Experiences in building the Indo WordNet – A WordNet for Hindi

Debasri Chakrabarti
Dipak Kumar Narayan
Prabhakar Pandey
Pushpak Bhattacharyya

Abstract

It is increasingly being understood by the practioners of information retrieval, natural language processing and knowledge engineers that a rich lexical knowledge base is the heart of any intelligent information processing system. The words of a language are extremely powerful units that bind together in extensive and unique ways to create a knowledge web. Indo WordNet – as we call the WordNet for Hindi – is an on-line lexical database. It is an attempt to build a lexical reference system for Hindi language. The design has been inspired by the famous English WordNet. For each word we find the synonym set, representing one lexical concept. These synonym sets are linked with other synonym sets through the well–known semantic relationships of hypernymy, hyponymy, meronymy, holonymy, antonymy and so on. The Indo WordNet has some unique features like graded antonymy and meronymy. It also addresses the unique Indian language phenomenon like causative, compound and conjunct verbs, both at the conceptual level and the implementation. It has an efficient underlying database design. The web interface for querying the Indo WordNet has been implemented using Php4 scripting language. The data entry interface, implemented using Java/Jfc is also simple and elegant.

Keywords: WordNet, Semantic Relations, Hindi language, Lexical database

1 Introduction

The Indo WordNet is a system for bringing together different lexical relationships between the Hindi words. It organizes the lexical information in terms of word meanings. It can be said to be a dictionary based on psycholinguistic principles [Fel98]. In any WordNet each word is organized into synonym sets or *synsets*. This is done to remove ambiguity in cases where a single word has multiple meanings. Thus, the synset serves as an identifying definition of lexical concepts. For example, the two synsets {घर, कमरा} {*house, room*}, {घर, गृह} {*house, home*} can serve as an unambiguous differentiator of these two meanings of {घर} {*house*} [Var98] [KK97].

The WordNet is actually a word sense network. A word sense node in this network is a *synset* which is regarded as a basic object in the WordNet. All word sense nodes are linked by various semantic relations, such as synonymy, *IS–A* (hypernymy/hyponymy), *PART–OF* (meronymy/holonymy) *etc.* These relations serve to organize the lexical knowledge base.

2 Semantic Relations

We have implemented all the common semantic relations like synonymy, hypernymy, hyponymy, meronymy, holonymy, troponymy, entailment *etc.* [Fel98], [JNPB01]. The concept of *gradation* is implemented for the noun, adjective, verb and adverb categories. Compounding of verbs is a typical Indian language phenomenon. This along with the structure of verbs and their categorization have been incorporated in the Indo WordNet.

The English WordNet describes seven types of part-whole relationship and have implemented three types of relations *e.g. component–object, member–collection* and *stuff–object*. EuroWordNet [Vos99] has limited the part-whole relations to five types *e.g. component–object, portion and the whole from which it has been detached, a place and a wider place which includes it, a set and their members* and *a thing and the substance it is made of.* The Indo WordNet implements part-whole relationship in eight different ways as shown in Table 2.

We have implemented three types of antonymy relations, *viz*, gradable antonym {गर्म– ठंडा} {*hot – cold*}, complementary antonym {जीवित– मृत, {*alive–dead*}, and converse antonym {लेना– देना} {*give–take*}.

Table 1: Antonyms can be formed on the basis of the following criteria

Size	छोटा– बड़ा, मोटा– पतला	big – small, thick – thin
Quality	अच्छा– बुरा, प्यार– घृणा	good – bad, love – hatred
Gender	बेटा– बेटी, माता– पिता	son – daughter, father – mother
State	शुरू– अंत	begining – end
Personality	राम– रावण, डेविड– गोलिएथ	Rama – Ravana, David – Goliath
Direction	पूर्व– पश्चिम, आगे– पीछे	east – west, front – behind
Action	लेना– देना, खरीद– बिक्री	give – take, buy – sell
Amount	कम– ज्यादा, हल्का– भारी	little – much, light – heavy
Place	दूर– पास	far – near
Time	दिन– रात, सुबह– शाम	day – night, morning – evening

Table 2: Meronymy Relationship

Component–object	माथा– शरीर	head – body
Member–Collection	पेड़– जंगल	tree – jungle
Feature-Activity	भाषण– समारोह	lecture – ceremony
Place-Area	दिल्ली– भारत	Delhi – India
Phase-State	जवानी– उम्र	youth – age
Portion-mass	ढेला– मिट्टी	lump – clay
Resource-process	कलम– लेखन	pen – writing
Position-Area	चिकित्सक– चिकित्सा	doctor – medical treatment

