Estimation of Word Net-Based Lexical Semantic Similarity Measure for Telugu Documents

Mrs. A. Kanaka Durga [1], Dr. A. Govardhan[2]
[1]*(Professor CSE & HOD-IT, SCETW, HYD, India)*
[2]*(Professor CSE & DE- JNTUH, HYD, India)*

ABSTRACT: *The estimation of lexical semantic relatedness has numerous applications in NLP. Several measures are available for the evaluation of lexical semantic relatedness. This paper presents two approaches for measuring semantic similarity/distance between words and concepts with the help of WordNet-Telugu. The edge-based approach of the edge counting scheme and the node-based approach of the information content calculation have been explored. In the field of concepts, the measure of Wu and Palmer has the advantage of being simple to implement and have good performances compared to the other similarity measures. The obtained results show that the Wu and Palmer approach presents a better performance in terms of relevance.*

Keywords *- NLP, Semantic Measure, Word Net-Telugu, Information content and Relevance*

I. INTRODUCTION

The features of synonymy and polysemy pose a challenge in the areas of Natural Language Processing and Information Retrieval. At the word level, humans have a problem in finding the interpretation of an ambiguous word and also find it difficult to duplicate the computational process.

The subject of words/terms relationships can be considered in the terminology of the information sources. Knowledge – free approaches depend on the corpus data and use less information. The computation of semantic distances are considered as a research subject and experimented in the areas of linguistics, Artificial Intelligence and data processing. The necessity to find semantic distance and semantic relatedness between two lexically denoted concepts is a complication that diffuses most of natural language processing. Specifically, the area of Information retrieval is heavily based on the similarity identification between the documents. The disadvantage with the similarity identification and semantic measure approaches is that these approaches concentrate on single words of a document and neglects ontological relationships that exist between the words. Measures of relatedness or distance are used in such applications as word sense disambiguation, determining the structure of texts, text summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection, and the automatic correction of word errors in text. It is known that semantic relatedness is a more general concept than similarity. Computational applications generally need relatedness rather than just similarity.

According to Church and Hanks[1989], Hindle[1990], Grefenstette[1992], word relationships are obtained from their co-occurrence distribution in a corpus. With the introduction of machine readable dictionaries, lexicons, thesauri, and taxonomies, these manually built pseudo-knowledge bases provide a natural framework for organising words or concepts into a semantic space. Kozima and Furugori [1993] measured word distance by adaptive scaling of a vector space generated from LDOCE (Longman Dictionary of Contemporary English). Morris and Hirst[1991] used Roget's thesaurus to detect word semantic relationships. With the recently developed lexical taxonomy WordNet[Miller 1990, Miller et al. 1990], many researchers have taken the advantage of this broad-coverage taxonomy to study word/concept relationships[Resnik 1995, Richardson and Smeaton 1995].

This paper explores two approaches to find the semantic similarity between objects in ontology. The first approach is the estimation of the similarity by the information content also known as the node based approach.The second approach is an evaluation of the similarity based on conceptual distance also called edge based approach.

The purpose of this paper is to estimate and compare the performance of proposed approaches of semantic relatedness for use in applications in natural language processing and information retrieval.

The remainder of the paper is organized as follows. Adaptation of WordNet-Telugu is explored in section 2 Section 3 speaks about similarity estimators. Section 4 and section 5 describes about results and analysis.

II. Word Net

Not all words are equally significant for representing the semantics of a document. In written language, some words carry lexical meaning and others functional whereas certain others with a gradience between the two. Other dominant group of word with lexical sense is verbs but they indicate actions and activities held by arguments (nouns) in a sentence and therefore not considered relevant for our purpose. Usually nouns or groups of noun words are the ones which are most representative of a document content. Therefore it is usually considered worth while to process the text of the documents in the collection to determine the terms to be used as indexical terms. It is understood that representing documents by sets of index terms leads to a rather imprecise representation of the semantics of the documents in the collection. For example, a term like ఆ, ఇది, లో, ని etc has no lexical meaning (hence functional entities) and might lead to the retrieval of various documents which are unrelated to the present user query. Using the set of all words in a collection to index its documents generates too much noise for the retrieval task. One way to reduce the noise is to reduce the set of words which can be used to refer to ie., to index documents. Thus, the processing of the documents in the collection might be viewed as a process of controlling the size of the vocabulary ie the number of distinct words used as a an index terms. The use of a controlled vocabulary leads to an improvement in retrieval performance.

