International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 8, No.6 (Nov-2019)

http://dx.doi.org/10.12785/ijcds/080602

Enhancing Ontology by Integrating Facts from Unstructured Data and Mapping with Linguistic Knowledge

Runumi Devi[1] and Deepti Mehrotra[1]

[1]*Amity University Uttar Pradesh, India*

Received 09 Jan 2019, Revised 15 Jul. 2019, Accepted 20 Aug. 2019, Published 1 Nov. 2019

Abstract: Ontology plays a crucial role in the growth of semantic web by providing a domain- knowledge resource comprehensible to both humans as well as computers. Knowledge resources can be of two types-taxonomical knowledge supporting description logic in the form of ontology representing a particular domain and linguistic knowledge. An approach is presented in this paper for constructing a third type of knowledge resource(Hybrid) that accommodates variations in the structure as per RDF triple extracted from unstructured corpus and subsequently semantically associating ontological concepts with their linguistic counter-part. The purpose of this knowledge base is to add linguistic expressiveness to conceptual knowledge and to accommodate the dynamic nature of knowledge. Semantic similarity and term coverage are computed between terms(concepts). A case study i.e enhancing a dengue ontology by incorporating RDF triple extracted from patients' case sheets and semantically associating concepts with WordNet by using similarity and relatedness measures is presented.

Keywords: Ontology, Dengue, Knowledge base, Semantic similarity, WordNet

1. Introduction

Ontologies are widely used in the semantic web as knowledge representation systems through explicit and formal specification of conceptualization. It is related to cognitive science and is considered as a sub-branch of Artificial Intelligence (AI). Although ontologies have similarities with modelling languages, but they differ from modelling language by supporting description logic. It enables restrictions to be applied on relationship and property characteristics that enhance the capability of ontology reasoning [1]. Semantic web is characterized not only as human and machine shareable knowledge but also in the multilingual aspects. This necessitates the development of linguistically motivated ontology representing a relevant part of the world in order to build a collaborative semantic web. Adding linguistic expressivity to conceptual knowledge represented by ontology, may be helpful while capturing the intended meaning of concepts and roles, thus facilitating shareability of the ontology [2]. Semantic association of domain ontology concepts with linguistic knowledge base is critical in improving information retrieval. On the other hand, as the knowledge evolves over time, ontologies on the semantic web are often likely to be subjected to augmentation [3]. Thus the challenging task that is required to be addressed is to enhance a knowledge base by accommodating new properties with values and semantically associating with linguistic knowledge base. The paper uses a dengue ontology designed for dengue patients as existing Knowledge Base (KB) to be enhanced and WordNet: semantic network as linguistic knowledge base for semantic association. Thus, a Hybrid knowledge base is developed by combining ontological knowledge base, instantial knowledge base and linguistic knowledge base. The objective of enhancing the ontology is achieved in two phases:

i) Implementation of an algorithm for augmenting knowledge base by incorporating new knowledge in the form of RDF triples, extracted from unstructured resources.

ii) Semantically associating the augmented knowledge base with linguistic counterpart based on identifying the degree of similarity between words using information derived from semantic networks.-WordNet.

Final output would be an enhanced ontology in terms of knowledge represented which is integrated not only with instantial knowledge base but also associated with linguistic knowledge base facilitating shareability and retrieval.

The organization of this article is as follows: Section 2 provides a background study on enhancement of knowledge representation, linguistic knowledge base-

E-mail: runumi@jssaten.ac.in, mehdeepti@gmail.com

WordNet, measures of similarity and relatedness and ontology evaluation measures. Approach to accommodate variations in the ontology structure with the help of a case study-Dengue ontology is discussed in section 3. Section 4 include the addition of linguistic expressivity to ontology. Evaluation considering semantic similarity and term coverage is provided in section 5 and section 6 include conclusion.

2. BACKGROUND STUDY

Information extraction is generally concerned with retrieving information from a particular domain. Thus information extraction will yield better result provided there is an explicit and formal specification of concepts, for a particular domain through an ontology that is subjected to augmentation and semantic association with WordNet. For example, a dengue ontology that defines concepts like Patients, Symptom, Treatment etc, can be useful for assisting information extraction related to dengue symptom, treatment. However, ontology conceptualizing a particular domain is dynamic in nature, which is one of the challenging problem and restrict retrieval process to only model checking [3]. On the other hand, associating ontological content with linguistic knowledge base may also present better solution for increasing reusability of the represented knowledge.

The recent overviews of present state-of-the-art of ontology and linguistic knowledge integration are found in [4] and [5]. Pazienza, M. T. and Stellato [4] have provided taxonomy alignment framework using WordNet for two ontologies- one is Baseball Ont and the other is Moses Italian. Basili R. et al. [5] build a hierarchy of linguistically motivated domain concept using MeSH (Medical Sub Headings) and WordNet. MeSH categories are mapped with WordNet senses by computing their conceptual density along with term coverage. However, no direct solution is found for addressing the requirements mentioned in the introduction section in either of the two works[4,5] as none has addressed ontology enhancement by addressing the dynamic nature of domain conceptualization using posteriori schema definition of RDF model together with semantic association with Linguistic knowledge base. In the remaining section, the literature is explored in four subsections namely, Lexical Database-WordNet, WordNet usage for Ontology Enhancement ,Techniques used for finding relatedness and various perspectives of ontology evaluation.

A. Lexical Database-WordNet

WordNet is developed by Princeton University as a lexical resource, facilitating linguistic research that can be viewed as a semantic network [6]. It offers a wide range of predefined concepts and relations by organizing concepts into similar words which is termed as synsets[7]. Gloss is associated with each concept for its decription that includes an example usage. Thus, the basic unit of WordNet is synonym sets (synsets) that contain compounds, phrasal verbs, collections and idiomatic phrases. Synsets are unordered sets of distinct words that correspond to concepts [8].It organizes lexical information in terms of meanings as a taxonomical hierarchy of meanings that includes semantic relations between synsets. Two kinds of relations are used in WordNet, one is lexical and the other one is semantic relation [9]. Lexical relations exist between word forms whereas semantic relations are between synsets. Similar to dictionary, in WordNet also, word meanings are mapped with word forms. There are different types of word forms e.g. polysemous and synonymous. Polysemous are word forms having more than one meaning mapped to one word form. On the other hand, synonymous are word forms having same meaning mapped to more than one word form. Synonymous word forms are grouped into sets called synsets defining the meaning that they share. WordNet is divided into four separate semantic networks- nouns, adjectives, adverbs and verbs and include following semantic relations:[10]

i) Synonymy ii) Antonymy iii) Hyponymy iv) Metonymy v) Troponymy and vi) Entailment

Synonymy and Antonymy are for Noun, Verb, Adj and Adv. Whereas Hyponymy and Meronymy are for Noun and Troponymy and Entailment are for Verb.

