EXPLOITING SEMANTIC SIMILARITY FOR INFORMATION RETRIEVAL

A dissertation submitted to the Jawaharlal Nehru University in
partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE AND TECHNOLOGY

BY

JAGENDRA SINGH

SCHOOL OF COMPUTER AND SYSTEMS SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI – 110067
JULY 2012

जवाहरलाल नेहरू विश्वविद्यालय

School of Computer & Systems Sciences

JAWAHAR LAL NEHRU UNIVERSITY

NEW DELHI-110067, INDIA

DECLARATION

I hereby declare that the dissertation entitled "**EXPLOITING SEMANTIC SIMILARITY FOR INFORMATION RETRIEVAL**", Submitted by me to the School of Computer and Systems Sciences, **Jawaharlal Nehru University**, New Delhi, in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in **Computer Science and Technology**, is a bona fide work carried out by me under the supervision of **Dr. Aditi Sharan**.

The matter embodied in this dissertation has not been submitted to any other university or institution for the award of any other degree or diploma.

JAGENDRA SINGH

M.TECH

SC&SS, JNU,

New Delhi-110067

जवाहरलाल नेहरू विश्वविद्यालय

School of Computer & Systems Sciences

JAWAHAR LAL NEHRU UNIVERSITY

NEW DELHI-110067, INDIA

CERTIFICATE

This is to certify that the dissertation entitled "**EXPLOITING SEMANTIC SIMILARITY FOR INFORMATION RETRIEVAL**", Submitted by **Jagendra Singh** to the School of Computer and Systems Sciences, **Jawaharlal Nehru University**, New Delhi, in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in **Computer Science and Technology**, is a bona fide work carried out by him under my supervision.

The matter embodied in this dissertation has not been submitted to any other university or institution for the award of any other degree or diploma.

Dr. Aditi Sharan

(Supervisor)	(Dean, SC&SS,)
SC&SS, JNU, New Delhi	JNU, New Delhi

Acknowledgement

I would like to gratefully acknowledge the enthusiastic supervision of **Dr. Aditi Sharan** during this work. This work wouldn`t have been possible without her constant support, valuable suggestions and comments during my whole tenure of this dissertation work. I feel privileged to work under her for my master`s dissertation. Apart from the academic guidance she has always been a great mentor of mine in encouraging me to be disciplined and well organized. I must surely say, she has given her best in providing me the infrastructure required, which led to the successful completion of my dissertation. I would take this opportunity to thank her once again for her esteemed support and I, from the bottom of my heart would like to wish her the best in all her future endeavors.

I wish to thank my colleagues Mr. Mayank Saini, Mr. Nitin Prajapati and my senior Mrs. Manju for creating a home like environment in our lab to keep the stress away. I would also like to thank my best critics Nitin Prajapati, Md. Sajid and Vipin Kumar for suggesting to remove errors in my dissertation. Thank you guys!!

Finally I would like to thank the whole faculty member of our department for clarifying my doubts throughout this work and last but not least, the JNU administration for creating such a secular and healthy environment amongst the students.

Dedicated to all, who fight to fulfill their dreams…

Abstract

Semantic similarity techniques are used to compute the semantic similarity (common shared information) between two concepts according to certain language or domain resources like ontologies, taxonomies, corpora, etc. Semantic similarity techniques constitute important components in most Information Retrieval and knowledge based systems. This is an important problem in Natural Language Processing and Information Retrieval Research and has received considerable attention in the literature. Several algorithmic approaches for computing semantic similarity have been proposed. However semantic similarity is challenging problem.

In this work, we have tried to study and compare various semantic similarity approaches in context of Hindi language. Further, we have tried to find out difficulties and challenges in implementing these approaches. To make the Hindi semantic similarity module, we have proposed semantic similarity finding approaches for Hindi text. These approaches use Hindi WordNet API, Database and Hindi corpus. We find semantic similarity of Benchmark data set with the help of our proposed similarity module. And for justification of our similarity results, we compare these results with human judgment results.

The proposed similarity modules addresses two major challenges, the most important challenge is to provide a method for Word Sense Disambiguation. Another challenge is that semantic similarity calculation is based on lexical ontology and on corpus. However proper tools for dealing with lexical ontology are not available, except English and there is also scarcity of properly organized corpus for different languages.

Table of Contents

Chapter 1

Introduction

1.1 Background and Motivation

Web contains very large amount of information, which is scattered and dynamic as well as diverse in terms of content and nature. The magnitude and complexity of electronically available information is increasing rapidly day by day. This raises questions concerning whether the ways we usually access information are scalable. Users often find it irritating to browse through long lists of answers to queries. Much of the available information is accessed by individuals who do not have very deep knowledge about the domain they are querying and thus they typically have difficulties formulating their information needs using the terminology of the domain. Traditional information retrieval use keyword based similarity matching between query and documents. Techniques based on keyword matching are constrained by attempting to match the user keyword to the source document and present information to the user with documents that lexically matched the user keyword. Keyword matching technique fails to retrieve semantically related document thus retrieve more irrelevant results [15]. Therefore, we need semantic matching based methods for handling this possible information overload, in terms of both quantity and quality. For semantic matching based method we need good semantic similarity measures.

Most of the research and development in information retrieval has been done in the development of either CLIR (cross language information retrieval) system or purely for English language. Considering knowledge based approaches like English WordNet has been extensively used for semantic similarity based information retrieval for English language. Very less work has been done for Hindi language although Hindi is third most spoken language of world. Overall aim of this work is to focus on using Hindi corpus and Hindi WordNet ontology to semantic similarity based information retrieval for Hindi text.

1.2 Information Retrieval: Traditional Approach

Information Retrieval is devoted to finding relevant set of document from a collection of documents. The area of information retrieval deals with the representation, categorization, storage and retrieval of information base objects. These information objects are typically text documents but can be any text, visual, audio information [36].

Representation and categorization is done to provide the user with easy access to the information needed. Retrieval is done in accordance with the model used by the system. A query posed to the system is typically a translation of the user's information need into a set of keywords or index terms that summarizes the information need. The goal of the retrieval system is then to retrieve information possibly ranked according to usefulness and relevance with respect to the user.

1.2.1 Information Retrieval System

In modern information retrieval system a user enters the query that describes the request of information and information retrieval system respond by identifying documents that are relevant to the user query.

Following figure explains working of Information Retrieval system:

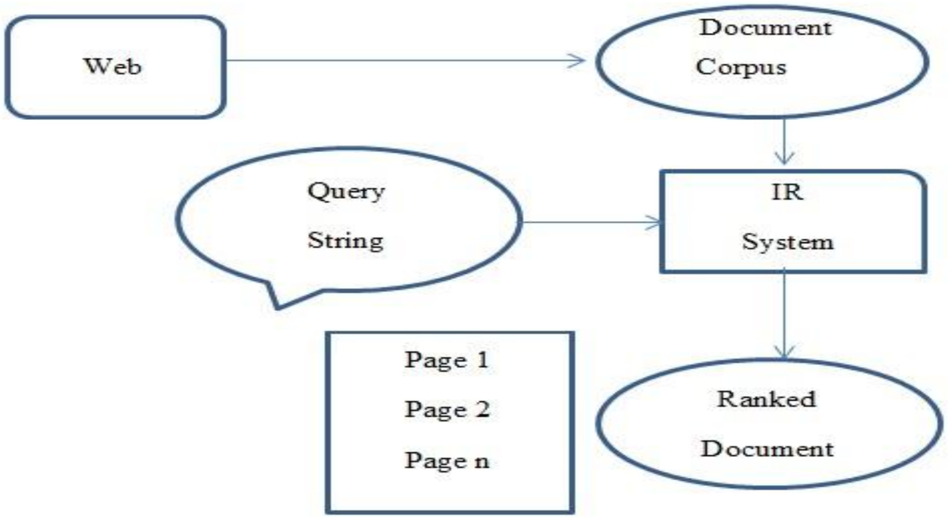


Fig. 1.1 Information Retrieval System [15].

As in above diagram showing the whole working of Information Retrieval system, there are numerous documents on web, user fires a query which is submitted to information retrieval system and Information Retrieval system retrieves all the document which is related to query.

Information Retrieval system is devoted to finding relevant documents. The whole idea of finding relevant documents depends on finding similarity between query and documents. Thus the performance of Information Retrieval system heavily depends on the similarity measure used by an Information Retrieval system to match query with documents.

1.2.2 Categorization of Information Retrieval Models

In this section we discuss various approaches for finding similarity between query and document.

1. Traditional or Classical Models.
2. Semantic Based Models.

Traditional or Classical Models

Traditional information retrieval models are based on keyword based similarity. In these approaches query and documents are represented in some standard model. Based on the nature of model various approaches are defined for finding similarities.

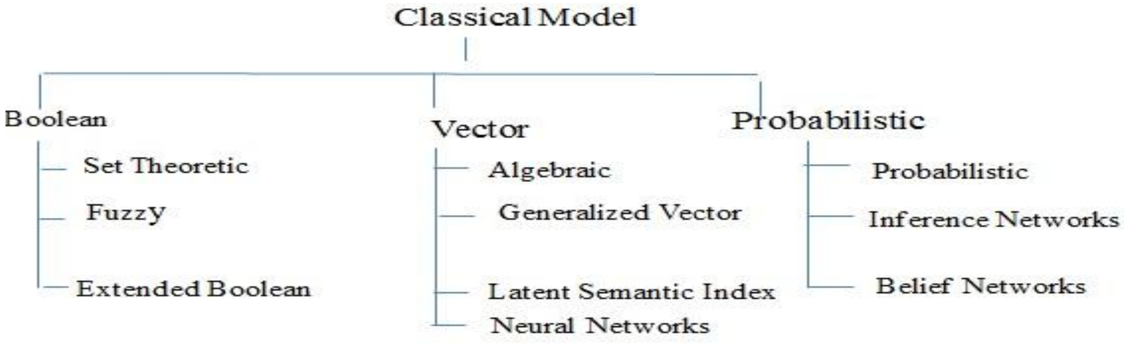


Fig. 1.2 Classification of classical IR model

The Boolean model is set theoretic, as both queries and documents are represented as sets of index terms. In the vector model, queries and documents are represented as vectors in a t-dimensional space, and degree of similarity is measured based on the distance between the vectors. The vector model is considered algebraic. The probabilistic model bases the modeling and representation of documents and queries on probability theory.

Each of the three basic models has been subject to further development over the years. The specializations of each model are illustrated in Figure 1.2.

1.2.2.1 The Boolean Model

The Boolean model is based on Boolean algebra and set theory. The basis for retrieval is a binary criterion, determining whether a document is relevant or not. The model is simple and has a neat formalism, but has a number of drawbacks. The first is that most users find it difficult to translate their information need into a Boolean expression and the second is that the use of non-graded similarity prevents good retrieval performance. Index terms are either present or not, meaning that the weight of term *i* in document *j* is binary, $w_{i,j} \epsilon \{0,1\}$. Queries are posed using the three logical connectives *and*, *or*, and *not*. A query is therefore essentially a Boolean expression, which can be represented as a disjunction of conjunctive vectors (i.e. in disjunctive normal form). Consider as an example Figure 2.2, where a query $q = ta \wedge (tb \vee - tc)$ can be written in disjunctive normal form as $q = (1, 1, 1) \vee (1, 1, 0) \vee (1, 0, 0)$, where each of the components is a binary weighted vector associated with the tuple (*ta, tb, tc*).

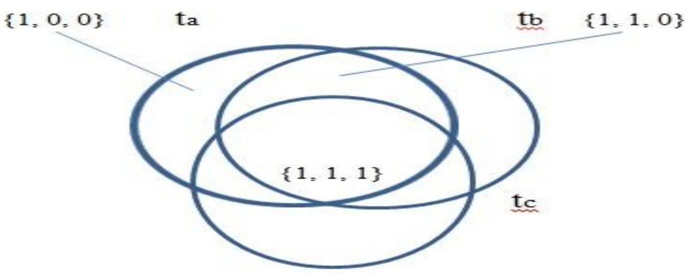


Fig. 1.3 Boolean IR Model

In the Figure three conjunctive components of a query is, *q = ta ^ (tb ∨- tc)*

The notion of partial match does not exist. For example, a document *dj* = (0, 1, 0) would be considered non relevant with respect to the query *q = ta ^ (tb ∨- tc).*

1.2.2.2 The Probabilistic Model

The classic probabilistic model, known as the binary independence model, was introduced in 1976 by Robertson and Sparck Jones [35]. The theory is essentially concerned with probabilistic methods for ranking search output, to maximize recall and minimize fallout, based on assumptions about term distributions and principles of output ordering" [35]. The idea is to use a matching function (matching coefficient) derived from the distribution of the index terms, or a subset of index terms, throughout the collection of documents, to rank documents in order of decreasing probability of relevance to a given user query. Central to the approach are two factors, an ordering principle and an in- dependence assumption. The ordering principle states that the probability of the relevance of a document should be calculated from the terms present in the document and from those absent in the document [35]. The independence assumption assumes independence between occurrences of different terms within both the set of relevant documents and the set of non-relevant documents.