One of the text transformations of document pre-processing is construction of term categorization structures such as a thesaurus, or extraction of structure directly represented in the text, for allowing the expansion of the original query with related terms. The main components of the thesaurus are its index terms, the relationships among the terms, and a layout design for these term relationships. The set of terms related to a given thesaurus term is mostly composed of synonyms and near-synonyms. Relationships can be induced by patterns of co-occurrence within documents.

Word Net at Princeton University has attempted to model the lexical knowledge of a native speaker of English, [Mill a, Mill b and Beck]. The system has the power of both an on-line thesaurus and on-line dictionary.

Word Net -Telugu: It gives a detailed database of semantic relationships between telugu words. It was developed by Dr. G.Uma Maheswara Rao and his research team from CALTS-HCU. Nouns, adjectives, verbs, and adverbs are grouped into about 36,000 synonym sets called Synsets containing about 85000 words. Information in WordNet is organized around logical groupings called synsets. Each synset consists of a list of synonymous word forms and semantic pointers that describe relationships between current synset and other synsets. A word form can be a single word or two or more words connected by underscores, referred as collocations.

III. EVALUATION:

To find the conceptual similarity of two words in a hierarchical semantic network, many ways are available. One such approach is (a) Node-Based (Information content) Approach i.e. to find the conceptual similarity and (b) Edge-Based Approach

Word-based Semantic Similarity:

Many methods have been considered previously to measure conceptual similarity between words Today's feature based similarity model, [Tver77: A. Tversky, “features of similarity”, Psychological Review, 84,(4), 1977, 327 – 352.], is arguably the most powerful similarity model to date. But a much richer knowledge base than available is required for applying to IR. A WordNet – derived knowledge base is helpful in its coverage of concepts, the number of semantic relation types connecting these concepts is considerably less than would be required for use by a feature based similarity model. Experimentation with two approaches to estimate semantic similarity is done.

The information based similarity estimator:

The information based approach to measuring semantic similarityis based on work carried out by Resnick, [{Resn93a, P.Resnik, “selection and Infomation: A class based Approach to lexical relationships”, PhD dissertation at the University of Pennsylvania. Also appears as technical report 93-42, November 1993.}, {Resn93b, P.resnik, “Semantic Classes and syntactic Ambiguity”, ARPA Workshop on Human Language Technology, Princeton, March, 1993.}]. Resnick views noun synsets as a class of words where the class is made up of all words in a synset as well as words in all directly or indirectly subordinate synsets. Conceptual similarity is considered in terms of class similarity. The similarity between two classes is approximated by the information content of the first class in the noun hierarchy that subsumes both classes. The information content of a class is approximated by estimating the probability of occurrence of the class in a large text corpus. In this paper, the probabilities of occurrence of classes were computed from a collection of 1000 noun occurrences from the text of the News papers and NCERT text books.

In a given multidimensional space, a node represents a unique concept consisting of a certain amount of information and an edge represents a direct association between two concepts, the similarity between two concepts is the extent to which they share information in common i.e. Super class in the hierarchy.

Information _Content (IC) = log(1/P(c))-------(5)

where IC is the information content of a concept / class C, P(c) is the probability of encountering an instance of concepts C. i.e. Monotonic as one moves up the hierarchy . As the node's probability increases, IC decreases. Given the monotonic feature of the IC value, the similarity of two concepts can be

$$\mathbf{Sim\,(C1,C2) = \max_{c\,\varepsilon\,sup(C1,C2)}\,[IC(c)] = \max_{c\,\varepsilon\,sup(C1,C2)}\,[-\log P(c)]} \text{ --------- } (6)$$

Where Sup(C1,C2) is the set of concepts that subsume both C and C2. i.e. The node is the lowest upper bound among those that subsume both C1 and C2.

In the case of multiple inheritances, where words can have more than one sense and hence multiple direct classes, word similarity can be

$$\mathbf{Sim(w_1,w_2) = \max_{c_1\,\varepsilon\,sen(w_1)\;c_2\,\varepsilon\,sen(w_2)}\,[\,Sim(C_1,C_2)]} \text{ -------------}(7)$$

where sen(w) denotes the set of possible senses for word w.

Calculation of Class Probabilities:

Class probabilities are used in the determination of the information content or specificity of WordNet classes. The specificity of a class can be defined in terms o its class probability as follows:

Specificity (C_i) = [-log(P(C_i))] , where P(**C_i**) is the class probability of class i.

