Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue12, December-2016

Study of Gujarati WordNet for Semantic and Lexical Relationships

Smt Rekha Manish Shah

Computer Engineering Department, Government Polytechnic, Ahmedabad.

Abstract*—*The Gujarati WordNet is a huge lexical database of Gujarati Language. Gujarati language is one of regional languages of India, mainly spoken in the Gujarat State. This paper mainly focuses on the structure of the Gujarati WordNet. By understanding the WordNet's structure, anyone can efficiently use it for natural language processing like WSD(Word Sense Disambiguation) and computational linguistics. Gujarati WordNet is created from Hindi WordNet or you may say it is an expanded form of Hindi WordNet. Hindi is the national language of India.

Keywords*—*Gujarati WordNet, synset, herpnym, hyponym, meronym, holonym, ontology

I. INTRODUCTION

WordNet is a very useful and huge lexical database of English Language developed by Princeton University. Afterwards many WordNets were developed all across the world for various languages. In India, Indian Institute of Technology, Bombay is the primary institute for developing WordNets for most Indian languages. Gujarati WordNet is developed from Hindi Wordnet using expansion approach. Dharmsinh Desai University was involved in the development of Gujarati Wordnet as a part of INDRADHANUSH project (http://www.cfilt.iitb.ac.in/gujarati/).

WordNet, a computational lexicon, is very useful for many applications like Machine Translation, Word Sense Disambiguation, Semantic Search etc. WordNet is very different from traditional dictionary. It organizes words as a graph so people can make its efficient use for Machine processing.

Gujarati WordNet is not just like a traditional dictionary. However it differs from traditional ones in many ways. For instance, words in this WordNet are arranged semantically instead of alphabetically. Synonymous words are grouped together to form synonym sets, or synsets. Each such synset therefore represents a single distinct sense or concept. Words that belong to several synsets are polysemous or ambiguous. Words with only one sense are said to be monosemous and therefore appear in only one synset. Currently Gujarati WordNet contains 101,614 total words which are organized into 39,490 synsets.

II. STRUCTURE OF WORDNET

The design of WordNet is inspired by Quillian's model of semantic memory(1968). WordNet consists of the network of semantic relations that link various lexicalized concepts. WordNet contains words from four parts–of–speech : nouns, verbs, adjectives and adverbs. Prepositions and conjunctions are not included. The Logical structure of WordNet is as follow:

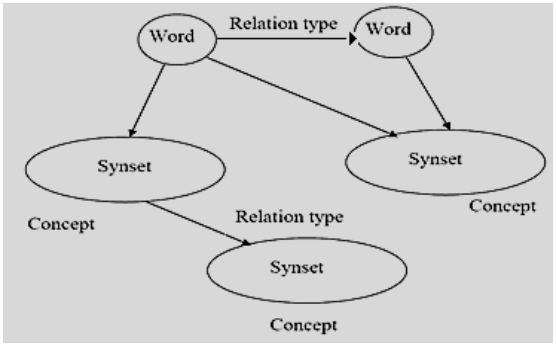

Figure 1. Logical Structure of WordNet

Each synset in WordNet has an associated definition or gloss. This consists of a short entry explaining the meaning of the concept represented by the synset. Many synsets also contain example sentences that show how the words in the synset may be used in Gujarati. Nouns in the WordNet have by far the longest glosses while the glosses of verbs, adjectives and particularly adverbs are quite short. Further, on average verbs have almost twice the number of senses as compared to nouns, adjectives and adverbs. The combination of short glosses and large number of senses make the verbs particularly difficult to disambiguate accurately.

WordNet defines a variety of semantic and lexical relations between words and synsets. Semantic relations define a relationship between two synsets. Lexical relations on the other hand define a relationship between two words within two synsets of WordNet. Thus a semantic relation between two synsets relates all the words in one of the synsets to all the words in the other synset, whereas a lexical relationship exists only between particular words of two synsets. In WordNet,

most relations do not cross part of speech boundaries, so most synsets and words are only related to other synsets and words that belong to the same part of speech.