Now we can derive a formula for the probability of a document, represented as a binary vector X, being relevant with respect to a query *qk*, by means of odds instead of probabilities and using Bayes' theorem**.**

$$P(a/b) = \frac{P(b/a) \times P(a)}{P(b)} \qquad \dots \dots \dots \dots \dots \dots \dots \dots .. Eq. 1.1$$

Recall that,

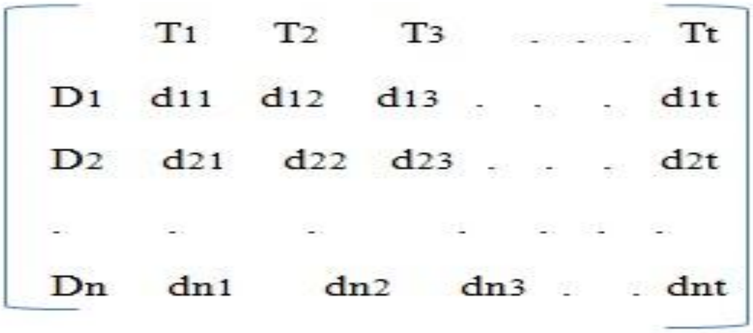
$$O(h) = \frac{P(h)}{P(\bar{h})} = \frac{P(h)}{1 - P(h)} \qquad \dots \dots \dots \dots \dots \dots \dots \dots . Eq. 1.2$$

We write as follows

$$O(R/qk,\bar{x})=\frac{P(R/qk,\bar{x})}{P(\bar{R}/qk,\bar{x})}\quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots..\dots..\ Eq.\ 1.3$$

1.2.2.3 Vector Space Model

In VSM both documents and query are represented as vectors of terms. Then a vector based approach is used to find similarity between queries and documents [14]. The model is based on the idea that, in some rough sense, the meaning of a document is conveyed by the words used. The documents and their distinct terms can be represented in vector form in following way:



$$\left[\begin{array}{cccccc} & T1 & T2 & T3 & \dots & Tt \\ D1 & d11 & d12 & d13 & \dots & d1t \\ D2 & d21 & d22 & d23 & \dots & d2t \\ . & . & . & . & \dots & . \\ Dn & dn1 & dn2 & dn3 & \dots & dnt \end{array}\right]$$

Fig. 1.4 Document-Term Vector

In above vector

Di = Set of documents from 1 to n

Ti = Unique terms in all the documents

dij = Weight of jth term in ith document (calculated depending on the frequency of each term in correspondence to each document.)

If a query is considered to be like a document, a similarity coefficient (SC) that measures the similarity between the document and a query can be computed. Documents whose content, as measured by the term in the document, correspond most closely to the content of query are judged to be most relevant. This model involves constructing a vector which represents the terms in the documents and choosing a method of measuring the closeness of any two vectors by considering the magnitude of the difference between two vectors.

A component of each vector is required to represent each distinct term in the collection, all queries and documents can be represented in two dimensional space. Each component in vector is assigned some weight-typically based on the frequency of the term as it occurs across the entire document collection. The idea is that a term that occurs infrequently should be given higher weight than a term that occurs frequently. Weight is computed using the Inverse Document Frequency (*idf*) corresponding to a given term.

To construct a vector that corresponds to each document, we consider some important definitions.

n = number of distinct terms in a document collection.

tfij= Number of occurrences of term tj in document di.

d*fj*= Number of documents which contain tj in document Di

idfj = log (d/dfj): where D is total number of documents (inverse document frequency).

The vector for each document has n components and contains an entry for each distinct term in the entire document collection. The components in the vector are filled with weight computed for each term in the document collection. The term in each document are automatically assigned weights based on how frequently they occur in the entire document collection and how often a term appears in a particular document? The weight of a term in a document increases the more often the term appears in one document and decreases the more often it appears in all other documents.

A weight computed for a term in a document vector is nonzero only if the term appears in the document. For a large document collection consisting of numerous small documents, document vectors are likely to contain mostly zeroes. Calculation of the weighting factor (*d*) for a term in document is defined as a combination of term frequency (*tf*), and inverse document frequency (idf*)*.

To compute the *jth* entry in the vector corresponding to document i, the following equation is used:

$$d_{i,j} = tf_{i,j} \times idf_j$$

The two important factors used in computing this coefficient are term frequency and inverse document frequency. When a document retrieval system is used to query a collection of documents with t terms, the system computes a vector *D (di1,di2,…dit)* of size t for each document. The vectors are field with term weights as described above. Similarly, a vector *Q (wq1, wq2…wqt)* is constructed for the terms found in the query.

A simple similarity coefficient *(SC)* between a query *Q* and a documents *Di* is defined by the product of the two vectors. Since a query vector is similar in length to a document vector, this same measure is often used to compute the similarity between two documents.

$$SC(Q, Di) = \sum_{j=1}^{t} Wqj \times dij$$

Several different means of comparing a query vector with a document vector have been implemented. The most common of these is the *cosine measures* where the distance between vectors *d*1 and *d*2 captured by the cosine of the angle *x* between them.

$$SC(Q, Di) = \frac{\sum_{j=1}^{t} Wqj \times dij}{\sqrt{\sum_{j=1}^{t} (dij)^2 \times \sum_{j=1}^{t} (Wqj)^2}} \quad \dots\dots\dots\dots\dots\ Eq. 1.4$$

Although the VSM model is a very simple count model still it has some limitations.

1.2.3 Limitation of Traditional Information Retrieval

Traditional Information Retrieval System has a number of limitations, some of them are-

1. In VSM, Long documents are poorly represented because they have poor similarity values (a smaller scalar product and a large dimensionality).

2. In VSM, Search keywords must precisely match document terms; words substring might result in a "false positive match".

3. In VSM, Semantic sensitivity; documents with similar context but different term Vocabulary wouldn't be associated, resulting in "false negative match".

4. In VSM, Scoring Phrases of words difficult.

5. The main disadvantage of the Boolean model is therefore that exact match may lead to retrieval of too few answers, or too many.

6. In VSM, Does not support Boolean queries as with the Boolean model.

7. In Boolean, Users find it difficult to construct effective Boolean queries for several reasons [34].

8. In Boolean, Only documents that satisfy a query exactly are retrieved.

9. In Boolean, It is difficult to control the number of retrieved documents.

10. In Boolean, It does not represent the degree of uncertainty or error due the vocabulary problem [11].

11. In Boolean, the Boolean model does not support the assignment of weights to the query or document terms.

12. Traditional IR systems judge precision and recall based on a match between index and query terms. This mode of operation is the 'best-match' principle [1]. However, this method is limited as it does not consider the contextual nature of human judgment [24].

1.3 Semantic Similarity

As discussed in previous section (section1.1) semantic similarity allows us to capture the meaning of the words. There is a lot of subjectivity in dealing with semantic similarity aspects. Therefore different people deal with semantic similarity in their own ways.

However, According to Lin's proposal, the semantic similarity between two topics in taxonomy is defined as a function of the meaning shared by the topics and the meaning of each of the individual topics [27]**.**

In another words, **Semantic similarity** is a concept whereby a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning / semantic content [17].

1.3.1 Role of Semantic Similarity for Information Retrieval

As we have analyzed traditional approaches based on VSM for information retrieval and found some of their pros and cons. The main problem with traditional approaches is that they only consider keyword based similarity and do not taken care of meaning and relationship between words. To overcome from such limitations a new approach of semantic similarity introduced which considers keyword based similarity as well as semantics of word and relationship between them.

To relate concepts or entities between different sources, the concepts extracted from each source must be compared in terms of their meaning (i.e. semantically). Semantic similarity offers the means by which this this goal can be realized.

There are two terms generally researchers used in context of information retrieval based on semantics, one is semantic relatedness and another is semantic similarity. The difference between these two is depending on the set of relationship between words they used to compare. We can say that semantic relatedness is a broader term than semantic similarity. Semantic similarity only considers Hypernym – hyponym relationship to find that how similar the two words are. While semantic relatedness, apart from Is-A hierarchy they also considers Meronymy, Antonymy, polysemy and some other functional associations. *For example,* car and truck are not similar while we are considering traditional information retrieval without semantic similarity but are semantic similar while we are using information retrieval based on semantic similar.

1.3.2 Challenges of Semantic Similarity

Semantic similarity is based on senses as well as keyword based matching of the words and each word has many senses depending on its uses. So this leads problem of word sense disambiguation. The most important challenge is to provide a method for Word Sense Disambiguation. Another problem is that semantic similarity calculation is based on lexical ontology and on corpus. However proper tools for dealing with lexical ontology are not available, except English and there is also scarcity of properly organized corpus for different languages.

1.4 Outline of Dissertation

As it appears more and more obvious that there is a great need of implementing semantic similarity for information retrieval system that can handle the challenges described in this chapter, we will start presenting some of the strengths and weakness of the traditional information retrieval model to identify their usefulness with respect to problem at hand. A possible way to overcome the limitations of the systems described is to include the growing body of available structured knowledge. One such type is the knowledge contained in ontologies, and in chapter 2, we will investigate some of the lexical ontology specially WorldNet based lexical ontology representation and modeling formalism described in the literature, so as to choose one suitable for our methodology. In chapter 3, we will discuss about ontology based semantic similarity measures with their strengths and weakness. Therefore after Chapter 4, describing the proposed work section wise followed by experiments and results in which we show comparison of proposed semantic similarity approaches in tables with respect to correlation with human judgments. Chapter 5 is all about the conclusion and future scope of this proposed research.

Chapter 2

Lexical Ontologies Based on WordNet

The purpose of this chapter is to present and describe Lexical ontologies, mainly focusing on WordNet. WordNet is the most prevalent approach to knowledge representation in the literature of Natural Language Processing, with respect to expressiveness and reasoning capabilities, and thereby motivates to take this as a choice of formalism adopted in the dissertation.

2.1 Ontology: An Overview

In philosophy, the word ontology denotes the special branch of metaphysics that deals with the study of being.

The principal area of metaphysical speculation is generally called ontology and is the study of the ultimate nature of being [12].

The branch of metaphysics that deals with the nature of being [13].

An essential aspect of ontology concerns with identifying what categories of being are fundamental. In the Aristotelian tradition, this investigation concerns the determination of the most fundamental senses in which things can be said to be.

Aristotle distinguishes the categories of Substance, Quantity, Quality, Relation, Place, Time, Posture, State, Action, and Passion, and uses them to classify anything that may be predicated about anything in the world [19]. Take as an example the statements \the vase is standing on the table" and\the vase is In the Middle Ages, the key issue in ontology was universals as opposed to individuals [19]. Universals are classes or concepts and individuals are black".

The first statement says something about the category of place, whereas the latter says something about the category of quality Instances of classes. The question was whether universals are actual things (realism), mere words (nominalism) or words used to denote concepts used to represent the individuals of a class (conceptualism). For a realist, the universal *dog* is an actual concept with an extension in the world, whereas a nominalist would say that it only has a verbal extension without external reality. A conceptualist would say that *dog* is a concept used by the mind to denote the set of all dogs in the world but does not correspond to an external reality. The latter was the most prominent view.

The preceding descriptions serve as an insight into the philosophical understanding of the concept of ontology and especially the very interesting discussion concerning the nature of being, which facilitates considerations about what denotes and distinguishes things.

Another definition of ontology concerns the more application-oriented interpretation, where we are still dealing with the essence of concepts, but where the primary objective is to provide a common understanding of the organization of concepts within a given domain.

Ontologies provide a structured way of describing knowledge. **According to Gruber it is a "shared specification of conceptualization" [13].** Practically we can say that ontology is a "formal, explicit specification of a shared conceptualization". Ontology can also be seen as data model consist of words and relationship among them. The basic building blocks of ontologies are concepts and relationships. Concepts or classes can be thought of as sets and appear as nodes in the ontology graph. Classes describe concepts in the domain. For example, a class of wines represents all wines. Specific wines are instances of this class. A class can have subclasses that represent concepts that are more specific than the super class.

For example, we can divide the class of all wines into red, white, and rose wines. Alternatively, we can divide a class of all wines into sparkling and non-sparkling wines. At the class level, we can say that Instances of the class Wine describing their flavor, body, sugar level, and the maker of the wine and so on.

Concepts in the ontology usually have a textual description defining them called glosses, although some ontology includes a formal definition in some kind of logic as well. In ontology, concepts are described by one or more terms. Note that each concept might have more than one term describing it and that a term need not match only one concept. For example, to describe the concept of bicycle the terms “bicycle” and bike can be used. However the term “bike” might also refer to the concept of motorcycle. Usually, ontologies include a single and unambiguous term for each concept. Relationships in ontology are represented as edges between two concepts. Most ontologies include is-a (hypenym-hyponym), has-a part-of (meronym-holonym), polysemy, antonym, synonym etc. relationships.

In practical terms, developing ontology includes

- Defining classes in the ontology.
- Arranging the classes in a taxonomic (subclass–super class) hierarchy.
- Defining slots and describing allowed values for these slots.
- Filling in the values for slots for instances.

We can then create a knowledge base by defining individual instances of these classes filling in specific slot value information and additional slot restrictions. In short, a commitment to a common ontology is a guarantee of consistency, but not completeness, with respect to queries and assertions using the vocabulary defined in the ontology.

In next section we will discuss different type of Wordnet Ontologies as what they include, what’s the criteria of performing reasoning, and their plus and minus point as compared to others.

2.2 WordNet: A Lexical Ontology

Lexical Ontology is ontology of concepts, used in Natural Language Processing (NLP). Serving as a lexical semantics, they provide means of mapping lexical concepts and reason about them. Lexical ontologies can be generalized or they may limit their scope to some domain. In further section we will analyze some general purpose lexical ontologies with brief explanation that how they reason between the words. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations [25].

2.2.1 Information from WordNet lexical Ontology

In ontology concepts are treated as classes generally described by one or more terms so a term need not only represents one concept in ontology, for example, to describe the concept of "*यान* " the term "*पोत*" can be used however "*पोत*" also refer to the concept of "*वाष्प पोत* " and "*पाल नाव* ". Usually, ontologies include a single and unambiguous term for each concept. Relationships are mostly of specific type and connect two or more concepts. Most ontologies includes various relationships to represent proximity among concepts, relationships such as Hypernym - Hyponym (car is-a vehicle), Meronym - Holonym (Earth is Part-of solar system).