Hyponym relation organizes nouns and verbs into a lexical hierarchy that can also be read as IS-A or IS-A-KIND-OF relation or subsumption relation. Adjectives represented by WordNet are divided into two categories: descriptive and relational. Antonymy is the basic semantic relation used for adjectives that express opposite values of an attribute. A cluster is formed by consisting of all antonyms (opposite pairs). However, all descriptive adjectives do not have antonyms. Thus, a similarity is defined indicating that the adjective that lack antonyms are similar in meaning with reference to the adjective that do not have antonyms. The relative objectives are connected to the nouns from which they are derived instead of forming an independent structure [11].

Unlike nouns and verbs, there is no hierarchical structure for adverbs as most are derived from adjectives by suffixation. However, semantic relation synonymy and antonymy are recognized for some adverbs.

B. Linguistic Knowledge base-WordNet for Ontology Enhancement

Nováček, V. et al. [3, 12] addressed the dynamic nature of knowledge and presented a framework that has taken into account the dynamics and data-intensive-ness in the domain of e-health and biomedicine applications. Toumouth, A. [9] addresses the problem of distinguishing domain specific sense from general purpose sense. The authors have presented a method of extracting Lexico-syntactic pattern like Noun_CJC_Noun from Ohsumed corpus where CJC can be {and, or, but}.Similarity between the pairs of nouns are found using the noun hierarchy of WordNet.

Dzikovska, M. O. et al. [13] presented a method that of developing spoken dialogue system in multiple domains that maintains two ontologies-one for language representation and the other is customized to domain. Broad-coverage language components are preserved across domains by defining mapping between ontologies.

Salahli, M. A.[14] presented an approach of measuring semantic relatedness between words based on determiner words sets. Montoyo, A. et al.[15] explored collaboration between (Word Sense Disambiguation)WSD methods-specification marks, a knowledge-based method and maximum entropy, a corpus-based method and shown that one can help the other to better perform the disambiguation process. Burgun, A., and Bodenreider, O.[16] compares knowledge representation using general terminological system(WordNet) and a domain specific system(UMLS). Slimani, T.[17] examines some of the most recognized semantic similarity measures in terms of accuracy, typology and key properties, used for estimating the similarity between concepts or terms. Bhatt, B. and Bhattacharyya, P. [18] links WordNets of Indian language with Suggested Upper Merged Ontology (SUMO)[19] that can be used for text processing applications. Koeva, S. et al.[20] merges WordNet concepts with Corpus Pattern Analysis semantic types. William D. Lewis [21] presents a method of automatic calculation of a metric-relative density measures within WordNet for measuring semantic similarity. Sanfilippo, A. P. et al. [22] defines a WordNet-based ontology offering a manageable set of concept classes along with integration of automated class recognition algorithm. Suchanek, F. M. et al. [23] presents a large Ontology-YOGO by extracting facts from Wikipedia's category system and infoboxes and subsequently combining with WordNet's taxonomic relations.

C. Measures for Similarity and relatedness of Concepts

The semantic similarity quantify the similarity between two concepts(terms) based on i)closeness of concepts in the taxonomical hierarchy, ii) sharing of information between concepts and iii) properties of concepts.[24,25] On the other hand, semantic relatedness measures are based on how two concepts are related. Similar concepts are semantically related by some kind of relationships whereas dissimilar entities are also related by some other kind of relationships.[7] The number of intervening terms decides the similarity between concepts. Semantically closer terms are close neighbors in the taxonomical network having fairly small distance measure whereas terms that are not conceptually closer have larger distance measure. Closeness of concepts consider the depth of the two given words along with the depth of the LCS (least common subsumes) while computing the similarity (distance) measure.[26]

Degree of similarity between terms (words) in a knowledge-based similarity is identified using the information derived from semantic network. Semantic similarity is a special type of relatedness measure where degree of concepts relationships are measured based on their location within an IS-A hierarchy. Since concepts are related in many ways apart from being similar, WordNet also supports has-part, is-made-of and is-an-attribute-of relations [27].Similarity measures are classified into following categories:

Path Based: consider the number of nodes along the shortest path between senses in the 'IS-A' hierarchies of WordNet for computing semantic relatedness. Distance measure was introduced by Rada et al.[28] which is used for finding closeness between two terms as defined in Eq.(1)

$$\mathrm{path(T1,T2)}=\frac{1}{\mathrm{shortest}(\mathrm{IS-A}(\mathrm{path}(\mathrm{T1},\mathrm{T2})))} \qquad (1)$$

Path and Depth based: Wu and Palmer[29], considered LCS as the most specific ancestor shared by terms and incorporated depth of LCS for computing semantic relatedness as defined in Eq.(2)

$$\mathrm{wup}(\mathrm{T1},\mathrm{T2})=\frac{2*(\mathrm{depth}(\mathrm{LCS}(\mathrm{T1},\mathrm{T2})))}{\mathrm{depth}(\mathrm{T1})+\mathrm{depth}(\mathrm{T2})} \qquad (2)$$

Where depth (T) is shortest IS-A path (root,T)
Leacock and Chodorow [30] proposed related measure by incorporating the depth of taxonomy as defined in Eq.(3)

$$\mathrm{lch}(\mathrm{T1},\mathrm{T2})=-\log\frac{\mathrm{slength}(\mathrm{T1},\mathrm{T2})}{2*\mathrm{max-depth}} \qquad (3)$$

Where slength is the length of the shortest path between the terms and max-depth is the maximum depth of the taxonomy.

Path and Information Content based: consider the amount of information content shared by terms along with path, where Information Content(IC) is defined as negative logarithm of the probability of terms as defined in Eq.(4)

$$\mathrm{IC}(\mathrm{Term})=-\log(\mathrm{p}(\mathrm{Term})) \qquad (4)$$

Resnik [31] incorporated IC for measuring semantic relatedness as defined in Eq. (5)

$$\mathrm{res}(\mathrm{T1},\mathrm{T2})=\mathrm{IC}(\mathrm{LCS}(\mathrm{T1},\mathrm{T2})) \qquad (5)$$

Jiang and Conrath[32] proposed to use IC for measuring distance between terms and relatedness is computed by taking the reciprocal of the distance as defined in the Eq.(6)

$$\mathrm{jcn}(\mathrm{T1},\mathrm{T2})=\frac{1}{\mathrm{jcn-distance}(\mathrm{T1},\mathrm{T2})}$$

$$\mathrm{jcn-distance}(\mathrm{T1},\mathrm{T2})=\mathrm{IC}(\mathrm{T1})+\mathrm{IC}(\mathrm{T2})-2*(\mathrm{IC}(\mathrm{LCS}(\mathrm{T1},\mathrm{T2}))) \qquad (6)$$

Lin[33] also incorporated IC both for LCS and individual terms and defined relatedness as given in Eq.(7)

$$\mathbf{lin}(\mathbf{T1},\mathbf{T2}) = \frac{\mathbf{2*IC(LCS(T1,T2))}}{\mathbf{IC(T1)+IC(T2)}} \qquad (7)$$

Relatedness between concepts are determined by considering the overlapping between their definitions, where overlap is the longest sequence of one or more consecutive terms.

Lesk [34] measured relatedness on the basis of extent of overlapping in the dictionary definitions. This method is further extended [35] with the addition of related words' definition for finding overlapping between words where WordNet is used as dictionary for retrieving word definition.