Each entry in the Gujarati WordNet consists of following elements:

1. **Synset ID** : Each synset has given an unique id. This id can be used to retrieve further information of the synset and related synsets/words from database.
2. **POS** : This indicates which part of speech this synset belongs to. E.g. Noun, Adjective, Adverb or Verb.
3. **Synonyms**: It is a set of similar words. For example, "નદી, સરિતા, તટિની, અગ્રુ, નિમ્નગા, નિઝરણી etc." (nadi, saritaa, tatini, aagru, nimnagaa, nirzarani) represents the concept of river as a large stream of water. The words in the synset are arranged according to the frequency of usage.
4. **Gloss:** It describes the concept of the sysnet. It consists of two parts:

   ***Text definition*:** It explains the concept denoted by the synset. For example, "પાણીનો પ્રાકૃતિક પ્રવાહ જે કોઇ પર્વતમાંથી નિકળીને નિશ્ચિત માર્ગે સમુદ્ર કે બીજા મોટા જળ પ્રવાહને મળે છે" (pani no prakrutik pravah je koi parvat mathi nikli ne nishchit maarge samudra ke bija mota jal pravah ne male chhe.) explains the concept of river as a large stream of water.

   ***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, "ગંગા, યમુના, સરસ્વતી, કાવેરી, સરયુ વગેરે ભારતની મુખ્ય નદીઓ છે" (Ganga, Yamuna, Saraswati, Kaveri, Sarayu vagere bhaarat ni mukhya nadio chhe.) gives the usage for the words in the synset representing river as a large stream.
5. **Gloss in Hindi** : It describes the concept in Hindi language as Hindi is the national language of India. This will help the person/machine to understand the concept in Gujarati language who has knowledge of Hindi language.
6. **Gloss in English**: It describes the concept of the given synset in English language
7. **Relations**: It displays the list of various relations as described below:

Table 1. Various relations in WordNet

Relation	Meaning
Hypernymy/Hyponymy	Is-A (Kind-Of)
Meronymy/Holonymy	Has-A (Part-Whole)
Entailment/Troponymy	Manner-Of (for verbs)

8. **Ontology:** It also displays the Position of the given word 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 house is shown below.

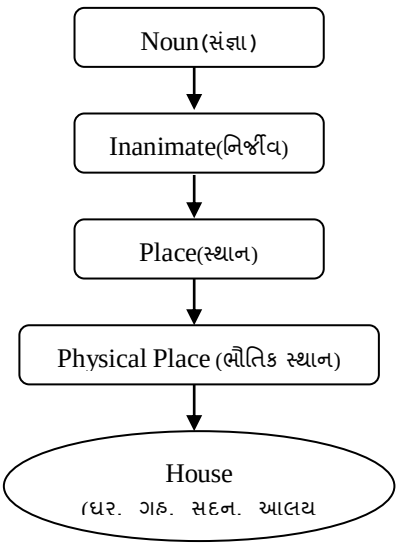

Figure 2. Ontology for the Synset

III. RELATIONSHIPS IN WORDNET

Following are the relationships available for various parts of speech in WordNet:

A. **All parts of speech**

Synonymy: This links words that have similar meanings, e.g. પુત્ર(putra)(son) and દીકરો(dikro)(son)

Antonymy: The opposite of synonymy, e.g સવાર(savaar)(morning) and સાંજ(saanj)(evening).

B. **Nouns**

Hyponymy & Hypernymy(is-a): This refers to a hierarchical relationship between synsets. For example, bird(પક્ષી) is a hypernym of falcon(બાજ) since every falcon(બાજ) is a bird(પક્ષી) (but not vice-versa) and falcon(બાજ) is a hyponym of bird(પક્ષી).

Meronymy & Holonymy(part/whole): It refers to a part/whole relationship. For example, પાંખ(pankh) (wing) is a meronym of પક્ષી(pakshi) (bird), since wing is a part of a bird and પક્ષી(pakshi) (bird) is Holonym of પાંખ(pankh) (wing). Holonyms can be of three types: Member–Of, Substance–Of and Part–Of. Conversely there are three types of meronyms: Has–Member, Has–Substance and Has–Part. Thus a { પાંખ(pankh) (wing)} is a part–of { પક્ષી(pakshi) (bird)}; { લોહી(lohi)(blood)} has–substance {ગ્લુકોજ glucose)}; and an {ટાપૂ(tapu)(island)} is a member–of an {દ્વીપમંડળ (dwipmandal) (archipelago) }.

