=”#bact”>
        <rdfs: label xml: Lang=”en”> Bacteria Rest</rdfs: label>
        <rdfs: comment xml: Lang=“en“>A set of infections bacteria
        </rdfs: comment>
        <owl: IntersectionOf RDF: ParseType=”Collection”>
            <owl: Class RDF: about=”#Bacteria”/>
            <owl: Restriction>
                <owl: onProperty RDF: resource=”#CAUSE”/>
                <owl: someValuesFrom RDF: resource=”#Infections”/>
            </owl: Restriction>
        <rdfs: SubClassOf RDF: resource=”#Bacteria”/>
    </owl: Class>
```

Note that in defining the new concept, it is interpreted as subclass of the subject concept in the association, the concept *Bacteria* in above case. *Defined role axioms* (ρ) is an interpretation and representation of a set of defined roles. Defined roles are created when binary association of two roles is interpreted. In this research, however, only taxonomic association is defined between roles, i.e. there are no non-taxonomic associations between roles. And thus, there are no defined non-taxonomic roles, which is an empty set axiom, $\rho = \{\}$. Thus, all defined concepts are interpreted in a similar fashion to produce a set of defined concept axioms (α). In order to generate these axioms, algorithm is developed that takes non-taxonomic associations as input and produce a set of defined concept axioms (α) as output where part of it is illustrated in the above listing.

Pseudo-algorithm – to generate a set of defined concept axioms

```
//let input = a set of non-taxonomic associations, T_non
//let output = a set of defined concept axioms, α
For each association t_non in T_non
      //define new concept axioms
      <owl: Class RDF: about="#new_concept_id">
            <rdfs: label xml: Lang="en"> new concept name</rdfs: label>
            <rdfs: comment xml: Lang="en">new concept description</rdfs:
            comment>
            <owl: IntersectionOf RDF: ParseType="Collection">
                  <owl: Class RDF: about="#subj_concpt_id"/>
                  <owl: Restriction>
                        <owl: onProperty RDF: resource="#relationship_type"/>
                        <owl: someValuesFrom RDF: resource="#obj_concpt_id"/>
                  </owl: Restriction>
            //create subsumption link
            <rdfs: SubClassOf RDF: resource="#subj_concpt_id"/>
      </owl: Class>
```

Finally, the primitive axiom (Ψ) is reformulated as three tuples as illustrated in:

$$\Psi = (\beta, \delta, \alpha) \tag{4.6}$$

Primitive Attribute Axioms (ψ) is an empty set axiom, {}, because attribute information is not acquired in the acquisition of concepts and roles. Consequently, $\psi \equiv \{\ \}$.

Concept Taxonomy Axioms (H_c) is an interpretation of concept associations related using an 'ISA' link. 'ISA' link is a subsumption relation and interpreted with an OWL DL primitive *owl: subClassOf*. This is exemplified with part of the output listings for a set of biomedical concepts: *Tissue, Anatomy*, *EmbryonicStructure, Cell* and *cellComponent*.

```
      <owl: Class RDF: about="#tisu">
            <rdfs: SubClassOf RDF: resource="#ANAT">
      </owl: Class>
      <owl: Class RDF: about="#emst">
```

```
      <rdfs: SubClassOf RDF: resource="#ANAT">
</owl: Class>
<owl: Class RDF: about="#cellComp">
      <rdfs: SubClassOf RDF: resource="#ANAT">
</owl: Class>
<owl: Class RDF: about="#cell">
      <rdfs: SubClassOf RDF: resource="#ANAT">
</owl: Class>
```

All subsumption relations (i.e. concepts linked by 'ISA') are axiomatized in a similar fashion and represented as a set of concept taxonomy axioms (H_c).

Role Taxonomy Axioms (H_r) is an interpretation role associations related using an 'ISA' link. 'ISA' is a subsumption relation and interpreted with an OWL DL primitive *owl: subPropertyOf*. This is exemplified with part of the output listings for a set of biomedical roles: *analysis*, *conceptually_related_to*, *assesses_effect_of* and *conceptally_part_of*.

```
<owl: ObjectProperty RDF: about="#analysis">
      <rdfs: SubPropertyOf RDF: resource="#conceptually_related_to">
</owl: ObjectProperty>
<owl: ObjectProperty RDF: about="#assesses_effect_of">
      <rdfs: SubPropertyOf RDF: resource="#conceptually_related_to">
</owl: ObjectProperty>
<owl: ObjectProperty RDF: about="#conceptally_part_of">
      <rdfs: SubPropertyOf RDF: resource="#conceptually_related_to">
</owl: ObjectProperty>
<owl: ObjectProperty RDF: about="#conceptally_related_to">
      <rdfs: SubPropertyOf RDF: resource="#associated_with">
</owl: ObjectProperty>
```

All role subsumption relations (i.e. roles linked with an 'ISA' link) are axiomatized in a similar fashion and represented as a set of role taxonomy axioms (H_r).

In order to generate the set of concept taxonomy axioms (H_c) and the set of role taxonomy axioms (H_r), an algorithm is introduced that accepts as an inputs a set of concepts (C_H) and a

set of roles (R_H) associated by an *isa* relationship. Finally, the algorithm produces a set of concept hierarchy axioms (H_c) and a set of role hierarchy axioms (H_r).

Pseudo-algorithm – to construct a set of concept and role taxonomy axioms, H_c and H_r

```
//input=set of concept (C_H) and role (R_H) associations linked by isa
//output=set of concept (H_c) and role (H_r) taxonomy axioms
For each hierarchy h in H_c or H_r
      If h is concept hierarchy
            //generate a set of concept taxonomy axioms, H_c
            <owl: Class RDF: about="#subj_concept_id">
                  <rdfs: SubClassOf RDF: resource="#obj_concept_id">
            </owl: Class>
      Elseif h is role hierarchy
            //generate a set of role taxonomy axioms, H_r
            <owl: ObjectProperty RDF: about="#subj_role_id">
                  <rdfs: SubPropertyOf RDF: resource="#obj_role_id">
            </owl: ObjectProperty>
      End if
End for
```

Non-Taxonomic Relation Axiom (Φ) is an interpretation of non-taxonomic concept associations, i.e. relations other than 'ISA' links. Each of these axioms is interpreted based on existential restrictions for unambiguous representations. For example, the non-taxonomic associations:

{(Anatomy, Spatially_related_to, Disorder), (BodyPartOrganOrOrganComponents (bpoc), location_of, pathologicFunctions (patf)), and (Eye (C0015392), location_of, Complication (C0009566))}

Are interpreted and represented as part of output listings shown below:

```
      <owl: Class Rdf: about="#ANAT">
            <Owl: restriction>
                  <owl: OnProperty RDF: Resource="#Spatially_related_to"/>
                  <owl: someValuesFrom RDF: resource="#DISO"/>
```

```
            <Owl: restriction>
      </owl: Class>
      <owl: Class Rdf: about="#bpoc">
            <Owl: restriction>
                  <owl: OnProperty RDF: Resource="#location_of"/>
                  <owl: someValuesFrom RDF: resource="#patf"/>
            <Owl: restriction>
      </owl: Class>
      <owl: Class Rdf: about="#Eye">
            <Owl: restriction>
                  <owl: OnProperty RDF: Resource="#location_of"/>
                  <owl: someValuesFrom RDF: resource="#Complication"/>
            <Owl: restriction>
      </owl: Class>
```

Generally, all non-taxonomic relations are axiomatized in a similar fashion and represented as a set of non-taxonomic axioms (Φ). In order to generate this set of axioms, an algorithm is introduced that accepts as an input the set of non-taxonomic concept associations and produces the set of non-taxonomic axioms (Φ).

Pseudo-algorithm – to generate a set of non-taxonomic axioms, Φ

```
//input= a set of non-hierarchical concept associations, C_non
//output= a set of non-taxonomic axioms, Φ
For each non-hierarchical concept association c_non in C_non
      //generate non-taxonomic axioms, Φ
      <owl: Class Rdf: about="#subj_concept_id">
            <Owl: restriction>
                  <owl: OnProperty RDF: Resource="#relationship_type"/>
                  <owl: someValuesFrom RDF: resource="#obj_concept_id"/>
            <Owl: restriction>
      </owl: Class>
End
```

Assertional Axioms (A) is an empty set axiom, {}, because individual entities are not included in the acquisition of biomedical artefacts. Thus, $A \equiv \{\ \}$.

4.5.3 Other Axioms

Equivalence and disjointness are another set of axioms, which are interpreted and represented in OWL DL ontology model. OWL DL provides *owl: equivalentClass* construct to define concept equivalence and *owl: equivalentProperty* construct to define role equivalence. Accordingly, biomedical concept and role equivalence and disjointness are defined and interpreted to produce a set of equivalence and disjoint axioms. The equivalence axiom of a non-leaf concept c is defined as the union of its children, $c_1, c_2, ..., c_n$ (see eq. 4.7). Similarly, the equivalence axiom of a non-leaf role r is defined as the union of its children $r_1, r_2, ..., r_n$ (see eq. 4.7). Thus, as presented in eq. 4.7, equivalence of a concept c or a role r is interpreted as the union of its children.

$$\begin{aligned} & c \equiv c_1 \cup c_2 \cup ... \cup c_n \equiv \bigcup_i^n c_i \\ & or \\ & r \equiv r_1 \cup r_2 \cup ... \cup r_n \equiv \bigcup_i^n r_i \end{aligned} \tag{4.7}$$

Where, n is the number of children, c_i and r_i are child of c and r respectively. As shown in eq. 4.8, OWL DL provides *owl: disjointWith* construct to define concept and role disjointness axioms. Similarly, the disjoint axiom of a concept c is a set to be disjoint with its sibling concepts $sc_1, sc_2, ..., sc_n$. Similarly, the disjoint axiom of a role r is a set to be disjoint with its sibling roles $sr_1, sr_2, ..., sr_n$.

$$\begin{aligned} & c \cap sc_1 \cap sc_2 \cap ... \cap sc_n \equiv \{\} \\ & or \\ & r \cap sr_1 \cap sr_2 \cap ... \cap sr_n \equiv \{\} \end{aligned} \tag{4.8}$$

Where, n is the number of siblings and sc_i or sr_i is a sibling of c or r respectively. For example, from the output listing, a concept *anatomy* (ANAT) has three children and four disjoint classes, then its equivalence and disjoint axioms are represented as:

```
<owl: Class RDF: about="#ANAT">
      <Owl: equivalentClass>
            <Owl: Class>
                  <owl: unionOf RDF: ParseType="Collection">
                        <rdfs: Class RDF: resource="#anst">
                        <rdfs: Class RDF: resource="#blor">
```

```
                        <rdfs: Class RDF: resource="#bpoc">
                    </owl: unionOf>
                </owl: Class>
            </owl: equivalentClass>
            <owl: disjointWith RDF: resource="#DISO">
            <owl: disjointWith RDF: resource="#PHEN">
            <owl: disjointWith RDF: resource="#ACTI">
            <owl: disjointWith RDF: resource="#CHEM">
        </owl: Class>
```

In a similar fashion, the role *physically_related_to* has three children and disjoint with three roles, then its equivalence and disjoint axioms are represented as:

```
        <owl: ObjectProperty RDF: about="#PHYRT">
            <Owl: EquivalentProperty>
                <Owl: Class>
                    <owl: unionOf RDF: ParseType="Collection">
                        <rdfs: Class RDF: resource="#part_of">
                        <rdfs: Class RDF: resource="#contains">
                        <rdfs: Class RDF: resource="#consist_of">
                    </owl: unionOf>
                </owl: Class>
            </owl: EquivalentProperty>
            <owl: disjointWith RDF: resource="#conceptually_related_to">
            <owl: disjointWith RDF: resource="#spatially_related_to">
            <owl: disjointWith RDF: resource="#functionally_related_to">
            <owl: disjointWith RDF: resource="#temporary_related_to">
        </owl: Class>
```

Furthermore, if a concept c has multiple parents, equivalence axiom is defined as an intersection of a set of parent concepts $c_1, c_2, ..., c_n$ as presented in eq. 5.9.

$$c \equiv c_1 \cap c_2 \cap ... \cap c_n$$
$$c \equiv \bigcap_i^n c_i \tag{4.9}$$

Where, n is the number of parent concepts and c_i is a parent of c. Note that, in this case, we do not have role equivalence axioms as there are no roles that have multiple parents in the ontology

structure. Similar to the above, such concept equivalence axioms are interpreted using *owl: equivalentClass* construct. For example, as part of the output listings, the concept *denervation* (*C0011307*) has two parent classes, *surgery* (*C0543467*) and *treatment* (*C0087111*), equivalence axiom is interpreted as the intersection of these concepts:

```
<owl: Class RDF: about="#C0011307">
      <Owl: equivalentClass>
            <Owl: Class>
                  <owl: IntersectionOf RDF: ParseType="Collection">
                        <rdfs: Class RDF: resource="#C0543467">
                        <rdfs: Class RDF: resource="#C0087111">
                  </owl: IntersectionOf>
            </owl: Class>
      </owl: equivalentClass>
</owl: Class>
```

To generate a set of concept or role equivalence and disjoint axioms, an algorithm is developed that inputs a set of concepts (C) or roles (R) and retrieves a set of parents (P), siblings (S) and childs (Ch) of each concept c or role r, and then produces a set of equivalent and disjoint axioms.

Pseudo-algorithm – concept or role equivalence and disjointness axiomatization:

```
//input= a set of concepts, C , in the ontology structure
//input= a set of roles, R , in the ontology structure
//output= a set of equivalence (ω) and disjointness axioms (ϖ)
For each concept, c_i, or role, r_i, in C or R
    Read concept, c_i or role, r_i, in C or R
    Retrieve c_i's or r_i's children(Ch), if not leaf node
        Define c_i's or r_i's equivalence axiom (ω), union of children
    Retrieve c_i's or r_i's parents(P), if not root node
        Define c_i's or r_i's equivalence axiom, intersection of parents
    Retrieve c_i's or r_i's siblings (S), in not root node
        Define c_i's or r_i's siblings as disjoint axioms (ϖ)
End
```

4.5.4 Ontological Knowledge

OWL DL ontology is a set of schema and assertional axioms expressed using OWL DL primitives and constructs. In this research, OWL DL ontology is a set of primitive, concept and role taxonomic, non-taxonomic, equivalence and disjoints axioms structured into a direct acyclic graph. Assertional (A) and attribute (ψ) axioms are interpreted as empty set axioms ($A = \{\ \}$ and ψ = {} respectively). Consequently, the OWL DL ontology (K_o) is defined as a set of schema axioms.

These set of schema axioms are integrated to model the formal biomedical ontology (K_o). In eq. 5.6, the set of primitive axioms (Ψ) is splited into three tuples for technical reasons. Furthermore, in section 5.5.3, two set of axioms, equivalence (ω) and disjointness (ϖ), are introduced for practical reasons. Thus, OWL DL ontology is a set of six axioms: a set of primitive axioms (Ψ), a set of concept hierarchy axioms (H_c), a set of role hierarchy axioms (H_r), a set of non-taxonomic axioms (Φ), a set of equivalent (ω) and disjoint (ϖ) axioms. Practically, the ontological knowledge (K_o) is represented as six tuples:

$$K_o = (\Psi, H_c, H_r, \Phi, \omega, \varpi) \tag{4.9}$$

The OWL DL ontology is, therefore, the logical integration of these set of axioms. To produce the logical integration and generate the OWL DL ontology, a main algorithm (implemented as main function) is developed that calls the different algorithms (implemented as sub-functions), which generates the different set of axioms, in an ordered manner. After generating the ontology header information, the main algorithm calls the algorithm that produces set primitive axioms (Ψ). Secondly, it calls the algorithm that produces concept (H_c) and role (H_r) taxonomy axioms. Thirdly, the main algorithm calls the algorithm that produces set of non-taxonomic axioms (Φ). Lastly, the equivalence and disjoint axioms are generated by calling the respective algorithms. Finally, the integrated set of OWL DL ontology axioms represented the ontological knowledge (K_o). Figure 4.15 illustrates a snapshot of the OWL DL ontology axioms as represented in the ontology (bioOntology.owl) file.

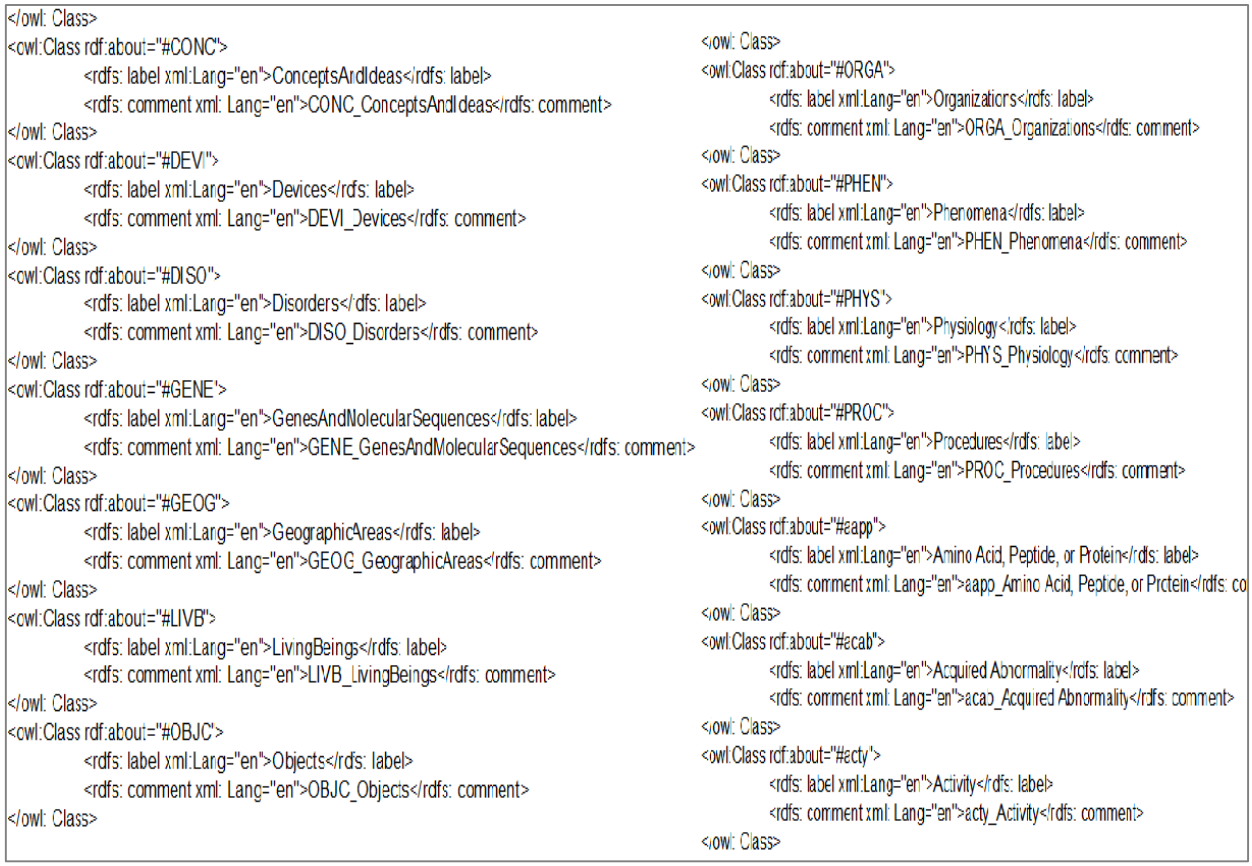

Figure 4.15 – Snapshot of the Ontology Axioms

4.5.5 Contributions and Challenges

Knowledge representation in graph formalism (e.g. conceptual ontology) is ambiguous and computationally less tractable [327]. For unambiguous representation of biomedical knowledge and computational tractability, a logic-based formalism, for example using OWL DL, is required to represent ontological knowledge. Consequently, a restriction-based interpretation is used for unambiguous representation of biomedical concepts and their associations, which results a set of ontological axioms.

In a restriction-based interpretation, concept association interpretation may be either universal-restriction-based or number restriction-based or existential restriction-based interpretations. In this research, number restriction-based interpretations are generalized into existential restriction-based interpretations. According to experimental analysis based on domain expert judgment, existential restriction-based interpretations of biomedical concept associations are the most

frequent (63%) as compared to universal restriction-based interpretations (11%). Biomedical concept association interpretation is therefore generalized to existential restriction-based interpretation. Thus, independent biomedical semantic interpretation is realized as compared to interpretations supported by knowledge engineers and ontology tools (e.g. protégé), which minimizes efforts, time and costs incurred in formal ontology learning.

The use of cardinality information for each restriction-based interpretation is left as research challenges yet. Knowing all the individuals in each concept enables to determine the number of individual in a concept related to another concept through the defined semantic predicates. The cardinality information, therefore, enables to determine universal restriction-based interpretation, number restriction-based interpretation and existential restriction-based interpretations of the biomedical concept associations independently.

4.6 Integrity and Consistency across Formalisms

The proposed framework passes through different representation formalisms to acquire ontological knowledge from biomedical texts. Initially, domain artefacts are encoded in natural language lexical, syntactic and semantic structures. These biomedical artefacts are acquired from biomedical text collections and structured into graph formalism and finally to logic formalism. In such transformation-based ontology acquisition, integrity and consistency across formalism is very crucial. Consequently, integrity and consistency depends on a set of formalisms M and computational theories, O, which are formulated as in the following.

$$M \cong M_i, M_{i+1}, ..., M_{n-1}, M_n$$
$$O \cong O_i, O_{i+1}, ..., O_{n-1}, O_n \qquad 4.10$$

In this study, four representation formalisms are defined: 1) natural language formalisms where the biomedical knowledge in the text is encoded; 2) a set of semantic triples/knowledge artefacts in which biomedical artefacts acquired from biomedical texts are represented; 3) the graph formalisms where the conceptual ontology structure is defined, e.g Direct Acyclic Graph; and 4) finally, the OWL DL formalism where the conceptual knowledge is expressed with OWL DL Axioms. The NLP computational theories are used to transform knowledge from biomedical texts to knowledge artefacts. Semantic and graph theories are used to prune and structure the

knowledge artefacts to graph structure and finally the OWL DL computational theories are used to formalize the conceptual knowledge.

In transforming knowledge from formalism to another, consistency is measured based on the number of semantic concepts, roles and their associations acquired from biomedical texts and how many of them are used to structure the conceptual ontology. Plus, how accurate is the interpretation of conceptual ontology into OWL DL axioms. Consequently, in measuring the process of acquiring biomedical artefacts, more than 1.5 million concepts and 6 million associations are extracted. Out of these figures, 1.2 concepts and 5.2 associations are used for conceptual ontology structuring. Although the result is promising, consistency across formalisms posed serious challenges in the perspective of attribute and cardinality information acquisition, and universal quantifier based interpretation in the proposed biomedical knowledge acquisition framework.

Chapter Five Evaluation of the Proposed Framework

The proposed ontology learning framework is instantiated and the results are illustrated. This chapter evaluates the correctness and quality of the structural aspect of the framework and presents evaluation results. Thus, evaluation is designed to assess the extent of approximating biomedical knowledge (i.e. correctness) and quality of the ontology structure design. As stated, while ontology structure is influenced with structural parameters, such as complexity, adaptability, schema potential and clarity, correctness is influenced with accuracy, completeness, conciseness and consistency of the acquisition process. Consequently, extent of approximating biomedical knowledge and quality of the structural design are determined by measuring the structural properties of the proposed framework.

In this research, while extent of approximation is evaluated by measuring precision and recall, coverage, relevance and inconsistencies of the acquisition process, ontology structure quality is evaluated by measuring Size of Vocabulary (SOV), Tree ImPurity (TIP), Edge Node Ratio (ENR) and Entropy of Graph (EOG). These properties enable to measure complexity of the ontology structure. Others are Number of Leaf nodes (NoL), Number of Root nodes (NoR), Average Depth of Inheritance Tree of Leaf Nodes (ADIT-LN), Number of External Classes (NEC) and Number of External Roles (NER). These graph properties enable to measure the adaptability of the ontology structure in terms of its cohesiveness and coupling. Relationship richness (RR), Attribute Richness (AR) and Inheritance Richness (IR) are also graph properties for measuring ontology schema potential.

A criteria-based approach is used to evaluate correctness and quality of the proposed framework for its less computational costs. Thus, eight criterions expressing different properties of an ontology structure are considered for evaluation. Each criterion, in turn, is measured with a combination of different metrics, each of which describing a property of the ontology structure (G_o).

5.1 Evaluation Criteria and Metrics

Exhaustive utilization of criteria for evaluating knowledge acquisition is very difficult and resource intensive. Consequently, eight criterions (accuracy, completeness, conciseness, consistency, complexity, adaptability, schema potential and clarity) are adapted to form a coherent and succinct set for evaluation. However, all of them may not perform equally; even some of them seem contradicting, for example conciseness and completeness. It is also note that none of these criteria can be directly measured and most of them may not be perfectly achieved.

A criterion is quantified using metrics, meaningful properties of ontology structure. For example, coverage and relevance are metrics to measure completeness and conciseness respectively, and precision and recall are metrics to measure accuracy. In criteria-based approach, for each criterion ϕ, there can be metrics $m_1, m_2 ... m_n$ whose values are $v_1, v_2, ..., v_n$ respectively. A criterion is measured based on metrics values $v_1, v_2, ..., v_n$. In this consideration, the following metrics are measured for the respective criterion.

Accuracy metrics determine extent of divergence from the background knowledge. It is measured as a total number of artefacts correctly acquired over the set of artefacts (K_p) (i.e. precision), plus, the total number of correctly acquired artefacts over all knowledge that should have been found (i.e. recall). Consequently, accuracy is equivalent to percentage values of precision and recall values of biomedical concepts, roles and their associations or the proportion of correctly extracted artefacts. More formally, precision (p) and recall (r) are computed as in the following formulations.

$$precision \quad (p) = \frac{Ex_{relevant}}{Ex_{relevant} + Ex_{non-relevant}} \tag{5.1}$$

Where, $Ex_{relevant}$ or $Ex_{non-relevant}$ are the number of relevant or non-relevant artefacts disambiguated from the bioMed text collection. Recall is also formulated as:

$$recall \quad (r) = \frac{Ex_{relevant}}{Ex_{relevant} + NEx_{relevant}} \tag{5.2}$$

Where, $NEx_{relevant}$ is the number of relevant artefacts but not disambiguated in the bioMed text collection (K_s). In this context, accuracy is considered as equivalent to correctly acquired relevant knowledge artefacts that are disambiguated. The pseudo-algorithm for computing accuracy is also as shown in the following.

$$Accuracy \cong \frac{Ex_{relevant} + NEx_{relevant}}{Ex_{relevant} + NEx_{relevant} + Ex_{non-relevant} + NEx_{non-relevant}} \tag{5.3}$$

Pseudo-algorithm – computing accuracy of correctly acquired knowledge artefacts

```
//input_1=number of extracted knowledge artefacts, K_p
//input_2=number of domain artefacts in the background knowledge, K_b
Ex_relevant =relevant number of knowledge artefacts in K_p
Ex_non-relevant =non-relevant number of knowledge artefacts in K_p
NEx_relevant =relevant knowledge artefacts in K_b, but not extracted
NEx_non-relevant =non-relevant knowledge artefacts in K_b
p = Ex_relevant/(Ex_relevant + Ex_non-relevant), r = Ex_relevant/(Ex_relevant + NEx_relevant)
Accuracy = (Ex_relevant.NEx_relevant)/(Ex_relevant + Ex_non-relevant + NEx_relevant + NEx_non-relevant)
```

Completeness metrics measure the proportion of acquired knowledge artefacts compared with over all artefacts in the background knowledge. That is, it measures how exhaustive is the framework to acquire knowledge artefacts from biomedical texts. It is computed as the total coverage of knowledge artefacts (K_p) over the total artefacts in the background knowledge (K_b). Coverage or completeness is therefore computed as:

$$Coverage = \frac{K_p}{K_b} \tag{5.4}$$

The formulation computes the coverage of biomedical concepts, roles and their associations. Coverage values of concepts, roles and their associations are analyzed to discuss the completeness of the framework. A simple algorithm for computing coverage, and thus, completeness is developed as follows.

Pseudo-algorithm – computing coverage and completeness

```
//inputs=number of acquired knowledge artefacts, K_p
//input=number of biomedical artefacts in the background knowledge, K_b
Coverage = |K_p| / |K_b|
Completeness = Coverage
```

Conciseness metrics determine whether irrelevant knowledge artefacts or redundant representations of semantics are defined. On the other hand, it determines the coverage of relevant biomedical artefacts and their associations. It is measured by computing coverage of relevant knowledge artefacts and their associations (K_r) over the set of acquired knowledge artefacts (K_p). Values of these knowledge artefacts are analyzed to discuss the conciseness of the acquisition process. More formally, conciseness (Γ) is computed as the ratio of the number of relevant knowledge artefacts involved in the ontologization (K_r) to the total number of disambiguated artefacts in the bioMed text collection (K_p), i.e.

$$\Gamma = \frac{K_r}{K_p} \tag{5.5}$$

This formulation computes the conciseness of biomedical concepts, roles and their associations. A simple algorithm is developed for computing conciseness as in the following.

Pseudo-algorithm – computing relevance, and thus, conciseness

```
//inputs=number of acquired knowledge artefacts, K_p
//output=conciseness, Γ
//relevant artefact=involved in the ontology structuring
K_r=relevant knowledge artefacts in K_p //relevant set of artefacts
Γ = |K_r| / |K_p| //conciseness of the framework/ontology
```

Consistency metrics measure how consistent the framework is in approximating the biomedical artefacts, knowledge. They measure number of inconsistent definitions and artefacts produced. Particularly, they compute partitioning and cyclic inconsistencies introduced in the acquisition

process. Consequently, inconsistency is computed as the proportion of cyclic redundancies and wrong partitioning. Thus, inconsistency is formulated as a ratio of counted inconsistencies (K_{inc}) (wrong partitioning and cycles) to the total knowledge artefacts (K_p).

$$Inconsiste \quad ncies = \frac{|K_{inc}|}{|K_p|} \tag{5.6}$$

This formulation computes inconsistencies of biomedical concepts, roles and their associations. The values are used to analyze and discuss the consistency of acquiring knowledge artefacts. An algorithm is developed to compute inconsistencies (K_{inc}) and consistencies (K_{con}).

Pseudo-algorithm – computes the inconsistencies

```
//input= extracted knowledge artefacts, K_p
//input=set of ontology hierarchies, H_c
For each hierarchy, h_c
      Get cyclic links, c_1
      Get wrong partitions, r_p , if any
      If c_1 or r_p exist
        Increment incon_counter by 1 or 2
        //1x, only when one inconsistencies happened, otherwise increment 2x
      End
End
Set inconsistency= incon_counter/K_p  //inconsistencies
Set consistency=(|K_p| – incon _ counter ) /|K_p|  //consistencies
```

C*omplexity Metrics* measure the computational efficiency of the ontology structure. Consequently, complexity is measured by computing the Size of Vocabulary (SOV)), the Edge-Node Ratio (ENR), Tree ImPurity (TIP) and Entropy of the ontology Structure (EOG). The larger metrics values, the more cognitive resources are required to understand and maintain the ontological knowledge and therefore greater complexity.

SOV is the total number of biomedical concepts (N_n) and roles (R_n) defined in the ontology structure (G_o). The larger SOV means, larger size of the ontology and longer time and effort required to build and maintain it. It is defined as the cardinality of named artefacts, N_n and R_n, in G_o as:

$$SVO = |N_n| + |R_n| \tag{5.7}$$

Where, N_n is number of defined concepts and R_n is number of defined properties or roles. ENR represent the connectivity density, number of relationships per concept. The larger ENR value means, greater complexity of the ontology structure (G_o).

$$ENR = \frac{|E|}{|N|} \tag{5.8}$$

Where, $|E|$ is the number of edges and $|N|$ is the number of nodes (concepts) in G_o. TIP measures how far the ontology inheritance hierarchy structure is deviated from a pure tree structure. It is computed as:

$$TIP = |E^{'}| - |N| + 1 \tag{5.9}$$

Where, $|E^{'}|$ is the number of *ISA* edges, and $|N|$ is the number of nodes (concepts) in the hierarchical structure. The greater the TIP value, the more ontology inheritance hierarchy deviates from a pure tree structure and greater complexity of the ontology structure. Note, we considered the top class ***bioMed*** and ***owl: Thing***. Each class node c with no explicit supperClass nodes have an edge added for it: (c, *rdfs: subClassOf, bioMed*) and (*bioMed rdfs: subClassOf, owl: Thing*), in which such additions ensure that TIP is always non-negative.

EOG is the entropy of the ontology structure (G_o), which measures how diverse the ontology structure is. Lower *EOG* indicates regular structure and lesser complexity. *EOG* is formulated as:

$$EOG = -\sum_i p(i) \log_2 p(i) \tag{5.10}$$

Where, $p(i)$ is probability of a node i having n edges, incoming and outgoing degrees. An algorithm is developed to compute the four complexity metrics as in the following.

Pseudo-algorithm – computes complexity of the ontology structure

```
//input= acquired knowledge artefacts, K_p
//input= the ontology structure, G_o
  N=unique number of concepts in G_o
  R=unique number of roles in G_o
  SOV = |N| + |R|    //vocabulary size
  E=number of edges in G_o    //r/ships among concepts
  ENR = |E|/|N|    //connectivity density
  E' = number of subsumption edges   //isa links
  TIP = |E'| - |N| + 1    //sets the tree impurity
  D= sum of degrees for all nodes in G_o
  For each node in G_o ,
      d=degrees of node i in G_o //both incoming and outgoing
      p(i)=d/D //probability of node i having d degrees
      EOG=-(EOG + p (i)*log2p (i)) //diversity of the ontology structure
  End
```

Adaptability Metrics measures the ease of use of the ontology for different contexts, possibly by extending it without the need to remove existing axioms. It is measured by computing the coupling and cohesiveness of the ontology structure (G_o). Coupling refers to the number of external classes and roles referenced, where as cohesiveness refers to the Number of Root (*NoR*) nodes, Number of Leaf (*NoL*) nodes and Average Depth of Inheritance Tree of Leaf Nodes (*ADIT-LN*).

Coupling is computed as the Number of External Classes (*NEC*) and Roles (*NER*) referenced in the ontologization, i.e.

$$Coupling \equiv |NEC| + |NER| \tag{5.11}$$

Cohesiveness is computed based on *NoR*, *NoL* and *ADIT-LN*. The *NoR* class is the number of root nodes explicitly defined in the ontology structure (G_o). It is computed as in the following where *n* is the number of root classes in G_o.

$$NoR\ (G_o) = \sum_{j=1}^{j=n} 1 \tag{5.12}$$

NoL is the number of leaf nodes (concepts) explicitly defined in the ontology structure (G_o). NoL is computed as in the following where *m* is the number of leaf nodes in G_o.

$$NoL(G_o) = \sum_{i=1}^{i=m} 1 \tag{5.13}$$

ADIT-LN is the ratio of the sum of depths of all paths to the total number of paths in the ontology structure (G_o). Depth of a path is the number of nodes in the path, and sum of depths is the total number of nodes in all paths of the ontology structure (G_o).Total number of paths is all the distinct paths from root node to each leaf node. ADIT-LN is computed as in the following where, D_j is the total number of nodes in the j[th] path, $1 \le j \le n$, and *n* is the total number of paths in G_o. An algorithm is developed to compute adaptability in terms of coupling and cohesiveness.

$$ADIT - LN\ (G_o) = \frac{\sum D_j}{n}, \forall D_j \tag{5.14}$$

Pseudo-algorithm – computes adaptability of the ontology structure (G_o)

```
//input= extracted knowledge artefacts, K_p
//input= the ontology structure, G_o
NEC= number of external concepts included in G_o //=0
NER= number of external roles included in G_o //=0
coupling = |NEC| + |NER| //referenced concepts and roles, =0
NoR=number of root concepts in G_o
NoL=number of leaf concepts in G_o
sumOfDepth=sum of depths of all paths in G_o
```

$Path_{tot}$ =sum of all paths in G_o

ADIT-LN = sumOfDepth/ $Path_{tot}$ //avg depth of inheritance tree of LNs

Clarity Metrics measure how effective the ontology communicates the intended meaning of each biomedical term, which refers to term senses. Hence, it focuses on the human readable descriptions of each term in the ontology, such as comments, labels or descriptions. Specifically, it computes the readability (RD) of each class c_i or role r_i in the ontology. It is defined as the average number of attributes of concepts and roles in the ontology. This is formulated as in eq.6.15 where an algorithm is developed to compute readability as in the following.

$$RD = |A, A = rdfs:comments| + |A, A = rdfs:label| + |A, A = rdfs:description| + ... \quad (5.15)$$

Pseudo-algorithm – computes clarity of the ontology structure (G_o)

//input=the ontology structure, G_o

nod_{tot} =number of concept nodes

$role_{tot}$ =number of unique edge links

For each node, *nd*, and role, *r*,
    attr_nod=number of attributes of nodes, nd
    attr_r=number of attributes of *edge links, r*
End

$avg\_rd\_node \equiv |attr\_nod| / |nod_{tot}|$ //readability of nodes/concepts

$avg\_rd\_edge \equiv |attr\_r| / |role_{tot}|$ //readability of edges/roles

Ontology schema metrics measure the potential of the ontology structure to represent knowledge. This is measured in terms of semantic richness. Semantic richness is defined using Relationship Richness (RR), Attribute Richness (AR) and Inheritance Richness (IR). RR reflects the diversity of relations in the ontology structure, which is a ratio of the number of relationships (R) to the sum of the number of subclasses (SC) plus the number of relationships in the ontology structure. It is computed as:

$$RR = \frac{|R|}{|R| + |SC|} \quad (5.16)$$

AR is the number of attributes (slots) defined for each class. Note that the more slots defined, more knowledge the ontology conveys. Attribute richness is defined as the average number of attributes (slots) per class. It is computed as the ratio of number of attributes for all classes ($|attr|$) to the number of classes ($|C|$).

$$AR = \frac{|att|}{|C|} \tag{5.17}$$

IR describes the distribution of information across different levels of the ontology inheritance tree, fan-out of parent classes. This is an indication of how well knowledge is grouped into different categories and subcategories in the ontology structure. IR is computed as the average number of subclasses per class, where the number of subclasses (C_l) for a class C_i is defined as $|H^c(C_l, C_i)|$ and $|C|$ is the number of concepts in the ontology structure (G_o).

$$IR = \frac{\sum_{C_i \in C} |H^c(C_l, C_i)|}{|C|} \tag{5.18}$$

Pseudo-algorithm – computes semantic richness of the ontology structure (G_o)