Hso [36] proposed relatedness as a path based measure where relations in WordNet are classified as having direction. For eg, IS-A relation is considered as having upward direction. Concepts are semantically closer if their WordNet synsets are connected by a path which is not too long and change of direction is not frequent. Relatedness is defined as given in Eq. (8)

$$\mathbf{hso}(\mathbf{T1},\mathbf{T2}) = \mathbf{C} - (\mathbf{path\_length}(\mathbf{T1},\mathbf{T2})) - \mathbf{K} * \mathbf{NCD}(\mathbf{T1},\mathbf{T2}) \qquad (8)$$

Where C, K are constants derived through experiments and NCD is the number of changes of direction in the path that the two terms T1 and T2.

There are other categories of relatedness measures in literature that include Feature-Based measure and Hybrid measure. Methods used under Feature-Based category consider relatedness between terms as a function of their properties i.e their definition or glosses. On the other hand Hybrid category methods combine the characteristics of the above discussed methods i.e Path based, Path and Depth based, Path and Information Content based [17].

It is observed that no single measure proved to be best however combination of various measures provide complementary results [37].

D. Measures of Ontology Evaluation

Hlomani, H. and Stacey [38] propose various matrices for evaluating correctness and quality of ontology. Measures proposed for correctness are accuracy, completeness(coverage), conciseness and consistency whereas for quality, measures are computational efficiency, adaptability and clarity. Completeness is described as measures for evaluating whether the domain of interest is aptly covered or not. Coverage is proposed for reflecting how well the ontology models the intended domain. Guarino [39] defined ontology as an approximate specification of particular domain and thus evaluation should use metrics reflecting the degree of such approximate specification. The author has also proposed coverage as evaluation metrics which is equated with precision and recall. Spiliopoulou, M et al. [40] propose appropriateness criteria referring to enhancement of ontology for a particular domain incorporating concepts that are appropriate either as individual or as group.

3. Enhancement of Ontology

Choosing the Ontology language is the first decision that needs to be taken while building an ontology. Most of the languages used, are based on eXtensible Markup Language (XML) enabling knowledge to be machine interpretable [41, 42]. Most frequently used languages are the Resource Description Framework (RDF), RDF Schema and the Ontology Web Language (OWL) [43, 6]. In the Semantic Web Layer technologies, RDF and OWL hold two separate layers. RDF specification tells us how to write a triple whereas OWL specification defines what should be written with RDF in order to have a valid ontology. Every OWL can be serialized to RDF/XML format.

A Dengue Patient Ontology capturing concepts like- Patients, Treatment, Symptoms, Diagnosis etc is considered as seed ontology to be enhanced. The taxonomy of concepts in the ontology can be defined as a pair

O:=(C,Δ), where C is a finite set of concepts

Δ is an order relation on C×C i.e $\Delta \subseteq C\times C$

We introduce ontology enhancement as an approach that incorporate instantial Knowledge Base containing dengue patients triples extracted from unstructured case sheets. Let T be the text corpus of patients case sheets and R (T) be the set of triples extracted from T. An ontology enhancement approach is a procedure which takes given triples R(T) corresponding to patient, an ontology taxonomy O and Lexical Database LD as input .The procedure incorporates R(patient)⊆R(T) into O resulting schema updation in the ontology and defines a semantic association mapping f such that :

$$\mathbf{f: Concept \rightarrow Term}$$

for Concept∈ {C,R} where R⊆Δ and Term∈synsets from LD.

A. Accommodating variations in the Ontology structure

RDF is considered as self-describing considering the fact that no separate schema is required for describing the data. However, schema definition is possible which can be defined later based on already existing instances(posteriori). The strength of RDF model is its ability to accommodate variations in the structure of data. For instance, schema may be updated and subsequently instance-related assertions may be added as per variations in the schema. In case of updation in the schema also, RDF retains the previous instances. Fig 2 illustrates the same schema as in Fig 1 with the difference that the instance model has an additional data property which was not defined in the original ontology.

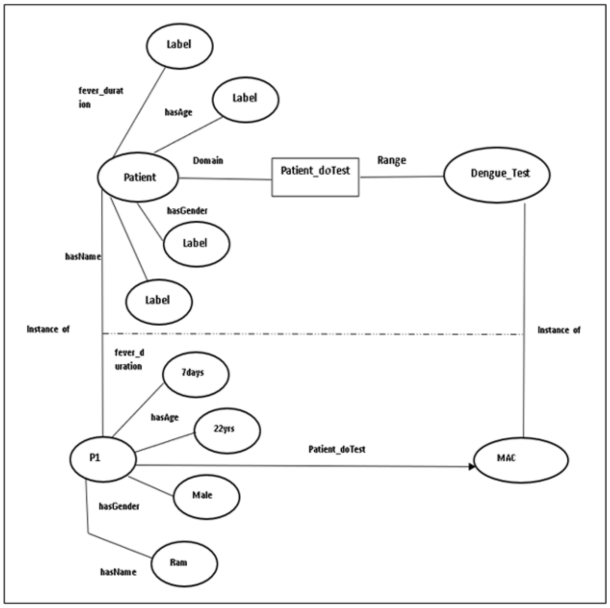


Figure 1. RDF Schema

As shown in RDF Schema of Fig.1 Patient has data property- hasName, hasAge, hasGender and fever_duration. However, in Fig.2 an enhanced schema is presented which is updated posterori as a result of new patient instance having data property suffering.

B. Enhancing Dengue Ontology by integrating Instantial Knowledge base

The dynamic nature of knowledge is addressed by following the Ontology Enhancement Module as demonstrated in Fig.3.The strength of RDF model as discussed in the earlier section is utilized in the Enhancement Module for accommodating structural variations. Our enhancement is an iterative process on one or more Dengue Patients' case sheets in the form RDF model. Patients information are incorporated into the initial Dengue knowledge base by including new triples as instance of Patients. Data property along with its values are thus reflected in the ontology based on each triple of the RDF model contributing to the enrichment process. Ontology is associated with linguistic knowledge base-WordNet, where synsets are found and semantic relatedness is computed between concepts and lemmas constituting the synsets.