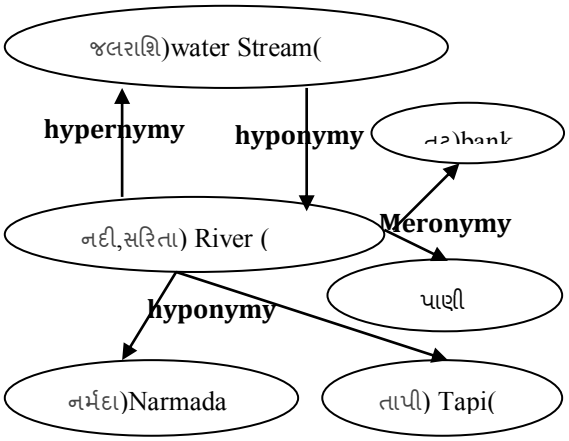

Figure 3. Noun Relationships

Attribute: The attribute relation is a semantic relation that links together a noun synset A with an attribute synset B when B is a value of A. For example, the noun synset {રંગ (rang)(colour)} is related to the adjective synsets {ઘેરો(ghero)(dark)} and {આછો(aachho)(light)} since both {light} and {dark} are "values" of {colour}.

C. **VERBS**

Troponymy: Troponymy is the semantic relationship of doing something in the manner of something else. For example, "ચાલવું(chalvu)(walk)" is a troponym of "ખસવું (khasavu)(move)" and "લંગડાવું(langdavu)(limp)" is a troponym of "ચાલવું(chalvu)(walk)."

Entailment: Entailment refers to the relationship between verbs where doing something requires doing something else. If you are shouting "રાડો પાડવી(rado padavi)" , you must be speaking "બોલવું(bolvu)" so speaking is entailed by shouting.

Cause: Two synsets A and B are related by cause relationship if A *causes* B. For example,{embitter} is related to {resent} because something that embitters causes one to resent.

Also–see: Also–see relationships can be either semantic or lexical in nature.

D. Adjectives and Adverbs

Similar to: This is a semantic relationship that links two adjective synsets that are similar in meaning, but not close enough to be put together in the same synset.

Pertainymy: This is a cross POS relationship. This lexical relationship relates adjectives to other adjectives and nouns, and relates adverbs to adjectives. An adjective A is related to another adjective or to a noun B if A pertains to B. For example the adjective bicentennial pertains to the adjective centennial which pertains to the noun century. For adverbs, pertainym relations link adverbs to adjectives that they pertain to. For example, the adverb animatedly pertains to the adjective animated. Note that this is the third relationship that crosses part of speech boundaries to relate adjectives and adverbs to nouns.

Also–see: semantic in nature

Attribute: The attribute relation is a semantic relation that links together an adjective/advern synset A with an attribute synset B when B is *a value of A.*

Participle of: This is unique to adjectives, and links adjectives to verbs. This is a lexical relationship.

E. Cross-POS relations

The majority of the WordNet's relations connect words from the same part of speech (POS). Thus, WordNet really consists of four sub-nets, one each for nouns, verbs, adjectives and adverbs, with few cross-POS pointers. Cross-POS relations include the "morphosemantic" links that hold among semantically similar words sharing a stem with the same meaning: protect (verb), protecive (adjective) , protection (noun). In many of the noun-verb pairs the semantic role of the noun with respect to the verb has been specified: {swimming_pool} is the LOCATION for {swim} and {tailor}is the AGENT of {sew}, while {sewing, shirt} is its RESULT.

IV. CONCLUSION

Gujarati WordNet can be served as very useful and informative online lexical system. It can be used in various NLP applications like Machine Translation, Word Sense Disambiguation etc.. It serves a useful tool for the person who wants to explore about the Gujarati language. The various relationships that exist in the Gujarati WordNet are very important to retrieve meaningful information from it.

REFERENCES

[1] English WordNet http://WordNet.princeton.edu

[2] Hindi WordNet from http://tdil-dc.in/indowordnet/

[3] Gujarati WordNet _http://tdil-dc.in/indowordnet/

[4] Sinha, Reddy, Bhattacharyya, An Approach towards Construction and Application of Multilingual Indo-WordNet http://www.cse.iitb.ac.in/~pb/papers/gwc06_IITB_IndoWN.pdf

[5] 5 papers on WordNet http://wordnet.princeton.edu/wordnet/related-projects/

[6] Preeti Yadav, Mohd. Shahid Husain "Study of HindiWord Sense Disambiguation Based on HindiWorldNet"

[7] B.A.Sharada Ph.D., Central Institute of Indian Languages ,Manasagangotri, Mysore "Exploring Hindi WordNet as a Lexical Interface and Subject Headings Tool in Library OPAC"

[8] Sudha Bhingardive, Rajita Shukla, Jaya Saraswati, Laxmi Kashyap, Dhirendra Singh and Pushpak Bhattacharyya , Indian Institute of Technology Bombay, India "Synset Ranking of Hindi WordNet"

[9] Poonam Panchal, Namrata Panchal, Harsh Samani, Department of Computer Engineering,KJSIEIT, "Development of Gujarati WordNet for Family of Words"