2.2.1.1 Generality vs. Specificity Concepts in WordNet Ontology

The order of term representation in ontology leads to the concept of generality vs. specificity which makes ontologies a most suitable tool to compute semantic similarity. In most taxonomic ontologies concept that are higher in hierarchy are more general than those lower in hierarchy. The generality and specificity quantifies the importance of term, specific term being more important. The importance of a term can be measured by information content represented by the node representing the term. The information content of a node in WordNet is –log the sum of all probabilities of all words in that class. The higher the probability of encountering an instance of a class, the lower its information content becomes, whereas classes containing rare words have high information content.

2.2.1.2 Information Content

Information Content is a measure of the specificity of concept means the information the concept expresses in taxonomy. It's a mathematical means of computation which computes, that how specific a concept is. In mathematical terms, for any concept c in the taxonomy, let p(c) be the probability of encountering an instance of concept c. following the standard definition from information theory, the information content of c is then –log p(c). p(c) is a monotonic function, it’s value decreases as one moved up in the hierarchy and increases when moves downside in hierarchy.

So, WordNet’s structure makes it a useful tool for computational linguistics and natural language processing. Various approaches have developed to find semantic similarity which mainly uses WordNet ontology for reasoning.

2.2.2 EuroWordNet

EuroWordNet is a system of semantic networks for European languages (Currently English, Dutch, Spanish, Italian, German, Czech and Estonian). Each language develops its own WordNet but they are interconnected with interlingual links stored in the Interlingual Index (ILI). Such a resource is beneficial for many fields of language processing including machine translation and cross lingual information retrieval, and produce insights into the lexical semantics of each language. In order to achieve the multi-lingual aspect, all WordNet’s are structured using the same building block of English version, i.e. employing the notion of synset and using the same lexical relations [28].

The ILI is a list of concepts with the sole purpose of linking synsets across languages. For example, consider the feline sense of the word cat in both the English and Spanish WordNet. Both synsets would be linked to the same concept in the ILI through a synonym equivalence relation. Another point to be considered here is that the concepts in the ILI are not distinguished by their syntactic category (noun, verb, adverb and adjective). So by being neutral in regards to syntactic category words from different languages and belonging to different category may be linked to the same ILI concept. For example, the Dutch have the verb *bankrukken* which is translated to English as *benchpress* which is noun but because they share the same meaning “a weight lifting exercise” they are linked to each other via the ILI.

The multi-lingual resource presents enormous potential for extending or completing each of the individual WordNets. It is possible to study specific linguistic phenomenon such as regular polysemy occurring in one language and see if it occurs in another. If it does not occur then one can investigate that whether the polysemy pattern in one language can be projected in another.

In brief, we can say that with the use of multi-lingual lexical knowledge base that can be used to support many cross language processing tasks.

2.2.3 EngWordNet

WordNet is a broad coverage of lexical network of English words, developed under the direction of George A. Miller at Princeton University. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), that each representing one underlying distinct lexical concept. Synsets are interlinked by means of variety of conceptual-semantic and lexical relations. In the first version of WordNet, the networks for the four different parts of Speech were not linked one another. The noun network of WordNet was the first to be richly developed [25].

Although in WordNet various relations (Hypernym - Hyponym, Meronym – Holonym, Synonym, Antonym, Polysemy) among word are incorporated but it is heavily grounded on its taxonomic structure that employs the IS-A (Hypernym - Hoponym) inheritance relation. WordNet can also be viewed as graph where synsets (concepts) are vertices and relationships are represented in form of edges. Each concept also combined with gloss which defines that concept. By far, noun synsets are the most dominant type of synsets, around 70% of synsets belongs to noun category. There are several type of relations used to connect the different type of synsets.

We can observe that the noun portion of WordNet makes use of many types of relations and exhibits the highest connectivity between its elements. The backbone of the noun network is Hypernym - Hyponym hierarchy which constitutes 65% of the relations connecting noun synsets. At the top of the hierarchy are 11 abstract concepts, termed unique beginners, such as entity and psychological feature. *The maximum depth of noun hierarchy is 20 nodes.* The nine types of relations are defined in noun sub network. Psycholinguistic researchers proved that these types of relations provide the best source of knowledge for similarity assessments. Another important characteristic of WordNet is that all possible senses of word are contained.

For example, the word bank has 18 senses in WordNet. An example of Hypernymy – Hyponymy (IS-A) relation extracted from WordNet ontology is given in following figure.

EngWordNet Lexical Structure:

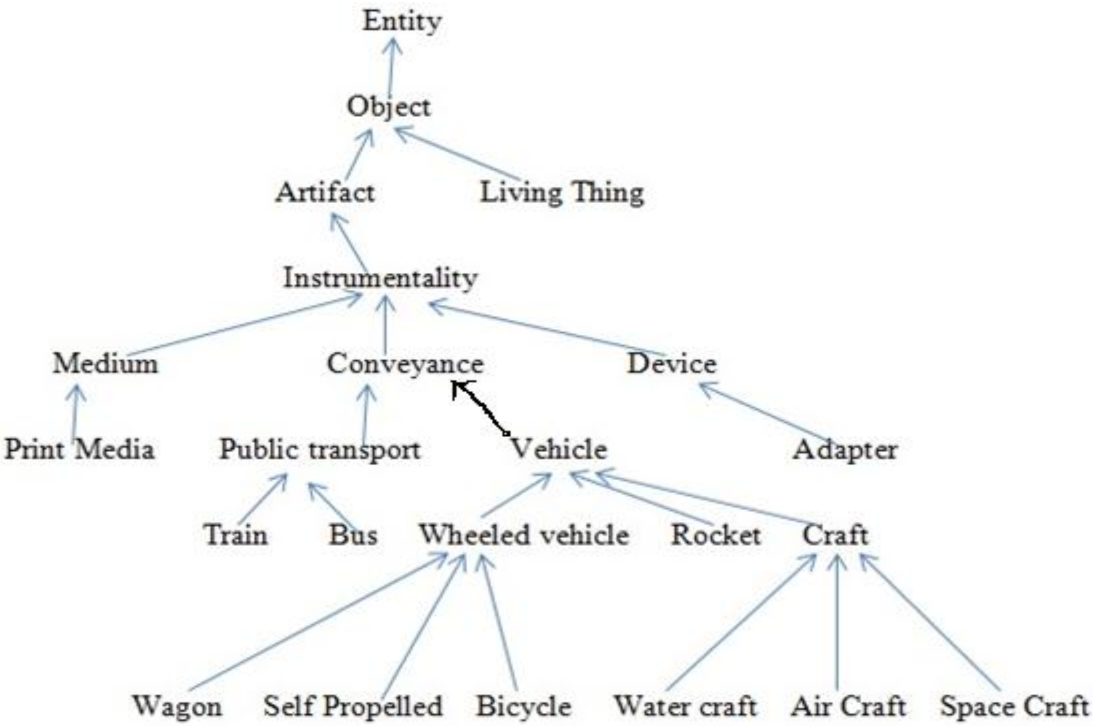


Fig. 2.1 English WordNet Hypernymy tree Structure.

2.2.4 Indo WordNet

Indo WordNet is a system of semantic networks for Indian languages (Currently Hindi, Gujrati, Bangali, Punjabi, Marathi, Tamil, Telugu, Nepali and Kashmiri). Each language develops its own WordNet but they are interconnected with interlingual links stored in the Interlingual Index (ILI). Such a resource is beneficial for many fields of language processing including machine translation and cross lingual information retrieval, and produce insights into the lexical semantics of each language. In order to achieve the multi-lingual aspect, all WordNet's are structured using the same building block of Hindi version, i.e. employing the notion of synset and using the same lexical relations [31].

2.2.4.1 Hindi WordNet

The Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Hindi WordNet is inspired by the famous English WordNet [31].

In the Hindi WordNet the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set, or synset, in the Hindi WordNet, representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings. Synsets are the basic building blocks of WordNet. The Hindi WordNet deals with the content words, or open class category of words. Thus, the Hindi WordNet contains the following category of words- Noun, Verb, Adjective and Adverb.

2.2.4.2 Description and Design of Hindi WordNet

Each entry in the Hindi WordNet consist synset, gloss and ontology which discuss blow.

Synset: It is a set of synonymous words. For example, "विद्यालय, पाठशाला, स्कूल" (vidyaalay, paaThshaalaa, skuul) represents the concept of school as *an educational institution*. The words in the synset are arranged according to the frequency of usage.

Gloss: It describes the concept. It consists of two parts:

Text definition: It explains the concept denoted by the synset. For example, “वह स्थान जहाँ प्राथमिक या माध्यमिक स्तर की औपचारिक शिक्षा दी जाती है” (vah sthaan jahaan praathamik yaa maadhyamik star kii aupachaarik sikshaa dii jaatii hai) explains the concept of school as *an educational institution.*

Example sentence: It gives the usage of the words in the sentence. Generally, the words in a synset are replaceable in the sentence. For example, “इस विद्यालय में पहली से पाँचवी तक की शिक्षा दी जाती है” (is vidyaalay men pahalii se paanchaviin tak kii shikshaa dii jaatii hai) gives the usage for the words in the synset representing school as *an educational institution.*

Position in Ontology: An ontology is a hierarchical organization of concepts, more specifically, a categorization of entities and actions. For each syntactic category namely noun, verb, adjective and adverb, a separate ontological hierarchy is present.

Each synset is mapped into some place in the ontology. A synset may have multiple parents. The ontology for the synset representing the concept school is shown in below given figure.

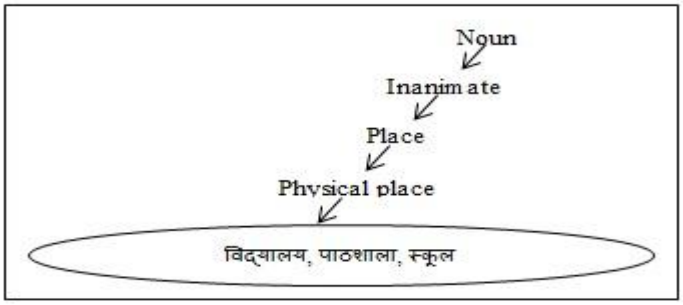


Fig. 2.2 Ontology for the synset of स्कूल

Relations in Hindi WordNet

A WordNet is a word sense network. A word sense node in this network is a synset which is regarded as a basic object in the WordNet. Each synset in the Hindi WordNet is linked with other synsets through the well-known lexical and semantic relations of *hypernymy,* *hyponymy, meronymy, troponymy, antonymy, entailment etc*. Semantic relations are between synsets and lexical relations are between words. These relations serve to organize the lexical knowledge base.

There are 16 relations in the Hindi WordNet. These relations are described below.

Hyponymy and Hypernymy: Hypernymy is a semantic relation between two synsets to capture super-set hood. Similarly, hyponymy is a semantic relation between two synsets to capture sub-set hood. The hyponymy relation is transitive and asymmetrical. Hypernymy is the reverse of hyponymy.

Example:

बेलपत्र, बेल-पत्र, बेलपत्ती, बिल्वपत्र (bel patr, bel-patr, belpattii, bilvapatr; *a leaf of a tree named bela*)

==> पत्ता, पात, पर्ण, पत्र, दल (pattaa, paat, parN, patr, dal; *leaf*)

Here, *बेलपत्र* (bel patra; *a leaf of a tree named bela*) is a kind of *पत्ता* (pattaa; *leaf*) means *पत्ता* (pattaa; *leaf*) is a hypernym and बेलपत्र (bel patra; *a leaf of a tree named bela*) is the hyponym.

Meronymy and Holonymy (Part-whole relation**):** It is a semantic relation between two synsets. If the concepts A and B are related in such a manner that A is one of the constituent of B, then A is the meronym of B and B is the holonym of A. The meronymy relation is transitive and asymmetrical. Holonymy is the reverse of meronymy. It is used to construct a *part-of* hierarchy.

Example:

जड़, मूल, सोर (jaR, muul, sor; *root*)

==> पेड़, वृक्ष, पादप, द्रुम, तरु, विटप, रूख, अघ्रिप, अग (peR, vriksh, paadap, drum, taruu, viTap, ruukh, ruuMkh, aghrip, ag; *tree*)

Here, जड़ (jaR; *root*) is the part of पेड़ (peR ; *tree*), meaning that जड़ (jaR; *root*) is the meronym of पेड़ (peR ; *tree*) and पेड़ (peR ; *tree*) is the holonym of जड़ (jaR; *root*).

Entailment: Entailment refers to a relationship between two verbs. Any verb A entails B, if the truth of B follows logically from the truth of A. The relation of entailment is unilateral, *i.e.*, it is one way relation.

Example:

खर्राटा लेना, नाक बजाना (kharraaTaa lenaa, naak bajaanaa; *snore*)

==> सोना (sonaa; *sleep*)

Antonymy: Antonymy is a relation that holds between two words that (in a given context) express opposite meanings. It is a lexical relation as it holds between two words and not the entire synset.

Example:

मोटा, स्थूलकाय (**moTaa**, sthuulkaay; *fat*)

==> **पतला**, दुबला, दुबला-पतला, छरहरा (**patlaa**, dublaa, dublaa-patlaa, charharaa; *thin*).

The words in bold face in the synset are in antonymy relation.

Gradation: Gradation is a lexical relation. It represents the intermediate concept between two opposite concepts. Figure 2 shows the gradation relation among three words.