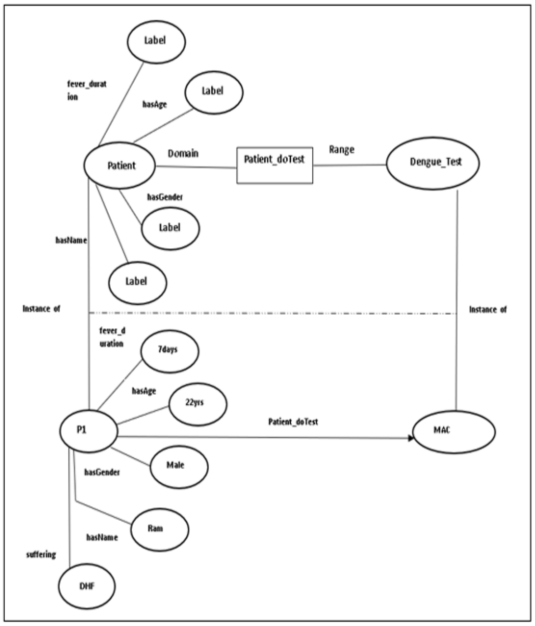


Figure 2. RDF Schema with additional property

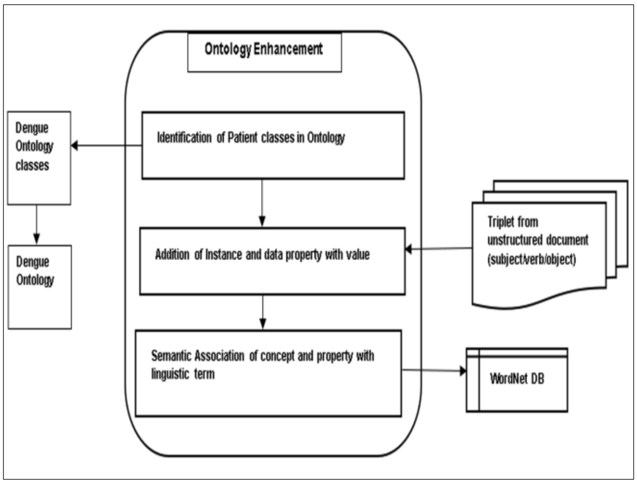

Ontology Enhancement
Dengue Ontology classes
Identification of Patient classes in Ontology
Dengue Ontology
Addition of Instance and data property with value
Triplet from unstructured document (subject/verb/object)
Semantic Association of concept and property with linguistic term
WordNet DB

Figure 3. Ontology Enhancement Module

A Dengue Ontology developed using Dengue Patients records of Swami Dayanand Hospital, C-Block, Dilshad Garden, Patparganj Industrial Area, Delhi-110095[46], is considered in our case study for enhancement which comprises Dengue_fever_Test, Dengue_fever_symptom, Diagnosis, Likely_To_Suffer,Patient, Symptom and Treatment as root level concepts.

Apart from various object properties between concepts Data property of Patient concept are fever_duration, hasAge, hasGender,hasName. Instantial Knowledge Base in RDF format is used that contains the information in the form of triplet extracted from unstructured text of patients' case sheets. This knowledge base provide various data property and their corresponding value of patients which is accommodated in existing ontology using the Onto_Integration algorithm as shown in Algorithm 1.

Algorithm 1: Ontology_Integration

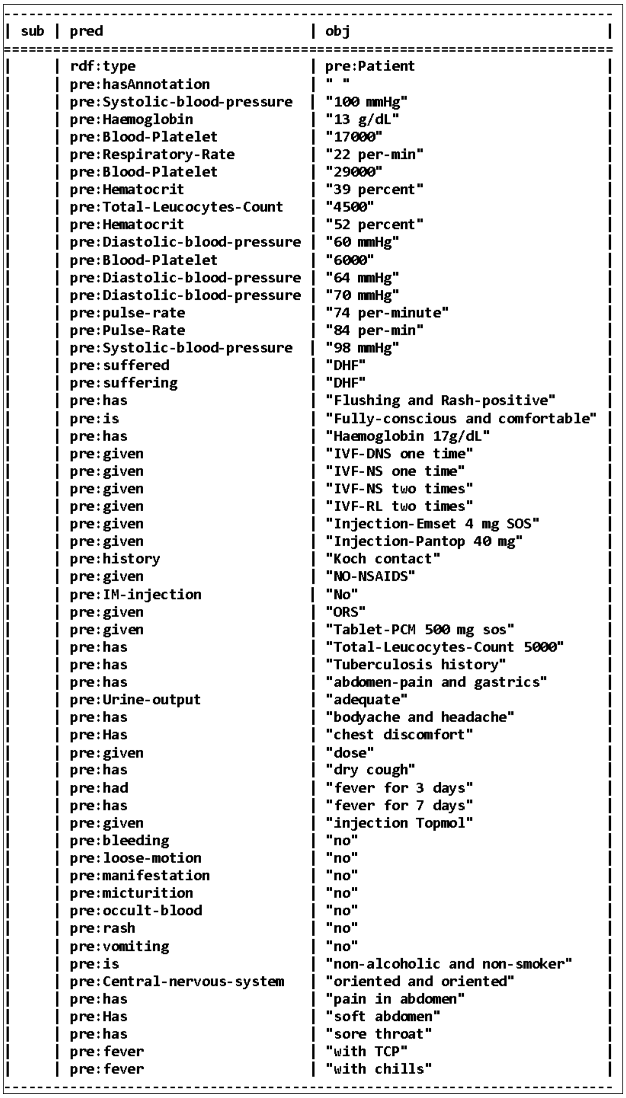
```
Input: subject, predicate, object from each triple of Patient RDF
knowledge base
OntModel : model
Output: OntModel incorporated with Data Property and
individual Patient
DatatypeProperty
pr=model.createDatatypeProperty(namespace+predicate);
ExtendedIterator it = ((OntModel) model).listClasses();
while it.hasNext() exists
        OntClass cls= (OntClass)it.next();
        if(cls.getLocalName().toString().contains("Patient"))
              pr.addDomain(cls);
              pr.addRange(XSD.xstring);
              Individual
              in1=cls.createIndividual(namespace+subject);
              in1.addProperty(pr,
              model.createTypedLiteral(object));
```

The algorithm traverses the ontology for Patient concept and subsequently data property is included based on the predicates from the extracted triples along with the data values as Literal. The Domain and Range of the data property is considered as Patient and String respectively. New patient individual is incorporated as instance of Patient with data property and its corresponding value.

A snapshot of SPARQL query results showing data property (pred) and value (obj) after incorporating patient's case sheets details is illustrated in the Fig 4. Thus existing ontology is enriched with the inclusion of more knowledge in the form of patients data property and its value while updating the RDF Schema and retaining the existing knowledge.

