Development of Gujarati WordNet for Family of Words

Poonam Panchal, Namrata Panchal, Harsh Samani
Department of Computer Engineering,KJSIEIT
Ayurvihar Complex, Everard Nagar, Sion, Mumbai 400022
Maharashtra, India

Abstract* — Gujarati is one of the 22 official languages of India. It is an Indo-Aryan language descended from Sanskrit. Gujarati wordnet is being built using expansion approach with Hindi as the source language. This paper describes experiences of building Gujarati wordnet. Paper discusses basic features of Gujarati language and evaluates suitability of Hindi language for expansion approach. Various issues related to synset linking using expansion approach and challenges related to language specific concepts[5] are also discussed This paper presents the preliminary analysis of Gujarati WordNet and the set of relevant computational tools. Although the design has been inspired by the famous English WordNet, and to certain extent, by the Hindi WordNet, the unique features of Gujarati WordNet are graded antonyms and meronym relationships, nominal as well as verbal compoundings, complex verb constructions. Gujarati WordNet would not only add to the sparse collection of machine-readable Gujarati dictionaries, but also will give new insights into the Gujarati vocabulary.

***Keywords*— WordNet, natural language processing(nlp), hyponym and meronym, morphology and ontology.**

I. INTRODUCTION

WordNets have emerged as a very useful resource for computational linguistics and many natural language processing applications. Since the development of Princeton WordNet (Fellbaum C., 1998), WordNets are being built in many other languages. Hindi WordNet (Narayan D. et al., 2002) was the first WordNet for the Indian languages. Based on Hindi WordNet, WordNets for 17 different Indian languages are getting built using the expansion approach. One such effort is Gujarati WordNet.

WordNet is a lexical database which comprises of synonym sets, gloss, position in relations. A synonym set in a WordNet represents some lexical concept (Miller, 1993). The gloss gives definition of the underlying lexical concept and an example sentence to illustrate the concept. 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. The WordNet also maintains semantic an lexical relations. Semantic relations are between synsets and lexical relations are between words. Semantic relations are Hyponymy, Hypernymy, Meronymy, Holonymy etc. Lexical relations are antonomy and such (Bhattacharyya, 2010). Thus we can say a WordNet is a dictionary plus a thesaurus and much more.

II. BACKGROUND AND MOTIVATION

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.

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.
Each entry in the Hindi WordNet consists of following elements:

1. 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.
2. 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 पहेली 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.

3. 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.

III. PROBLEM STATEMENT

Based on the English and Hindi WordNet we are developing Gujarati WordNet. Gujarati WordNet contains Gujarati words used in a family‘s day to day life. It groups words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synsets or their members.

IV. REVIEW OF LITERATURE

A. The Lexical Matrix

Lexical semantics begins with a recognition that a word is a conventional association between a lexicalized concept and an utterance that plays a syntactic role. This definition of ‘word’ raises at least three classes of problems for research. First, what kinds of utterances enter into these lexical associations? Second, what is the nature and organization of the lexicalized concepts that words can express? Third, what syntactic roles do different words play? Although it is impossible to ignore any of these questions while considering only one, the emphasis here will be on the second class of problems, those dealing with the semantic structure of the English lexicon.[1] Since the word word‘‘ is commonly used to refer both to the utterance and to its associated concept, discussions of this lexical association are vulnerable to terminological confusion. In order to reduce ambiguity, therefore, word form will be used here to refer to the physical utterance or inscription and word meaning to refer to the lexicalized concept that a form can be used to express. Then the starting point for lexical semantics can be said to be the mapping between forms and meanings (Miller, 1986). A conservative initial assumption is that different syntactic categories of words may have different kinds of mappings. Mappings between forms and meanings are many:many—some forms have several different meanings, and some meanings can be expressed by several different forms. Two difficult problems of lexicography, polysemy and synonymy, can be viewed as complementary aspects of this mapping. That is to say, polysemy and synonymy are problems that arise in the course of gaining access to information in the mental lexicon:a listener or reader who recognizes a form must cope with its polysemy; a speaker or writer who hopes to express a meaning must decide between synonyms.[1,2]

B. Relations in WordNet

1. Synonymy: From what has already been said, it should be obvious that the most important relation for WordNet is similarity of meaning, since the ability to judge that relation between word forms is a prerequisite for the representation of meanings in a lexical matrix. According to one definition two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made. By that definition, true synonyms are rare, if they exist at all. A weakened version of this definition would make synonymy relative to a context: two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value. For example, the substitution of plank for board will seldom alter truth values in carpentry contexts, although there are other contexts of board where that substitution would be totally inappropriate.[1] Note that the definition of synonymy in terms of substitutability makes it necessary to partition WordNet into nouns, verbs, adjectives, and adverbs. That is to say, if concepts are represented by synsets, and if synonyms must be interchangeable, then words in different syntactic categories cannot be synonyms (cannot form synsets) because they are not interchangeable.