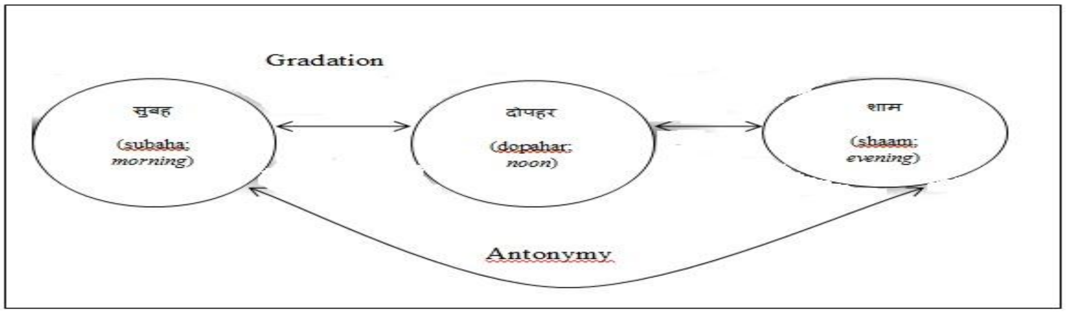


Fig. 2.3 Gradation relation

Causative: In Hindi, there is a convention of forming causation by making morphological change in the base verb. The Causative relation links the causative verbs and the base verbs and show interdependency between them.

Example:

खाना (khaanaa ; *eat*)

==> खिलाना (khilaanaa; *to make someone to eat*)

Troponymy: Troponym denotes a specific manner elaboration of another verb. It shows manner of an action*, i.e.*, X is a troponym of Y if *to X* is *to Y* in some manner.

Example:

मुस्कुराना,मुस्कराना,मुस्काना (muskuraanaa, muskaraanaa, muskaanaa; *smile*)

==> हँसना,विहँसना (hansnaa, vihansnaa *laugh*)

Cross parts of speech linkage: Following relations are between the synsets of different parts of speech.

Ability Link: This link specifies the inherited features of a nominal concept. This is a semantic relation.

Example:

मछली,मच्छी,मत्स्य,मीन,माही (machlii, macchii, matsya, miin, maahii; *fish*)

==> तैरना, पैरना, पौंरना, पौरना, हेलना (tairnaa, pairnaa, pauMrnaa, paurnaa; *swim*).

Capability Link: This link specifies the acquired features of a nominal concept. This is a semantic relation.

Example:

व्यक्ति,मानस,शख़्स,शख्स,जन (vyakti, maanas, sakhs, jan; *person*)

==> तैरना,पैरना,पौंरना,पौरना,हेलना (tairnaa, pairnaa, pauMrnaa, paurnaa; *swim*)

Function Link: This link specifies the function of a nominal concept. This is a semantic relation.

Example:

अध्यापक,शिक्षक,आचार्य,गुरु,मास्टर (adhyaapak, shikshak, aacaarya, guru; teacher)

==> पढ़ाना,शिक्षा देना (paRhaanaa, shikshaa denaa; *teach*)

Linkage between nominal and adjectival concepts

Attribute: This denotes the properties of noun. It is a linkage between noun and an adjective. This is a semantic relation.

Example:

पक्षी,चिड़िया,पंछी,खग,परिंदा,विहंग,विहंगम,पखेरू,विहग (pakshii, ciRiyaa, panchi, khag, parindaa,vihanga, vihangam, pakheru, vihaga; *bird*)

==> पंखदार,पाँखदार,पंखयुक्त (pankhdaar, paankhdaar, pankhyukt; *having wings*)

Modifies Noun: Certain adjectives can only modify certain nouns. Such adjectives and nouns are linked in the Hindi WordNet by the relation *Modifies Noun*.

Example:

सुपात्र,सत्पात्र,अच्छा पात्र (supaatra, satpaatra, acchaa paatra, *eligible*)

==> **व्यक्ति**,मानस,शख़्स,शख्स,जन,बंदा,बन्दा (vyakti, maanas, sakhs, jan; *person*)

Linkage between adverbial and verbal concepts

Modifies Verb: Certain adverbs can only go with certain verbs. *Modifies Verb* is a relation to show connection between such words.

Example:

कभी,किसी समय (kabhii, kisii samay; *sometimes*)

==> काम करना,कार्य करना (kaam karnaa, kaarya karnaa; *to work*)

Derived From: This relation specifies the root form from which a particular word is derived. This relation can go from noun to adjective or vice versa, noun to verb and adjective to verb and aims to handle derivational morphology. This is a lexical relation.

Example:

क्रमशः,क्रमानुसार,यथाक्रम,सिलसिलेवार,बारी-बारी से,क्रमवार (kramashaH, kramaanusaar, yathaakram, silsilevaar, baarii-baarii se, kramvaar; *step by step*)

==> क्रम,सिलसिला,शृंखला,अनुक्रम,अनुक्रमणिका (kram, silsilaa, shrinkhalaa: series)

Hindi WordNet Lexical Structure:

Ontological representation of Hindi WordNet given in below figure.

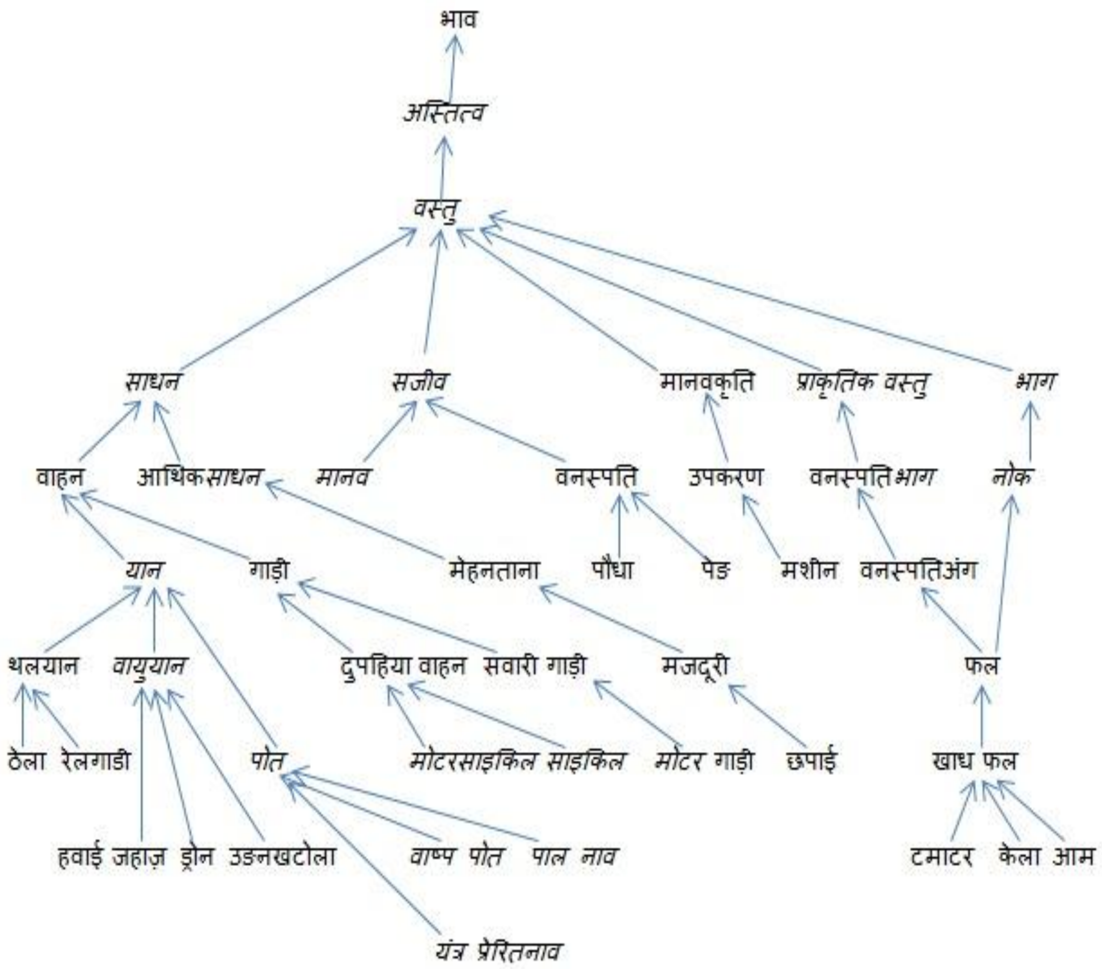
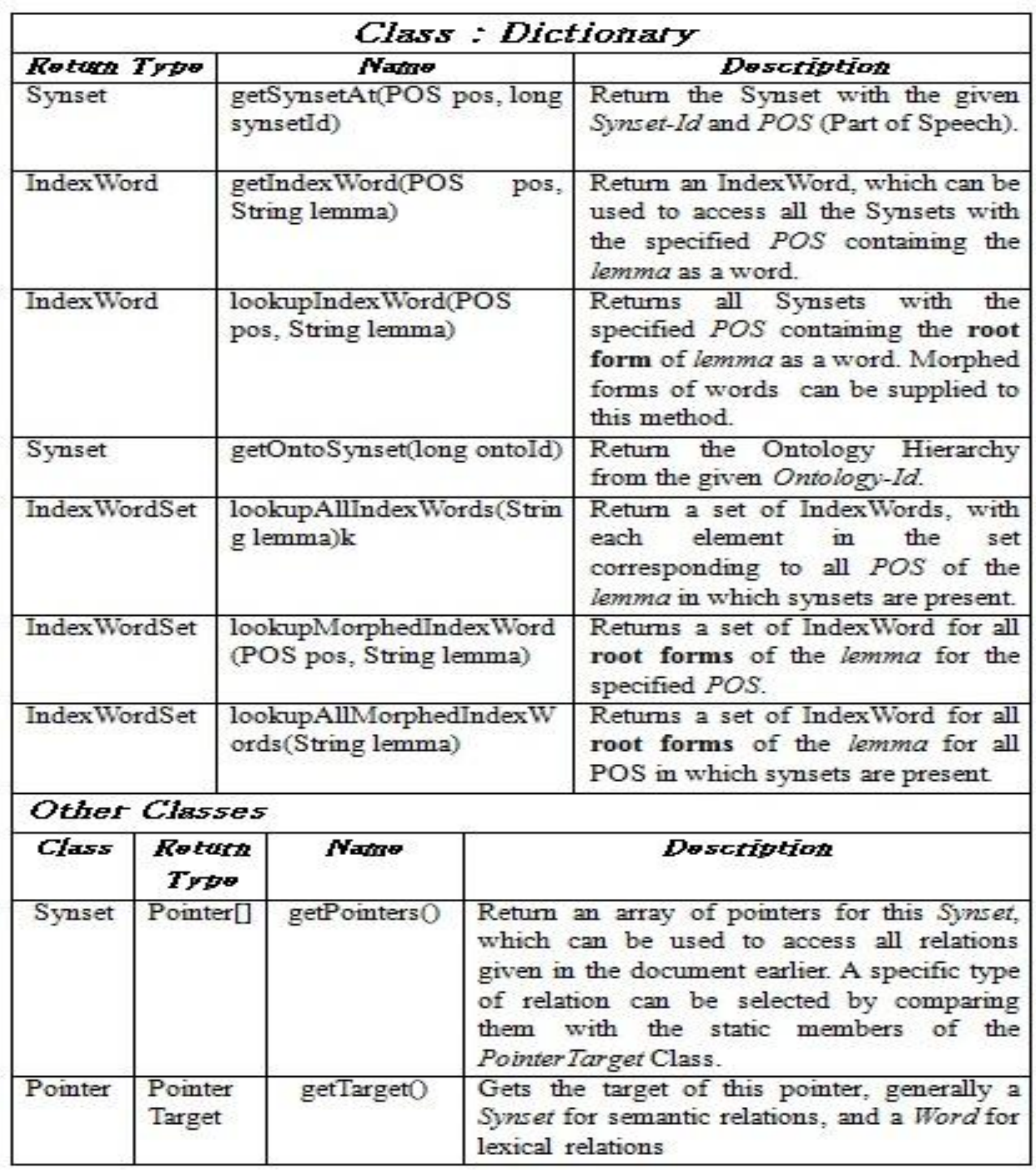


Fig. 2.4 Hindi WordNet Hypernymy Tree Overview

This figure shows the overall structure of Hindi WordNet from the figure we can observe that at the upper level the concept are generalize but when we go to lower level specification of concept increase. Each concept has ontologies according to their senses if concept has single sense then it has single ontology but if it has more than one senses than it will have more than one ontologies. For example- फल as a noun is used in two senses, one as खाध फल and other as *भाग (नोक)*. So it has two ontologies.

Hindi WordNet: The Application Programming Interface (API)

The Hindi Wordnet data can be accessed by using APIs () written in Java called JHWNL or Java Hindi WordNet Library. These APIs allow searching of synsets containing a particular word and accessing the relations of the synsets. The most important functions in the API are described below [31]:



Class : Dictionary

Return Type	Name	Description
Synset	getSynsetAt(POS pos, long synsetId)	Return the Synset with the given *Synset-Id* and *POS* (Part of Speech).
IndexWord	getIndexWord(POS pos, String lemma)	Return an IndexWord, which can be used to access all the Synsets with the specified *POS* containing the *lemma* as a word.
IndexWord	lookupIndexWord(POS pos, String lemma)	Returns all Synsets with the specified *POS* containing the **root form** of *lemma* as a word. Morphed forms of words can be supplied to this method.
Synset	getOntoSynset(long ontoId)	Return the Ontology Hierarchy from the given *Ontology-Id*.
IndexWordSet	lookupAllIndexWords(Strin g lemma)k	Return a set of IndexWords, with each element in the set corresponding to all *POS* of the *lemma* in which synsets are present.
IndexWordSet	lookupMorphedIndexWord (POS pos, String lemma)	Returns a set of IndexWord for all **root forms** of the *lemma* for the specified *POS*.
IndexWordSet	lookupAllMorphedIndexW ords(String lemma)	Returns a set of IndexWord for all **root forms** of the *lemma* for all POS in which synsets are present

Other Classes

Class	Return Type	Name	Description
Synset	Pointer[]	getPointers()	Return an array of pointers for this *Synset*, which can be used to access all relations given in the document earlier. A specific type of relation can be selected by comparing them with the static members of the *PointerTarget* Class.
Pointer	Pointer Target	getTarget()	Gets the target of this pointer, generally a *Synset* for semantic relations, and a *Word* for lexical relations

Although in WordNet various relations (Hypernym – Hyponym, Meronym– Holonym, Synonym, Antonym, Polysym) among word are incorporated but it is heavily grounded on its taxonomic structure that employs the IS-A (Hypernym - Hoponym) inheritance relation. WordNet can also be viewed as graph where synsets (concepts) are nodes and relationships are represented in form of edges. Each concept also combined with gloss which defines that concept. By far, noun synsets are the most dominant type of synsets, around 70% of synsets belongs to noun category. There are several type of relations used to connect the different type of synsets.