C. *Semantic association of ontological concepts with Linguistic counter part*

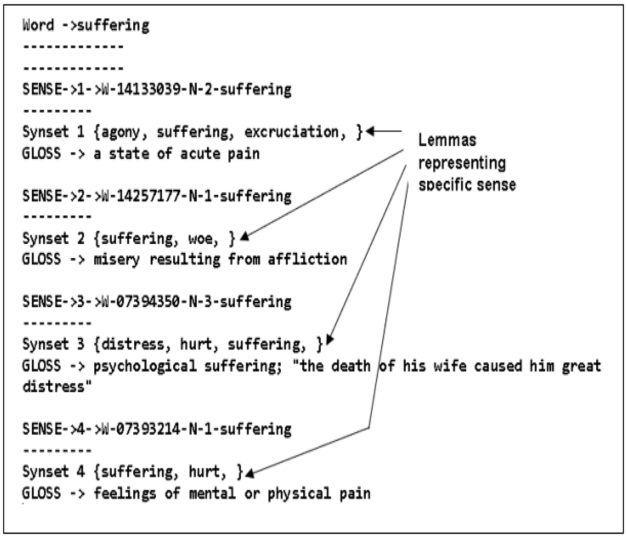
Semantic association can be considered as a mapping that connects resources from two different KB. In this study, concepts, individuals and properties, both object and data properties of the integrated ontology, have been selected to be enriched with information from WordNet as it is observed that they are informative. The association maps the semantics of the concepts, properties (data, object) and instances with a specific WordNet concepts based on their relatedness. The semantic association selects the synsets for concepts, individuals as well as properties (object, data) that better expresses the meaning [44]. Any ontological concept C (either class or property) is associated with either a singleton or with a multiword synsets S by a linguistic label L. Linguistic senses for concept C corresponding to the label L may or may not exist depending on the presence or absence technical term in the linguistic knowledge base. Finding Synonym and computing similarities are performed for each concepts both root level as well as subclasses along with their instances, data properties and object properties. As discussed in the Ontology enhancement module discussed in section 3, concepts are retrieved from the integrated ontology and stored in respective ArrayList for semantic association. Stemming, a heuristic process for removing derivational affixes, is performed only on data property



```
-----------------------------------------------------------------------
| sub | pred                           | obj                                 |
=======================================================================
|     | rdf:type                       | pre:Patient                         |
|     | pre:hasAnnotation              | " "                                 |
|     | pre:Systolic-blood-pressure    | "100 mmHg"                          |
|     | pre:Haemoglobin                | "13 g/dL"                           |
|     | pre:Blood-Platelet             | "17000"                             |
|     | pre:Respiratory-Rate           | "22 per-min"                        |
|     | pre:Blood-Platelet             | "29000"                             |
|     | pre:Hematocrit                 | "39 percent"                        |
|     | pre:Total-Leucocytes-Count     | "4500"                              |
|     | pre:Hematocrit                 | "52 percent"                        |
|     | pre:Diastolic-blood-pressure   | "60 mmHg"                           |
|     | pre:Blood-Platelet             | "6000"                              |
|     | pre:Diastolic-blood-pressure   | "64 mmHg"                           |
|     | pre:Diastolic-blood-pressure   | "70 mmHg"                           |
|     | pre:pulse-rate                 | "74 per-minute"                     |
|     | pre:Pulse-Rate                 | "84 per-min"                        |
|     | pre:Systolic-blood-pressure    | "98 mmHg"                           |
|     | pre:suffered                   | "DHF"                               |
|     | pre:suffering                  | "DHF"                               |
|     | pre:has                        | "Flushing and Rash-positive"        |
|     | pre:is                         | "Fully-conscious and comfortable"   |
|     | pre:has                        | "Haemoglobin 17g/dL"                |
|     | pre:given                      | "IVF-DNS one time"                  |
|     | pre:given                      | "IVF-NS one time"                   |
|     | pre:given                      | "IVF-NS two times"                  |
|     | pre:given                      | "IVF-RL two times"                  |
|     | pre:given                      | "Injection-Emset 4 mg SOS"          |
|     | pre:given                      | "Injection-Pantop 40 mg"            |
|     | pre:history                    | "Koch contact"                      |
|     | pre:given                      | "NO-NSAIDS"                         |
|     | pre:IM-injection               | "No"                                |
|     | pre:given                      | "ORS"                               |
|     | pre:given                      | "Tablet-PCM 500 mg sos"             |
|     | pre:has                        | "Total-Leucocytes-Count 5000"       |
|     | pre:has                        | "Tuberculosis history"              |
|     | pre:has                        | "abdomen-pain and gastrics"         |
|     | pre:Urine-output               | "adequate"                          |
|     | pre:has                        | "bodyache and headache"             |
|     | pre:Has                        | "chest discomfort"                  |
|     | pre:given                      | "dose"                              |
|     | pre:has                        | "dry cough"                         |
|     | pre:had                        | "fever for 3 days"                  |
|     | pre:has                        | "fever for 7 days"                  |
|     | pre:given                      | "injection Topmol"                  |
|     | pre:bleeding                   | "no"                                |
|     | pre:loose-motion               | "no"                                |
|     | pre:manifestation              | "no"                                |
|     | pre:micturition                | "no"                                |
|     | pre:occult-blood               | "no"                                |
|     | pre:rash                       | "no"                                |
|     | pre:vomiting                   | "no"                                |
|     | pre:is                         | "non-alcoholic and non-smoker"      |
|     | pre:Central-nervous-system     | "oriented and oriented"             |
|     | pre:has                        | "pain in abdomen"                   |
|     | pre:Has                        | "soft abdomen"                      |
|     | pre:has                        | "sore throat"                       |
|     | pre:fever                      | "with TCP"                          |
|     | pre:fever                      | "with chills"                       |
-----------------------------------------------------------------------
```

Figure 4. Data Property and value

concepts using Snowball Stemmer algorithm as class and individual concepts are already in their base form [45]. Synonym sets (Synsets) and their corresponding wordforms are retrieved for each type of concepts including class, individual, data property and object property using WordNet Linguistic Database as shown in the algorithm 2. NOUN POS of WordNet is used for retrieving Synsets that contains one or more lemmas representing specific sense of specific word. As illustrated in the Fig.5, each synsets comprising multiple lemmas for the concept “suffering” represent same sense or meaning.



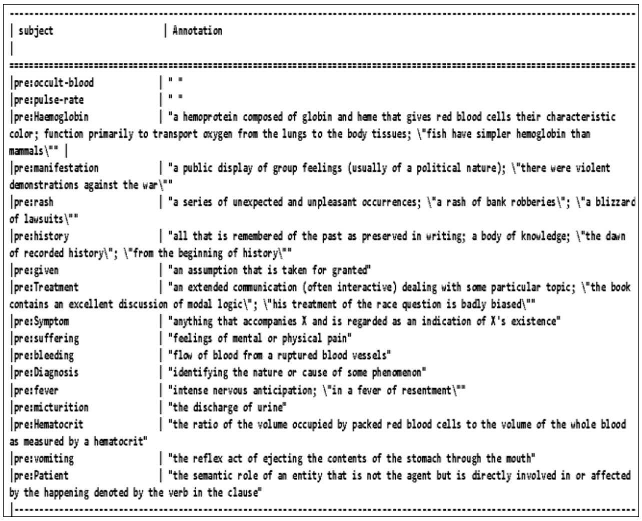
```
Word ->suffering
-------------
-------------
SENSE->1->W-14133039-N-2-suffering
---------
Synset 1 {agony, suffering, excruciation, }
GLOSS -> a state of acute pain

SENSE->2->W-14257177-N-1-suffering
---------
Synset 2 {suffering, woe, }
GLOSS -> misery resulting from affliction

SENSE->3->W-07394350-N-3-suffering
---------
Synset 3 {distress, hurt, suffering, }
GLOSS -> psychological suffering; "the death of his wife caused him great
distress"

SENSE->4->W-07393214-N-1-suffering
---------
Synset 4 {suffering, hurt, }
GLOSS -> feelings of mental or physical pain
```