2.1 *Gradation*

Gradation is also a type of semantic relation. The Indo WordNet deals elaborately with the interpretation of Gradation in different domains as shown in Table 3. Gradation is possible only when there are two opposite terms, for *example,* in the diagram shown in Fig 1, the two terms A and B which are in contrast with each other, open a possibility for other terms which may fall between them. We can call A and B as positive and negative respectively between which there is a neutral term, which can be called *Intermediate–term.* Intermediate–term is the core of gradation. In the example shown in Fig 1, {नम} {*damp*} is the Intermediate–term which is the blend of {सूखा} {*dry*} as well as {गीला} {*wet*}. A term which may also occur between (+1) and (0), is called Pre-Intermediate term which has been denoted by (+1/2), for *example* {शुष्क} {*arid*}. On the otherhand, there is a term between (0) and (-1) called Post-Intermediate term, like {तर} {*moist*}. Here it is noteworthy that Pre and Post Intermediate terms can also be antonymous to each other, as in {शुष्क–तर} {*arid–moist*}, but the antonym of Intermediate term is not possible.

Figure 1: Gradation levels

A (+1)	(+1/2)	(0)	(-1/2)	B (-1)
सूखा	शुष्क	नम	तर	गीला
	↓	|	↓	
	Pre-Intermediate	↓	Post-Intermediate	
		Intermediate		

Table 3: Gradation Relationship

State	बचपन, जवानी, बुढ़ापा	childhood, youth, old–age
Size	बड़ा, मँझला, छोटा	large, medium, small
Light	उजाला, धुँधला, अँधेरा	bright, foggy, dark
Gender	मर्द, नपुंसक, औरत	male, eunuch, female
Temperature	गरम, गुनगुना, ठंडा	hot, lukewarm, cold
Colour	गोरा, साँवला, काला	fair, wheatish, dark
Time	दिन, गोधूलि, रात	day, dusk, night
Quality	अच्छा, सामान्य, खराब	good, ordinary, bad
Action	सोना, ऊँघना, जागना	sleep, drowsy, awake
Manner	तेजी से, मध्यम गति से, धीरे–धीरे	speedly, medium, slowly

2.2 *Verbs in the Indo WordNet*

Hindi verbs are varied in nature. The Indo WordNet tries to give much semantic and syntactic information about the Hindi verbs. An effort has been made to examine the entire verb lexicon by classifying them according to their structure and semantic domain [Sin85] [Bah97].

2.2.1 Simple Verb {सरल धातु क्रिया}

Simple verbs are the verbs made of one root verb. For *example*, {खाना} {*to eat*}, {सोना} {*to sleep*}.

2.2.2 Compound Verb {संयुक्त क्रिया}

These verbs are made up from either another verb or other part–of–speech. Accordingly they can be classified as

- Verbs made with Nouns : for *example*, {आरम्भ करना} {*to begin*}, {प्रहार सहना} {*to bear a blow*}.
- Verbs made with Adjectives : for *example*, {मीठा लगना} {*to taste sweet*}, {उत्पन्न होना} {*to be born*}.
- Verbs made with Adverbs : for *example*, {दूर हटना} {*to go away*}, {आगे बढ़ना} {*to go forward*}.
- Verbs made with Verbs : for *example*, {खा चुकना} {*to have eaten*}, {जा बैठना} {*to go and sit*}.

2.2.3 Combination Verb {सामासिक क्रिया}

These verbs are made up of two such root verbs which are related to each other from the point of view of meaning. For *example*, {पढ़ना– लिखना} {*to read and write*}, {खाना–पीना} {*to eat and drink*}.

2.2.4 Onomatopoeic Verb {अनुकरणात्मक क्रिया}

These verbs are made from the Onomatopoeic words. For *example*, {खटखटाना} {*to knock*} is derived from the sound {खटखट}, and {गुनगुनाना} {*to hum*} is made from the sound {गुनगुन}.

2.2.5 Conjunct Verb {यौगिक क्रिया}

These are made up of two root verbs where the sense of {कर} {*action*} is hidden. For *example*, {ले जाना} {*to take away*} it means {ले कर जाना} {*to take away and go*}.

2.3 *Causative Verbs* {प्रेरणार्थक क्रिया}

A verb that denotes causing something to happen is a causative verb. According to the structure, causative verbs are of two types, one is the {प्रथम प्रेरणार्थक क्रिया} *first causative verb* and the other is the {द्वितीय प्रेरणार्थक क्रिया} *second causative verb* [Sin85]. For *example*, {सुलाना} {*to make somebody sleep*}, {सुलवाना} {*to make somebody sleep through the effort of third person*}. {माँ बच्चे को सुला रही है} {*the mother is putting the baby to sleep*}, {माँ बच्चे को सीता से सुलवा रही है} {*the mother is having Sita to put the baby to sleep*}.