All about the ontology, lexical ontology and WordNet has been described in this chapter. In the next chapter, our focus will be how we can use this ontological information in semantic similarity calculation between words?

Chapter 3

Lexical Ontology Based Semantic Similarity

3.1 Semantic Similarity Approaches: An Overview

This chapter describes different measures of semantic similarity, all of which are based on semantic networks. As we have analyzed traditional approaches based on VSM for information retrieval and found some of their pros and cons. The main problem with traditional approaches is that they only consider keyword based similarity and do not take care of meaning and relationship between words. To overcome from such limitations, semantic similarity approaches have been introduced. These approaches consider keyword based similarity as well as semantics of word and relationship between them. There are two terms generally researchers use in context of semantics, one is semantic relatedness and another is semantic similarity. The difference between these two is depending on the set of relationship between words they use to compare. We can say that semantic relatedness is a broader term than semantic similarity [2].

Semantic similarity generally considers Hypernym – hyponym relationship to find that how similar the two words are. While semantic relatedness, apart from IS-A hierarchy also considers Meronym, Antonym, polysemy and some other functional associations. The inverse of semantic relatedness is semantic distance, the term which is generally used by many researchers as alternative.

As the quantification of lexical semantic similarity has many applications in Natural Language Processing (NLP), so different approaches and measures based on these approaches have been proposed. In our thesis we will mainly discuss semantic similarity approaches, which exploit characteristics of ontology to compute similarity. Throughout this thesis we will use only WordNet ontology for all measures of semantic similarity.

After a thorough research in area of linguistics it is inferred that language semantics are mostly captured by nouns and noun phrases therefore most of the retrieval methods are based on noun representation, so generally nouns are compared to measure semantic similarity. Next section describes various categories of semantic similarity approaches

3.2 Ontology Based Semantic Similarity Approaches

Semantic similarity methods are broadly classified in four main categories depending on which component and properties of ontology they use to determine similarity value.

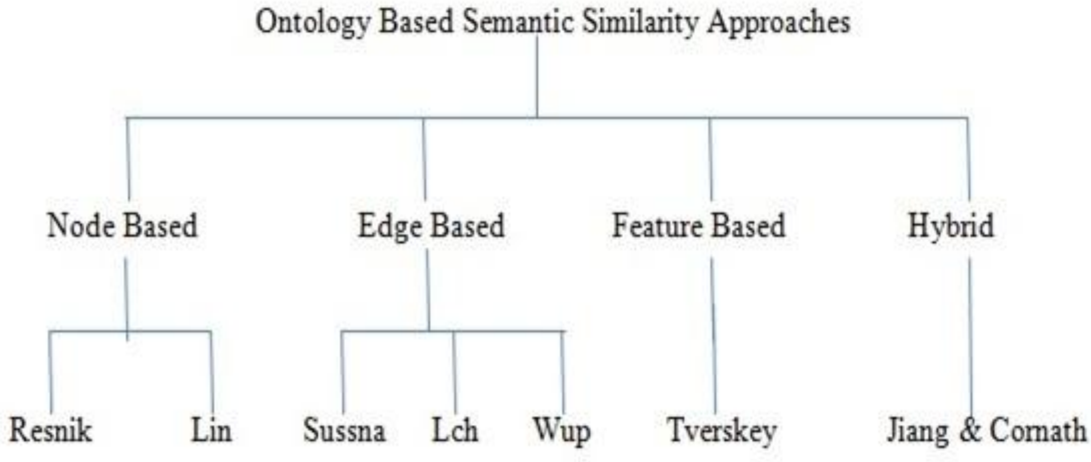


Fig. 3.1 Similarity Approaches

In this chapter we will discuss all the approaches and measures based on these approaches, thoroughly. Before discussing all approaches of finding semantic similarity, we will explain some important aspect regarding ontology on which different approaches are based and invariably used in finding similarity by many researchers.

Depth: - Depth of a node is the length of the path between global root and the word in taxonomical tree.

Local Network Density: - Local density is a function of the number of nodes that occupy a particular region within the WordNet semantic space. There is a problem with this feature that is how one determines that what constitutes a region over which a density measure should be calculated. The choice region should not be too small and two large.

Link Strength: - This is measured by the closeness between a specific child node and its parent node against those of its siblings.

Length: - The length of the shortest path in WordNet from synset Ci to Cj (measured in terms of number of edges between both) denoted by length (Ci, Cj).

Least Common Subsumer (LCS):-LCS is the lowest subsumer which subsumes both the concepts C1 and C2, denoted by LCS (C1, C2).

We shall start our discussion by explaining Edge based approach first.

3.2.1 Edge Based Approach

One of the most natural approaches to determine semantic similarity using ontology is to explore its graphical representation and identify similarity with path length between the concepts. The easiest approach is to compute number of edges in the shortest path between two concepts. “*The shorter the path from one node to another, the more similar they are*” . This measure is simple but it is not sufficient to represent conceptual distance between those concepts. We present an example to justify this fact. We consider word pairs <rocket, helicopter> and <rocket, bicycle>. Referring to our fig. 2.1 the minimum path length from rocket to helicopter is 5 and the minimum path length form rocket to bicycle is 3, but we should not say that rocket is much similar to bicycle than helicopter.

This kind of problem occurs with this approach because it heavily relies on the notion that links in the taxonomy represent uniform distances, which is typically not true. To overcome such kind of limitations some weight must be assigned to the edges depending on the structural characteristics of like: depth, density, link type and strength of link. The weights are reduced as one goes farther down the network since conceptual distance shrinks. The weight is also reduced in a dense part of network since edges in a dense part are considered to represent smaller conceptual distance therefore several measures are available, based on improving original edge count approach by assigning weight to edges. We will mainly discuss three approaches based on the edge based measure (weighted).

Sussna’s Measure

In 1993 **Sussna** gave concept of *depth-relative scaling*, his approach of scaling is based on his observation that sibling-concepts deep in taxonomy appear to be more closely similar to one another than those higher-up. His method construes each edge in the WordNet noun network as consisting of two directed edges representing inverse relations. Each relation *r* has a weight or a range [minr ; maxr] of weights associated with it.

For example, hypernym, hyponym, holonym, and meronym have weights between 1 (minr) to 2ki power 4 (maxr). The weight of each edge of type r from some node C1 is reduced by factor that depends on the number of edges. The distance between two adjacent nodes C1 and C2 is then the average of the weights on each direction of the edge, scaled by the depth of the nodes. Here r and r’ are relation and inverse relations respectively.

By using Sussna’s measure weight for each edge is computed as follows:

$$Wt(c1 \rightarrow r) = max_r - \frac{max_r - min_r}{edge_r\,(c1)} \quad \dots\dots\dots\dots\dots..\ Eq.\ 3.1$$

The distance between two adjacent nodes c1 and c2 is then the average of the weights on each direction of the edge, scaled by the depth of nodes:

$$\mathrm{dist(c1, c2)} = \frac{\mathrm{wt(c1 \rightarrow r) + wt(c2 \rightarrow r')}}{\mathrm{2 \times max\{depth(c1),\ depth(c2)\}}} \quad \dots\dots\dots\dots.\ Eq.\ 3.2$$

Where r is the relation that holds between C1 and C2 and r’ is its inverse (relation from c2 to c1). Finally the semantic distance between two arbitrary nodes Ci and Cj is the sum of the distances between the pairs of adjacent nodes along the shortest path connecting them.

Above distance can be converted in terms of similarity by using following formula:

$$sim(c1, c2) = \frac{1}{dist(c1, c2)}$$

Wup-Palmer Measure

Wup defines the measure based on path length and also introduced scaled metrics as follows [22]:

$$Sim_{wup}(c1, c2) = \left[\frac{2 \times depth(LCS(c1,\ c2))}{depth(concept1) + depth(concept2))}\right] \ldots Eq.\ 3.3$$

Formula For two identical concepts Wu-Palmer measure gives similarity score 1. Note that the common thing between above measures is that they both will always return a value greater than 0, so will not be effected by sparse data problems.

Leacock-Chodorow

This method defines a similarity measure based on the Shortest path d (c1, c2) between two concepts and scaling that value by twice the Maximum depth of the hierarchy, and then taking the logarithm to smooth the resulting score [11].

$$Sim_{lch}(c1, c2) = max\left[-log\left[\frac{length(c1, c2)}{2D}\right]\right] \quad \ldots\ldots\ldots.Eq.\ 3.4$$

Where D is the maximum depth (i.e. 20 in case of WordNet-3.0) note that in practice, we add 1 to both d(c1,c2) and 2d to avoid log(0) when shortest path length is0.

3.2.2 Node Based Approach

Node Based Approaches make use of information content and of lexical taxonomy hierarchy. All node based similarity measures are corpus based measure of the specificity of the concept. In this approach words (concepts) are treated as classes and the basis of comparing two classes for similarity purposes is the Information Content of first subsumer (subsume both the classes) class of these two. The information content can be determined by estimating the probability of occurrence of the class in large text corpus. Shannon defined a measure of the information content of a message:

$$IC(mi) = -log(P(mi))$$

As we observe these approaches we might see that the IC is decreasing as we move from leaves to the higher level in hierarchy that indicates the more generality of the concept and denotes more specificity otherwise. There are three similarity measures based on information content these are i) Resnik (1995) ii) Lin (1998).

Resnik's Measure

The similarity of concepts is viewed as similarity of the class. Hence, the semantic similarities between two classes can be computed as follows [21]:

$$Sim_{Res}(c1, c2) = \max_{ci}\left[log\,\frac{1}{P(Ci)}\right] \qquad \ldots\ldots\ldots\ \ldots\ldots\ldots\ Eq.3.5$$

As above the factor of Ci is the set of classes dominating both c1 and c2, p (Ci)is the probability of Ci and$\left[log\,\frac{1}{P(Ci)}\right]$is the information content of class Ci(subsumer class). Note that in above formula the max function is used because of multiple inheritances in the taxonomy. There might be more than one subsumer for each concept.

In order to compute p(Ci) firstly we have to define word(c) and class(w). Where word(c) is a set of words in all directly or indirectly subordinate classes of class c. For example words (cloister) consist of religious residence, convent, abbey, friary, and monastery.

Class (w) denotes the set {c | w ε words(c)} which includes all the classes in which the word w is contained regardless of its particular sense. For example word(orange)comes in temple orange, bitter orange, sweet orange, jaffa orange, naval orange, Valencia orange. By using these definitions we compute the frequency of the class as-

$$Freq(Ci) = \sum\nolimits_{w \in words(w)} \frac{1}{|class(w)|} \times Freq(w)$$

Where freq(w) is the frequency of occurrence of word w in a large text corpus. The class probabilities can be estimated from such a distribution using maximum likelihood estimation (MLE).

$$P(C) = \frac{Freq(C)}{N}$$

Where N is defined as total size of sample or$\sum_{C'} Freq(C')$.

The reason of taking negative logarithm is that if the probability of a concept to appear is high than that concept becomes less informative, it leads the notion that infrequent words are more informative than frequent ones.

One very important aspect in resnik's experiments that is the probabilities of concepts in the taxonomy were estimated from noun frequencies gathered from one million words Brown Corpus of American English. The key characteristic of his counting method is that an individual occurrence of any noun in the corpus was counted as an occurrence of each taxonomic class containing it.

Lin [1998]

Lin has also proposed a similarity measure based on information content. Lin utilizes two type of the information to compute similarity; one is the ratio between the amounts of the information needed to state commonality between the concepts to be compared and second is the information needed to fully describe these concepts. The formula below makes better understanding of this [4]:

$$Sim_{lch}(C1, C2) = max\left[\frac{2 \times logp(LCS(c1, c2))}{\log p(a) + \mathrm{logp}(b)}\right] \quad \dots\dots\dots.. \ Eq.\ 3.6$$

The Lin Measure gives score from 0 to 1.

4.2.3 Hybrid Approach

An approach which may use combination of two or more approaches discussed above. Jiang & Cornath gave a hybrid approach to measure similarity.

Jiang & Cornath's Measure

Jiang & Cornath further proposed a model which combines edge based notion and information content [5]. The measure uses information content as decision factor with edge based measure of similarity. This method uses various properties of edge counting like depth, link strength and density. In particular more emphasis is given on link strength of an edge which links parent node to child node, where link strength of a child link is proportional to the conditional probability of encountering an instance of child concept Ci given an instance of its parent concept *p*: *P*(*Ci*/ *p*) . Here *p*: (*P*(*Ci*/ *p*)) is calculated as-

$$P\left(\frac{Ci}{p}\right) = \frac{P(Ci \cap P)}{P(p)} = \frac{P(Ci)}{P(p)}$$

Link strength (LS) is computed by taking negative logarithm of above formula.

$$u1 = LS(Ci, p) = -\log\left(P\left(\frac{Ci}{p}\right)\right) = IC(Ci) - IC(p)$$

By observing above formula we can state that link strength is simply the difference between information content of child concept and its parent concept. Further Jiang has also considered density (u2), depth (u3), and link type (u4) along with link strength (u1) (as discussed above) of the edge to compute edge weight. Formula to compute these weights are:

For density:-

$$u2 = \left[\beta + (1 - \beta \frac{\bar{E}}{E(p)})\right]$$

For depth:

$$u3 = \left[\frac{d(p) + 1}{d(p)}\right] power\ \alpha$$

And For link type:-

$u4 = T(C, p)$

Here d (p) is depth of node p in taxonomy, E is average density of whole taxonomy, E (p) is local density. The α and β are the factors which control the degree of contribution to which extent the node depth and density factor contribute to the edge weigh computation.