Figure 5. Synsets with respective Gloss



```
--------------------------------------------------------------------------------------------------
| subject                    | Annotation
|
==================================================================================================
|pre:occult-blood            | " "
|pre:pulse-rate              | " "
|pre:Haemoglobin             | "a hemoprotein composed of globin and heme that gives red blood cells their characteristic
color; function primarily to transport oxygen from the lungs to the body tissues; \"fish have simpler hemoglobin than
mammals\"" |
|pre:manifestation           | "a public display of group feelings (usually of a political nature); \"there were violent
demonstrations against the war\""
|pre:rash                    | "a series of unexpected and unpleasant occurrences; \"a rash of bank robberies\"; \"a blizzard
of lawsuits\""
|pre:history                 | "all that is remembered of the past as preserved in writing; a body of knowledge; \"the dawn
of recorded history\"; \"from the beginning of history\""
|pre:given                   | "an assumption that is taken for granted"
|pre:Treatment               | "an extended communication (often interactive) dealing with some particular topic; \"the book
contains an excellent discussion of modal logic\"; \"his treatment of the race question is badly biased\""
|pre:Symptom                 | "anything that accompanies X and is regarded as an indication of X's existence"
|pre:suffering               | "feelings of mental or physical pain"
|pre:bleeding                | "flow of blood from a ruptured blood vessels"
|pre:Diagnosis               | "identifying the nature or cause of some phenomenon"
|pre:fever                   | "intense nervous anticipation; \"in a fever of resentment\""
|pre:micturition             | "the discharge of urine"
|pre:Hematocrit              | "the ratio of the volume occupied by packed red blood cells to the volume of the whole blood
as measured by a hematocrit"
|pre:vomiting                | "the reflex act of ejecting the contents of the stomach through the mouth"
|pre:Patient                 | "the semantic role of an entity that is not the agent but is directly involved in or affected
by the happening denoted by the verb in the clause"
|-------------------------------------------------------------------------------------------------
```

Figure 6. Gloss as Annotation property

Gloss associated with each synset for each concept of the enhanced knowledge base is also retrieved using Wordnet. Similarity measures are computed between ontology Concepts (both class and data properties types) and lemmas of corresponding WordNet synsets. For a particular ontology concepts, lemma that results highest similarity value are considered as the most related linguistic term.

Gloss of the highest related linguistic term is associated as Annotation property for the ontology concepts. A sample SPARQL result showing Gloss association as Annotation property to concepts is presented in Fig.6.

The Annotation property is included only for those ontology concepts (class type and data property type) whose synsets are available in the Linguistic KB-WordNet where subject refer to concept and Annotation refer to the Gloss of the most related synset.

Algorithm 2: Ontology Concepts Extraction

```
Input: Ontology model, WordNet database

Method:
Declare class as ArrayList[String]
Declare instance as ArrayList[String]
Declare dprop as ArrayList[String]
Declare oprop as ArrayList[String]

ExtendedIterator[OntClass]
it←model.listHierarchyRootClasses();
While it.hasNext() exists
   OntClass cls← (OntClass)it.next();
   append cls to class
    if cls.hasSubClass() exists
   OntClass clasub ←model.getOntClass(URI+sclass);
    for Iterator i← clasub.listSubClasses() to i.hasNext()
             OntClass c ← (OntClass) i.next();
             append c to class
             ExtendedIterator subinstance←c.listInstances();
             While subinstance.hasNext() exists
             Individual subinstance1←
                          (Individual)subinstance.next();
             append subinstance1 to instance;
             end while
    end for
   end if
   ExtendedIterator pinstance ← ((OntClass)cls).listInstances();
    While pinstance.hasNext() exists
     Individual pinstance1← (Individual)pinstance.next();
     append pinstance1 to instance;
     ExtendedIterator dp ←
     ((OntModel)model).listDatatypeProperties();
     While dp.hasNext() exists
          DatatypeProperty p ← (DatatypeProperty) dp.next();
           if (p.isDatatypeProperty() && p.getDomain()!=null
           &&p.getRange()!=null)
           append p to dprop;
     end while
     ExtendedIterator objp ←
     ((OntModel) model).listObjectProperties();
      While objp.hasNext() exists
       ObjectProperty o← (ObjectProperty)objp.next();
```

```
        if (o.isObjectProperty() && o.getDomain()!=null
         &&o.getRange()!=null)
          append o to oprop;
        end while
    end while
end while
SYNONYM_FIND(class,database);
SYNONYM_FIND(instance,database);
SYNONYM_FIND(dprop,database);
SYNONYM_FIND(oprop,database);
```

Synonyms retrieved from the respective synsets of WordNet where ontology concepts are one of the lemmas of the synsets are illustrated in the procedure of algorithm 3

Algorithm 3: Synonym_Find

```
SYNONYM_FIND (ArryList al[String], WordNet database)

Input: ArrayList al for each of Class, Instance, Dataproperty
           and Objectproperty.
           Lexical database WordNet

Method:
for i←0 to al[size]
   Synset[] synsets = database.getSynsets(al.get(i));// synonym
sets
   if (synsets.length > 0)
      create ArrayList wf[String]
                  for j ←0 to synsets[length]
                          String[] wordForms =
                          synsets[j].getWordForms();// set of words
                          For k ←0 to wordForms[length]
                             append wordForms[k] to wf
                          end
                 end
      end
end
```

4. Evaluation-Semantic Similarity and Ontology Coverage:

Three(03) different methods-path, lch, lin refered in Eq.(1), Eq.(3) and Eq.(7) based on aforementioned criteria of measuring relatedness are used for finding similarity of concepts of various types such as class, instance, data property and object type with their WordNet counterpart which is presented in Table 1.On the other hand, depending on the presence of technical term, synsets may or may not be present in the linguistic knowledge base for particular ontology concept.

Each of the three methods measures relatedness based on path, path and depth and information count. Our objective of combining the measures is to capitalize their strength so that semantic association provides complementary information for quantifying the degree of relatedness. Measures are combined by normalizing the individual score using z-score normalization so that they are placed on the same scale. Fifty six (56) distinct concept-term pairs that include concepts of each type are selected for finding association with their linguistic counterpart.

Normalization is a technique used for scaling where a new range is used from existing one range. So to maintain large variation in scale of relatedness measures, Z-score normalization technique is used for relatedness value computed by each of the three methods which is defined in Eq (9)[47]

$$\mathbf{Z-score} = \frac{\mathbf{x-mean}}{\mathbf{stddev}} \qquad (9)$$

Z-score normalized relatedness values of 03(three) complementary methods are presented in Table 1.

Correlation of different relatedness measures w.r.t each other for the respective concept-term pairs showing similar property by each method is presented in Table 2.