In order to define the probability of a class we must first define Words(c) and Class (w). Words(c) are defined as the set of words in all directly or indirectly subordinate classes of the class c. For example Words (దేవాలయము) consists of గుడి, దేవాలయము, మందిరము, ఆలయము and దేవాలము. Classes (w) represents the set {c|w € words(c)}, i.e., this includes all the classes in which the word w is contains, regardless of the particular sense of w. From these two definitions we can define the frequency of a class as:

classes (w) = {c|w € words(c)}------------------ (1)

Resnik class **/ concept frequency formula:**

$$\mathbf{freq(c) = \sum_{w \in words(c)} freq(w)} \text{ ----------------------}(2)$$

Richardson and Smeaton (1995) proposed a slightly different calculation by considering the number of word senses factor:

$$\mathbf{freq(c) = \sum_{w \in words(c)} [freq(w)\,] \,/\, [|classes(w)|]} \text{--------}(3)$$

Finally,the class/concept probability can be computed using maximum likelihood estimation(MLE) :

P(c) = freq(c) / N -----------------------------------(4)

Where N is defined as **Σ** freq(c'), i.e. the total size of the sample.

Node based approach is used the Info_content to find the conceptual similarity. The similarity between two concepts is obtained by the degree of sharing information. Edge based approach is based only on the hierarchy or the edge distances. The difficulty with this approach is that the taxonomy arcs represents uniform distances i.e. all the semantic links may have the same weight.

Edge-Based Approach:

The principle of Wu and Palmer of similarity measure is based on the edge counting method can be stated as follows:

Given an ontology ε formed by a set of nodes and a root node(R). C1 and C2 represent two ontology elements for which similarity is measured. The idea of similarity is based on the distance N1 and N2 which separates nodes C1 and C2 from the root node and the distance N which separates the closest common ancestor CS of C1 and C2 from the node R. The similarity measure of Wu and Palmer is defined by the following expression:

Simwp = [2.N] / [N1 + N 2] --------------------- (8)

The similarity values obtained by Wu and Palmer show that the neighbour concepts C2 and C3 are more similar than the concepts C1 and C2 located in the same hierarchy, which is problematic and inadequate within the semantic information retrieval. Adopt a new approach, which is represented by the following formula:

Simibk(C1,C2) = [2.N / N1 + N2] * PF(C1,C2) –(9)
where PF(C1,C2) = (1- λ)(Min(N1,N2) – N) + [1/(λ(|N1- N2| + 1))]
Let PF (C1, C2) be the penalization factor of two concepts C1 and C2 placed in the neighbourhood.
The coefficient λ is a Boolean value indicating ' 0 ' or '1', with ' 0' indicating two concepts in the same hierarchy and '1' indicating two concepts in neighbourhood respectively. Min (N1, N2) represent the minimum value between C1 and C2.

In the formula of PF (C1, C2) '1' is added outside the absolute value for the distance between C1 and C2 i.e., 1 / (|N1- N2|) + 1), because otherwise there could be a division by ' 0 ' in case N1 – N2.

This measure is advantageous because it leads to a lower similarity value for close concepts compared to concepts in the same hierarchy.

Comparison of node-based Information Content model developed by Resnik (1992) and the basic edge-based counting model was performed. For consistency, semantic similarity measures will be used rather than the semantic distance measures.

The noun portion of the WordNet-Telugu was selected as the taxonomy to compute the similarity between concepts. It contains about 66,000 nodes (synsets). The frequencies of concepts were estimated using noun frequencies.

IV. FIGURES AND TABLES

Table-1: Calculation of class probability. Info_content and specificity

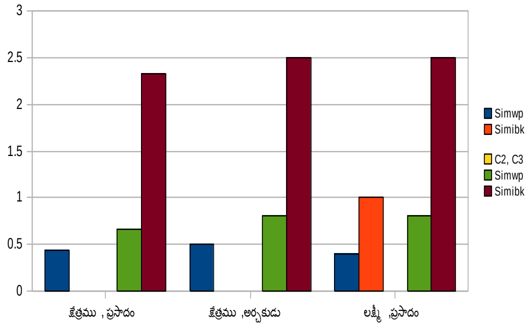
Word w	Freq of word in text	Classes (w)	Σ freq (w)	freq(c)	P(c)	Info_Content= log(1/P(c))	Speificity(C) = -log(P(C))
మండలం	2	3	2	0.666666667	0.000856898	3.067070871	3.067070856
గ్రామము	6	8	14	1.75	0.002249357	2.647941611	2.647941548
జ్వాలా	1	8	1	0.125	0.000160668	3.794070612	3.794069584
క్షేత్రము	3	2	3	1.5	0.001928021	2.71488824	2.714888338
జలపాతము	1	3	3	1	0.001285347	2.890979612	2.890979597
ప్రసాదము	1	2	1	0.5	0.000642674	3.19200927	3.192009593
అర్చకుడు	1	4	3	0.75	0.00096401	3.015918461	3.015918334
ఇతిహాసము	2	3	2	0.666666667	0.000856898	3.067070871	3.067070856
మసీదు	1	1	1	1	0.001285347	2.890979612	2.890979597
లక్ష్మి	1	9	1	0.111111111	0.000142816	3.845223135	3.845222106
కుంకుమ	2	6	2	0.333333333	0.000428449	3.368100866	3.368100852
గుండము	3	4	3	0.75	0.00096401	3.015918461	3.015918334