By combining all above weights viz: u1, u2, u3 and u4 Jiang proposed formula for edge weight computation as follows:

$$wt(C,p) = u1 \times u2 \times u3 \times u4$$

Now the overall distance between two nodes is computed by taking the summation of edge weights along the shortest path.

$$Dist(w1,w2) = \sum_{c\epsilon\{path(c1,c2)-LSuper(c1,c2)\}} wt(C, Parent(C))$$

In the above formula (c1, c2) is set of all nodes connecting c1 and c2, where c1 and c2 are the senses of w1 and w2 respectively and Lsuper is the LCS of c1 and c2. J&C also suggest the formula for distance computation in which only the link strength property of edge is considered. For achieving this they change the weighting scheme stated above by modifying the values for α and β and T(c, p) as 0, 1 and 1 respectively. So the simplified function for distance computation is now:-

$$Dist(w1,w2) = IC(C1) + IC(C2) - 2 \times IC(LSuper(c1,c2))$$

As semantic distance and semantic similarity have inverse relationship in terms of quantitative measurements i.e. if semantic distance between two concepts is larger, then these concepts are less similar. Hence, WordNet::Similarity implementation scoring is implemented as-

$$Sim\_jcn(c1,c2) = \frac{1}{Dist\_jcn(c1,c2)} \quad \dots\dots\dots\dots\dots\dots \; Eq.\ 3.7$$

Note that in WordNet::Similarity implementation for Jcn formula only link strength of edge between child and parent node is considered instead of considering other structural properties of taxonomy. The J&C and Lin’ s formula solves the problem existed with Resnik’s measure i.e. when we compare two identical concepts by using resnik formula then one will obtain information content of this class not the maximum similarity value 1 while if we compute similarity of identical concepts by using Lin or J&C formula then they yields simlin(c1, c1)=1 and distjcn(c1, c1)=0 respectively.

3.2.4 Feature Based Approach

These methods measure the similarity between concepts as a function of their glosses in WordNet or based on their relationships to other similar terms in the taxonomy. [Tverskey] proposed a method using this approach.

Tversky's Measure

Tvs's proposed an abstract model of similarity which takes into account the features that are common to two concepts and also the differentiating features which are specific to each. A function Ψ(c) yields the set of features relevant to c. Tvs proposes following function to compute Semantic similarity [11]:

$$S(c1, c2) = \frac{x}{f(c1 \cap c2)} \cdot \frac{y}{f(c1 - c2)} \cdot \frac{z}{f(c2 - c1)} \quad \dots\dots\dots\dots \; Eq.3.8$$

Where the f() denotes the salience of a set of features of concepts and x, y and z are parameters, used to focus differences among different components. According to Tvs's model similarity is not symmetric. For e.g. Sim$_{tvr}$(c1,c2) = Sim$_{tvr}$(c2,c1).The reason being is that the subject tends to focus on one subject than the other. As we can observed that above formulation is not framed in information theoretic terms. But we can assume that a parallel may be established that will lead to a new similarity function.

In this chapter we have tried to explain all the semantic similarity measures available based on node based, edge based approaches, hybrid approach and feature based approaches. In next chapter we will have more insight into these in terms of their comparative study with respect of suitability framework of different applications and in comparison with human judgments.

Chapter 4

Proposed Work & Experimental Results

As we discussed in chapter 1, little amount of work has been done for Hindi language in area of information retrieval and extraction although Hindi is third most spoken language of world. That’s why my focus is on developing modules and technique for Hindi language especially semantic similarity modules that may be used to extract valuable information from Hindi text in order to make Hindi language more flexible for the computers and Hindi spoken society.

Objectives of our work are as follows:

1-To develop a general framework for calculating semantic similarity between Hindi words pair using Hindi WordNet API.

2- To make a Comparative study of semantic similarity measures with human judgment similarity for Hindi text.

4.1 General Framework Developed for Semantic Similarity

Measuring the semantic similarity between words is an important component in various tasks of information retrieval such as text summarization, word sense disambiguation, searching engine, plagiarism checker, relation extraction, community mining, document clustering, and automatic metadata extraction. Despite the usefulness of semantic similarity measures in these applications, accurately measuring semantic similarity between two words (or entities) remains a challenging task.

We propose a framework to estimate semantic similarity between Hindi words using Hindi WordNet ontology. In our work, we have used Hindi WordNet API’s provided by IIT, Bombay. All about the API’s used in our framework has been explained in chapter 2. All semantic similarity measuring approaches used here has been discussed in chapter 3.

Overall structure of proposed framework is given in below diagram.

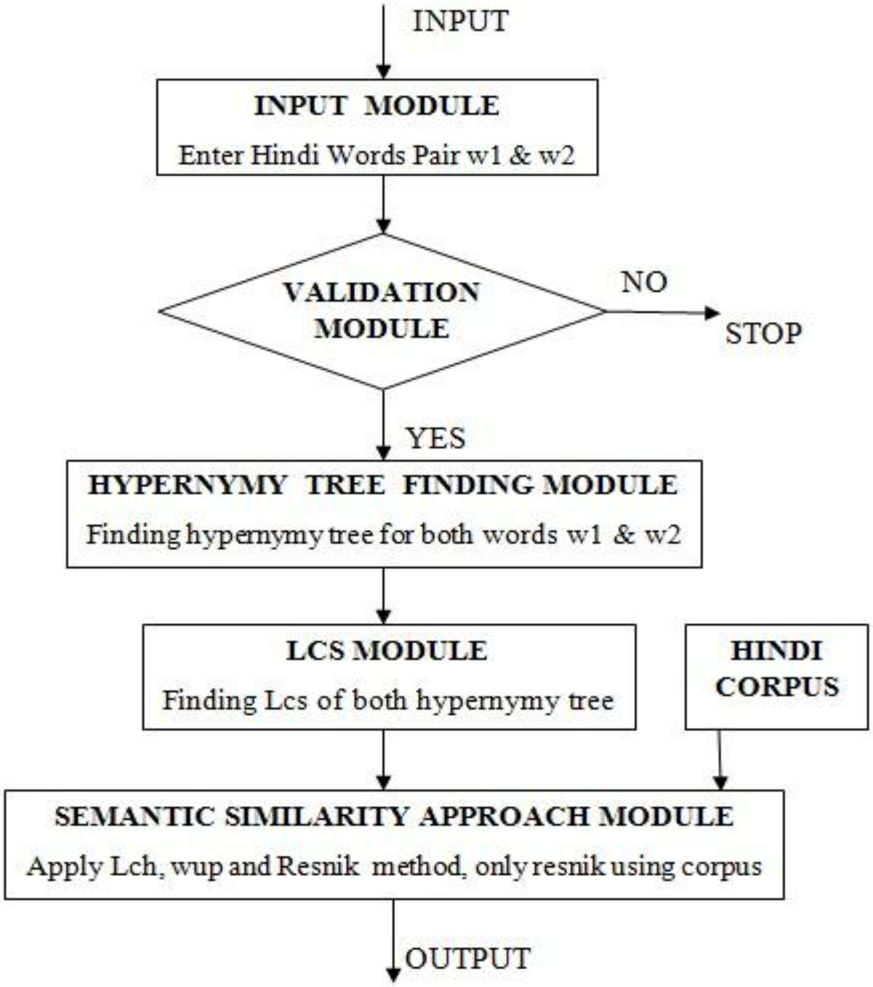

INPUT
INPUT MODULE
Enter Hindi Words Pair w1 & w2
VALIDATION MODULE
NO
STOP
YES
HYPERNYMY TREE FINDING MODULE
Finding hypernymy tree for both words w1 & w2
LCS MODULE
Finding Lcs of both hypernymy tree
HINDI CORPUS
SEMANTIC SIMILARITY APPROACH MODULE
Apply Lch, wup and Resnik method, only resnik using corpus
OUTPUT

Fig. 4.1 Architecture of Proposed Semantic Similarity Framework

In above framework, input module take Hindi words pair as an input then validation module checks that input words pair is present in Hindi WordNet taxonomy or not. If words pair present then it goes to hypernymy tree finding module otherwise terminate here. Hypernymy trees of input words pair goes to lcs module as an input which find common least parent in both tree. Similarity approach concern module gets lcs of words pair from lcs module and use it according to similarity measures. Similarity module calculates semantic similarity as an output.

4.2 The Benchmark Data Set

In order to judge the efficiency of semantic similarity measures, we need Benchmark data set, so our first focus is on: how to get the benchmark data set? And how we can use this data set in our work?

Human Similarity judgment Approaches for Benchmark Data Set

Studies of human synonymy judgments was performed by Rubinstein and Goodenough [37] in 1965,Rubinstein and Goodenough asked two groups totaling 51 subjects to perform synonymy judgments on 65 pairs of nouns. These experiments did not evaluate similarity measures. They solely tried to obtain experimental corroboration for the hypothesis that the proportion of words common to the contexts of word A and to the contexts of word B is a function of the degree to which A and B are similar in meaning [37]. A part of the experiments was to ask humans to judge the similarity between pairs of words, and it was these similarity judgments that later became the basis for evaluation of measures as described below.

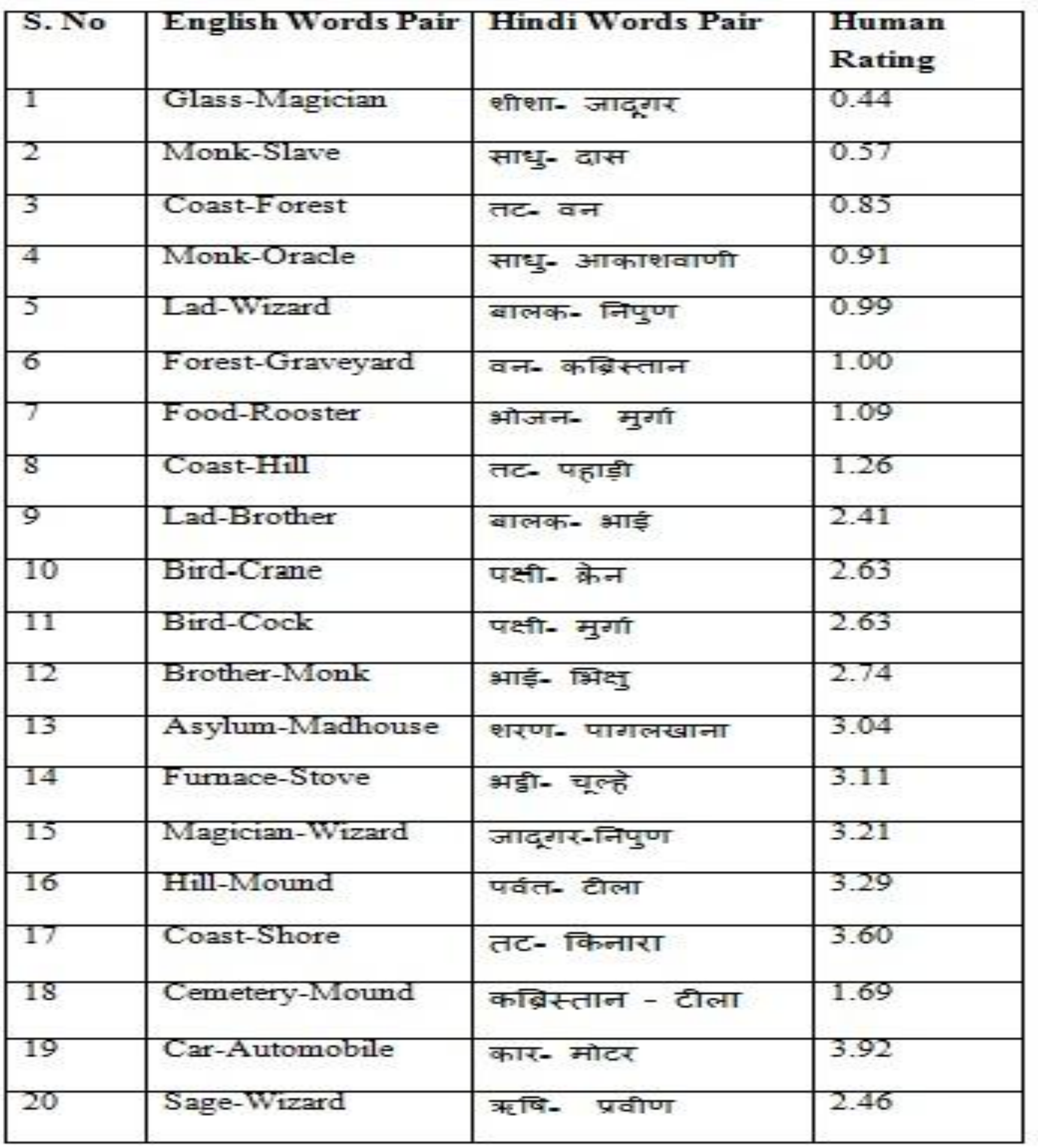
The purpose was to obtain judgments on how similar in meaning one word was to another. Each subject was given the following instructions.

1. After looking through the whole deck, order the pairs according to their similarity of meaning" so that the slip containing the pair exhibiting the greatest amount of similarity of meaning" is at the top of the deck and the pair exhibiting the least amount is at the bottom.