The proposed ontology enhancement and semantic association is an iterative process on one or more patient's case sheets information in the form of RDF model. Its initial input is a seed ontology to be enhanced and the

TABLE 1. SEMANTIC ASSOCIATION OF ONTOLOGY AND WORDNET

Concept-Term Pairs	lch-score	lin-score	Path-score
Pair-1	0.57	0.58	0.58
Pair-2	0.56	0.58	0.58
Pair-3	0.56	0.58	0.58
Pair-4	0.56	0.58	0.58
Pair-5	0.56	0.58	0.58
Pair-6	0.56	0.58	0.58
Pair-7	0.56	0.58	0.58
Pair-8	0.56	0.58	0.58
Pair-9	-1.81	-1.73	-1.74
Pair-10	-1.81	-1.73	-1.74
Pair-11	0.56	0.58	0.58
Pair-12	0.56	0.58	0.58
Pair-13	0.56	0.58	0.58
Pair-14	0.56	0.58	0.58
Pair-15	0.56	0.58	0.58
Pair-16	0.56	0.58	0.58
Pair-17	0.56	0.58	0.58
Pair-18	0.56	0.58	0.58
Pair-19	0.56	0.58	0.58

Concept-Term Pairs	lch-score	lin-score	Path-score
Pair-20	-1.81	-1.73	-1.74
Pair-21	0.56	0.58	0.58
Pair-22	0.56	0.58	0.58
Pair-23	0.56	0.58	0.58
Pair-24	0.56	0.58	0.58
Pair-25	0.56	0.58	0.58
Pair-26	0.56	0.58	0.58
Pair-27	-1.81	-1.73	-1.74
Pair-28	0.56	0.58	0.58
Pair-29	0.56	0.58	0.58
Pair-30	0.56	0.58	0.58
Pair-31	0.56	0.58	0.58
Pair-32	-1.81	-1.73	-1.74
Pair-33	-1.81	-1.73	-1.74
Pair-34	-1.29	-1.61	-1.61
Pair-35	-1.81	-1.73	-1.74
Pair-36	-0.91	-1.50	-1.51
Pair-37	0.56	0.58	0.58
Pair-38	0.56	0.58	0.58
Pair-39	0.56	0.58	0.58
Pair-40	0.56	0.58	0.58
Pair-41	0.56	0.58	0.58
Pair-42	0.56	0.58	0.58
Pair-43	0.56	0.58	0.58
Pair-44	0.56	0.58	0.58
Pair-45	-1.81	-1.73	-1.74
Pair-46	0.56	0.58	0.58
Pair-47	0.56	0.58	0.58
Pair-48	0.56	0.58	0.58
Pair-49	0.56	0.58	0.58
Pair-50	0.56	0.58	0.58
Pair-51	0.56	0.58	0.58
Pair-52	-1.81	-1.73	-1.74
Pair-53	0.56	0.58	0.58
Pair-54	-1.81	-1.73	-1.74
Pair-55	-1.81	-1.73	-1.74
Pair-56	-1.81	-1.73	-1.74

TABLE 2. CORRELATION OF RELATEDNESS MEASURES FOR CONCEPT-TERM PAIR SCORE

Correlation	lch-score	lin-score
lch-score	1	
lin-score	0.9953822	1

Correlation	lin-score	Path-score
lin-score	1	
Path-score	1	1

Correlation	lch-score	Path-score
lch-score	1	
Path-score	0.9953822	1

final one should be an enriched ontology which is complete towards integration of instantial KB in association with linguistic KB. Enhanced ontology is evaluated based on the completeness perspective which measures the amount of domain of interest being conceptualized. Four random case sheets in the form of RDF model are used for instantial integration and semantic association.

Some of the ontology metrics used for evaluation are as below:
Number of Classes
Number of Instances
Number of Data property
Number of Object Property
Number of Annotation Property

TABLE 3. ONTOLOGY METRIC VALUE

Metric	Case-sheet1	Case-sheet2	Case-sheet3	Case-sheet4
Classes	10	10	10	10
Instances	178	178	178	178
Data Property	20	19	18	32
Object Property	8	8	8	8
Annotation Property	10	9	8	15
Total	226	224	222	243

Before enhancement Number of classes, Instances, Data property and Object property were 10, 177, 4 and 8 respectively. The extent of concept coverage after integration and semantic association on the ontology based on 04(four) case sheets triples has increased number of data property and annotation property as presented in Table 3.

The completeness of the enhanced ontology on the basis of concept coverage i.e inclusion of relevant concept towards conceptualization of intended domain w.r.t data property and annotation property is computed for each case sheet and is given in Fig. 7

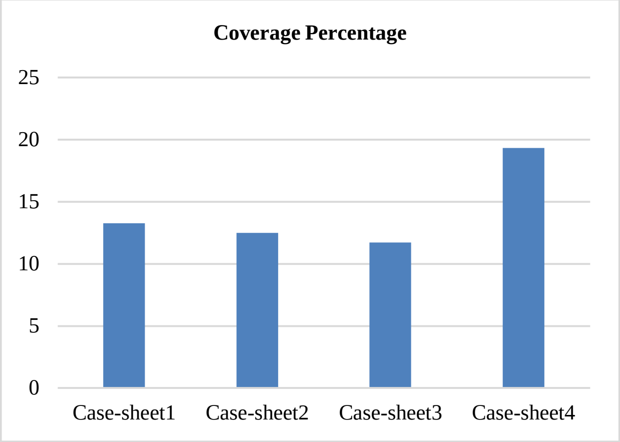


Figure 7. Case-sheet wise concept coverage

5. Conclusion

The discussed approach provides a hybrid network by enhancing a knowledge base accommodated with new instances, properties and semantically associating ontology with linguistic knowledge base-WordNet, that will facilitate sentence disambiguation apart from information retrieval. On the other hand, interpreting the meaning of concepts and their properties is really feasible only in presence of association of ontology to textual units. Semantic association of ontology and the linguistic knowledge base-WordNet is done by retrieving synonym sets from WordNet for various types of ontology concepts including class, instance and property type. The extent of relatedness of ontology concept and synonym set members are computed and placed on the same scale by normalizing the relatedness score. Lemmas with highest relatedness values are considered as most similar and respective gloss is associated as annotation property with the concept. Three(03) different relatedness measures are applied which are complementary to each other :

i) lch takes into consideration the depths of two synsets in the taxonomies along with the depth of LCS while measuring relatedness

ii) Lin consider information content for computing relatedness between concepts. iii) Path count the number of nodes along the shortest between word senses in the IS-A hierarchies for measuring relatedness. Moreover, the three methods are also analyzed on the basis of their z-score which showed positive correlation between lch-score and lin-score, lin-score and Path, lch-score and Path-score. i.e if one method is measuring concept-term pairs are related then other two complementary methods also measure them as related. Further study can be performed on extending ontology association with other computer readable information sources like thesauri, online dictionaries etc instead of WordNet and can also be scaled to include other languages.

References

[1] Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches.

[2] Onyshkevych, B. A., & Nirenburg, S. (1991, June). Lexicon, ontology, and text meaning. In Workshop of SIGLEX (Special Interest Group within ACL on the Lexicon) (pp. 289-303). Springer, Berlin, Heidelberg

[3] Nováček, V., Laera, L., Handschuh, S., & Davis, B. (2008). Infrastructure for dynamic knowledge integration—Automated biomedical ontology extension using textual resources. Journal of biomedical informatics, 41(5), 816-828.

[4] Pazienza, M. T., & Stellato, A. (2006, May). Linguistic Enrichment of Ontologies: a methodological framework. In Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006).

[5] Basili, R., Vindigni, M., & Zanzotto, F. M. (2003, October). Integrating ontological and linguistic knowledge for conceptual information extraction. In Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on (pp. 175-181). IEEE.

[6] Warin, M., Oxhammar, H., & Volk, M. (2005). M.: Enriching an ontology with wordnet based on similarity measures. In In: MEANING-2005 Workshop.