నగ	2	5	2	0.4	0.000514139	3.288919451	3.288919606
ఉత్సవము	6	2	6	3	0.003856041	2.413858357	2.413858342
చందనము	3	5	3	0.6	0.000771208	3.112828474	3.112828347
హారము	1	9	1	0.111111111	0.000142816	3.845223135	3.845222106
గోపురము	3	2	3	1.5	0.001928021	2.71488824	2.714888338
దేవాలయము	7	5	13	2.6	0.003341902	2.47600629	2.476006249
కొండ	8	5	14	2.8	0.003598972	2.443821532	2.443821566
మఠం	2	0	2	0	0	0	1

Table 2: Experimental results comparing Simibk measure to the Wu Palmer measure Simwp

C1, C2	Simwp	Simibk	C2, C3	Simwp	Simibk
క్షేత్రము, ప్రసాదం	0.44	0	ప్రసాదం,హారము	0.66	2.33
క్షేత్రము,అర్చకుడు	0.5	0	అర్చకుడు,హారము	0.8	2.5
లక్ష్మి ,ప్రసాదం	0.4	1	ప్రసాదం, చందనము	0.8	2.5

Fig 4: Comparative histogram of the effectiveness of our measure compared to the Wu Palmer measure



Couple of concepts Vs relevance

V. Conclusion

The objective of the two approaches is finding the semantic similarity from various ways. The edge-based method is more natural, where as the node-based information content approach is more philosophical. Both have inherent strength and weakness.

The relevance of edge-based measure compared to the Wu and Palmer measure is localized at the level of two concepts located in a hierarchy from which the subsuming concept is different. As the distance between the direct subsuming concepts increases, lower similarity values are obtained. A comparison of relevance of this measure compared to the Wu Palmer measure.

The measure of Wu and Palmer has is simple to implement and have good performances compared to the other similarity measures. The information content method requires less information on the detailed structure of taxonomy. It is still dependent on the skeleton structure of the taxonomy. The drawback of the edge based approach is dependent on the ontology construction and also adopting IS-A ontology, which has some disadvantages such as a similarity value of two elements of an ontology contained in the neighbourhood exceeds the value of similarity of two concepts contained in the same hierarchy. Some of the technical terms are not available since Word Net-Telugu is still under construction. Hence, in this paper a hybrid approach which combines both node-based and edge-based approaches is proposed.

ACKNOWLEDGEMENTS

Word Net-Telugu was developed and provided by Dr. G.Uma Maheswara Rao and his research team from CALTS-HCU, Hyderabad.

References:

1. [Mill a]: George A. Richard Beckwith, Christiane felbaum, Derek Gross, and Katherine Miller, "Introduction to WordNet: An On-line Lexical Database", International Journal of Lexicography, Vol. 3, No. 4, 1990, 235-244.
2. [Mill a]: George A.Miller, "Nouns in WordNet: A lexical Inheritance system", International Journal of Lexicography, Vol. 3, No. 4, 1990, 245-264.
3. [Resn b]: P. Resnik, "Semantic Classes and Syntactic Ambiguity", ARPA workshop on Human Language Technology, Princeton, March, 1993.
4. [Tver] : A. Tversky, " Features of Similarity", Psychological Review, 84, (4), 1977, 327-352.
5. [Ben]: T.Slimani, B. Ben Yaghlane, and K. Mellouli : " A New Similarity Measure based On Edge Counting", World Academy of Science, Engineering and Technology 23, 2006.
6. [Wu]: Z. Wu and M.Palmer,: " Verb Semantics and Lexical Selection", In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp 133-138, 1994.
7. [Resn]: P.Resnik: " Semantic similarity in a Taxonomy: An Information based measure and its application to problems of ambiguity in natural language", Journal of Artificial Intelligence research, 11,pp. 95-130, 1999.
8. [Rada]: R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and application of a metric on semantic nets", IEEE Transaction on systems, Man, and Cybernetics, pp. 17-30, 1989.