```
//input=the ontology structure, G_o
//output= RR, AR and IR
RR=number of non-isa edges in G_o
SC=number of isa edges in G_o
RR = |R|/(|R| + |SC|)  //relationship richness
attr_tot=number of attributes of concept nodes in G_o
N_nod=number of concept nodes in G_o
AR = |attr _ tot|/|N _ nod|  //attribute richness
C= N_nod //number of concept nodes in G_o
For each concept, c_i, in G_o and  c_i not leaf concept//BFA
    sib_count=number of siblings of c_i
    tot_count= tot_count + sib_count
End
```

$IR = |tot\_count|/|C|$ //inheritance richness

5.2 Evaluation Results

Analyzing the metrics values enabled to have insights about correctness and quality of the acquired ontology. Thus, this section provides the experimental setup and the analysis of metrics values obtained from computations to understand the characteristics of the ontology and its structure.

5.2.1 Experimental Setup

A set of biomedical artefacts, such as concepts, predicates (roles) and their associations (semantic predictions), are acquired from the bioMed text collection. This set of artefacts and their structuring approximated the biomedical knowledge as stated in the previous chapters. For reasons of computational efficiency, we have considered only relevant 986,340 concepts, 682 predicates and 5,221,148 concept associations to structure the ontology for evaluation experimentation. The size of knowledge artefacts identified for experimentation is shown in table 5.1. The first three columns are inventories of artefacts acquired from the bioMed text collection, and the rest three columns are inventories of biomedical artefacts identified as relevant.

Table 5.1 - Size of Acquired Knowledge Artefacts

#Concepts	*#Roles*	*#Associations*	*#R. Concepts*	*#R. Roles*	#R. Associations
1,447,169	682	5,820,062	986,340	682	5,221,148

Python programming language, in windows and Linux environment, is used to write necessary codes (scripts) for experimentation. Accordingly, eight experiments are performed corresponding to each criterion: accuracy, completeness, conciseness, consistency, complexity, adaptability, schema potential and clarity. Experiments are performed by computing metrics values for each criterion and analyzing the results of computations. Consequently, experimental results shown in section 5.2.2 and 5.2.3 below illustrates the correctness of the proposed framework. Whereas,

experiments performed in sections 5.2.4, 5.2.5 and 5.2.6 illustrates the quality of the proposed framework.

5.2.2 Measuring Accuracy

Accuracy is the amount of correctly acquired knowledge artefacts from the biomedical texts. Thus, accuracy of knowledge artefact (concept, role and their association) is shown in table 5.2.

Table 5.2 - Accuracy of the Knowledge Artefacts

Artefacts	*Ex_relev*	*Ex_non_relev*	*Nex_relev*	*Nex_non-relev*	*Precision*	*Recall*	*F-measure*	*Accuracy*
Concept	986,340	460,829	96,328	223	**0.68**	**0.91**	**0.78**	**0.70**
Role	682	13	56	17	**0.98**	**0.92**	**0.95**	**0.96**
Association	5,692,148	598,914	1,123,658	3,213	**0.90**	**0.84**	**0.87**	**0.92**

As shown in the table, accuracy of concept extraction is 70%, role extraction is 96% and that of concept association extraction is 92%. This implies that extraction of concept associations, ones the ontological predicates and concepts are set, shows noticeable effectiveness. Similarly, both concept and semantic predicate extraction is illustrates good effectiveness.

5.2.3 Measuring Completeness, Conciseness and Consistencies

As previously pointed out, completeness, conciseness and consistencies are measured by computing the coverage, relevancy coverage and inconsistencies respectively of the acquired ontological knowledge. The result of computing these metrics is shown in table 5.3.

Table 5.3 - Completeness, Conciseness and Consistency

Artefacts	*Coverage*	*Relevance*	**Inconsistency**	**Completenes**	**Concisness**	**Consistency**
Concept	0.82	0.68	0.00	**0.82**	**0.68**	**1.00**
Role	0.93	1.00	0.00	**0.93**	**1.00**	**1.00**
Association	0.59	0.90	0.00	**0.59**	**0.90**	**1.00**

In the table, coverage and relevant coverage are equivalent to completeness and conciseness respectively. This is because, completeness is the amount of biomedical artefacts acquired from the domain text as compared with the background knowledge. But, there is no partitioning and cyclic inconsistencies found in the ontology structure. These results inconsistencies to be zero, this, in turn, results consistency to be one.

5.2.4 Measuring Complexity

Efficiency of learned ontology is evaluated by measuring the complexity of its structure. Thus, complexity is measured by computing four major metrics, namely SOV (size of vacabulary), ENR (edge-node ratio), TIP (tree impurity) and EOG (graph entropy). Accordingly, the computational results of these metrics are shown in table 5.4, considering relevant artefacts.

Table 5.4 – Complexty Metrics

Metrics	SOV	ENR	TIP	EOG
Values	986,448	18.73	145,246	226,698.00

The vocabulary size is 986,448, showing large ontology size. The edge-node ratio (18.73) shows the connectivity density, on average about 18 edges per node. The tree impurity (145,246) shows large divation of the ontology structure from pure tree structure. Finaly, the ontology structure has large graph entropy. All these demonistrate complex ontology structure.

5.2.5 Measuring Adaptability

In a similar fashion, adaptability is measured based on cohesiveness and coupling metrics. While cohesiveness is measured by computing NoR and NoL nodes and ADIT-LN of the ontology structure, coupling is measured by computing NEC and NER. Consequently, the result is shown in table 5.5.

Table 5.5 – Adaptability Metrics

Sub-Dims	NoR	NoL	ADIT-LN	NEC	NER
Cohesiveness	1	545,246	23	-	-
Coupling	-	-	-	0	0

In the table, NoR is 1 which means that the ontology has only one root node and highly cohesive. Furthermore, NEC and NER are not defined (hyphen, '-') for cohesiveness, and NoL, NoR and ADIT-LN are not defined (hyphen, '-') for coupling. Values of NEC and NER are zero, which means that there are no external classes and roles the proposed framework imported in the learning process, and thus, the ontology produced is higly decoupled.

5.2.6 Measuring Schema Potential and Readability

Ontology schema potential and clarity are also evaluated using Semantic Richness (SR) and readability (RD). While semantic richness is measured by computing relationship richness (RR), Attribute Richness (AR) and Inheritance Richness (IR), readability (RD) is measured by computing the average numbre of class level attributes. Consequently, the result of ontology shema potential metrics values are shown in table 6.6.

Table 5.6 – Ontology Shema Potential metrics

OSMs	RR	AR	IR
Values	0.8014	3	13,253

In the table, AR is three which means that aeverage number of attributes for each class is 3. Consequently, readability becomes 3, which measures the clarity of the ontology structure.

5.3 Measuring Correctness and Quality

5.3.1 Correctness

Correctness is an extent of approximating biomedical knowledge by the domain model, proposed framework. It is measured based on a set of criteria related to the property of the ontology structure (G_o), where each criterion is, in turn, expressed with a set of values computed for each metrics. As shown in table 5.2 in section 5.2.2, accuracy is found to be 70%, 96% and 92% for concepts, roles and associations respectively.

Completeness, conciseness and consistency are also shown in table 6.3 in section 6.2.3. Completeness is 59%, 93% and 82% for semantic associations, roles and concepts respectively. Conciseness is 68%, 100% and 90% for concepts, roles and semantic associations respectively, where as consistency is 100% as there is no cyclic and partitioning redundancies found in the acquisition process. These results showed better extent of acquiring biomedical knowledge from biomedical text.

5.3.2 Quality

In this study, quality is measured related to the ontology structure design efficiency. This is related to the computational complexity, coupling and cohesiveness, schema potential and clarity of the ontology structure. Each of these criteria is measured using a set of metrics, whose values are computed based on their formulations. Thus, a set of metrics values are analyzed for discussions. Sections 5.2.4, 5.2.5 and 5.2.6 illustrate results for each criterion, such as complexity, adaptability, schema potential and clarity.

Consequently, in table 5.4, measurements for SOV (=986,448), ENR (=18.73), TIP (=145,246) and EOG (=226,698) showed larger values. Larger values for each metrics, more complex are the ontology structure, and thus, higher complexity the ontology has. In table 5.5, the measured values for cohesiveness and coupling revealed that the ontology is more cohesive and decoupled. The ontology has only one root node and larger number of leaf nodes (545,246) and depth of inheritance tree (23). The ontology doesn't used external classes and roles, and hence, coupling is zero, decoupled.

Table 5.6, in section 5.2.6, also illustrated that the ontology structure has higher potential to hold factual knowledge in the domain. This is revealed with larger values for semantic richness metrics: relationship richness (RR =0.80), attribute richness (AR=3) and inheritance richness (IR=13,253). The measured values of readability (RD=3) is also showed an average value in which it shows good clarity of each class in the ontology.

5.4 Comparisons with PIKES Framework

After a through review of PIKES knowledge extraction framework, we understood that the PIKES framework has two major phases[1]: a deep linguistics analysis and instance level knowledge integration from multiple sentences. The linguistics analysis uses NLP tools (POS tagging, named entity recognition, dependency parsing, co-referencing and semantic role labeling) and results mention graphs in which every mention is characterized with attribute and relationship properties [329]. The instance level knowledge integration, knowledge distillation, aggregates instances from different sentences in the mention graph and builds instance level knowledge, forming RDF triplets [329]. The ultimate result of PIKES framework therefore is a linked data, similar to DBpedia database. Although linked data is a potential of assertional knowledge, it lacks ontological knowledge structuring and representation.

PIKES uses deep linguistics analysis tools (semantic role labeling, dependency parsing, co-referencing, named entity recognition and POS tagging) and results a set of RDF triplets as mention graphs. Lastly, based on lexical knowledge, such as nombank, probank, verbnet, framebase, pikes distilled the mention graph to knowledge graph. This shows lexical relationships in the form of graph, but not the world knowledge encoded either explicitly or implicitly in the discourse in an ontological formalism. For example, co-reference is a lexical relationship but not ontological relationship. In description logic concept, it lacks the Tbox, corresponding to ontological schema knowledge. Furthermore, there is no clear hierarchical (taxonomic) structuring and classes of individuals/mentions, the major benefit of ontological knowledge representation, where reasoning services are possible. PIKES deals with individual/assertional knowledge (Abox) but the taxonomies are not explicitly structured and a lot of cyclic representations exists that results inferencing impossible and intractable. Our framework deals with the ontology schema knowledge (Tbox) with ontological formalisms, having clear subsumption knowledge or taxonomies. But, as demonstrated in the limitation, the framework lacks knowledge about individuals, assertional knowledge.

Consequently, empirical comparison between the two frameworks, Pikes and proposed framework, are impossible. This is because of that: 1) Pikes is not an ontological knowledge

[1] http://pikes.fbk.eu/index.html

learning framework, but rather a set of instances and their association in RDF triplets. 2) The proposed framework is an ontological knowledge learning framework with defined taxonomies and Direct Acyclic Graph Structure. This supports intelligibility and interoperability services anticipated from ontological knowledge representation.

However, the major limitations of the proposed framework (cardinalities and attributes information, and domain adaptability) can be supplemented with the results of Pikes framework. The Pikes framework produces a set of assertional knowledge (instances and their associations) and these can be used by the proposed ontology acquisition framework to populate its schema ontology, referred as ontology population. This in turn avoids the limitations on universal quantifier based interpretation in the formal representation of our framework. Furthermore, although Pikes framework is domain independent, our framework is dependent to a specific domain where background knowledge exists. Pikes could support our framework domain adaptability using adaptive learning techniques.

5.5 Findings of the Study

Semantic disambiguation includes phrase segmentation, phrase and semantic association disambiguation. While existing tools (MetaMap) are applied for phrase segmentation and disambiguation, an enhanced semantic processing program is used for predicting semantic associations between biomedical concepts. SemRep program used only 54 semantic predicates for associatiating biomedical concepts as a result most biomedical concepts are left alone. But, enhanced-semRep program adresses this problem by devloping an algorithm that look for valid associations in the background knowledge for each pair of concepts in each sentence. This enhances the predicted semantic assoction from 5 million (only semRep program) to 7 million (enhanced semRep program). This enhanced the extent of approximating biomedical knowledge from free biomedical texts. Thus, the degree of approximating biomedical knowledge is significantly improved in which it is measured with a set of criteria and their metrics.

The set of four criterion are selected and measured by computing their metrics. The first criterion is accuracy, which is measured by computing precision, recall and F-measure. Completeness and conciseness are also two criteria measured by computing coverage and relevant coverage

respectively. Consistency is the fourth criterion, which is measured by computing inconsistencies in the acquisition process. In this work, however, only two aspects of inconsistencies, cyclic and partitioning redundencies, where both of them are found to be zero, are considered. Consequently, an improved degree of biomedical knowledge approximation is dimonstrated.

Existing ontology learning frameworks structure and interpret domain knowledge with the help of domain experts and ontology engineers. This study, however, accomplishes these tasks without the involvement of domain experts and ontology engineers, i.e. automatically. Automated ontology structuring and interpretation could minimize time, effort and costs incurred for ontology learning. Background knowledge (e.g. UMLS) is used to suggest conceptualization, and structuring of the anticipated ontology structure. The proposed framework, thus, uses this knowledge for modeling and structuring biomedical artefacts into a direct acyclic graph. The framework also uses OWL DL primitives and constructs for interpreting (axiomatizing) the conceptual structure (G_o).

Existing ontology learning frameworks use ontology tools for interpreting semantic associations. For independent ontologization, restriction-based interpretation is used for axiomatizing the conceptual structure. To decide which restrictions to apply for each semantic association interpretation, experimental analysis has been made to get insights about interpretation trends. After analyzing the trends, existential restriction-based interpretation is used and implemented. Consequently, in this work, independent interpretation is practiced instead of using ontology tools and engineers.

In ontology learning, one of the major challenges is that texts are represented in natural language formalism whereas ontologies are represented with structured language formalism, for example OWL DL. Ontologies are expected to represent the possible semantics represented or implied in the domain texts so that any semantics in the domain text must also exist in an ontology acquired from the text. This requires consistent and integrated ontology acquisition methods and techniques across formalisms, from natural language to structured ontology representation. Such ontology acquisition has provided less attention in the existing ontology learning frameworks. In this work, however, emphasis is provided to keep natuaral language semantics using scenario based interpretations suggested by a prior expectations. The prior expectations represent domain

semantics, which could appear in the text collection. Consequently, the conceptual disambiguation models and structures biomedical artefacts (concepts, roles and their associations) produced by the semantic disambiguator into a conceptual structure, conceptual ontology. Once the conceptual ontology (G_o) is constructed, it is interpreted for its formal representation based on OWL DL primitives and constructs. A restriction-based interpretation is applied on concepts, roles and their associations for better approximation of the representation. In this way, an integrated consistency across representation formalisms are achieved.

In chapter two, table 2.1 summarized common and recent knowledge acquisition frmaeworks and discribed them using ten (10) charactersitics features. In the table, scalability is in the order of thansands (<5,000 concepts) where as in the proposed work it is in the order of millions (≈ 4 million concepts and their associations). The components of existing frmaeworks are also developed for different interests at different times. This produced problems of integration among components of a particular framework. In the proposed work, however, all components are designed and developed for ontology learning at the same time, and thus, integrity is not an issue.

Existing ontology learning frmaeworks are data-driven, and thus, reaching consensus among domain experts and users are very difficult, and even never achievable, with data-driven methods. To ease this problem, the proposed work used an already consensus-reached semantics (the UMLS), and referred it as a knowldege-based method. Consequently, each semantics interpretation is suggested by already consensus reached semantics. The proposed framework is independent from ontology engineers and domain experts (manual works), and thus, geared towards fully automatic acquisition. In the existing frameworks, domain relevance is determined using TFIDF or its derivatives (e.g. KFIDF). These techniques are designed for discriminating documents rather than artefacts, and thus, highly inefficient and ineffective in ontology learning. In the proposed work, domain relevance is determined using the backgrpound knowledge, which is pre-built and consensus-reached biomedical knowledge.

Exhaustive domain aretefact learning is very difficult in the existing ontology learning frmaeworks in which scalability is very limited. The proposed framework, however, is highly scalable and exhuastive in acquiring biomedical artefacts (concepts, roles and their associations).

Consequently, over 4 million concepts and 7 million associations are acquired from the bioMed text collection. Furthermore, ontology structuring is manual in the existing ontology learning frmaeworks but automatic in the proposed framework. The manual structuring is supported with knowledge base, commonly Wordnet. Wordnet, however, is a lexical knowledge comprised of lexical information rather than semantic information. Plus, it is a shallow knowledge about language charactersitics than the domain semantics. Interpretations of domain artefacts are also manual, supported with ontology tools (e.g. protégé). But in this work, interpretation is automatic using ontology language primitives and constructs (e.g. OWL DL). Generally, while existing ontology learning frameworks are supported with ontology tools (e.g. ontoEdit), this framework doesn't require it.

Chapter Six Conclusions and Way Forwards

Knowledge acquisition systems have become increasingly emerging technologies. They have been set as the most emphasized research directions recently by the scientific communities, and driven with the emergency of semantic web and intelligent information processing. One of the core components of these systems is prior expectation, which represent the world of intended applications. Ontology is semantically-rich representation formalism, which enables interoperability, integrity, intelligibility and knowledge sharing among domain applications. Consequently, it is a core component of semantic web applications, particularly in the context of integration and information sharing.

In this context, several investigations have been forwarding ways of acquiring and representing domain knowledge from heterogeneous sources, namely from biomedical texts. The challenge is that acquisition of such knowledge required intensive expert efforts, time and huge investments. Even with intensive resource mobilization, most constructed ontologies are either very shallow (e.g. Cyc and Galen) or very narrow (e.g. GO), which have poor coverage, organization and formalization. These ontologies are not deployable in practical application scenarios. In order to strengthen knowledge acquisition and ontology construction, literatures have recommended the need to design methodologies and frameworks that enable to acquire and represent ontologies from their sources at large scale with minimal expert involvement, time and investment, reducing the knowledge acquisition bottleneck.

In order to contribute in this direction, this research investigates the development of ontology acquisition framework from biomedical texts. The proposed framework is designed, constructed and instantiated using biomedical text collection and background knowledge (UMLS). The instantiated framework is also evaluated for its correctness and quality using a set of criteria. The evaluation result showed better extent of approximation and quality of the ontology structure.

The framework formulates the ontology (K_o) as six tuples. $K_o = (\Psi, \psi, H_c, H_r, \Phi, A)$, where, Ψ is a set of primitive axioms, ψ is a set of primitive attribute axioms, H_c is a set of concept taxonomy axioms, H_r is a set of role taxonomy axioms, Φ is a set of non-taxonomy axioms and A is a set of assertional axioms. However, instantiation of this formulation, attribute and

assertional axioms are not disambiguated from the biomedical text. On the other hand, another two axioms are introduced, equivalent (ω) and disjoint (ϖ) axioms. As a result, the biomedical ontology is formulated and instantiated as six tuples, $K_o = (\Psi, H_c, H_r, \Phi, \omega, \varpi)$.

6.1 Conclusion

The objective of the research is to design and construct an ontology acquisition framework from biomedical texts, which enables to reduce effort, time and costs incurred in ontology learning. Although the framework requires further development for its maturity, it is possible to consider it as comprehensive ontology acquisition framework towards automatic ontology learning from unstructured biomedical knowledge sources. It is designed in a way that can easily be adapted and generalized into other unstructured sources, such as video, audio, image and video. As far as domain knowledge base is available, the framework can be generalized to multimedia, multimodal and multilingual unstructured sources in a domain. In a similar fashion, the proposed framework can also be adapted and generalized easily to other domains (e.g. agriculture and tourism) as far as the knowledge base is available.

The construction of the framework is also rigorous in dealing with different representation formalisms and their inconsistencies, Natural Language (NL), Conceptual Language (CL) and Formal Language (FL). Acquiring and representing knowledge in a consistent manner across these formalisms is very challenging and difficult in which most authors in the field point as a major cause of ontology learning bottlenecks. Thus, the proposed framework identified possible inconsistencies across different formalisms and points its limitations and possible extensions. While the framework has limitations to disambiguate individuals and entity attribute information, it has also drawbacks for interpreting restricted semantics, such as existential, universal and number restrictions based interpretations. Consequently, in instantiating the framework, an approximated interpretation, based on existential restriction, is implemented. Since it is more relaxed than others and most of the biomedical concept associations seem to lay in this category of restrictions.

Consequently, development and design of the proposed framework are based on three core semantic modeling and interpretations, which are designed as generic as possible. These are

semantic disambiguation, conceptual disambiguation and formal interpretation. A careful design and development of these models lead to consistent integration and cooperation of them for achieving ontology acquisition in biomedicine.

6.1.1 Semantic Disambiguation

Semantic disambiguation is comprised of phrase segmentation, phrase disambiguation and semantic proposition disambiguation. Phrase segmentation is performed with the help of specialist minimum commitment parser, which disambiguated syntactic structures. Syntactic analysis of each text fragment is performed using surface semantic analyzer, phrase segmentor. In disambiguating biomedical concepts, situation-specific scenarios in the text fragments are interpreted as suggested by scenarios in the background knowledge. Consequently, concept disambiguation is instantiated using MetaMap program. Semantic proposition disambiguation interprets semantic associations between concepts and semantic roles. A semantic processing (enhanced-semRep) program is used to instantiate and interprets semantic associations. This produces a conceptual structure based on graph representation formalism.

The semantic disambiguation is generic to analyze biomedical texts by chunking into argument and relational phrases and interpreting them with the support of background knowledge. That is, situation-specific scenarios of text fragments suggest knowledge instances, where as knowledge base scenarios suggest interpretation of each scenario in the text fragment. In a similar fashion, semantic association is interpreted with the help of semantic indicators, which suggest semantic links in the background knowledge. Indicator rules suggest interpretation of semantic indicators and background knowledge suggests interpretation of semantic associations in the text.

The semantic disambiguator is a crucial component of the framework, which has used a shallow phrase segmentor, the specialist minimum commitment parser. But, this parser is not exhaustive for segmenting argument and relational phrases. Concept and semantic association disambiguators have also their own drawbacks. They are less accurate to disambiguate biomedical concepts, roles and their associations. For example, enhanced-semRep used only 682 roles out of 736, which is very limited and affects the accuracy and completeness of approximating biomedical domain. They have also limitations to disambiguate and interpret entity attributes and individuals from free biomedical texts. The consequence is a limitation of

restricted interpretation (e.g. lack of cardinality information) in the formal representation of ontology.

Generally, instantiation of the semantic disambiguation generated more than four million biomedical concepts and seven million concept associations, semantic propositions. This set of biomedical artefacts approximated the biomedical knowledge from bioMed text collection. The extent of approximation is evaluated by measuring the correctness of the proposed framework using a set of criterion. A set of metrics are measured to evaluate each criterion and the results are very appreciable. For example, concept identification: 70% accuracy; 82% completeness; 68% conciseness; and 100% consistency. These results showed that the extent of domain approximation (correctness) is in a promising correctness, good extent of biomedical knowledge approximation.

6.1.2 Conceptual Disambiguation

The semantic association disambiguation leads to the design and development of the conceptual structure, which is instantiated with a Direct Acyclic Graph (DAG). Conceptual disambiguation constructed the conceptual design structure (G_o). Thus, biomedical knowledge is modeled into two knowledge layers based on granularities of domain artefacts. The upper knowledge layer categorizes and partitions the domain into sub-domain categories, such as *Anatomy* and *chemicalsAndDrugs* and semantic categories, such as *Disease or Syndrome* and *Pathologic Functions*. This layer is instantiated with 15 sub-domain categories and 135 semantic categories. Semantic categories are narrower than sub-domain categories in their semantics scope and they are disjoint partitions of sub-domain categories. These layers built the upper ontology and they are reused from existing upper biomedical ontologies.

Each of the 135 semantic categories is partitioned into more fine-grained biomedical concepts. More than 4 million biomedical concepts and 682 relationships are learned from the biomedical texts and included in the lower ontology structure (G_l). A set of fine-grained concepts, roles and their associations are acquired from biomedical texts by instantiating the semantic disambiguation model. After this, they are integrated to upper ontology structure. The integration results the over all biomedical ontology structure as a DAG (G_o).

Quality of the ontology structure is evaluated with a set of criteria and their metrics. Complexity, adaptability, ontology schema potential and clarity are used to assess the quality of the ontology structure (G_o). Quality measurement showed complex ontology structure with metrics values of 986,448 for vocabulary size, 18.73 for connectivity density, 145,246 for tree impurity and 226, 698 for graph entropy. The schema potential metrics values are also 0.80 for relationship richness, on the average 3 for attribute richness, and 13,253 for inheritance richness. Ontology clarity showed promising readability, which is found to be 3 attributes on average. Average attribute is computed as the ratio of the total number of concept attributes to that of the number of concepts in the ontology structure.

6.1.3 Ontologization

The semantic and conceptual disambiguation constructed the conceptual ontology (G_o), taxonomic and non-taxonomic structures. For unambiguous understanding and interpretation, the conceptual structure is interpreted and represented with OWL DL primitives and constructs. Consequently, the shared conceptualization of biomedical knowledge is formalized to a set of OWL DL axioms, from which we referred as OWL DL ontology (K_o).

The interpretation is limited to model the effectiveness and efficiency of OWL DL primitives and constructs. Expressivity and inferencing are also limited with the functionalities of OWL DL primitives and constructs. In this context, formal interpretation of the conceptual structure is logically formulated into six tuples, $K_o = (\Psi, \psi, H_c, H_r, \Phi, A)$, where Ψ is a set of primitive axioms, ψ is a set of primitive attribute axioms, H_c is a set of concept taxonomy axioms, H_r is a set of role taxonomy axioms, Φ is a set of non-taxonomic relation axioms and A is a set of assertional axioms. Technically, there are empty set of assertional (A) and attribute (ψ) axioms but another set of axioms are introduced, equivalent (ω) and disjoint (ϖ) axioms. Thus, the instantiated (K_o) is reformulated as $K_o = (\Psi, H_c, H_r, \Phi, \omega, \varpi)$.

In the axiomatization, root concepts are defined as the union of its children using *owl: unionOf* construct, where all of them are set as disjoint using *owl: disjointWith* construct. In addition, concepts in the taxonomic hierarchy are structured with a transitive relationship, *ISA*. The non-

hierarchical structure is anchored on non-isa relationships. Pair of concepts associated with this set of roles is interpreted using existential restrictions and the interpreted classes are set as anonymous classes whose type is the subject class. For example, a semantic association "*Bacteria CAUSE Infection*" is interpreted as *Bacteria* $\cap \exists$*CAUSE.Infection*. This defined a class of Bacteria, which can cause at least one infection whose type is also a Bacteria, i.e. a set of infections Bacteria. Consequently, such anonymous class of bacteria is set to have a type of Bacteria.

6.1.4 Evaluation

The proposed framework is evaluated using criteria-based approach to measure the extent of approximation and quality of the ontology structure. Each criterion represented a set of structural properties, where each property is measured using a set of metrics. The metrics values are obtained computationally. Eight criteria are chosen for measurement out which four of them measured extent of approximation and the rest are used to measure quality of the ontology structure.

To measure extent of approximation, five metrics are chosen, namely precision, recall, coverage, relevance and inconsistency. Accuracy is measured in terms of precision and recall and the result enabled to judge better accuracy. Completeness and conciseness are measured in terms of coverage and relevance, and the results are also good coverage. However, we considered the presence of cycles and partitioning errors as source of inconsistencies and found none. For details of the results, please see sections 6.2.2 and 6.2.3 above.

To measure quality, thirteen metrics are chosen and computed for measurement. Structural complexity is measured based on SOV, ENR, TIP and EOG and judged to be complex (see sec. 6.2.4). Adaptability is measured with coupling and cohesiveness, where their metrics are NEC, NER, NoR, NoL, ADIT-LN, and judged to be good (see sec. 6.2.5). Ontology schema potential is measured with semantic richness properties such as RR, AR and IR (see sec. 6.2.6). Lastly, clarity is measured with readability (RD) metrics. Both shows better results and judged to be good and encouraging. In general, according to the evaluation results shown in section 6.3, the correctness and quality of the proposed framework is inline with the original expectation.

6.1.5 Contribution

The contribution of this resaerch work is the construction of rigrous and integrated ontology acquisition framework from biomedical texts. The specifities of the framework are rigor, shared knowledge, scalability, independent and automatic ontology acquisition in the biomedical domain. Scalability is in the order of millions (>4 million concepts and their associations). The components of the frmaework are highly integrated for consistent and semantically relevenant acquisition. The challenges of consensus reaching (shared knwoeldge acquisition) is better minimized using shared domain semantics in the background knowledge. This enabled ontology learning independent from ontology engineers and domain experts (manual works), and thus, geared towards fully automatic acquisition. Domain relevance is determined using the semantics in the background knowledge, which is pre-built and consensus-reached set of domin artefacts rather than using TFIDF.

The proposed framework is highly scalable and exhuastive in acquiring biomedical artefacts (concepts, roles and their associations). Consequently, over 4 million concepts and 7 million associations are acquired from the bioMed text collection. Ontology structuring is automatic supported by structures in the bckground knowledge. Interpretation is autoamatic using ontology language primitives and constructs (e.g. OWL DL). Thus, while existing ontology learning frameworks are supported with ontology tools (e.g. ontoEdit), this framework doesn't require it.

Generally, this study is the first attempt to fully automate ontology learning from text collection at larger scale and rigourously focusing on a shared understanding of the domain semantics. Firstly, it disambiguates biomedical artefacts at large scale; secondly, the framework structures biomedical concepts and their associations into ontology structures; Finally, formally interprets the conceptual ontology. The framework does these activities independently from knowledge engineers and ontology tools, which results reduction of time, effort and cost in conventional ontology learning.

6.2 Way Forward

This research designed and developed a generic and rigorous ontology acquisition framework as a first step to automate ontology learning from biomedical texts. Even though the proposed

framework is applicable to biomedical domain, its practicality remained dependent on the technicality of instantiating the three semantic disambiguations and interpretations: semantic disambiguation, conceptual disambiguation and ontologization. Furthermore, the unstructured knowledge sources and background knowledge coverage have also their on contribution for practicality of the framework. In this context, we pleased to urge further investigations as future extensions of the framework. Consequently, the following points should be noted to enrich and mature the ontology learning framework.

- The use of very dense domain knowledge source collection. In the experimentation of the framework, it is found that several biomedical artefacts presented in the background knowledge aren't found in the bioMed text collection. This leads us to notice the knowledge source collection missed many of the biomedical entities from the knowledge base, which means that the collection is sparse.

- The use of semantically rich domain knowledge base. Several biomedical entities in the bioMed text collection are not found in the domain knowledge base. This leads us to further observe that there are many biomedical entities, which are not found in the background knowledge. Thus, we urge the development and use of highly rich background knowledge.

- The use of highly accurate concept and semantic association disambiguation programs. MetaMap and enhanced-semRep are semantic disambiguation programs. MetaMap disambiguated biomedical concepts with certain limitations as shown in the literature. But, further investigations to extend MetaMap for individual and attribute information as well as to minimize its drawback are a requirement. Furthermore, currently enhanced-semRep disambiguated 682 biomedical roles out of the 736 defined in the background knowledge. As a result, we have observed that the recall and hence accuracy of biomedical roles and semantic associations can be enhanced by considering more semantic relationships. Thus, further extension of the semantic processing program at least to include the 736 roles in the background knowledge is required for still better ontology acquisition framework design and development.

- Consistent interpretation of conceptual semantics. As presented with the instantiation of the ontology acquisition framework, it didn't disambiguated semantic restrictions for each

semantic association. To this end, we forced to interpret all associations using existential restrictions. This might have bad consequences to some associations, which are either universal or number restrictions. Thus, the interpretation model should also include universal and number restriction based interpretations.

- Exhaustive evaluation of the knowledge acquisition framework. It was evaluated on the structural dimension aspects only in terms of correctness and quality. However, the functional and usability dimensions are also equally measures quality and correctness of the framework. Thus, the proposed framework should also be assessed from these perspectives.

- Use of hybrid method for ontological knowledge acquisition. Data-driven and knowledge driven methods can be used to acquire richer ontological knowledge from biomedical texts. For example Pikes framework can be integrated with our concept-driven method to acquire both schema and individual knowledge about a domain.

- Deploying the knowledge acquisition framework for ontology-based applications. It means that applying application-based evaluation approach for the proposed framework. Application-based evaluation might investigate accurately the effectiveness and efficiency of the proposed knowledge acquisition framework and in turn generates the ontological knowledge.

- Domain adaptivity. The proposed framework can be made domain adaptive using adaptive learning techniques. Pikes framework can be used to enhance the domain adaptiveness of the proposed framework

References

[1]. Institute for Alternative Futures, "The Bio-monitoring Futures Project: Final Report and Recommendations," Robert Wood Johnson Foundation, Rep. 06-14, 2006.

[2]. N. Wickramasinghe and E. Geisler, "Encyclopedia of Healthcare Information Systems," Medical Information Science Reference, pp. 1-10, 2008.

[3]. C. Bock, L. Carnahan, S. Fenves, M. Gruninger, V. Kashyap, B. Lide, J. Nell, R. Raman and R. D. Sriram, "Healthcare Strategic Focus Area: Clinical Informatics," National Institute of Standards and Technology, U.S, NISTIR 7263, 2005.

[4]. A. Kokinaki and I. Chouvarda and N. Maglaveras, "Integrating SCP-ECG Files and Patient Records: An Ontology Based Approach," Information Technology Applications in Biomedicine, Greece, 2006, pp. 1-7,

[5]. A. Ryan, "Towards Semantic Interoperability in Healthcare: Ontology Mapping from SNOMED CT to HL7 Version 3," AOW 2006 Proceedings of the second Australasian workshop on Advances in ontologies, 2006, pp. 69-74.

[6]. E. Ko, H. Lee and J. Lee, "Ontology and CDSS Based Intelligent Health Data Management in Healthcare Server," International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, World Academy of Science: Engineering & Technology, 2007, pp. 119-123.

[7]. E. S. Berner, "Hospital Based Decision Support," in Clinical Decision Support Systems – Theory and Practice, 2nd edition, Birmingham, Springer, Chapter 3,4, 5, 2006.

[8]. K. Verlaenen, W. Joosen and P. Verbaeten,"Arriclides: An Architecture Integrating Clinical Decision Support Models," Proceedings of the 40th Annual Hawaii International Conference on System Sciences, 2007, pp. 1–10.

[9]. P. Marcheschi, A. Mazzarisi, S. Dalmiani and A. Benassi, "ECG Standards for the Interoperability in Patient EHRS in Italy," Computers in Cardiology, 2006, pp.549−552.

[10]. M. D. Silveira, N. Guelfi, J.D Baldacchino, P. Plumer, M. Seil and A. Wienecke,"A Survey of Interoperability in E-Health Systems: The European Approach," International Conference on Health Informatics and Health Informatics, 2008, pp. 172-175.

[11]. W. J. Graben and M. G. Deftsch,"Ontologies and Their Application in HER," eHealth2008–Medical Informatics Meets ehealth, 2008, pp.1-4.

[12]. G. B. Laleci and A. Dogac,"A Semantically Enriched Clinical Guideline Model Enabling Deployment in Heterogeneous Healthcare Environment," IEEE Transaction on Information Technology in Biomedicine, 2009, pp. 263-273.

[13]. R. Cornet, "Clinical Terminology in Practical Use for Recording and Researching Reasons for Admission in Intensive Care," J1B. Academic Medical Center, 2006, pp. 1-7.

[14]. S. Gnanambal and M. Thangaraj, "Research Directions in Semantic Web on Healthcare," (IJCSIT) International Journal of Computer Science and Information Technologies, 2010, pp.449-453.

[15]. C. Snae and M. Brueckner,"Personal Health Assistance Service Expert System," International Journal of Social, Human Science and Engineering, 2007, pp. 197-200.

[16]. D. Han and et al, "An Evolving Mobile E-Health Service Platform," International Conference in Consumer Electronics, ICCE 2007 Digest of Technical Papers, 2006, pp. 475-485.

[17]. F. B. Nardon and L. A. Moura," Knowledge Sharing and Information Integration in Healthcare Using Ontologies and Deductive Databases." MEDINFO 2004, 2004, pp. 62-66.

[18]. D. Han, I. Ko and S. Park," Ontology Based Context Modeling and Reasoning for U-Healthcare," Institute for Infocomm Research, 2002, pp. 1-5.

[19]. J. Dang and A. Hedayati,"An Ontological Knowledge Framework for Adaptive Medical Workflow" Journal of Biomedical Informatics, 2008, pp.829-836.

[20]. T. Ah-Hwee, 'Text Mining: The State of Art and the Challenges,' In proceedings of PAKDD'99 Workshop on Knowledge discovery from Advanced Databases', 1999, pp.71-76.

[21]. W. Pik and M. Gahegan, "Beyond ontologies: Toward situated representations of scientific knowledge," International Journal of Human-Computer Studies, 2007, pp. 659–673.

[22]. G. Geleijnse and J. Korst, "Learning Effective Surface Text Patterns for Information Extraction," The 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Workshop on Adaptive Text Extraction and Mining, 2006, pp.1-8.

[23]. C. Brewster, F. Ciravegna and Y. Wilks, "Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance," In Proceedings of the SIGIR Semantic Web Workshop, 2003, pp. 1-8.

[24]. R. Studer, V. R. Benjamin and D. Fensel, "Knowledge Engineering: Principles and Methods," Data Knowledge Engineering, 1998, pp.161-197.

[25]. P. D. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," In Proceedings of the 12th European Conference on Machine Learning (ECML, 2001), pp.491-502.

[26]. D. Lin, "Automatic Retrieval and Clustering of Similar Words," In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL), 1998, pp.768–774.

[27]. M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," In Proceedings of the 14th International Conference on Computational Linguistics (COLING), 1992, pp. 539–545.

[28]. K. Ahmad and H. Fulford, "Knowledge Processing: Semantic Relations and their Use in Elaborating Terminology," Technical Report, University of Surrey, 1992.

[29]. M. Berland and E. Charniak, “Finding Parts in Very Large Corpora,” In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), 1999, pp.57–64.

[30]. J. Lyons, “Semantics: Volume 1,” Cambridge University Press, U.S, New York, Chapters 6 - 9, 1977.

[31]. J. Volker, P. Hitzler and P. Cimiano, “Acquisition of OWL DL Axioms from Lexical Resources,” In Proceedings of the European Semantic Web Conference (ESWC), 2007, pp. 670–685.

[32]. Z. Vendler, “Verbs and Times,” The Philosophical Review, Cornell University, 1957, pp.143–160.

[33]. B. Comrie, “Aspect: Introduction to the Study of Verbal Aspect and Related Problems,” Cambridge University Press, 1st edition, 1976.

[34]. W. Van De Velde and A. Aamodt, “Machine Learning Issues in CommonKADS,” KADS-II Project Deliverable D2.11, Rep. D2.11, 1992.

[35]. Y. Sure, S. Staab and R. Studer,”Methodology for Development and Employment of Ontology-Based Knowledge Management Applications,” Sigmod Record, 2002, pp.18–23.

[36]. D. Vrandecic, S. Pinto, C. Tempich and Y. Sure, “The Diligent Knowledge Processes,” Journal of Knowledge Management, 2005, pp.85–96.

[37]. M. Fernandez, A. G6mez-Prez and N. Juristo, “METHONTOLOGY: From Ontological Art Towards Ontological Engineering,” In Proceedings of the AAAI Spring Symposium on Ontological Engineering, 1997, pp. 33–40.

[38]. A. Gomez-Perez and et al,”Evaluation of Taxonomic Knowledge on Ontologies and Knowledge bases,” Proceedings of the Banff Knowledge Acquisition for Knowledge-Based Systems, 1999, pp.1-19.

[39]. A. Gangemi, C. Catenacci, M. Ciaramita and J. Lehmann, "A Theoretical Framework for Ontology Evaluation and Validation,"Proceedings of SWAP 2005, the 2nd Italian Semantic web Workshop, 2005, pp. 1-16.

[40]. J. Pak and L. Zhou, "A Framework for Ontology Evaluation," Lecture Notes in Business Information Processing, 2011, pp. 10-18.

[41]. J. Brank and et al, "A Survey of Ontology Evaluation Techniques," Proceeding of the Conference on Data Mining and Data Warehouses Sikdd2005, 2005, pp.166-169.

[42]. T. Mitchell, "Machine Learning," Int'I edition, McGraw Hill, Chapters 2, 10, 1997.

[43]. A.K. Jain and R.C. Dubes, "Algorithms for Clustering Data," Prentice Hall College Division, U.S, 1988.

[44]. R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 1993, pp. 207–216.

[45]. A. Maedche and S. Staab, "Discovering Conceptual Relations from Text," In Proceedings of the 14th European Conference on Artificial Intelligence (ECAI), 2000, pp.321–325.

[46]. P. Cimiano, A. Hotho and S. Staab,"Learning Concept Hierarchies from Text Corpora Using Formal Concept Analysis," Journal of Artificial Intelligence Research (JAIR), 2005, pp. 305–339.

[47]. N. Lavrac and S. Dzeroski, "Inductive Logic Programming: Techniques and Applications," Prentice Hall, Ellis Horwood Series in Artificial Intelligence, 1994.

[48]. S. Staab and R. Studer, "Ontology Learning," in Handbook of Ontologies, 2nd ed., chapter 3, 2004, pp.173–190.

[49]. K. E. Campbell, D. E. Oliver and E. H. Shortliffe, "The Unified Medical Language System: Toward a Collaborative Approach for Solving Terminological Problems," JAMIA, 1998, pp.12-16.