This theory of verbs has been implemented in the Indo WordNet by introducing three new semantic relations called *compound, causative,* and *conjunct*. Thus, *for example*, there is a bidirectional *compound* relation between {पढ़ना} {*to read*} and {लिखना} {*to write*}, a unidirectional *causative* relation from {सुलाना} {*to make somebody sleep*} to {सुलवाना} {*to make somebody sleep through the effort of third person*} and a bidirectional *conjunct* relation between {दूर} {*far*} – {हटना} {*to move*}.

3 Database

The design of the Indo WordNet is based on the fact that basic relations or lexical links are between synonym sets. The lexical database is stored in MySQL database [MyS00]. The synsets and the semantic links between them are grouped according to different grammatical categories like *noun, adjective, verb, adverb* and are stored in separate tables. For all categories the word and its polysemy are stored in the table *tbl_synset_data*. Table 4 gives the description of the database tables. Each column heading gives the

Table 4: **Tables in the Database**

tbl_synset_data	**tbl_noun_data**	**tbl_verb_data**	**tbl_adj_data**	**tbl_adv_data**
synset_id	synset_id	synset_id	synset_id	synset_id
head_word	hypernyms	hypernyms	antonyms	antonyms
synset	hyponym	troponyms	similar	derived_from
gloss	meronyms	entailments	attributes	gradation
onto_node_id	antonyms	gradation	gradation	
category	gradation	causative		
		compound		
		conjunct		

table-name in the database and the values below them give the *field-names* in that *table*.

4 Conclusions, Observations and Results

We have described in this paper our experiences with the building of a WordNet for Hindi which we call the Indo WordNet. After thoroughly examining the unique features of Hindi as a language, we have incorporated these features in the Indo WordNet. All the basic semantic relations like synonymy, hypernymy, hyponymy, meronymy, holonymy, troponymy, entailment *etc.* have been included. Additionally, the Indo WordNet gives finer shades of antonymy and meronymy relations. Compounding is a typical phenomenon of the Indian languages. Different types of nominal and verbal compounds are found in most of the Indian languages. The Indo WordNet implements compounding which is useful for the automatic analysis of Hindi.

The system also attempts to integrate Ontology by linking synsets to the ontological nodes. The integration of lexicon, WordNet and Ontology is expected to create an extremely rich and vast lexical resource which will support the extensive work on automatic analysis and generation of Hindi going on in IIT Bombay. This is consistent with the current thinking on solving the difficult problems like word sense disambiguation in natural language understanding, machine translation and information retrieval systems.

The data entry graphical user interface has been designed using Java/Jfc. It provides a user friendly method to insert the synsets and the semantic relations into the database. We have implemented an *organizer* utility to normalize and verify the consistency of the database. We have used ISCII font [1] for

[1] *http://www.tdil.gov.in/standards.htm*

data storage and DV-TT-Yogesh font [2] for the display of data in Devanagari text. The web interface to the Indo WordNet has been implemented. The site address is *http://www.cse.iitb.ac.in/meru/wn/webhwn/*. The web interface uses *Php4.0* scripting language to process the search query. At present the Indo WordNet contains over 7000 synsets. The distribution according to the POS is 60% noun, 25% adjective, 10% verb, 5% adverb synsets.

When completely implemented the Indo WordNet will be a mile stone in the computational processing of Indian languages. The basic design and implementation strategy can be applied to almost all Indian languages because of their unified underlying structure.

Acknowledgement

This research is supported by a grant from Ministry of Information Technology, Government of India, New Delhi.

Authors

Debasri Chakrabarti, Dipak Kumar Narayan, Prabhakar Pandey and Pushpak Bhattacharyya[3], Computer Science and Engineering Department, Indian Institute of Technology – Bombay, India.

References

[Bah97] Hardev Bahari. *Vyavaharik Hindi Vyakaran Tatha Rachna.* Lokbharti Prakashan, Allahabad, India, 1997.

[Fel98] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.

[JNPB01] Sanjay Kumar Jha, Dipak K. Narayan, Prabhakar Pandey, and Pushpak Bhattacharya. A Wordnet for Hindi. In *Workshop on Lexical Resourses in Natural Language Processing.* Hyderabad, India, January 2001.

[KK97] Arvind Kumar and Kusum Kumar. *Samantar Kosh Vol. 1 and 2.* National Book Trust, India, 1997.

[MyS00] MySQL. *Database for Linux.* http://www.mysql.com/, 2000.

[2] *http://www.mit.gov.in/hindi/loadfonts.htm*

[3] pb@cse.iitb.ac.in

[Sin85] Suraj Bhan Singh. *Hindi ka Vakyatmaka Vyakarana*. Sahitya Sahakar, Delhi, India, 1985.

[Var98] Ramchandra Varma. *Pramanika Hindi Kosh*. Lokabharati Prakashan, Allahabad, 1998.

[Vos99] Piek Vossen, editor. *EuroWordNet General Document*. University of Amsterdam, http://www.hum.uva.nl/~ewn, 1999.