2.Antonymy*:* Another familiar relation is antonymy, which turns out to be surprisingly difficult to define. The antonym of a word x is sometimes not-x, but not always. For example, rich and poor are antonyms, but to say that someone is not rich does not imply that they must be poor; many people consider themselves neither rich nor poor. Antonymy, which seems to be a simple symmetric relation, is actually quite complex, yet speakers of English have little difficulty recognizing antonyms when they see them. Antonymy is a lexical relation between word forms, not a semantic relation between word meanings. For example, the meanings {rise, ascend} and {fall, descend} may be conceptual opposites, but they are not antonyms; [rise/fall] are antonyms and so are [ascend/descend], but most people hesitate and look thoughtful when asked if rise and descend, or ascend and fall, are antonyms. Such facts make apparent the need to distinguish between semantic relations between word forms and semantic relations between word meanings. Antonymy provides a central organizing principle for the adjectives and adverbs in WordNet, and the complications that arise from the fact that antonymy is a semantic relation between words are better discussed in that context.[1]

3. Hyponymy: Unlike synonymy and antonymy, which are lexical relations between word forms, hyponymy/hypernymy is a semantic relation between word meanings: e.g., {maple} is a hyponym of {tree}, and {tree} is a hyponym of {plant}. Much attention has been devoted to hyponymy/hypernymy (variously called subordination or superordination, subset/superset, or the ISA relation).A concept represented by the synset {x, x , ...} is said to be a hyponym of the concept represented by the synset {y, y , ...} if native speakers of English accept sentences constructed from such frames as An x is a (kind of) y. The relation can be represented by including in {x, x , . . .} a pointer to its superordinate, and including in {y, y , . . .} pointers to its hyponyms. Hyponymy is transitive and asymmetrical (Lyons, 1977, vol. 111), and, since there is normally a single superordinate, it generates a hierarchical semantic structure, in which a hyponym is said to be below its superordinate. Such hierarchical representations are widely used in the construction of information retrieval systems, where they are called inheritance systems (Touretzky, 1986): a hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that

superordinate. For example, maple inherits the features of its superordinate, tree, but is distinguished from other trees by the hardness of its wood, the shape of its leaves, the use of its sap for syrup, etc. This convention provides the central organizing principle for the nouns in WordNet.[1]

4 Meronymy: Synonymy, antonymy, and hyponymy are familiar relations. They apply widely throughout the lexicon and people do not need special training in linguistics in order to appreciate them. Another relation sharing these advantages—a semantic relation—is the part-whole (or HASA) relation, known to lexical semanticists as meronymy/ holonymy. A concept represented by the synset {x, x , . . .} is a meronym of concept represented by the synset {y, y , . . .} if native speakers of English accept sentences constructed from such frames as A y has an x (as a part) or An x is a part of y. The meronymic relation is transitive (with qualifications) and asymmetrical (Cruse, 1986), and can be used to construct a part hierarchy (with some reservations, since a meronym can have many olonyms). It will be assumed that the concept of a part of a whole can be a part of a concept of the whole, although it is recognized that the implications of this assumption deserve more discussion than they will receive here. These and other similar relations serve to organize the mental lexicon.[1]They can be represented in WordNet by parenthetical groupings or by pointers (labeled arcs) from one synset to another. These relations represent associations that form a complex network; knowing where a word is situated in that network is an important part of knowing the word‘s meaning. It is not profitable to discuss these relations in the abstract, however, because they play different roles in organizing the lexical knowledge associated with different syntactic categories.[1]

C. Techniques for WordNet

1. Part Of Speech Tagging: We present a new part-of-speech tagger that demonstrates the following ideas:
(i) explicit use of both preceding and following tag contexts via a dependency network representation,
(ii) broad use of lexical features, including jointly conditioning on multiple consecutive words,
(iii) effective use of priors in conditional loglinear models, and
(iv) fine-grained modeling of unknown word features.

Using these ideas together, the resulting tagger gives a 97.24%accuracy on the Penn TreebankWSJ,an error reduction of 4.4% on the best previous single automatically learned tagging result. Almost all approaches to sequence problems such as part of speech tagging take a unidirectional approach to conditioning inference along the sequence. Regardless of whether one is using HMMs, maximum entropy conditional sequence models, or other techniques like decision trees, most systems work in one direction through the sequence (normally left to right, but occasionally right to left, e.g., Church (1988)). There are a few exceptions, such as Brill‘s transformationbased learning (Brill, 1995), but most of the best known and most successful approaches of recent years have been unidirectional. Most sequence models can be seen as chaining together the scores or decisions from successive local models to form a global model for an entire sequence. Clearly the identity of a tag is correlated with both past and future tags‘ identities. However, in the unidirectional (causal) case, only one direction of influence is explicitly considered at each local point. In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context— i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc. Once performed by hand, POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags. POS-tagging algorithms fall into two distinctive groups: rule-based and stochastic. E. Brill's tagger, one of the first and widely used English POS-taggers, employs rule-based algorithms.