[50]. D. Ayele, J. Chevallet, M. Meshesha and G. Mitikie, ”Enhancing Semantic Relation Quality of UMLS Knowledge Sources,”, MEDES '12 Proceedings of the International Conference on Management of Emergent Digital Ecosystems, 2012, pp.59-66.

[51]. D. Ayele, J. Chevallet, M. Meshesha and G. Mitikie, “Constructing Reference Semantic Predictions from Biomedical Knowledge Sources,” Proceedings of COLING 2012: Technical Papers, 2012, pp. 133–148.

[52]. C. Fellbaum, “WordNet: An Electronic Lexical Database,” Cambridge, MA: MIT Press, 1998.

[53]. E. Beisswanger, “BioTop: An Upper Domain Ontology for the Life Sciences,” IOS Press, 2008, pp.1-7.

[54]. F. Freitas, S. Schulz and E. Moraes, “Survey of Current Terminologies and Ontologies in Biology and Medicine,” RECIIS – Elect. J. Commun. Inf. Innov Health, 2009, pp.7-18.

[55]. A. T. McCray, “Aggregating UMLS Semantic Types for Reducing Conceptual Complexity,” MEDINFO, 2001, pp.216-220.

[56]. L. T. Vizenor, O. Bodenreider and A. T. McCray, “Auditing Associative Relations across Two Knowledge Sources,” Journal of Biomedical Informatics, 2009, pp.426-439.

[57]. H. Erdogan, “Exploiting UMLS Semantics for Checking Semantic Consistency among UMLS Concepts,” MEDINFO, 2010, pp.749-753.

[58]. A. Aronson and F. Lang,”An Overview of MetaMap: Historical Perspective and Recent Advances,” JAMIA, 2010, pp.229-236.

[59]. C. B. Ahlers, M. Fiszman, D. Demner-Fushman, F. Lang and T. C. Rindflesch, “Extracting Semantic Predications from Medline Citations for Pharmacogenomics,” Pacific Symposium on Biocomputing, 2007, pp.1-12.

[60]. A. Lozano-Tello and A. Gómez-Pérez, “Ontometric: A Method to Choose the Appropriate Ontology,” J. Datab. Mgmt, 2004, pp.1–18.

[61]. T. Berners-Lee and et al, “The Semantic Web,” Scientific American, 2001, pp.30–37.

[62]. S. Mukundan, “Spinning the Semantic Web,” MIT Press, Cambridge, 2003.

[63]. T. R. Gruber, “A Translation Approach to Portable Ontology Specifications,” Knowledge Acquisition, 1993, pp.199-221.

[64]. L. Prusak, “Where Did Knowledge Management Come From?” IBM Systems Journal, 2001, pp.1002-1007.

[65]. T. H. Davenport, “Some Principles of Knowledge Management,” ResearchGate, 1998.

[66]. R. Dieng, O. Corby, A. Giboin and M. Ribiere, “Methods and Tools for Corporate Knowledge Management,” International Journal of Human-Computer Studies, 1999, pp.567-598.

[67]. H. J. Hendricks and D. J. Vriens, “Knowledge-Based Systems and Knowledge Management Friends or Foes?” Information and Management, 1999, pp.113–125.

[68]. J. Sinclair, “Look Up: an Account of the COBUILD Project in Lexical Computing,” HarperCollins, COBUILD, 1987.

[69]. T. Winograd, “Understanding Natural Language,” Academic Press, New York, 1972.

[70]. A. Maedche and S. Staab, “Mining Ontologies from Text,” Knowledge Acquisition, Modeling and Management, Proceedings of the 12th International Conference, Lecture Notes in Computer Science, 2002, pp.189-202.

[71]. B. Moulin and D. Rousseau, “Automated Knowledge Acquisition from Regulatory Texts,” IEEE Expert, 1992, pp.27-35.

[72]. K. A. Ericsson and H. A. Simon, “Protocol Analysis: Verbal Reports as Data,” MIT Press, Cambridge, 1996.

[73]. J. H. Boose and J. M. Bradshaw, “Expertise Transfer and Complex Problems: Using AQUINAS as A KA Workbench for KBS,” Knowledge Acquisition for Knowledge-Based Systems, Academic Press, 1988.

[74]. J. Blythe and S. Ramachandran, “Knowledge Acquisition Using an English-Based Method Editor,” In Proc. 1999 Knowledge Acquisition Workshop, 1999.

[75]. K. L. McGraw and K. Harbison-Briggs, “Knowledge Acquisition: Principles and Guidelines,” Prentice-Hall, 1989.

[76]. Y. A. Wilks, B. L. Slator and M. Guthrie, “Electronic Words Dictionaries, Computers and Meaning,” MIT Press, Cambridge, 1996.

[78]. J. F. Sowa, “Knowledge Representation,” Brooks/Cole Thomson Learning, Pacific Grove, 2000.

[79]. P. M. Roget, “Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition,” Longman, London, 1852.

[80]. M. Masterman, “The Thesaurus in Syntax and Semantics,” Mechanical Translation, 1957.

[81]. K. S. Jones, “Synonymy and Semantic Classification,” Edinburgh University Press, Edinburgh, Scotland, Scotland, 1986.

[82]. L. Newman, “Descartes’ Epistemology,” The Stanford Encyclopedia of Philosophy, 2000.

[83]. D. Yarowsky, “Word-Sense Disambiguation Using Statistical Models of Roget’s Categories Trained on Large Corpora,” In Proceeding of COLING-92, 1992, pp. 454–460.

[84]. C. T. Yu and G. Salton, "The Effectiveness of the Thesaurus Method in Automatic Information Retrieval," National Science Foundation, Washington DC, Ottawa, Rep. 75-261, 1975.

[85]. Y. Qiu and H. Frei, "Concept Based Query Expansion," In Proc. 16th Int. Conf. Research and Development in Information Retrieval, ACM SIGIR, 1993, pp.160–169.

[86]. C. J. Crouch and B. Yang, "Experiments in Automatic Statistical Thesaurus Construction," Proceedings of the 15th Annual International Conference in Information Retrieval, 1992, pp. 77–88.

[87]. E. M. Voorhees, "Using Wordnet for Text Retrieval," Wordnet: an Electronic Lexical Database, MIT Press, 1998, pp. 1-6.

[88]. J. Kekäläinen, "The Effects of Query Complexity, Expansion and Structure on Retrieval Performance in Probabilistic Text Retrieval," PhD Dissertation, Department of Information Studies, University of Tampere, Tampere, 1999.

[89]. E. Sormunen, J. Kekalainen, J. Koivisto and K. Jarvelin,"Document Text Characteristics Affect the Ranking of the Most Relevant Documents by Expanded Structured Queries," Journal of Documentation, 2000, pp.358–376.

[90]. P. Clough and M. Stevenson, "Cross-Language Information Retrieval Using EuroWordNet and Word Sense Disambiguation," ECIR Lecture Notes in Computer Science, 2004, pp.327–337.

[91]. N. Guarino, "Some Ontological Principles for Designing Upper Level Lexical Resources," In Proceedings of the 1st International Conference on Language Resources and Evaluation, 1998, pp. 1-8.

[92]. P. Morville, "Little Blue Folders," http://argus-acia.com/, accessed March, 2016.

[93]. J. Ellman, "Corporate Ontologies as Information Interfaces," IEEE Intelligent Systems, 2004, pp.79–80.

[94]. P. Vossen, “EuroWordNet: Building a Multilingual Database with Wordnets for European Languages,” The ELRA Newsletter, 1998, pp. 7-10.

[95]. A. Burgun and O. Bodenreider, “Comparing Terms, Concepts and Semantic Classes in Wordnet and the UMLS,” Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources, 2001, pp. 77-82.

[96]. C. Fellbaum, U. Hahn, B. Smith, “Towards New Information Resources for Public Health--From Wordnet to MedicalWordNet,” Journal of Biomedical Informatics, 2006, pp. 321-332.

[97]. N. F. Noy and D. L. McGuiness, “Ontology Development 101 a Guide to Creating Your First Ontology”, Stanford Knowledge Systems Laboratory, 2001, pp. 1-25.

[98]. M. Huhns and L. Stevens, “Personal Ontologies,” IEEE Internet Computing, 1999, pp.85–87.

[99]. IEEE, “IEEE P1600.1 Standard Upper Ontology,” IEEE, 2002.

[100]. J. A. Bateman, R. Henschel and F. Rinaldi, “Generalized Upper Model 2.0: Documentation,” ResearchGate, 1995, pp. 1-8.

[101]. S. Beale, S. Nirenburg and K. Mahesh, “Semantic Analysis in the Mikrokosmos Machine Translation Project,” In Proceedings of Symposium on Natural Language Processing, 1996, pp. 1-11.

[102]. D. B. Lenat and R. V. Guha, “Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project,” Elsevier, 1993, pp. 95-104.

[103]. J. Chaffee and S. Gauch, “Personal Ontologies for Web Navigation,” In CIKM ’00: Proceedings of the Ninth International Conference on Information and Knowledge Management, 2000, pp. 227–234.

[104]. Y. Kalfoglou, J. Domingue, E. Motta, M. Vargas-Vera and S. B. Shum, "Myplanet: an Ontology-Driven Web-Based Personalized News Service," In Proceedings IJCAI 2001 workshop on Ontologies and Information Sharing, 2001, pp. 1-10.

[105]. P. H. Winston, "Artificial Intelligence," 3rd edition, Addison-Wesley, 1993.

[106]. M. Davies, "Knowledge (explicit and implicit), "International Encyclopedia of the Social and Behavioral Sciences, 2001, pp. 8126–8132.

[107]. M. Polanyi, "The Tacit Dimension," Doubleday, New York, 1966.

[108]. C. Mantzavinos, "Individuals, Institutions and Markets," Cambridge University Press, 2001.

[109]. C. Mantzavinos, D. C. North and S. Shariq, "Learning, Institutions and Economic Performance," Perspectives on Politics, 2004, pp.75–84.

[110]. M. A. E. Dummett, "The Logical Basis of Metaphysics," The William James lectures, Harvard University Press, 1991.

[111]. D. C. Dennett, "Styles of Mental Representation," Proceedings of the Aristotelian Society, 1983, pp.213–226.

[112]. J. A. Fodor, "Psychosemantics: The Problem of Meaning in the Philosophy of Mind," MIT Press, 1983.

[113]. D. B. Lenat, R. Guha, K. Pittman, D. Pratt and M. Shepherd, "Cyc Toward Programs with Common Sense," Communications of the ACM, 1990, pp. 30-49.

[114]. F. Koessler, "Common Knowledge and Interactive Behaviors: A Survey," European Journal of Economic and Social Systems, 2000, pp.271–308.

[115]. R. Navigli, P. Velardi and A. Gangemi, "Ontology Learning and Its Application to Automated Terminology Translation," IEEE Intelligent Systems, 2003, pp.22–31.

[116]. R. J. Marzana, "Building Background Knowledge for Academic Achievement: Research on What Works in Schools (Professional Development)," 1st edition, Association for Supervision & Curriculum Development, 2004.

[117]. P. Vanderschraaf and G. Sillari, "Common Knowledge," The Stanford Encyclopedia of Philosophy, 2005.

[118]. D. Lewis, "Convention, a Philosophical Study," Harvard University Press, 1969.

[119]. H. H. Clark and C. R. Marshall, "Definite Reference and Mutual Knowledge," Elements of Discourse Processing, Cambridge University Press, 1981, pp.10–63.

[120]. K.C. Stevens, "The Effect of Background Knowledge on the Reading Comprehension of Ninth Graders," Journal of Reading Behavior, 1980, pp.151–154.

[121]. F. Dochy, M. Segers and M. M. Buehl, "The Relation between Assessment Practices and Outcomes of Studies the Case of Research on Prior Knowledge," Review of Educational Research, 1999, pp.145–186.

[122]. C. Emmott, "Narrative Comprehension a Discourse Perspective," Oxford University Press, 1999.

[123]. S. Thor, S. G. Anderson, A. Tomlinson and J. B. Thomas, "A Lim-Homeodomain Combinatorial Code for Motor-Neuron Pathway Selection," Nature, 1997, pp.76–80.

[124]. J. A. Langer, "Examining Background Knowledge and Text Comprehension," Reading Research Quarterly, 1984, pp.468–481.

[125]. E. Agirre, O. Ansa, E. Hovy and D. Martinez,"Enriching Very Large Ontologies Using the WWW," ECAI Workshop on Ontology Learning and CEUR Workshop Proceedings, 2000.

[126]. M. Banko and E. Brill, "Scaling to Very Large Corpora for Natural Language Disambiguation," In Proceedings of the 39th Annual Meeting and 10th Conference of the

European Chapter of the Association for Computational Linguistic (ACL 2001), 2001, pp.26–33.

[127]. J. Knowlson, “Universal Language Schemes in England and France 1600-1800,” University of Toronto Press, 1975.

[128]. M. M. Slaughter, “Universal Languages and Scientific Taxonomy in the Seventeenth Century,” Cambridge University Press, 1982.

[129]. R. Davis and D. Lenat, “Knowledge-Based Systems in Artificial Intelligence: AM and Teiresias,” McGraw-Hill, New York, 1982.

[130]. L. Eshelman, D. Ehret, J. McDermott and M. Tan, “MOLE: A Tenacious Knowledge Acquisition Tool,” Knowledge-Based Systems: Vol. 2: Knowledge Acquisition Tools for Expert Systems, Academic Press, 1987, pp.41-54.

[131]. J. Diederich, I. Ruhmann and M. May, “KRITON: A Knowledge Acquisition Tool for Expert Systems,” Knowledge-Based Systems: Vol. 2: Knowledge Acquisition Tools for Expert Systems, Academic Press, 1987, pp. 29 - 40.

[132]. E. Motta, M. Eisenstadt and K. Pitman, “Support for Knowledge Acquisition in the Knowledge Engineer’s Assistant (KEATS),” Expert Systems, 1988, pp.6-28.

[133]. E. Motta, T. Rajan, J. Domingue and M. Eisenstadt, “Methodological Foundations of Keats, The Knowledge Engineer’s Assistant,” Current Trends in Knowledge Acquisition, 1990, pp. 289-301.

[134]. M. L. G. Shaw and B. R. Gaines, “KITTEN: Knowledge Initiation and Transfer Tools for Experts and Novices,” Knowledge-Based Systems: Vol. 2: Knowledge Acquisition Tools for Expert Systems, Academic Press, 1988, pp. 251-280.

[135]. U. Hahn, M. Klenner and K. Schnattinger, “Automated knowledge acquisition meets metareasoning: incremental quality assessment of concept hypotheses during text understanding.” Proceedings of 9th European Knowledge Acquisition Workshop (EKAW-96), 1996b, 1996, pp. 131-146.

[136]. U. Hahn and K. Schnattinger, "Towards Text Knowledge Engineering." In Proc. 15th National Conference on Artificial Intelligence (AAAI-98), MIT Press, 1998, pp. 524-531.

[137]. M. Craven and J. Kumlien, "Constructing Biological Knowledge Bases by Extracting Information from Text Sources," Proc Int Conf Intell Syst Mol Biol, 1999, pp.77–86.

[138]. C. Friedman and G. Hripcsak, "Natural Language Processing and Its Future in Medicine," Acad Med, 1999, pp. 890–895.

[139]. P. Spyns, "Natural Language Processing in Medicine: An Overview," Methods Inf Med, pp.1996, pp.285–301.

[140]. R. Basili, M. T. Pazienzaa and P. Velardib, "An Empirical Symbolic Approach to Natural Language Processing," Elsevier Science B.V, 1996, pp. 59-99.

[141]. S. B. Johnson, A. Aguirre, P. Peng and J. Cimino, "Interpreting Natural Language Queries using the UMLS," Proceeding of Annu Symp Comput Appl Med Care, 1993, pp.294–298.

[142]. R. Grishman, S. Huttunen and R. Yangarber, "Information Extraction for Enhanced Access to Disease Outbreak Reports," Journal of Biomedical Informatics, 2002, pp.236-46.

[143]. T. C. Rindflesch and M. Fiszman, "The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text," Journal of Biomedical Informatics, 2003, pp.462–477.

[144]. M. K. Bergman, "The Deep Web: Surfacing Hidden Value," The Journal of Electronic Publishing, 2001, pp. 1-7.

[145]. A. Wright, "In Search of the Deep Web," URL http://www.salon.com/tech/feature/2004/03/09/deep_web/index.html, accessed in August, 2015.

[146]. P. R. Bowden, P. Halstead and T. G. Rose, "Extracting Conceptual Knowledge from Text Using Explicit Relation Markers," Proceeding 9th European Knowledge Acquisition Workshop (EKAW-96), 1996, pp. 147-162.

[147]. U. Reimer, "Automatic Acquisition of Terminological Knowledge from Texts," Proc. 9th European Conference on Artificial Intelligence (ECAI-90), 1990a, 1990, pp. 547-549.

[148]. U. Reimer, "Automatic Knowledge Acquisition from Texts: Learning Terminological Knowledge via Text Understanding and Inductive Generalization," In KAW'90 - Proc. of the Workshop on Knowledge Acquisition for Knowledge-Based Systems, 1990b, 1990, pp.1-16.

[149]. U. Hahn and K. Schnattinger, "An Empirical Evaluation of a System for Text Knowledge Acquisition," In Proceeding of European Knowledge Acquisition Workshop (EKAW-97), 1997, pp.129-144.

[150]. R. Hull and F. Gomez, "Automatic Acquisition of Historical Knowledge from Encyclopedic Texts," In Proc. Knowledge Acquisition Workshop, 1998, pp. 1-18.

[151]. M. Lebowitz, "The Use of Memory in Text Processing," Communications of the ACM, 1988, pp.1483-1502.

[152]. M. Lebowitz, "Researcher: An Overview," In Proc. National Conference on Artificial Intelligence, 1983a, 1993, pp.232-235.

[153]. A. Goel, K. Mahesh, J. Peterson and K. Eiselt, "Unification of Language Understanding, Device Comprehension and Knowledge Acquisition," In Proceeding of 10th Knowledge Acquisition Workshop, 1996, pp. 1-5.

[154]. M. Lebowitz, "Generalization from Natural Language Text," Cognitive Science, 1983b, 1983, pp.1-40.

[155]. D. Faure and C. Nédellec, "Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM," Proceedings of the 11th European Workshop, 1999, pp.329-334.

[156]. G. de Chalendar and B. Grau, “SVETLAN' or How to Classify Words Using Their Context,” Proceedings of the 12th European Knowledge Acquisition Workshop (EKAW-2000), 2000, pp.203-216.

[157]. J. F. Delannoy, C. Feng, S. Matwin and S. Szpakowicz, “Knowledge Extraction from Text: Machine Learning For Text-To-Rule Translation,” In Proceeding of the Workshop on Machine Learning Techniques and Text Analysis, European Conference on Machine Learning (ECML-93), 1993, pp. 1-157.

[158]. S. Lapalut, “Text Clustering to Help Knowledge Acquisition from Documents,” Proc. Ninth European Knowledge Acquisition Workshop (EKAW-96), 1996a, 1996, pp.115-130.

[159]. S. Lapalut, “How to Handle Multiple Expertises from Several Experts: A General Text Clustering Approach,” Proc. 2nd Knowledge Engineering Forum (KEF'96), 1996b, 1996, pp. 1-6.

[160]. S. Szpakowicz, “Semi-Automatic Acquisition of Conceptual Structure from Technical Texts,” International Journal of Man-machine Studies, 1990, pp.385-397.

[161]. B. Biébow and S. Szulman, “Acquisition and Validation: From Text to Semantic Network,” Proceeding of 7th European Knowledge Acquisition Workshop-EKAW-93, 1993, pp. 427-446.

[162]. B. Biébow and S. Szulman, “TERMINAE: A Linguistic-Based Tool for the Building of Domain Ontology,” Knowledge Acquisition, Modeling and Management, Proc. 11th European Workshop, Lecture Notes in Computer Science, 1999, pp. 49-66.

[163]. R. Lu and C. Cao, “Towards Knowledge Acquisition from Texts,” Current Trends in Knowledge Acquisition, 1990, pp. 289-301.

[164]. A. Mikheev and S. Finch, “A Workbench for Acquisition of Ontological Knowledge from Natural Text,” In Proceedings of the 7th conference of the European Chapter for Computational Linguistics (EACL'95), 1995, pp.194-201.

[165]. N. Aussenac-Gilles, B. Biebow and S. Szulman, “Revisiting Ontology Design: A Methodology Based On Corpus Analysis,” Knowledge Acquisition, Modeling and Management, Proc. Twelfth European Knowledge Acquisition Workshop (EKAW-2000), Lecture Notes in Computer Science, 2000, pp. 172-188.

[166]. D. R. Swanson and N. R. Smalheiser, “An Interactive System for Finding Complementary Literatures: a Stimulus to Scientific Discovery,” Artificial Intelligence, 1997, pp.183-203.

[167]. P. Spyns, “Natural Language Processing in Medicine: An Overview,” Methods Inf Med, 1996, pp.285–301.

[168]. R. H Baud, C. Lovis, A. M. Rassinoux and J. R. Scherrer, “Alternative Ways for Knowledge Collection, Indexing and Robust Language Retrieval,” Methods Inf Med, 1998, pp.315-326.

[169]. B. L. Humphreys, D. A. Lindberg, H. M. Schoolman and G. O. Barnett, “The Unified Medical Language System: Informatics Research Collaboration,” J Am Med Inform Assoc., 1998, pp.1-11.

[170]. M. B. Amaral, A. Roberts and A. L. Rector, “NLP Techniques Associated with the OpenGALEN Ontology for Semi-Automatic Textual Extraction of Medical Knowledge: Abstracting and Mapping Equivalent Linguistic and Logistic Constructs,” Proc AMIA Symp 2000, 2000, pp.76–80.

[171]. C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino and S. B. Johnson, “A General Natural-Language Text Processor for Clinical Radiology,” J Am Med Inform Assoc, 1994, pp.161–74.

[172]. C. Friedman, “A Broad-Coverage Natural Language Processing System,” Proc AMIA Symp2000, 2000, pp.270–274.

[173]. C. Friedman, H. Liu, L. Shagina, S. Johnson and G. Hripcsak, “Evaluating the UMLS as A Source of Lexical Knowledge for Medical Language Processing,” Proc AMIA Symp 2001, 2001, pp.189–193.

[174]. C. A. Knirsch, N. L. Jain, A. Pablos-Mendez, C. Friedman and G. Hripcsak, “Respiratory Isolation of Tuberculosis Patients Using Clinical Guidelines and an Automated Clinical Decision Support System,” Infect Control Hosp Epidemiol, 1998, pp.94–100.

[175]. A. Rassinoux, J. C. Wagnerl, C. Lovis1, R. H. Baud1, A. Rector and J. Scherrerl, “Analysis of Medical Texts Based On a Sound Medical Model,” Proc Annu Symp Comput Appl Med Care, 1995, pp.27–31.

[176]. A.L. Rector and W. A. Nowlan, “The GALEN Project,” Comput Methods Programs Biomed, 1994, pp.75–8.

[177]. P. Haug, S. Koehler, L. M. Lau, P. Wang, R. Rocha and S. Huff, “A Natural Language Understanding System Combining Syntactic and Semantic Techniques,” Proc Annu Symp Comput Appl Med Care, 1994, pp.247–51.

[178]. B. Rosario, M. A. Hearst and C. Fillmore, “The Descent of Hierarchy and Selection in Relational Semantics,” Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Association for Computational Linguistics, 2002, p. 247–254.

[179]. M. Fiszman, W. W. Chapman, D. Aronsky, R. S. Evans and P. J. Haug, “Automatic Detection of Acute Bacterial Pneumonia from Chest X-Ray Reports,” J Am Med Inform Assoc 2000, 2000, pp.593–604.

[180]. M. Fiszman and P. J. Haug, “Using Medical Language Processing to Support Real-Time Evaluation of Pneumonia Guidelines,” Proc AMIA Symp 2000, 2000, pp.235–239.

[181]. M. L. Gundersena, P. J. Hauga, T. A. Pryora, R. V. Breea, S. Koehlerb, K. Bauerc and B. Clemonsc, “Development and Evaluation of a Computerized Admission Diagnoses Encoding System,” Comput Biomed Res 1996, 1996, pp.351–372.

[182]. L. M. Christensen, P. J. Haug and M. Fiszman, “MPLUS: A Probabilistic Medical Language Understanding System,” In: Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Association for Computational Linguistics, 2002, pp. 29–36.

[183]. P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet and J. F. Boisvieux, "A Multi-Lingual Architecture for Building a Normalized Conceptual Representation from Medical Language," Proc Annu Symp Comput Appl Med Care 1995", 1995, pp.357–361.

[184]. P. Zweigenbaum, J. Bouaud, B. Bachimont, J. Charlet and J.F. Boisvieux, "Evaluating a Normalized Conceptual Representation Produced from Natural Language Patient Discharge Summaries," Proc AMIA Annu Fall Symp, 1997, pp.590–594.

[185]. U. Hahn, M. Romacker and S. Schulz, "How Knowledge Drives Understanding—Matching Medical Ontologies with the Needs of Medical Language Processing," Artif Intell Med 1999, 1999, pp.25–51.

[186]. U. Hahn, M. Romacker and S. Schulz, "MEDSYNDIKATE—Design Considerations for an Ontology-Based Medical Text Understanding System," Proc AMIA Symp 2000, 2000, pp.330–334.

[187]. M. Romacker S. Schulz and U. Hahn, "Streamlining Semantic Interpretation for Medical Narratives," Proc AMIA Symp 1999, 1999, pp.925–929.

[188]. A.R Aronson, "Effective Mapping of Biomedical Texts to the UMLS Metathesaraus: The MetaMap Program," Proceeding of AMIA Symposium, 2001, pp.17-21.

[189]. T. C. Rindflesch, C. A. Bean, and C. A. Sneiderman, "Argument Identification for Arterial Branching Predications Asserted in Cardiac Catheterization Reports," Proc AMIA Symp 2000, 2000, pp.704–708.

[190]. T. C. Rindflesch, L. Tanabe and J. N. Weinstein, "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature," Pac Symp Biocomputing 2000, 2000, pp.517–528.

[191]. T. C. Rindflesch, J. V. RAJAN and L. HUNTER, "Extracting Molecular Binding Relationships from Biomedical Text," in Proceedings of the 6th Applied Natural Language Processing Conference, Association for Computational Linguistics, 2000, pp.188–195.

[192]. D. Cutting, J. Kupiec, J. Pedersen and P. Sibun, “A Practical Part-Of-Speech Tagger,” In Proceedings of the Third Conference on Applied Natural Language Processing, 2002, pp. 133–140.

[193]. P. M. Kamde and S. P. Algur,”A Survey on Web Multimedia Mining,” The International Journal of Multimedia & Its Applications (IJMA), 2011, pp.72-84.

[194]. P. Mishra, N. Padhy and R. Panigrahi, “The Survey of Data Mining Applications and Feature Scope,” International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2012, pp.43-58.

[195]. R. J. Mooney and R. Bunescu, “Mining Knowledge from Text Using Information Extraction,” SIGKDD Explorations, 2005, pp.3-10.

[196]. S. R. Bhat, A. K. Tripathy, D. George, R. Jose and R. Pinto, “Advanced Knowledge Extraction from WebPages Using Natural language Processing,” International Journal of Emerging Technology and Advanced Engineering, 2012, pp.96-101.

[197]. S. Jusoh and H. M. Alfawareh, “Techniques, Applications and Challenging Issue in Text-Mining,” IJCSI International Journal Computer Science Issues, 2012, pp.431-436.

[198]. D. C. Wimalasuriya and D. Dou, “Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches,” Journal of Information Science, 2010, pp.306-323.

[199]. T. Ah-Hwee, “Text Mining: The State of Art and the Challenges," In proceedings PAKDD'99 Workshop on Knowledge discovery from Advanced Databases, 1999, pp.71-76.

[200]. C. Chang, M. Kayed, M. R. Girgis and K. Shaalan, “A Survey of Web Information Extraction Systems,” IEEE Transactions on Knowledge and Data Engineering, 2006, pp. 1-18.

[201]. F. Pugeault, “Knowledge Extraction from Texts: Method for Extracting Predicate-Argument Structures from Texts,” IRIT-CNRS, 1994, pp.1039-1043.

[202]. D. Consoli, “A New Framework to Extract Knowledge by Text Mining Tools,” The Journal of Knowledge Economy & Knowledge Management, 2010, pp.165-177.

[203]. J. Cimino and G. Barnett, “Automatic Knowledge Acquisition from MEDLINE,” Methods of Information in Medicine, 1993, pp.120-130.

[204]. H. Dai, Y. Chang, R. T. Tsai and W. Hsu, “New Challenges for Biological Text-Mining in the Next Decade,” Journal of Computer Science and Technology, 2010, pp.169-179.

[205]. G. Mariscal, Ó. Marbán and C. Fernández, “A Survey of Data Mining and Knowledge Discovery Process Models and Methodologies,” The Knowledge Engineering Review, 2010, pp.137–166.

[206]. P.D. Turney, “Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL,” In Proceedings of the 12th European Conference on Machine Learning (ECML), 2001, pp.491-502.

[207]. M. F. Moens, “Information Extraction: Algorithms and Prospects in Retrieval Context,” The Information Retrieval Series, Springer, 2006, pp. 1-4.

[208]. E. Riloff, “Automatically Constructing a Dictionary for Information Extraction Tasks,” Proceedings of the Eleventh National Conference on Artificial Intelligence, 1993, pp. 811–816.

[208]. A. M. Harabagiu and S. J Maiorano, “Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction,” In IEEE Transactions on Knowledge and Data Engineering, 1995, pp. 713-724.

[209]. E. Agichtein and L. Gravano,”Snowball: Extracting Relations from Large Plain-Text Collections,” In Proceedings of the 5th ACM International Conference on Digital Libraries, 2000, pp. 1-10.

[210]. A. Maedche, G. Neumann and S. Staab, “Bootstrapping an Ontology-Based Information Extraction System,” In studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web, Springer, 2002, pp.1-25.

[211]. L. K. McDowell1 and M. Cafarella, “Ontology-Driven Information Extraction with OntoSyphon,” In International Semantic Web Conference, 2006, pp. 428–444.

[212]. J. Aitken, “Learning Information Extraction Rules: An Inductive Logic Programming approach,” In Proceedings of the 15th European Conference on Artificial Intelligence, 2002, pp.355–359.

[213]. B. Yildiz and S. Miksch, “OntoX – A Method for Ontology-Driven Information Extraction,” In ICCSA, Springer, 2007, pp. 660–673.

[214]. C. Biemann, “Ontology Learning from Text: A Survey of Methods,” LDV Forum 20, 2005, pp.75–93.

[215]. M. A. Hearst, “Automated Discovery of Wordnet Relations,” In WordNet: An Electronic Lexical Database, 2nd ed., MIT Press, 1998.

[216]. P. Cimiano and J. Volker,”Towards Large-Scale, Open-Domain and Ontology-Based Named Entity Classification,” In Proceedings of Recent Advances in Natural Language Processing (RANLP), 2005, pp.166–172.

[217]. H. Tanev and B. Magnini, “Weakly Supervised Approaches for Ontology Population,” In EACL the Association for Computer Linguistics, 2006, pp.17-24.

[218]. M. Fleischman, “Fine Grained Classification of Named Entities,” In Proceedings of the 19th International Conference on Computational Linguistics, 2002, pp.1-7.

[219]. Y. Li, K. Bontcheva and H. Cunningham, “SVM Based Learning System for Information Extraction,” In Deterministic and Statistical Methods in Machine Learning, 2005, pp. 319–339.

[220]. M. Minsky, “A Framework for Representing Knowledge,” Massachusetts Institute of Technology, Cambridge, MA, USA, Rep. ADA011168, 1974.

[221]. J.S Aikins, “A Representation Scheme Using both Frames and Rules,” In Rule-Based Expert Systems, Addison-Wesley, 1984, pp.424-440.

[222]. P. J. Brachman, “What IS-A is and isn’t: An Analysis of Taxonomic Links in Semantic Networks,” Computer 16, 1983, pp.30-36.

[223]. D. G. Bobrow and T. Winograd, “An Overview of KRL, a Knowledge Representation Language,” Cognitive Science, 1977, pp. 3-46.

[224]. S. Russell and P. Norvig, “Artificial Intelligence – A Modern Approach,” 3rd edition, Prentice Hall, 1995.

[225]. P. J Brachman and J. G Schmolze, “Overview of KL-ONE knowledge Representation System,” Cognitive Science, 1985, pp.171-216.

[226]. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi and P. F. Patel-Schneider, “The Description Logic Handbook: Theory, Implementation and Application,” 2nd edition, Cambridge University Press, 2010.

[227]. B. Motik and B. Parsia, “OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax,” 2nd edition, W3C Recommendation, Rep. 11, 2012.

[228]. I. Horrocks, P. F. Patel-Schneider and F. V. Harmelen, “From SHIQ and RDF to OWL: The Making of a Web Ontology Language,” Journal of Web Semantics, 2003, pp.7–26.

[229]. V. Haarslev and R. Möller, “RACER System Description,” In Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2001), 2001, pp.701–705.

[230]. I. Horrocks, U. Sattler and S. Tobies, “Practical Reasoning for Very Expressive Description Logics,” Journal of the Interest Group in Pure and Applied Logic, 2000, pp.239–264.

[231]. E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur and Y. Katz, “Pellet: An OWL DL Reasoner,” Web Semantics: Science, Services and Agents on the World Wide Web, 2007, pp. 51-53.

[232]. D. Tsarkov and I. Horrocks, “Fact++ Description Logic Reasoner: System Description.” In Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006), 2006, pp.292–297.

[233]. R. Stevens, I. Horrocks, C. Goble and S. Bechhofer, “Building a Reasonable Bioinformatics Ontology Using OIL,” In Proceedings of the IJCAI-2001 Workshop on Ontologies and Information Sharing, 2001, pp.81–90.

[234]. A. Gómez-Pérez, “Evaluation of Taxonomic Knowledge in Ontologies and Knowledge Bases,” In Proceedings of the 12th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, 1999, pp. 1-19.

[235]. A. Gómez-Pérez, “Some Ideas and Examples to Evaluate Ontologies,” In Proceedings of the Eleventh Conference on Artificial Intelligence for Applications, 1995, pp.299–305.

[236]. A. Gómez-Pérez, “Ontology Evaluation,” International Handbooks on Information Systems, 2004, pp.251–274.

[237]. Y. Sure, “Why Evaluate Ontology Technologies? Because It Works!” IEEE Intelligent Systems, 2004, pp.1541-1672.

[238]. N. Noy, “Evaluation by Ontology Consumers,” IEEE Intelligent Systems, 2004, pp.1541-1672.

[239]. J. Hartmann, P. Spyns, A. Giboin, D. Maynard, R. Cuel, M. C. Suárez-Figueroa and Y. Sure, “Methods for Ontology Evaluation,” Knowledge Web Deliverable, D1.2.3, 2004.

[240]. A. Lozano-Tello and A. Gómez-Pérez, “ONTOMETRIC: A Method to Choose the Appropriate Ontology,” Journal of Database Management (JDM), 2004, pp. 1-43.

[241]. H. Yao, A. M. Orme and L. Etzkorn, “Cohesion Metrics for Ontology Design and Application,” Journal of Computer Science, 2005, pp.107-113.

[242]. C. Welty and N. Guarino, "Supporting Ontological Analysis of Taxonomic Relationships," Data and Knowledge Engineering, 2001, pp. 51-74.

[243]. P. Spyns, “EvaLexon: Assessing Triples Mined from Texts,” Semantic technology and Research Lab, Brussels, Rep. STAR-2005-09, 2005.

[244]. R. Porzel and R. Malaka, “A Task-Based Approach for Ontology Evaluation,” Proceeding of ECAI 2004, 2004, pp.1-11.

[245]. W. Daelemans and M. L. Reinberger, “Shallow Text Understanding for Ontology Content Evaluation,” IEEE Intelligent Systems, 2004, pp.1541-1672.

[246]. A. Gangemi, C. Catenacci, M. Ciaramita and J. Lehmann, “Modeling Ontology Evaluation and Validation,” In Proceedings of the 3rd European Semantic Web Conference, 2006, pp.140-154.

[247]. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, “Gene Ontology: Tool for the Unification of Biology,” The Gene Ontology Consortium., 2000, pp.25–29.

[248]. S. C. J. Lam, D. Sleeman and W. Vasconcelos, “ReTAX+: A Cooperative Taxonomy Revision Tool,” In Proceedings of AI-2004 Conference, 2004, pp.64–77.

[249]. H. Knublauch, R. W. Fergerson, N. F. Noy and M. A. Musen, “The Protégé OWL Plugins: An Open Development Environment for Semantic Web Applications,” In the Semantic Web – ISWC 2004, 2004, pp. 229–243.

[250]. E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, Y. Katz, “Pellet: An OWL DL Reasoner,” Description Logics, CEUR Workshop Proceedings, 2004, pp.1-26.

[251]. H. Alani, S. Harris, and B. O’Neil, “Ontology Winnowing: A Case Study on the AKT Reference Ontology,” In Proceedings of International Conference on Intelligent Agents, Web Technology and Internet Commerce-IAWTIC’2005, 2005, pp.185-199.

[252]. H. Alani, S. Harris and B. O'Neil, "Winnowing Ontologies Based On Application Use," In Proceedings of 3rd European Semantic Web Conference (ESWC), 2006b, 2006, pp. 1-6.

[253]. N. Guarino and C. Welty, "Evaluating Ontological Decisions with OntoClean," Commun. ACM, 2002, pp. 61–65.

[254]. A. Gangemi, N. Guarino and A. Oltramari, "Conceptual Analysis of Lexical Taxonomies the Case of Wordnet Top-Level," In Proceedings of the International Conference on Formal Ontology in Information Systems, 2001, pp.285–296.

[255]. A. Oltramari, A. Gangemi, N. Guarino and C. Masolo, "Restructuring WordNets Top Level: The OntoClean approach," In Proceedings of the Workshop OntoLex, Ontologies and Lexical Knowledge Bases, 2002, pp.1-10.

[256]. Y. Wilks, "Ontotherapy or How to Stop Worrying about What There Is," Third International Conference on Language Resources and Evaluation, 2002, pp.1-18.

[257]. L. Wittgenstein, "Philosophical Investigations," 4th edition, Blackwells, Oxford, 1953.

[258]. W. V. Quine, "Two Dogmas of Empiricism," The Philosophical Review, 1951, pp.20–43.

[259]. H. Putnam, "Is Semantics Possible?" Meta-philosophy, 1970, pp.187–201.

[260]. I. R. Horrocks, "Using Expressive Description Logic: Fact or Fiction?" Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning (KR-98), 1998, pp. 636–649.

[261]. V. Haarslev and R. Möller, "Description of the Racer System and Its Applications," Working Notes of the 2001 International Description Logics Workshop (DL-2001), 2001, pp.1-10.

[262]. J. S. Luciano, "PAX of Mind for Pathway Researchers," Drug Discovery Today, 2005, pp.937–942.

[263]. R. Davis, H. Shrobe and P. Szolovits, “What Is a Knowledge Representation? “ AI Magazine, 1993, pp.17–33.

[264]. C. Brewster, O. Iria, F. Ciravegna and Y. Wilks, “The ontology: Chimaera or Pegasus,” In Proceedings of the Dagstuhl Seminar Machine Learning for the Semantic Web, 2005, pp. 1-6.

[265]. P. Velardi, R. Navigli, A. Cucchiarelli and F. Neri, “Evaluation of Ontolearn: a Methodology for Automatic Learning of Domain Ontologies,” IOS Press, 2003, pp. 1-6.

[266]. G. Grefenstette, “The World Wide Web as a Resource for Example-Based Machine Translation Tasks,” In Translating and the Computer 21: Proceedings of the 21st International Conference on Translating and the Computer, 1999, pp. 1-12.

[267]. R. Porzel and R. Malaka, “A Task-Based Framework for Ontology Learning, Population and Evaluation,” Ontology Learning from Text: Methods, Evaluation and Applications Frontiers in Artificial Intelligence and Applications Series, IOS Press, 2005, pp. 107-177.

[268]. W. Wahlster, “Smartkom: Symetric Multimodality in an Adaptive and Reusable Dialog Shell,” Proceedings of the Human Computer Interaction Status Conference, 2003, pp.47–62.

[269]. G. Grefenstette, “Explorations in Automatic Thesaurus Discovery,” Kluwer, Amsterdam, 1994.

[270]. C. Brewster, H. Alani and A. Dasmahapatra, “Data Driven Ontology Evaluation,” In Proceedings of the International Conference on Language Resources and Evaluation (LREC-04), 2004, pp. 1-4.

[271]. J. Brank, D. Mladenić and M. Grobelnik, “Golden Standard Based Ontology Evaluation Using Instance Assignment,” In: Proc. of The Eon 2006 Workshop, 2006, pp. 1-8.

[272]. G. Grefenstette, “Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches,” Acquisition of Lexical Knowledge from

Text: Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the Association for Computational Linguistics, 1993, pp. 1-12.

[273]. A. Maedche and S. Staab, “Ontology Learning,” In S. Staab & R. Studer (eds.) Handbook on Ontologies, Springer, 2004.

[274]. K. Dellschaft and S. Staab, “On How to Perform a Gold Standard Based Evaluation of Ontology Learning,” In Cruz et al. [2006], 2006, pp.228–241.

[275]. W.M. Rand, “Objective Criteria for the Evaluation of Clustering Methods,” Journal of the American Statistical Association, 1971, pp.846–850.

[276]. M. Gruninger and M. S. Fox, “Methodology for the Design and Evaluation of Ontologies,” In International Joint Conference on Artificial Intelligence (IJCAI95), Workshop on Basic Ontological Issues in Knowledge Sharing, 1995, pp.1-10.

[277]. L. Obrst, W. Ceusters, I. Mani, S. Ray and B. Smith, “The Evaluation of Ontologies,” In C. J. Baker and K.-H. Cheung, editors, Revolutionizing Knowledge Discovery in the Life Sciences, chapter 7, 2007, pp. 139–158.

[278]. K. Supekar, “A Peer-Review Approach for Ontology Evaluation,” In Proceedings of the International Protégé Conference, 2005, pp. 1-5.

[279]. M. S. Fox, M. Barbuceanu, M. Gruninger and J. Lin, “Organization Ontology for Enterprise Modeling,” Simulating organizations: Computational Models of Institutions and Groups, AAAI/MIT Press, 1998, pp.131-152.

[280]. A. G´omez-P´erez, “Towards a Framework to Verify Knowledge Sharing Technology,” Expert Systems with Applications, 1996, pp.519–529.

[281]. K. W. Fung and O. Bodenreider, “Knowledge Representation and Ontologies,” Clinical Research Informatics, Springer-Verlag, 2012, pp.255-275.

[282]. G. Palermo, “The Ontology of Economic Power in Capitalism: Mainstream Economics and Marx,” Cambridge Journal of Economics, 2007, pp.539–561.

[283]. G. L. Zuniga, “Ontology of Economic Objects,” American Journal of Economics and Sociology, 1999, pp. 299-312.

[284]. L. Zhou, “Ontology Learning: State Of The Art and Open Issues,” Springer Science and Business, 2007, pp.241–252.

[285]. A. Lozano-Tello, A. Gómez-Pérez and E. Sosa, “Selection of Ontologies for the Semantic Web,” SOSA, 2004, pp. 414-416.

[286]. R. Witte, “Flexible Ontology Population from Text: The OwlExporter,” In: Int. Conf. on Language Resources and Evaluation (LREC), 2010, pp. 3845-3850.

[287]. M. Hazman, S. R. El-Beltagy and A. Rafea, “A Survey of Ontology Learning Approaches,” International Journal of Computer Applications, 2011, pp. 0975 - 8887.

[288]. A. Maedche and S. Staab, “Semi-automatic Engineering of Ontologies from Text,” Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering, 2000, pp.1-8.

[289]. M. Atzmueller and P. Kluegl and F. Puppe,”Rule-Based Information Extraction for Structured Data Acquisition using TextMarker,” In Proc. of the LWA-2008 (KDML Track), 2008, pp.1–7.

[290]. M. E. Califf and R. J. Mooney, “Bottom-up Relational Learning of Pattern Matching Rules for Information Extraction,” Journal of Machine Learning Research, 2003, pp.177–210.

[291]. P. Kluegl, M. Atzmueller, T. Hermann and F. Puppe, “A Framework for Semi-Automatic Development of Rule-based Information Extraction Applications,” Track on Knowledge Discovery and Machine Learning, 2009, pp.56-59.

[292]. T. S. Kuhn, “The Essential Tension: Tradition and Innovation in Scientific Research?” in the Essential Tension,” Chicago University Press, Chapter 9, 1977.

[293]. L. Chiticariu and et al, “Rule-Based Information Extraction Is Dead! Long Live Rule-Based Information Extraction System,” Proceedings of the 2013 Conference on Empirical methods in natural language Processing, 2013, pp. 827-832.

[294]. J. Zhu, Z. Nie, X. Liu, B. Zhang and J. Wen, “StatSnowball: a Statistical Approach to Extracting Entity Relationships,” ACM, 2009, pp. 101-110.

[295]. Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang and M. Ishizuka, “Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web,”Proceeding of the 47th Annual meeting of the ACL and the 4th IJCNLP of the AFNLP, 2009, pp. 1021-1029.

[296]. G. Hripcsak, C. Friedman, P. O. Alderson, W. DuMouchel, S. B. Johnson, and P. D. Clayton, “Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing,” Ann Intern Med, 1995, pp.681–688.

[297]. R. Yangarber and R. Grishman, “Machine Learning of Extraction Patterns from Annotated Corpora: Position Statement,” In Proceedings of Workshop on Machine learning in Information Extraction, 2001, pp. 76-83.

[298]. A. Jimeno-Yepes and A. A. Aronson, “Knowledge-based and Knowledge-lean Methods Combined in Unsupervised Word Sense Disambiguation,” IHI '12 Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 2012, pp.733-736.

[299]. B. T. McInnes, “Supervised and Knowledge-based Methods for Disambiguating Terms in Biomedical Text using the UMLS and MetaMap,” PhD Dissertation, University Of Minnesota, Minnesota, 2009.

[300]. A. M¨adche and R. Volz, “The Text-To-Onto Ontology Extraction and Maintenance System,” In Workshop on Integrating Data Mining and Knowledge Management, collocated with the 1st International Conference on Data Mining, 2001, pp. 1-5.

[301]. E. Bozsak, M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, G. Stumme, Y. Sure, J. Tane, R. Volz and V. Zacharias, “KAON – Towards a Large Scale Semantic Web,” In Proceedings of the Third International Conference on E-

Commerce and Web Technologies (EC-Web), Springer Lecture Notes in Computer Science, 2002, pp. 304-313.

[302]. P. Cimiano and J. V¨olker, "Text2onto – A Framework for Ontology Learning and Data-Driven Change Discovery," Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), Lecture Notes in Computer Science, 2005, pp. 227–238.

[303]. P. Buitelaar, D. Olejnik and M. Sintek, "A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis," In Proceedings of the 1st European Semantic Web Symposium (ESWS), 2004, pp.31–44.

[304]. C. Biemann, "Ontology Learning from Text: A Survey of Methods," LDV Forum, 2005, pp.75–93.

[305]. M. Y. Dahaba, H. A. Hassanb, A. Rafeab, "TextOntoEx: Automatic Ontology Construction from Natural English Text," Expert Systems with Applications, 2008, pp.1474–1480.

[306]. C. Brewster and K. O'Hara, "Knowledge Representation with Ontologies: The Present and Future, "IEEE Intelligent Systems, 2004, pp.72–81.

[307]. M. Missikoff, "The Usable Ontology: An Environment for Building and Assessing Domain Ontology," In ISWC 2002, 2002, pp.39–53.

[308]. P. Suraweera, A. Mitrovic and B. Martin, "Widening the Knowledge Acquisition Bottleneck for Constraint-based Tutors," International Journal of Artificial Intelligence in Education, 2010, pp. p137-173.

[309]. R. Navigli and P. Velardi, "Learning Domain Ontologies from Document Warehouses and Dedicated Websites," Computational Linguistics, 2004, pp.151–179.

[310]. R. Serban, A. T. Teije, F. V. Harmelen, M. Marcos, C. Polo-Conde, "Extraction and Use of Linguistic Patterns for Modeling Medical Guidelines," Artificial Intelligence in Medicine, 2007, pp.137-149.

[311]. S. Sahay, B. Li, E. V. Garcia, E. Agichtein and A. Ram, "Domain Ontology Construction from Biomedical Text," International Conference on Artificial Intelligence, CSREA Press, 2007, pp. 1-7.

[312]. S. Wu and W. Hsu, "SOAT A Semi-Automatic Domain Ontology Acquisition Tool from Chinese Corpus," In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), 2002, pp. 1-5.

[313]. X. Jiang and A. Tan, "CRCTOL: A Semantic-Based Domain Ontology Learning System," Journal of the American Society for Information Science and Technology, 2010, pp.150–168.

[314]. A. T. McCray, "Upper-Level Ontology for the Biomedical Domain," Comparative and Functional Genomics, 2003, pp. 80-84.

[315]. A. Burgun and O. Bodenreider,"Mapping the UMLS Semantic Network into General Ontologies," Proceedings of AMIA Annual Symposium 2001, 2001, pp. 86-90.

[316]. P. Cimiano, A. Pivk, L. Schmidt-Thieme and S. Staab, "Learning Taxonomic Relations from Heterogeneous Sources of Evidence," In Ontology Learning from Text: Methods, Evaluation and Applications, Frontiers in Artificial Intelligence, 2005, pp. 1-6.

[317]. L. Smith, T. Rindflesch and W. J. Wilbur, "MedPost: a Part-Of-Speech Tagger for Biomedical Text," Bioinformatics, 2004, pp. 2320-2321.

[318]. A. T. McCray, S. Srinivasan and A. C. Browne, "Lexical Methods for Managing Variation in Biomedical Terminologies," Proc 18th SCAMC, 1994, pp.235-239.

[319]. D. L. McGuinness and F. V. Harmelen, "OWL Web Ontology Language Overview," W3C Recommendation, Rep. 27, 2009.

[320]. D. Vrandecic, "Ontology Evaluation," In S. Staab and R. Studer, editors, Handbook on Ontologies, 2nd edition, Springer, 2009, pp. 293–313.

[321]. N. Guarino, “Toward a Formal Evaluation of Ontology Quality,” IEEE Intelligent Systems, 2004, pp78–79.

[322]. L. Al-Safadi, R. Alomran and F. Almutairi, “Evaluation of MetaMap Performance in Radiographic Images Retrieval,” Research Journal of Applied Sciences, Engineering and Technology, 2013, pp. 4231-4236.

[323]. H. Zhang, Y. Li and H. B. K. Tan, “Measuring Design Complexity of Semantic Web Ontologies,” The Journal of Systems and Software, 2010, pp.803–814.

[324]. M. A. Sicilia, D. Rodrнguez, E. Garcна-Barriocanal and S. Sбnchez-Alonso, “Empirical Findings on Ontology Metrics,” Expert Systems with Applications, 2012, pp.6706–6711.

[325]. M. Poprat, E. Beisswanger and U. Hahn, “Building a BIOWORDNET by Using WORDNET’s Data Formats and WORDNET’s Software Infrastructure - A Failure Story,” Software Engineering, Testing and Quality Assurance for Natural Language Processing, Association for Computational Linguistics, 2008, pp.31–39.

[326]. H. Hlomani and D. Stacey, “Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey,” IOS Press, 2014, pp. 1-11.

[327]. G. Bisson, C. Nedellec and L. Canamero, “Designing Clustering Methods for Ontology Building - The Mo'K Workbench,” In Proceedings of the ECAI Ontology Learning Workshop, 2000, pp.13-19.

[328]. A. T. McCray, A. R. Aronson, A. C. Browne, T. C. Rindflesch, A. Razi, S. Srinivasan, “UMLS Knowledge for Biomedical Language Processing,” Journal of medical Library of Associations, 1993, pp. 184-194.

[329]. F. Corcoglioniti, M. Rospocher, and A P. Aprosio, F. B. Kessler. “2-phase Frame-based Knowledge Extraction Framework.” SAC '16 Proceedings of the 31st Annual ACM Symposium on Applied Computing, 2016, pp.354-361.

Appendix

A. Instantiation

1. Before Cleaning

<?xml version="1.0" encoding="UTF-8"?>
<article filename="10.1007_s12178-007-9000-5.xml" doi="10.1007/s12178-007-9000-5" url=""
><fulltext>Scapular winging is a rare debilitating condition that leads to limited functional activity of the upper extremity. It is the result of numerous causes, including traumatic, iatrogenic, and idiopathic processes that most often result in nerve injury and paralysis of either the serratus anterior, trapezius, or rhomboid muscles. Diagnosis is easily made upon visible inspection of the scapula, with serratus anterior paralysis resulting in medial winging of the scapula. This is in contrast to the lateral winging generated by trapezius and rhomboid paralysis. Most cases of serratus anterior paralysis spontaneously resolve within 24   months, while conservative treatment of trapezius paralysis is less effective. A conservative course of treatment is usually followed for rhomboid paralysis. To allow time for spontaneous recovery, a 6 – 24   month course of conservative treatment is often recommended, after which if there is no recovery, patients become candidates for corrective surgery. Introduction Scapular winging is a rare, but potentially debilitating condition that can affect the ability to lift, pull, and push heavy objects, as well as to perform daily activities of living, such as brushing one ’ s hair and teeth and carrying grocery bags [1]. Cosmetically, some patients may be distressed by pronounced winging [2]. Disrupting scapulohumeral rhythm, scapular winging contributes to loss of power and limited flexion and abduction of the upper extremity and can be a source of considerable pain [3 – 8]. A condition due to a number of etiologies, most cases are due to lesions of the long thoracic and spinal accessory nerves that innervate the serratus anterior and trapezius muscles, respectively. Rarely, it may also be due to a lesion to the dorsal scapular nerve that innervates the rhomboid muscles. These scapular muscles (Fig.   1) contribute to keeping the medial border of the scapula protracted against the posterior thoracic wall, and denervation or paralysis of any of these muscles results in the winging of the medial border of the scapula as it lifts off the thoracic wall. In addition, the scapula may translate medially or laterally along the posterior thoracic wall due to unopposed muscle contraction of the other functioning scapular muscles, a distinction known as medial (serratus anterior paralysis) or lateral (trapezius or rhomboid paralysis) winging (Table   1). Scapular winging may present in a variety of clinical contexts, and may be due to traumatic- or sports-related injury [2 , 4 , 5 , 8 – 22], iatrogenic injury [1 , 2 , 12 , 15 , 16 , 19 , 23 – 27], or spontaneous in nature [6 – 8 , 11 , 27 – 30]. Here we discuss incidence and demographics, pertinent anatomy, the nature of the injury, diagnostic tests, and treatment options for scapular winging due to serratus anterior, trapezius, and rhomboid muscle paralysis./fulltext></article>>

2. After Cleaning