2. Assign a value between 4.0 and 0.0 to each pair {the greater the similarity of meaning", the higher the number. You may assign the same value to more than one pair.

Miller and Charles [24] repeated Rubinstein and Goodenough's original experiment on a subset of 30 noun pairs from the original list of 65 pairs. Miller and Charles obtained a correlation between their experiment and the study performed by Rubinstein and Goodenough of 0.97 [24].

Therefore, Miller-Charles ratings are considered as a reliable benchmark for evaluating semantic similarity measures. The M.C&R.G Benchmark data set is based on English. But our work is on Hindi and there is no standard Benchmark data set for Hindi. Based on NLP expert's advice, we translate Standard English Benchmark data into Hindi according to most usable sense to make Hindi Benchmark data set. In our experiment we consider only 20 words pair of M.C. Benchmark Data Set.



S. No	English Words Pair	Hindi Words Pair	Human Rating
1	Glass-Magician	शीशा- जादूगर	0.44
2	Monk-Slave	साधु- दास	0.57
3	Coast-Forest	तट- वन	0.85
4	Monk-Oracle	साधु- आकाशवाणी	0.91
5	Lad-Wizard	बालक- निपुण	0.99
6	Forest-Graveyard	वन- कब्रिस्तान	1.00
7	Food-Rooster	भोजन- मुर्गा	1.09
8	Coast-Hill	तट- पहाड़ी	1.26
9	Lad-Brother	बालक- भाई	2.41
10	Bird-Crane	पक्षी- क्रेन	2.63
11	Bird-Cock	पक्षी- मुर्गा	2.63
12	Brother-Monk	भाई- भिक्षु	2.74
13	Asylum-Madhouse	शरण- पागलखाना	3.04
14	Furnace-Stove	भट्ठी- चूल्हे	3.11
15	Magician-Wizard	जादूगर-निपुण	3.21
16	Hill-Mound	पर्वत- टीला	3.29
17	Coast-Shore	तट- किनारा	3.60
18	Cemetery-Mound	कब्रिस्तान - टीला	1.69
19	Car-Automobile	कार- मोटर	3.92
20	Sage-Wizard	ऋषि- प्रवीण	2.46

Table 4.1 Benchmark Data Set with Hindi Translation

4.3 Proposed Algorithms and Similarity Modules

As discussed in Chapter 3, some methods have been developed for calculating semantic similarity. Corresponding modules, implementing some of these methods are available for English Language. However such modules are not available for Hindi. In our work we have designed algorithm for calculating semantic similarity values using some of these methods. This algorithm will be very helpful in understanding the details of semantic similarity methods and provide a base for implementing a system that can calculate semantic similarity. Further we have developed similarity modules in Java for implementing these methods.

Algorithm_1(Based on Lch):

```
Input: Enter two Hindi words pair w1 and w2.
Output: Numeric value of Semantic similarity between w1 and w2.

Step 1: Input Hindi words w1 and w2.
Step 2: Words validation.
        If words valid, Enter Step 3
        Else words not valid, Enter Step 7.
Step 3: Find hypernymy of w1 and w2 using WordNet API from Hindi WordNet Taxonomy.
Step 4: Find common hypernymy parent node of both words, which is called least common subsume (LCS).
Step 5: Count Numbers of edges between w1 and w2 through LCS.
Step 6: Apply Lch semantic similarity measuring (Eq. 3.4) methods (Lch method using max depth D of Hindi WordNet taxonomy which is fix and equal to 12).
Step 7: Stop.
```

Algorithm_2 (Based on Wup):

Proposed second semantic similarity algorithm for Hindi text based on Wup that is nearby similar to Algorithm_1 (Based on Lch) with some minor changes.

Step 1: All Steps except **step 6** are same as Algorithm_1 (Based on Lch).

Step 6: Apply Wup (Eq. 3.3) semantic similarity measuring methods (Consider individual depth of words w1 and w2 in Hindi WordNet taxonomy instead of fixed taxonomy depth).

Algorithm_3 (Based on Resnik):

Input: Enter two Hindi words pair w1 and w2.

Output: Numeric value of Semantic similarity between w1 and w2.

Step 1: Input Hindi words w1 and w2.

Step 2: Words validation.

If words valid, Enter **Step 3**

Else words not valid, Enter **Step 7**.

Step 3: Find hypernymy of w1 and w2 using WordNet API from Hindi WordNet Taxonomy.

Step 4: Find common hypernymy parent node of both words, which is called least common subsume (LCS).

Step 5: Find information content of least common subsume (LCS) node (discussed in section 3.2.2 in Chapter 3).

Step 6: Apply Rasnik similarity measures (Eq. 3.5).

Step 7: Stop.

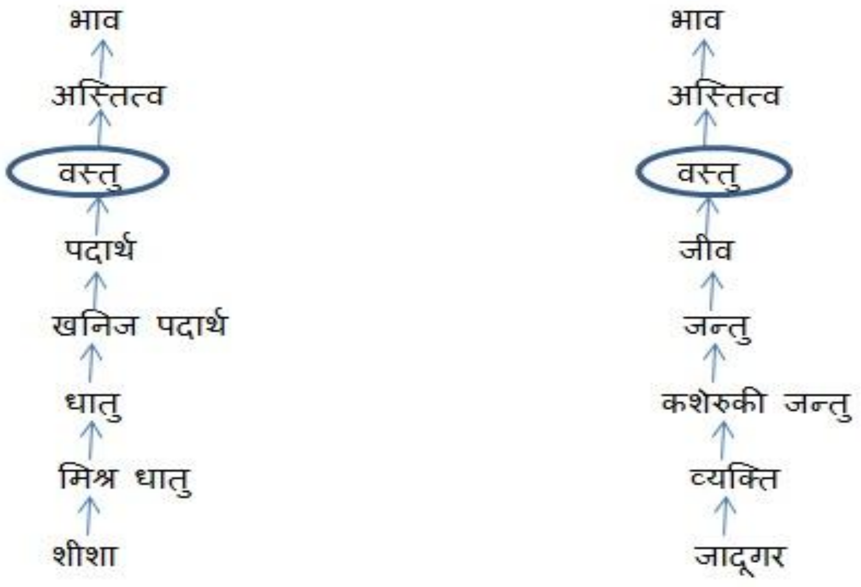
A Rasnik measure is based on corpus but there is no publicly available standard Hindi corpus. According to expert advice we use Hindi corpus designed by Kendriya Hindi Sansthan Agra.

Examples:

Example shows working of all algorithms.

Example- शीशा- जादूगर

Hypernymy of words w1 and w2 using WordNet APIis given below diagram.



(w1) (w2)

Fig. 4.2 Hypernymy relationship

Algorithm_1 (Based on Lch):

The common least subsumer(LCS) of both w1 and w2 words is **वस्तु**.

The number of edges between two words through LCS is **10**.

Now we apply Leacock & Chodorow semantic similarity approach to find similarity-

$$Sim_{lch}(w1, w2) = max\left[-log\left[\frac{length(w1, w2)}{2D}\right]\right]$$

Where D is the maximum depth (i.e. 12 in case of Hindi WordNet-1.2).

Length (w1,w2) is shortest path between word **w1**and word **w2**.

D= 12, and length (शीशा- जादूगर) = 10

Now by applying Lch similarity measures-

$$\boldsymbol{Sim_{lch}}(\text{शीशा} - \text{ जादूगर}) = \mathbf{0.98}$$

This similarity range varies from 0 to 4.

Algorithm_2 (Based on Wup):

The common least subsumer (LCS) of both w1 and w2 words is **वस्तु**.

The number of edges between two words through LCS is **10**.

Now we apply Wup-Palmer Measure semantic similarity approach to find similarity-

$$Sim_{wup}(w1, w2) = \left[\frac{2 \times depth(LCS(w1,\ w2))}{depth(w1) + depth(w2)}\right]$$

Here for given example- depth (LCS(w1, w2))= 3, depth(w1) = 8, depth(w2) = 8.

Now by applying Wup similarity measures-

$$\boldsymbol{Sim_{wup}}(\text{शीशा} - \text{ जादूगर}) = \mathbf{1.48}$$

This similarity range varies from 0 to 4.

Algorithm_3 (Based on Resnik):

The common least subsumer (LCS) of both words w1 and w2 is **वस्तु**.

Now we apply Resnik Measure semantic similarity approach to find similarity-

$$Sim_{Res}(c1, c2) = \max_{ci}\left[log\,\frac{1}{P(LCS(w1, w2))}\right]$$

Where $log\,\frac{1}{P(LCS(w1,w2))}$ is the information content (IC) of LCS.

From Hindi corpus-

P(LCS(w1, w2) = P(वस्तु) = .000139964

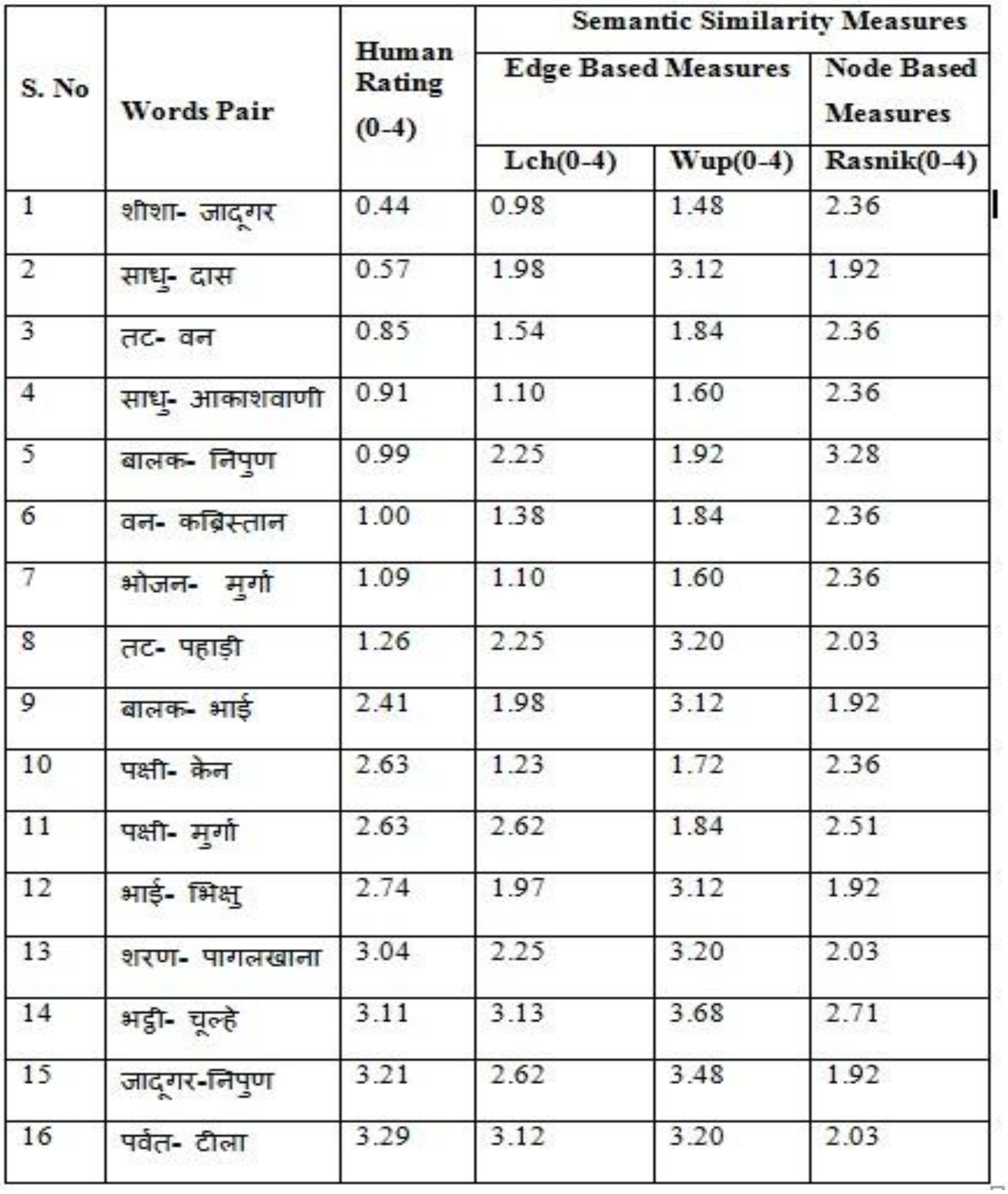
So $\boldsymbol{Sim_{Res}}(\text{शीशा} - \text{ जादूगर})$= **2.36**

This similarity range varies from 0 to 4.

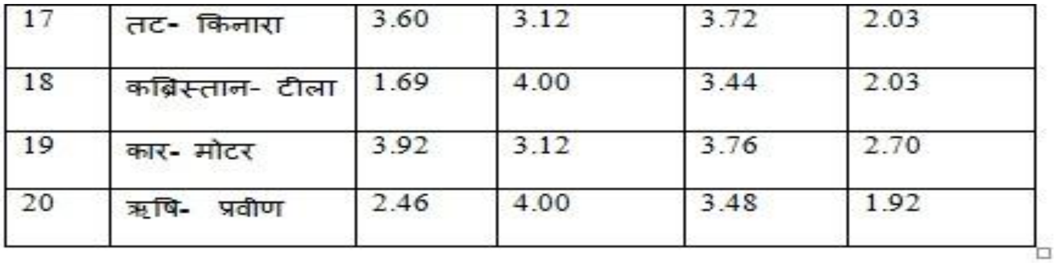
As above, the semantic similarity between 20 Hindi words pair is given below table.

4.4 Experiments and Results

Experimental results based on proposed semantic similarity approaches (discuss in above section 4.3) with respect to human similarity judgments are given below table.