Books:

1. **[**Baeza**]:** R. Baeza-Yates, B. Riberio-Neto, "Modern Information retrieval", ACM Press; Addison-Wesley: New York; Harlow, England: Reading, Mass., 1999

Chapters in Books: 2

Theses:

1. [Resn a]: P. Resnik, "Selection and Information: A Class based Approach to Lexical Relationships", PhD dissertation at the University of Pennsylvania. Also appears as Technical Report 93-42, November 1993.
2. [Ray]: Ray Richardson and Alan F. Smeaton, " Using WordNet in a Knowledge-Based Approach to Information retrieval", Dublin University.
3. [Rich]: R.Richardson, "A Semantic based Approach to Information Processing", Ph.D. thesis School of Computer Applications, Dublin University,1994.

Proceedings Papers:

1. [Jay] : Jay J. Jiang, David W. Conrath : " Semantic Similarity Based on Corpus statistical and Lexical Taxonomy", In proceedings of International conference Research on Computational Linguistics (ROCLING X), 1997, Taiwan.
2. [Resn]: P.resnik(1995), " Using Information Content to evaluate semantic similarity in taxonomy", In proceedings of 14th international JointConference on Artificial Intelligence,Montreal, 1995.

Appendix:

The Telugu Text

అహోబిలం కర్నూలు జిల్లా ఆళ్లగడ్డ ***మండలా***నికి చెందిన ***గ్రామం జ్వాలా*** నరసింహ నరసింహస్వామి ***క్షేత్రము*** దగ్గర భవనాశని అనే ***జలపాతము*** ఉంది పానకాలస్వామికి పానకం బెల్లం పంచదార చెరకు అభిషేకం చేస్తే అభిషేకం చేసిన పానకంలో సగం పానకాన్ని స్వామి త్రాగి మిగిలిన సగాన్ని మనకు ***ప్రసాదం*** గా వదిలిపెడతాడుట శివరాత్రి రోజున వంద మంది ***అర్చకుల***తో మహాలింగార్చన జరుపుతారు ***ఇతిహాసాల*** ప్రకారం కావలసిన వరాలను తీర్చే దైవం కాబట్టి అన్న వరం అన్నవరం దేవుడు అంటారు ప్రారంభించారు శ్రీ రాజరాజేశ్వర స్వామి దేవాలయ ప్రాంగణంలో ఏళ్ళ నాటి ***మసీదు*** ఉన్నది శ్రీదేవి భూదేవి మూర్తులకు ***లక్ష్మీ***నామార్చన జరుపుతారు అన్నపూర్ణ ఆలయంలో నిత్యం ***కుంకుమార్చన*** జరుగుతుంది మహానంది ఆలయములో బ్రహ్మ విష్ణు రుద్ర ***గుండాలు*** కలవు గోపన్న దేవునికి రకరకాల ***నగలు***

చింతాకుపతకం పచ్చలపతకం మొదలైనవి చేయించాడు స్వామిలోని వేడిని చల్లార్చడానికి ప్రతీరోజు ***చందనం*** తో పూత పూస్తుంటారు ఇప్పటికీ ఒక పచ్చల ***హారం*** ఆలయంలో ఉంది ఊరిమధ్యనగల భీమేశ్వర ఆలయం ప్రాకారాలతో గాలి ***గోపురా***లతో శివభక్తితో ధీటుగా నిలచింది లక్ష్మీ నరసింహ స్వామి ***దేవాలయము*** చాలా అందంగా శిల్పకళలతో విలసిల్లుతుంది.