```
<?xml version="1.0" encoding="UTF-8"?>
<document docid="1">
Scapular winging is a rare debilitating condition that leads to limited functional activity of the upper extremity. It is the result of numerous causes including traumatic iatrogenic and idiopathic processes that most often result in nerve injury and paralysis of either the serratus anterior trapezius or rhomboid muscles. Diagnosis is easily made upon visible inspection of the scapula with serratus anterior paralysis resulting in medial winging of the scapula. This is in contrast to the lateral winging generated by trapezius and rhomboid paralysis. Most cases of serratus anterior paralysis spontaneously resolve within 24 months while conservative treatment of trapezius paralysis is less effective. A conservative course of treatment is usually followed for rhomboid paralysis. To allow time for spontaneous recovery a 6 24 month course of conservative treatment is often recommended after which if there is no recovery patients become candidates for corrective surgery. Introduction Scapular winging is a rare but potentially debilitating condition that can affect the ability to lift pull and push heavy objects as well as to perform daily activities of living such as brushing one hair and teeth and carrying grocery bags. Cosmetically, some patients may be distressed by pronounced winging. Disrupting scapulohumeral rhythm  scapular winging contributes to loss of power and limited flexion and abduction of the upper extremity and can be a source of considerable pain. A condition due to a number of etiologies most cases are due to lesions of the long thoracic and spinal accessory nerves that innervate the serratus anterior and trapezius muscles respectively. Rarely, it may also be due to a lesion to the dorsal scapular nerve that innervates the rhomboid muscles. These scapular muscles contribute to keeping the medial border of the scapula protracted against the posterior thoracic wall  and denervation or paralysis of any of these muscles results in the winging of the medial border of the scapula as it lifts off the thoracic wall. In addition  the scapula may translate medially or laterally along the posterior thoracic wall due to unopposed muscle contraction of the other functioning scapular muscles a distinction known as medial serratus anterior paralysis or lateral trapezius or rhomboid paralysis winging. Scapular winging may present in a variety of clinical contexts and may be due to traumatic or sports related injury iatrogenic injury or spontaneous in nature. Here we discuss incidence and demographics pertinent anatomy the nature of the injury diagnostic tests and treatment options for scapular winging due to serratus anterior trapezius and rhomboid muscle paralysis.
</document>
```

3. Concept Disambiguation