[7] Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. Database, 2018(1), bay101.

[8] Smith, B., & Fellbaum, C. (2004, August). Medical WordNet: a new methodology for the construction and validation of information resources for consumer health. In Proceedings of the 20th international conference on Computational Linguistics (p. 371). Association for Computational Linguistics.

[9] Toumouth, A., Lehireche, A., Widdows, D., & Malki, M. (2006, March). Adapting WordNet to the Medical Domain using Lexicosyntactic Patterns in the Ohsumed Corpus. In Computer Systems and Applications, 2006. IEEE International Conference on. (pp. 1029-1036). IEEE.

[10] Lin, F., & Sandkuhl, K. (2008, September). A survey of exploiting wordnet in ontology matching. In IFIP International Conference on Artificial Intelligence in Theory and Practice (pp. 341-350). Springer, Boston, MA.

[11] Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health

[12] W3C, OWL 2 web ontology language, world wide web consortium(W3C), 2017

[13] Dzikovska, M. O., Allen, J. F., & Swift, M. D. (2003, August). Integrating linguistic and domain knowledge for spoken dialogue systems in multiple domains. In Proc. of IJCAI-03 Workshop on Knowledge and Reasoning in Practical Dialogue Systems.

[14] Salahli, M. A. (2009). An approach for measuring semantic relatedness between words via related terms. Mathematical and Computational Applications, 14(1), 55-63.

[15] Montoyo, A., Suárez, A., Rigau, G., & Palomar, M. (2005). Combining knowledge-and corpus-based word-sense-disambiguation methods. Journal of Artificial Intelligence Research, 23, 299-330.

[16] Burgun, A., & Bodenreider, O. (2001, June). Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In Proceedings of the NAACL’2001 Workshop,“WordNet and Other Lexical Resources: Applications, Extensions and Customizations (pp. 77-82).

[17] Slimani, T. (2013). Description and evaluation of semantic similarity measures approaches. arXiv preprint arXiv:1310.8059.

[18] Bhatt, B., & Bhattacharyya, P. (2011, December). IndoWordNet and its linking with ontology. In Proceedings of the 9th International Conference on Natural Language Processing (ICON-2011)

[19] Niles, I., & Pease, A. (2001, October). Towards a standard upper ontology. In Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001 (pp. 2-9). ACM.

[20] Koeva, S., Dimitrova, T., Stefanova, V., & Hristov, D. Mapping WordNet Concepts with CPA Ontology.

[21] William D. Lewis. (2002, January). Measuring Conceptual Distance Using WordNet: the design of a metric for measuring semantic similarity. Working papers in Linguistics, Language in Cognitive Science.

[22] Sanfilippo, A. P., Tratz, S. C., Gregory, M. L., Chappell, A. R., Whitney, P. D., Posse, C., ... & White, A. M. (2006). Ontological annotation with wordnet (No. PNNL-SA-54089). Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

[23] Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. Web Semantics: Science, Services and Agents on the World Wide Web, 6(3), 203-217.

[24] Hussain, A. (2012). Textual Similarity. Bachelor Thesis. Kongens Lyngby: University of Denmark.

[25] Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G., & Milios, E. E. (2005, November). Semantic similarity methods in wordNet and their application to information retrieval on the web. In Proceedings of the 7th annual ACM international workshop on Web information and data management (pp. 10-16). ACM.

[26] Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. Expert Systems with Applications, 42(4), 2264-2275.

[27] Meng, L., Huang, R., & Gu, J. (2014). Measuring semantic similarity of word pairs using path and information content. Int. J. Future Gener. Commun. Netw, 7(3), 183-194.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[28] Rada R, Mili H. Bicknell E., Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybernet 1989:19(1):17-30.

[29] Wu Z, Palmer M. Verbs semantic and lexical selection. In:Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics, Las Cruces, NM:1994. P.133-8.

[30] Leacock C, Chodorow M., Combining local context and WordNet similarity for word sense identification. In:Fellbaum C, editor. WordNet: an electronic lexical database, Cambridge,MA:MIT Press:1998. P.265-83.

[31] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada:1995. P.448-53.

[32] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th Research on Computational Linguistics International Conference, Tapei, Taiwan:1997.p.19-33

[33] Lin D. An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning, San Francisco, CA:1998.p.296-304

[34] Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In:Proceedings of the 5th Annual International Conference on Systems Documentation, Toronto, Canada:1998.p. 24-6

[35] Banerjee, S., & Pedersen, T. (2002, February). An adapted Lesk algorithm for word sense disambiguation using WordNet. In *International conference on intelligent text processing and*

[36] Hirst and D. St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, WordNet: An electronic lexical database, pages 305–332. MIT Press.

[37] McInnes, B. T., & Pedersen, T. (2015). Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs. *Journal of biomedical informatics*, *54*, 329-336.

[38] Hlomani, H., & Stacey, D. (2014). Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semantic Web Journal*, *1*(5), 1-11

[39] N. Guarino. Toward a formal evaluation of ontology quality. IEEE intelligent Systems, 19(4):78–79, 2004.

[40] Spiliopoulou, M., Schaal, M., Müller, R. M., & Brunzel, M. (2005). Evaluation of ontology enhancement tools. In *Semantics, Web and Mining* (pp. 132-146). Springer, Berlin, Heidelberg.

[41] Munir, K., & Anjum, M. S. (2017). The use of ontologies for effective knowledge modelling and information retrieval. Applied Computing and Informatics.

[42] Harold, E. R. (2004). Effective XML: 50 specific ways to improve Your XML. Addison-Wesley Professional.

[43] G. Klyne, J. Carroll, Resource Description Framework(RDF): Concepts and Abstract Syntax, W3C Recommendation 10 February 2004, RDF Core Working Group,2004.

[44] Schmidt, D., Basso, R., Trojahn, C., & Vieira, R. Indirect Matching of Domain and Top-level Ontologies using OntoWordNet.

[45] Moral, C., de Antonio, A., Imbert, R., & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval. *Information Research: An International Electronic Journal*, *19*(1), n1.

[46] Swami Dayanand Hospital, Dilshad Garden Delhi110095, India, sdnhospital.edu.in/

[47] Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Runumi Devi., is a research scholar in Amity University, Noida, India and working as an assistant professor at JSS Academy Of Technical Education, Noida, U.P, India. She did her Masters in Computer Applications from Jorhat Engineering College, Assam, India and Master of Technology in Computer Science and Engineering from GGSIPU, Delhi, India. She has broad research interest in Data Mining, Design and Analysis of Algorithm and Knowledge Engineering.

Dr. Deepti Mehrotra did Ph.D. from Lucknow University and currently she is working as Professor in Amity school of Engineering and Technology, Amity University, Noida, earlier she worked as Director of Amity School of Computer Science, Noida, India.. She has more than 20 year of experience in research, teaching and content writing. She had published more than 100 papers in international refereed Journals and conference Proceedings. She is editor and reviewer for many books, referred journal and conferences. She is regularly invited as resource person for FDPs and invited talk in national and international conference. She enjoys guiding research scholar and has many Ph.D. and M.Tech students.