Example: મારું નામ હર્ષ છે
(My name is Harsh)
In this sentence:

હર્ષ- નામ (noun)

મારું- વિશેષણ (adjective)

નામ - ક્રિયાવિશેષણ (adverb)

છે - નિર્ણાયક (determinant)

2 Inflection: In grammar,inflection or inflexion is the modification of a word to express different grammatical categories such as tense, mood, voice, aspect, person, number, gender and case. The inflection of verbs is also called conjugation, and the inflection of nouns, adjectives and pronouns is also called declension. An inflection expresses one or more grammatical categories with a prefix, suffix or infix, or another internal modification such as a vowel change. For example, the Latin verbducam, meaning "I will lead", includes the suffix -am, expressing person (first), number (singular), and tense (future). The use of this suffix is an inflection. In contrast, in the English clause "I will lead", the word lead is not inflected for any of person, number, or tense; it is simply the bare form of a verb. The inflected form of a word often contains both a free morpheme (a unit of meaning which can stand by itself as a word), and a bound morpheme (a unit of meaning which cannot stand alone as a word). For example, the English word cars is a noun that is inflected for number, specifically to express the plural; the content morpheme car is unbound because it could stand alone as a word, while the suffix -s is bound because it cannot stand alone as a word.These two morphemes together form the inflected word cars[3]. Words that are never subject to inflection are said to be invariant; for example, the English verb must is an invariant item: it never takes a suffix or changes form to signify a different grammatical category. Its categories can be determined only from its context.

Example: વિશ્વસનીય

વિશ્વસ– Root word

નીય- Inflection added

3.Stemmer: In linguistic morphology and information retrieval, stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form— generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root, Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation. Stemming programs are commonly referred to as stemming algorithms or stemmers. Stemmer means stripping off affixes (suffix, prefix, infix, circumfix). In linguistic morphology and information retrieval, **stemming** is the process for reducing inflected (or sometimes derived) words to their stem, base or root form— generally a written word form. linguistics, **morphology** is the identification, analysis, and description of the structure of a given language's morphemes Example:

1. ચાલવું (Walking)

ચાલ – Stem or the root word

વું - Suffix added to the stem

2. અપૂ ર્ણ(Incomplete)

પૂ ર્ણ- Stem or the root word

અ - Prefix added to the stem

4.Morphological Parsing: The process of breaking down an inflected word into morphemes and the inflections is called as morphological parsing. Morphology is the study of the way words are built up from smaller meaning bearing units called as morphemes. An important class of lexical relations are the morphological relations between word forms. Initially, interest was limited to semantic relations; no plans were made to include morphological relations in WordNet.

As work progressed, however, it became increasingly obvious that if WordNet was to be of any practical use to anyone, it would have to deal with inflectional morphology. For example, if someone put the computer's cursor on the word trees and clicked a request for information, WordNet should not reply that the word was not in the database. A program was needed to strip off the plural suffix and then to look up tree, which certainly is in the database. This need led to the development of a program for dealing with inflectional morphology. Although the inflectional morphology of English is relatively simple, writing a computer program to deal with it proved to be a more complex task than had been expected. Verbs are the major problem, of course, since there are four forms and many irregular verbs. But the software has been written and is presently available as part of the interface between the lexical database and the user.

Example: છોકરીઓ in this

છોકરી – Morpheme

ઓ - inflection

V. CONCLUSION

Existence of Hindi wordnet and similarity between Hindi and Gujarati languages helped development of Gujarati wordnet. Also, the resources like 'Bhagavad-Go-Mandal' and 'Gujarati Lexicon' were found to be very useful in synset development process. Synset categorization further simplifies the synset linking process. It is observed that most of the top level concepts are common and easily linked. The concepts that vary across languages are specific to culture and tradition of the people. Mostly these are noun concepts and do not have hyponymy. Many of these are singleton synsets that appear very low in the wordnet concept hierarchy. The future work is to identify and link language specific and in-family concepts. It is also required to develop lexical relations and for the suitability of semantic relations of Hindi wordnet for Gujarati language[5].

VI. REFERENCES

[1] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller, ―Introduction to WordNet: An On-line Lexical Database‖, CSL Report 43, Princeton University Cognitive Science Laboratory,1990 (Revised August, 1993).

[2] D. Narayan, D. Chakrabarty, P. Pande, P. Bhattacharyya, ―An Experience in Building the Indo-WordNet – A WordNet for Hindi‖, In Proceedings of the First International Conference on Global WordNet (GWC 02), Mysore, India, 2002.

[3] M. Sinha, M. Reddy, P. Bhattacharyya, ―An Approach Towards Construction and Application of Multilingual Indo-WordNet, In Proceedings of the 3rd Global Wordnet Conference (GWC 05), Jeju Island, Korea, 2006.

[4] Arindam Chatterjee, Salil Rajeev Joshi, Mitesh M. Khapra, Pushpak Bhattacharyya, ―Introduction to Tools for IndoWordNet and Word Sense Disambiguation**",** Department of Computer Science and Engineering, Indian Institute of Technology Bombay Powai, Mumbai-400076 Maharashtra, India.

[5] http://www.cse.iitb.ac.in/~pb/papers/gwc12-gujarati-wn.pdf ‖Introduction to Gujarati wordnet (GCW12) IIT Bombay, Powai, Mumbai-400076 Maharashtra, India.