```
Processing 00000000.tx.1: Anterior impingement is a common problem in dancers occurring
primarily secondary to the repetitive forced ankle dorsiflexion inherent in ballet.

Phrase: Anterior impingement
Meta Mapping (694):
   694   Anterior [Functional Concept]
Phrase: a common problem in dancers
Meta Mapping (696):
   593   Common (Common (qualifier value)) [Quantitative Concept]
   760   Problem [Finding]
   593   dancers (Dancer (occupation)) [Professional or Occupational Group]
Meta Mapping (696):
   593   common (Common Specifications in HL7 V3 Publishing) [Intellectual Product]
   760   Problem [Finding]
   593   dancers (Dancer (occupation)) [Professional or Occupational Group]
Meta Mapping (696):
   593   Common (shared attribute) [Functional Concept]
   760   Problem [Finding]
   593   dancers (Dancer (occupation)) [Professional or Occupational Group]
Phrase: occurring
Meta Mapping (966):
   966   Occur (Occur (action)) [Activity]
Meta Mapping (966):
   966   OCCUR (Occurrence) [Temporal Concept]
Phrase: primarily secondary to the repetitive forced ankle dorsiflexion
Meta Mapping (706):
   770   Secondary to [Temporal Concept]
   578   Forced (Force) [Phenomenon or Process]
   604   ankle dorsiflexion (Dorsiflexion of foot) [Organism Function]
Meta Mapping (706):
   770   Secondary to [Temporal Concept]
   578   Forced [Functional Concept]
   604   ankle dorsiflexion (Dorsiflexion of foot) [Organism Function]
Phrase: inherent in ballet.
Meta Mapping (623):
   623   ballet [Daily or Recreational Activity]
Processing 00000000.tx.2: Symptoms generally occur progressively and may respond to
conservative treatment including addressing biomechanical faults that contribute to the
problem.
Phrase: Symptoms generally
Meta Mapping (861):
   861   Symptoms [Sign or Symptom]
Meta Mapping (861):
   861   symptoms (Symptoms aspect) [Functional Concept]
Phrase: occur
Meta Mapping (1000):
  1000   Occur (Occur (action)) [Activity]
Meta Mapping (1000):
  1000   OCCUR (Occurrence) [Temporal Concept]
Phrase: respond to conservative treatment
Meta Mapping (770):
   770   Treatment (Administration procedure) [Therapeutic or Preventive Procedure]
Meta Mapping (770):
   770   Treatment (Biomaterial Treatment) [Conceptual Entity]
Meta Mapping (770):
   770   Treatment (Therapeutic procedure) [Therapeutic or Preventive Procedure]
Meta Mapping (770):
   770   Treatment (Treating) [Functional Concept]
Meta Mapping (770):
   770   TREATMENT (Treatment Epoch) [Research Activity]
Meta Mapping (770):
   770   treatment (therapeutic aspects) [Functional Concept]
```

4. Proposition Disambiguation before Post-processing