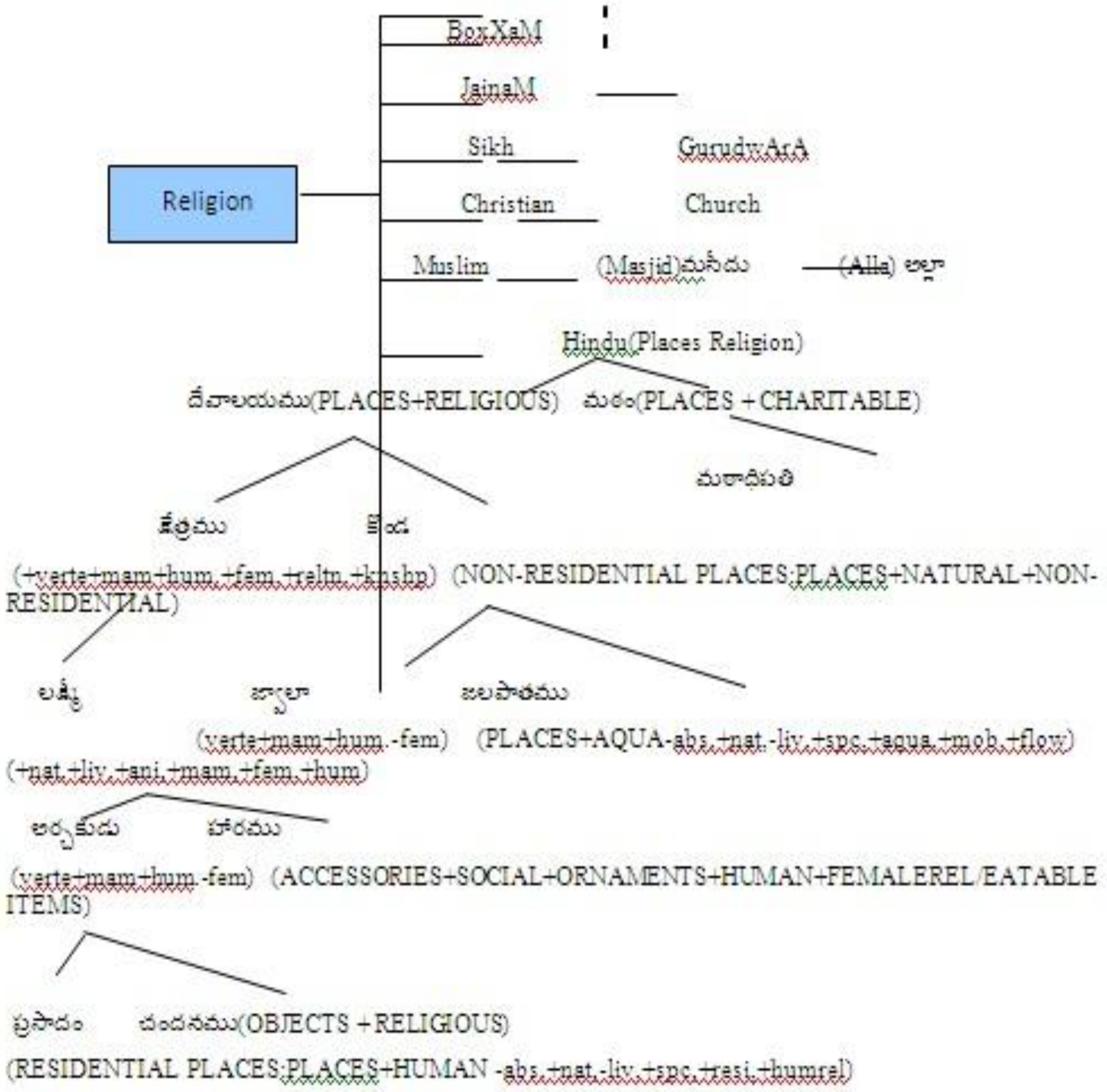

BoxXaM
JainaM
Sikh
GurudwArA
Religion
Christian
Church
Muslim
(Masjid)మసీదు
~~(Alla)~~ అల్లా
Hindu(Places Religion)
దేవాలయము(PLACES+RELIGIOUS)
మఠం(PLACES + CHARITABLE)
మఠాధిపతి
క్షేత్రము
కొండ
(+verte+mam+hum.+fem.+reltn.+knshp)
(NON-RESIDENTIAL PLACES;PLACES+NATURAL+NON-RESIDENTIAL)
లక్ష్మి
జ్వాలా
జలపాతము
(verte+mam+hum.-fem)
(PLACES+AQUA-abs.+nat.-liv.+spc.+aqua.+mob.+flow)
(+nat.+liv.+ani.+mam.+fem.+hum)
అర్చకుడు
హారము
(verte+mam+hum.-fem)
(ACCESSORIES+SOCIAL+ORNAMENTS+HUMAN+FEMALEREL/EATABLE ITEMS)
ప్రసాదం
చందనము(OBJECTS +RELIGIOUS)
(RESIDENTIAL PLACES;PLACES+HUMAN -abs.+nat.-liv.+spc.+resi.+humrel)

Fig: Extract from WordNet illustrating Lexical Inheritance.