S. No	Words Pair	Human Rating (0-4)	Semantic Similarity Measures		
			Edge Based Measures		Node Based Measures
			Lch(0-4)	Wup(0-4)	Rasnik(0-4)
1	शीशा- जादूगर	0.44	0.98	1.48	2.36
2	साधु- दास	0.57	1.98	3.12	1.92
3	तट- वन	0.85	1.54	1.84	2.36
4	साधु- आकाशवाणी	0.91	1.10	1.60	2.36
5	बालक- निपुण	0.99	2.25	1.92	3.28
6	वन- कब्रिस्तान	1.00	1.38	1.84	2.36
7	भोजन- मुर्गा	1.09	1.10	1.60	2.36
8	तट- पहाड़ी	1.26	2.25	3.20	2.03
9	बालक- भाई	2.41	1.98	3.12	1.92
10	पक्षी- केन	2.63	1.23	1.72	2.36
11	पक्षी- मुर्गा	2.63	2.62	1.84	2.51
12	भाई- भिक्षु	2.74	1.97	3.12	1.92
13	शरण- पागलखाना	3.04	2.25	3.20	2.03
14	भट्टी- चूल्हे	3.11	3.13	3.68	2.71
15	जादूगर-निपुण	3.21	2.62	3.48	1.92
16	पर्वत- टीला	3.29	3.12	3.20	2.03



17	तट- किनारा	3.60	3.12	3.72	2.03
18	कब्रिस्तान- टीला	1.69	4.00	3.44	2.03
19	कार- मोटर	3.92	3.12	3.76	2.70
20	ऋषि- प्रवीण	2.46	4.00	3.48	1.92

Table 4.2 Experimental Results

Above experimental semantic similarity results of our proposed approach can be justified by comparing with human judgment similarity rating with the help of correlation coefficient discussed in next section.

4.5 Comparison of Measures

The actual comparison between similarity measures and human similarity judgments is calculated in terms of correlation between the two sets of ratings, thereby giving us a qualitative assessment of the different measures correlation with human similarity judgments. This in turn is an indication of the measure usefulness in, for example, an information retrieval task.

Correlation coefficient

Correlation is a technique for investigating the relationship between two quantitative, continuous variables. In other words, correlation coefficient (r) is a measure of the strength of the association between the two variables [38].

In our experiment, three correlation coefficients are commonly used for making comparison study between human judgment and similarity measures.

These are -

1. Pearson (linear) correlation.

2. Spearman (rank) correlation.

3. Kendall’s tau (also based on ranks).

Linear Correlation Coefficient

In such a case, linear correlation coefficient can be used instead Ranking based approaches which disregard differences between the particular predicted and actual scores.

Pearson's correlation determines the degree to which two variables have a linear relationship, and takes the actual value of observations into account [38].

Ranking Based Approaches

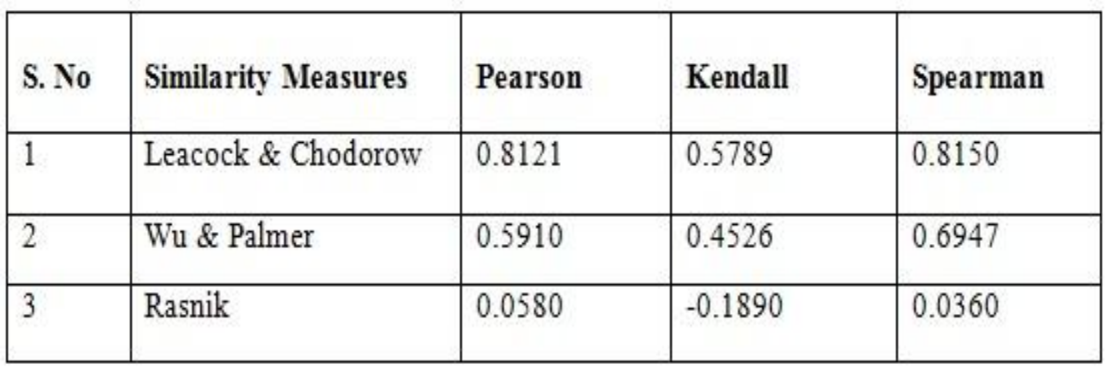
Rank correlations make no assumptions about the type of relationship between the two lists of scores (predictor scores and retrieval effectiveness scores). Both score lists are converted to lists of ranks where the highest score is assigned rank 1 and so on. Then, the correlation of the ranks is measured. Two commonly used rank based correlation coefficients are Spearman and Kendall tau.

Spearman'scorrelation coefficient is calculated based on the rank positions of observations. It therefore measures the degree to which a monotonic relationship exists between the variables [38].

Kendall'stau is also calculated from rank information, but in contrast to Spearman's coefficient is based on the relative ordering of all possible pairs of observations. Kendall's τ is sensitive to all differences in ranking [38].

Moreover, each of the correlation coefficients can be used to conduct a hypothesis test to determine whether there is a significant relationship between the two variables, up to a specified level of confidence. The closer the absolute value of the coefficient is to 1, the stronger the correlation, with a value of zeros indicating that there is no relationship between the variables.

Correlation coefficient between similarity measures and human judgments by pearson, kendall and spearman is given below table.



S. No	Similarity Measures	Pearson	Kendall	Spearman
1	Leacock & Chodorow	0.8121	0.5789	0.8150
2	Wu & Palmer	0.5910	0.4526	0.6947
3	Rasnik	0.0580	-0.1890	0.0360

Table 4.3 Correlation Coefficient

From analysis given in above table, we found that **Lch Semantic Similarity** measures performing well among three above discussed semantic similarity measures in our experiment.

Chapter 5

Conclusion and Future Scope

5.1 Conclusion

As indicated in the beginning of the Dissertation, overall purpose of this work is to devise a semantic similarity approach for Hindi text that can incorporate and utilize the potentially valuable information contained in Hindi WordNet ontology. In this Dissertation we have designed algorithm and developed java modules (for Hindi words) for three popularly used semantic similarity methods: Resenik, LCH and WUP. We experimented with this proposed Semantic Similarity Measurement Methods and evaluated their performance on Benchmark data set. Based on comparison between our experimental results and standard results with the help of correlation coefficient, we find that Lch based similarity measure is performing well among three semantic similarity measures in our experiment.

5.2 Future Work

Following is a listing of the areas for future research that we have identified during the course of the work reported.

Many of the similarity measures presented in this dissertation are based on assumptions and intuitive aspects that could be combined with the weighted shared nodes approach. We shall specially emphasize use of the information content based measures as it could be a possible advantage to combine the weighted shared nodes approach with concept of probability estimates based on corpus statistics.

As stated earlier, the motivation for introducing weighting of nodes is that there is a need to nuance degree of relatedness since we want two concepts to be more similar if they have an immediate subsuming concept than if they only share an attribute. The nodes reachable by semantic relations are therefore not as important as the nodes reachable by the concept inclusion relation. We could further more argue that for all nodes reachable, if a node occurs often in a corpus it is either highly general or very polysemy, which in both cases mean that it is less defining. These aspects can be incorporated in enhancing already existing or developing new semantic similarity methods.

Another topic of interest would be the further use of Sense Disambiguation techniques in our model. At this time, our model makes the assumption that the user is searching for the most common sense of the entered sense. Sense Disambiguation would help the model to understand which sense of the entered term user is searching for and choose this sense in order to make the calculation needed. By adding an initial Sense Disambiguation step, we believe that our model will become more widely applicable.

References

1. Leacock and Chodorow. "Combining local context and WordNet similarity for word sense identification". In Fellbaum 1998, pp. 265-283.
2. Resnik P. "Using information content to evaluate semantic similarity". In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995 pages 448-453, Montreal.
3. Patwardhan S., Banerjee S. and Pedersen T. "Using Semantic Relatedness for Word Sense Disambiguation". In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, 2002, Mexico City.
4. D. Lin, "An Information- Theoretic definition of similarity", In Proceedings of the 15th International Conference on Machine Learning, 1998.
5. J. Jiang and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", In Proceedings of the International Conf. on Research in Computational Linguistics, Taiwan 1998.
6. William D. Lewis, "Measuring Conceptual Distance Using WordNet: The design of a Metric for Measuring Semantic Similarity" DASFAA 2007, LNCS 4443, Pp.115-126, 2007 Springer verlag Berlin Heidelberg.
7. Y. Li, Z.A. Bandar and D. Mclean, "An Approach for Measuring Semantic Similarity between Words using multiple Information sources", IEEE Trans. On Knowledge and Data Engineering July-Aug, 2003.
8. Li, Y. McLean, D. Bandar, Z. A., O'Shea, J. D., and Crockett, K., "Sentence similarity based on semantic nets and corpus statistics", IEEE Transactions on Knowledge and Data Engineering Vol. 18, No. 8, pp. 1138–1150, 2006.
9. H. Zhuge, "Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning", IEEE Transactions on Knowledge and Data

Engineering, vol.21, no.6, 2009, pp.785-799.

10. H. Zhuge and Y.Sun, “The Schema Theory for Semantic Link Network, Future GenerationComputer Systems, vol. 26, no. 3, 2010, pp.408-420.
11. Tverskey, “Features of Similarity “, Psychological Review, 84(4): 327-352, 1977.
12. Tom Gruber, “Ontologies, web 2.0 and beyond”, Presentation to Ontology Summit April 24, 2007.
13. Gruber, “Towards Principles for the Design of Ontologies Used for Knowledge Sharing”, International Journal of Human and Computer Studies, 43(5/6), 907-928, 1995.
14. Banerjee**,** Satanjeev and Ted Pedersen, “Extended gloss overlaps as a measure of semantic relatedness.” In Proceeding of the 18th International Joint Conference on AI, Pages 805-810, 2003.
15. R.B.-Yates,. “Modern Information Retrieval. Addison Wesley Longman (1999)”.
16. Mandala, R. Takenobu, T. Hozumi, “The Use of WordNet in Information Retrieval”. In COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, CA (1998) page 469–477.
17. Mihalcea, R., Corley, C., Strapparava, C. “Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity”. In American Association for Artificial Intelligence (AAAI’06), Boston (2006).
18. Richardson, R., Smeaton, A., “Using WordNet in a Knowledge-Based Approach to Information Retrieval”. Techn. Report Working Paper: CA-0395, Dublin City University, Dublin, Ireland (1995).
19. Li, Y., Bandar, Z.A., McLean, D., “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources”. IEEE Trans. on Knowledge and Data Engineering 15(4) (2003) page 871–882.
20. Leacock, Chodorow M., “Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet”. In Fellbaum, An Electronic Lexical

Database. MIT Press (1998) page 265–283.

21. Resnik, O., “Semantic Similarity in Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language”. Journal of Artificial Intelligence Research 11 (1999) page 95–130.

22. Z. Wu and M. Palmer. “Verb Semantics and Lexical Selection”. In Annual Meeting of the Associations for Computational Linguistics (ACL’94), Pages 133-138, Las Cruces, News Mexico, 1994.

23. Verelas, E. Voutsakis, P. Raftopoulou, “Semantic Similarity Methods in WordNet and Their Application to IR on the Web”, WIDM’05 ACM Press, New York, NY, 10-16, 2005.

24. Miller, G., Charles W., “Contextual Correlates of Semantic Similarity”. Language and Cognitive Processes 6 (1991) 1–28

25. WordNet –“A lexical database for the English language”, http://wordnet.princeton.edu

26. Dang Chenghua, “WordNet Based Document Summarization”, Proceedings of 7th WSEAS Conf. On Applied Computer and Applied Computational Science ACACOS, Hang Zhou, China, Apr. 2008.

27. Kang, B., “Exploiting Concept Clusters for Content-based Information Retrieval”, Information Sciences, Volume 170, Issues 2-4, 2005, pages 443–462.

28. Sarma S.K., Brahma B., Gogoi M. and Ramchiary M. B., “A Wordnet for Bodo Language: Structure and Development”, Global Wordnet Conference (GWC10), Mumbai, India, 2010.

29. Mohanty R., Bhattacharyya P., Pande P., Kalele S., KhapraM. and Sharma A., “Synset Based Multilingual Dictionary: Insights, Applications and Challenges”, Global Wordnet Conference (GWC08),2008, Sieged, Hungary.

30. Chakrabarty D., Sarma V. and Bhattacharyya P., “Complex Predicates in Indian Language Wordnets”, Lexical Resources and Evaluation Journal, 2007.

31. Bhattacharyya P., Fellbaum C. and Vossen P. (eds.), “Principles, Construction and Application of Multlingual Wordnets”, Proceedings of the 5th Global Wordnet Conference, Mumbai, Narosa Publishing House, India, 2010.

32. Naskar S. K. , “Word Sense Disambiguation Using Extended Word Net”, Proceedings of the International Conference on Computing Theory and Applications (ICCTA’07), Published by IEEE Computer Society, 2007.

33. Bing Liu. , Web Data Mining. Springer publication, 2007.

34. Yurii Palkovskii, Alexei Belov, Iryna Muzyka, “Using Wordnet based semantic similarity measurement in external Plagiarism Detection”, Notebook for PAN at CLEF 2011.

35. Sparck Jones, Karen; Robertson, Stephen E. “Relevance Weighting of Search Terms”, Journal of the American Society for Information Science, Volume 27, Number 3, 2007, pp. 129.

36. Van Rijsbergen, C. J. “Information Retrieval”, 2nd Edition, Department of Computer Science, University of Glasgow, 1979.

37. Rubinstein, Goodnough, “Contextual Correlates of Synonymy”, Communications of the ACM, Volume 8, Number 10, 1965.

38. Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In Proceedings of the 16th ACM Conference on Research and Development in Information Retrieval (SIGIR’93), (pp. 329–338).

39. Hindi WordNet- “A lexical database of Hindi language”
http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php