```
<?xml version="1.0" encoding="UTF-8"?>
<SemRepAnnotation>
<Utterance id="U00000000.tx.1" section="tx" number="1" text="&lt;?xml version=&quot;1.0&quot;
encoding=&quot;UTF-8&quot;?&gt; &lt;article filename=&quot;10.1007_s12178-007-9000-5.xml&quot;
doi=&quot;10.1007/s12178-007-9000-5&quot; url=&quot;&quot;&gt;&lt;fulltext&gt;Scapular winging
is a rare debilitating condition that leads to limited functional activity of the upper
extremity.">
  <Entity id="U00000000.tx.1.E1" cui="C0333052" name="Version" semtypes="ftcn" text="version"
score="861" begin="7" end="13" />
  <Entity id="U00000000.tx.1.E2" cui="C1555005" name="UTF-8" semtypes="inpr" text="UTF-8"
score="1000" begin="31" end="35" />
  <Entity id="U00000000.tx.1.E3" cui="C1281575" name="Entire scapula" semtypes="bpoc"
text="Scapular" score="836" begin="139" end="146" />
  <Entity id="U00000000.tx.1.E4" cui="C0043189" name="Wing" semtypes="bpoc" text="winging"
score="836" begin="148" end="154" />
  <Entity id="U00000000.tx.1.E5" cui="C0522498" name="Rare" semtypes="qlco" text="rare"
score="802" begin="161" end="164" />
  <Entity id="U00000000.tx.1.E6" cui="C0348080" name="Condition" semtypes="qlco"
text="condition" score="802" begin="179" end="187" />
  <Entity id="U00000000.tx.1.E7" cui="C0439801" name="Limited" semtypes="ftcn" text="limited"
score="851" begin="203" end="209" />
  <Entity id="U00000000.tx.1.E8" cui="C0205245" name="Functional" semtypes="ftcn"
text="functional" score="851" begin="211" end="220" />
  <Entity id="U00000000.tx.1.E9" cui="C0439167" name="% activity" semtypes="qnco"
text="activity" score="851" begin="222" end="229" />
  <Entity id="U00000000.tx.1.E10" cui="C1140618" name="Upper Extremity" semtypes="bpoc"
text="upper extremity" score="1000" begin="238" end="252" />
  <Predication id="U00000000.tx.1.P1">
   <Subject maxDist="0" dist="0" entityID="U00000000.tx.1.E4" relSemType="bpoc" />
   <Predicate type="PART_OF" indicatorType="MOD_HEAD" begin="139" end="154" />
   <Object maxDist="0" dist="0" entityID="U00000000.tx.1.E3" relSemType="bpoc" />
  </Predication>
 </Utterance>
 <Utterance id="U00000000.tx.2" section="tx" number="2" text="It is the result of numerous
causes, including traumatic, iatrogenic, and idiopathic processes that most often result in
nerve injury and paralysis of either the serratus anterior, trapezius, or rhomboid muscles.">
  <Entity id="U00000000.tx.2.E1" cui="C1274040" name="result" semtypes="ftcn" text="result"
```

5. Semantic Proposition Disambiguation with Enhanced algorithm

abnormal_cell_affected_by_chemical_or_drug	alternatively_used_for	attributed_continuous_with
abnormality_associated_with_allele	alternatively_used_for	Attributed continuous with
access_device_used_by	analyzed_by	attributed_part_of
access_of	analyzes	Attributed part of
active_ingredient_of	Analyzes	attributed_regional_part_of
active_metabolites_of	anatomic_structure_has_location	Attributed regional part of
activity_of_allele	anatomic structure has location	biological_process_has_associated_location
actual_outcome_of	anatomic_structure_is_physical_part_of	biological process has associated location
adheres_to	anatomic structure is physical part of	biological_process_has_initiator_chemical_or_drug
Adheres to	anatomy_originated_from_biological_process	biological process has initiator chemical or drug
adjacent_to	anatomy originated from biological process	biological_process_has_initiator_process
Adjacent to	approximately_mapped_from	biological process has initiator process
adjectival_form_of	Approximately Mapped from	biological_process_has_result_anatomy
Adjectival form of	approximately_mapped_to	biological process has result anatomy
adjustment_of	Approximately Mapped to	biological_process_has_result_biological_process
Adjustment of	arterial_supply_of	biological process has result biological process
afferent_to	Arterial supply of	biological_process_has_result_chemical_or_drug
alias_of	articulates_with	biological process has result chemical or drug
Alias_of	Articulates with	biological_process_involves_chemical_or_drug
allele_absent_from_wild-type_chromosomal_location	associated_disease	biological process involves chemical or drug
allele absent from wild-type chromosomal location	Associated disease	biological_process_involves_gene_product
allele_associated_with_disease	associated_finding_of	biological process involves gene product
allele associated with disease	Associated finding of	biological_process_is_part_of_process
allele_has_abnormality	associated_genetic_condition	biological process is part of process
allele has abnormality	Associated genetic condition	biological_process_results_from_biological_process
allele_has_activity	associated_morphology_of	biological process results from biological process
allele has activity	Associated morphology of	biomarker_type_includes_gene_product
allele_in_chromosomal_location	associated_procedure_of	biomarker type includes gene product
allele in chromosomal location	Associated procedure of	biomarker_type_includes_gene
allele_not_associated_with_disease	associated_with_malfunction_of_gene_product	biomarker type includes gene
allele not associated with disease	associated with malfunction of gene product	bounded_by
allele_plays_altered_role_in_process	associated_with	Bounded by
allele plays altered role in process	Associated with	bounds
allele_plays_role_in_metabolism_of_chemical_or_drug	attaches_to	Bounds
allele plays role in metabolism of chemical or drug	attributed_constitutional_part_of	branch_of
allelic_variant_of	Attributed constitutional part of	Branch of
Allelic Variant of	attributed_continuous_with	british_form_of

6. Proposition Disambiguation after Post-processing

C1096593:CAUSES:C0917801
C0184661:TREATS:C0030193
C0011307:TREATS:C0442726
C0087111:METHOD_OF:C0543467
C0037004:LOCATION_OF:C1457887
C0087111:TREATS:C0522224
C0037047:LOCATION_OF:C1285497
C0817096:LOCATION_OF:C1306645
C0543467:TREATS:C0748691
C0580841:PROCESS_OF:C0030705
C0021153:ISA:C0543467
C0175677:COEXISTS_WITH:C0012691
C0196542:TREATS:C0030705
C0037004:LOCATION_OF:C0029365
C0034542:CAUSES:C0152180
C0037004:LOCATION_OF:C0271548
C0522224:PROCESS_OF:C0030705
C0021153:PRECEDES:C0021153
C0037763:CAUSES:C0030193
C0037949:PART_OF:C1281575
C0021400:ISA:C0042769
C0205076:LOCATION_OF:C0181620
C1140618:LOCATION_OF:C0580846
C0043189:PART_OF:C0030705
C1280230:LOCATION_OF:C0522224
C0000905:LOCATION_OF:C0175677
C0543467:ADMINISTERED_TO:C0030705
C0522224:ISA:C1457887
C0000905:LOCATION_OF:C0221198
C0434255:CAUSES:C1265748
C0026845:PART_OF:C1281575
C1281575:LOCATION_OF:C1285497
C0037004:LOCATION_OF:C0018670
C1281575:LOCATION_OF:C0935623
C1281575:LOCATION_OF:C0522224
C0006086:TREATS:C0237401
C0332835:PART_OF:C1305735
C0029423:PART_OF:C1281575
C1280976:LOCATION_OF:C0522224
C0027530:LOCATION_OF:C0225006
C0196878:TREATS:C1457887
C0817096:LOCATION_OF:C0221198
C0020164:LOCATION_OF:C0018670
C0037004:LOCATION_OF:C1306645
C0543467:TREATS:C0030705
C0185473:METHOD_OF:C0185470
C0240953:PROCESS_OF:C1524106
C0196878:ISA:C0543467
C0087111:PRECEDES:C0021153
C0434255:COEXISTS_WITH:C0332667
C1279046:LOCATION_OF:C0005558
C1293130:TREATS:C0175677
C0205166:ISA:C1444754
C1281575:PART_OF:C1281575
C1281575:LOCATION_OF:C0181620
C0522224:PROCESS_OF:C0237401
C0037004:LOCATION_OF:C0030193
C0221198:PROCESS_OF:C0030705
C0037004:LOCATION_OF:C0234238
C0748691:PROCESS_OF:C0030705
C0026845:PART_OF:C0036277
C0434255:ISA:C0043251
C0518031:PROCESS_OF:C0030705
C0019552:LOCATION_OF:C0018563

7. Other axioms

```
<owl:Class rdf:about="#ANAT">
    <rdfs:label xml:lang="en">Anatomy</rdfs:label>
        <owl:equivalentClass>
            <owl:Class>
                <owl:unionOf rdf:parseType="Collection">
                    <rdfs:Class rdf:resource="#anst"/>
                    <rdfs:Class rdf:resource="#blor"/>
                    <rdfs:Class rdf:resource="#bpoc"/>
                    <rdfs:Class rdf:resource="#bsoj"/>
                    <rdfs:Class rdf:resource="#bdsu"/>
                    <rdfs:Class rdf:resource="#bdsy"/>
                    <rdfs:Class rdf:resource="#cell"/>
                    <rdfs:Class rdf:resource="#celc"/>
                    <rdfs:Class rdf:resource="#emst"/>
                    <rdfs:Class rdf:resource="#ffas"/>
                    <rdfs:Class rdf:resource="#tisu"/>
                </owl:unionOf>
            </owl:Class>
        </owl:equivalentClass>
        <owl: disjiontWith rdf:resource="#DISO"/>
        <owl: disjiontWith rdf:resource="#PHEN"/>
        <owl: disjiontWith rdf:resource="#ACTI"/>
        <owl: disjiontWith rdf:resource="#CHEM"/>
    </owl:Class>
<owl:ObjectProperty rdf:about="#PHYRT">
    <rdfs:label xml:lang="en">physically_related_to</rdfs:label>
        <owl:equivalentProperty>
            <owl:ObjectProperty>
                <owl:unionOf rdf:parseType="Collection">
                    <rdfs:ObjectProperty rdf:resource="#part_of"/>
                    <rdfs:ObjectProperty rdf:resource="#contains"/>
                    <rdfs:ObjectProperty rdf:resource="#consists_of"/>
                    <rdfs:ObjectProperty rdf:resource="#connected_to"/>
                    <rdfs:ObjectProperty rdf:resource="#interconnects"/>
                    <rdfs:ObjectProperty rdf:resource="#branch_of"/>
                    <rdfs:ObjectProperty rdf:resource="#tributary_of"/>
                    <rdfs:ObjectProperty rdf:resource="#ingredient_of"/>
                </owl:unionOf>
            </owl:ObjectProperty>
        </owl:equivalentProperty>
        <owl: disjiontWith rdf:resources="#conceptually_related_to"/>
        <owl: disjiontWith rdf:resources="#spatially_related_to"/>
        <owl: disjiontWith rdf:resources="#functionally_related_to"/>
        <owl: disjiontWith rdf:resources="#conceptually_related_to"/>
    </owl:Class>
```

List of Publications

1. D. Ayele, C. Chevallet, M. Meshesha and G. Mitikie. ”Constructing Reference Semantic Predictions from Biomedical Knowledge Sources.” Proceedings of COLING 2012: Technical Papers, pp. 133–148, 2012.

2. D. Ayele, C. Chevallet, M. Meshesha and G. Mitikie. ”Enhancing Semantic Relation Quality of UMLS Knowledge Sources.” MEDES '12 Proceedings of the International Conference on Management of Emergent Digital Ecosystems, pp.59-66, 2012.

Declaration Sheet

Declaration Sheet

This dissertation is my original work, has not been presented for a degree in any other Universities and all sources of material used for the dissertation have been duly acknowledge.

Demeke Asres Ayele

As an advisor and co-advisors, we confirm to the best of our knowledge.

Jean-Pierre Chevallet (Prof)	Million Meshesha (PhD)	Getnet Mitikie (Prof)