IndoWordnet

Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
E-mail: pb@cse.iitb.ac.in

Abstract

India is a multilingual country where machine translation and cross lingual search are highly relevant problems. These problems require large resources- like wordnets and lexicons- of high quality and coverage. Wordnets are lexical structures composed of synsets and semantic relations. Synsets are sets of *synonyms.* They are linked by semantic relations like *hypernymy (is-a), meronymy (part-of), troponymy (manner-of) etc.* IndoWordnet is a linked structure of wordnets of major Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan families. These wordnets have been created by following the expansion approach from Hindi wordnet which was made available free for research in 2006. Since then a number of Indian languages have been creating their wordnets. In this paper we discuss the methodology, coverage, important considerations and multifarious benefits of IndoWordnet. Case studies are provided for Marathi, Sanskrit, Bodo and Telugu, to bring out the basic methodology of and challenges involved in the expansion approach. The guidelines the lexicographers follow for wordnet construction are enumerated. The difference between IndoWordnet and EuroWordnet also is discussed.

1. Introduction

Wordnets have emerged as crucial resources for Natural Language Processing (NLP). Wordnets are lexical structures composed of synsets and semantic relations (Fellbaum, 1998). Synsets are sets of *synonyms.* They are linked by semantic relations like *hypernymy (is-a), meronymy (part-of), troponymy (manner-of) etc.* The first wordnet in the world was built for English at Princeton University [1] . Then followed wordnets for European Languages: Eurowordnet[2] (Vossen, 1998). Since 2000, wordnets for a number of Indian languages are getting built, led by the Hindi wordnet[3] (Narayan *et. al.*, 2001) effort at Indian Institute of Technology Bombay[4] (IITB).

In wordnet creation, the focus shifts from words to concepts. For example, सूर्य (Sun), पृथ्वी (Earth), जल, पानी (Water) *etc.* are very common concepts. After selecting a concept, all the words standing for that concept are stored as the set of synonymous words.

In what follows we first describe the general methodology used in wordnet construction in section 2. The points made therein are substantiated through a case study of Hindi and Marathi wordnets construction in section 3. Section 4 is on the process details of IndoWordnet construction. Section 5 describes the experiences of a few Indian languages in expanding from Hindi wordnet. Section 6 enumerates some guiding principles of IndoWordnet construction. Section 7 is on difference between IndoWordnet (IWN) and EuroWordnet (EWN). Section 8 concludes the paper and points to future directions.

2. General methodology for wordnet creation

The foundation of wordnet construction is *relational semantics* (Cruse, 1986). Words and concepts can be looked upon as forming entries in a structure called the **Lexical Matrix.** Table 1 illustrates this. In the lexical matrix rows represent *word meanings* and columns the *forms*. For example, in Table 1, the column F_2 shows different meanings of *bank*, *i.e.*, the polysemy of *bank*, while the rows M_1 and M_2 show different synonyms of *bank*.

Word Meanings ↓	Word Forms → F_1	F_2	F_3	…	F_k
M_1	*depend*	*bank*	*Rely*		
M_2		*bank*		*embankment*	
M_3					
…		*bank*			
M_n					

Table 1: Lexical matrix showing the word *bank*

It is clear from the presence of other words in the same row (*e.g., depend* in M_1 and *embankment* in M_2) what these meanings or senses are. This is the principle of relational semantics. Words when put together in a common set disambiguate each other. Such sets are known as synsets.

There are three principles the synset construction process must adhere to. **Minimality** principle insists on capturing that minimal set of the words in the synset which uniquely identifies the concept. For example *{family, house}* uniquely identifies a concept (*e.g. "he is from the house of the King of Jaipur"}*. **Coverage** principle then stresses on the completion of the synset, *i.e.*, capturing ALL the words that stand for the concept expressed by the synset (*e.g., {family, house, household, ménage} completes the synset*). Within the synset the words should be ordered according their frequency in the corpus. **Replaceability** demands that the most common words in the synset, *i.e.*, words towards the beginning of the synset should be able to replace one another in the example sentence associated with the synset.

Wordnets are constructed by following either the *merge approach* or the *expansion approach* (Vossen, 1998). In the former- which can be said to be wordnet construction from *first principles*- exhaustive sense repository of each word is first recorded. Then the lexicographers constructs a synset for each sense, obeying the above three principles. In the expansion approach, the synsets of the wordnet of a given source language L_S are provided. Each synset is carefully studied for its meaning. Then the words of the

[1] http://www.wordnet.princeton.edu
[2] http:// http://www.illc.uva.nl/EuroWordNet/
[3] http://www.cfilt.iitb.ac.in/wordnet/webhwn
[4] http://www.iitb.ac.in

target language L_T, representing that meaning are collected and put together in a set in frequency order.

2.1 Comparing *merge* and *expansion* approaches to wordnet building

Both the merge and expansion approaches have their advantages and disadvantages. In the former, there is no *distracting influence* of another language, which happens when the lexicographer encounters culture and region specific concepts of the source language. The quality of the wordnet is good, provided the synset maker is well versed with the nuances of the language. But the process is typically slow. In the latter approach, the whole wordnet making process is well guided in the sense of following the synsets of the source language. Also it has the advantage of being able to borrow the semantic relations of the given wordnet. This saves an enormous amount of time. However, the lexicographer oftentimes is distracted by synsets standing for highly culture and region specific concepts. Also common is the problem of not finding the target language's *"own concepts"*.

One finds the predominance of the expansion approach in the wordnet building community. Many concepts are common across languages. Creating synsets for these *universal concepts* should be the first step in the construction of any wordnet. If a language has already done this job, it makes sense to leverage from this work. This fact and the fact of being able to borrow the semantic relations from the source language tilt the balance in favour of the expansion approach. If the source and target languages happen to have strong kinship relationship, the expansion approach becomes all the more attractive, since distracting influences of culture and region specific concepts is minimal in this case.

In the next section, we present a case study to elucidate the above ideas.

3. A case study: creation of Hindi wordnet (HWN) and Marathi wordnet (MWN)

We follow Chakrabarty *et. al.* (2007) in this section. We have, for long, been engaged in building lexical resources for Indian languages with focus on Hindi and Marathi (http://www.cfilt.iitb.ac.in). The Hindi and Marathi wordnets (HWN and MWN) and the Hindi Verb Knowledge Base (HVKB) (Chakrabarty *et. al*., 2007) have been given special attention. The wordnets more or less follow the design principles of the Princeton Wordnet for English while paying particular attention to language specific phenomena (such as *complex predicates*) whenever they arise.

While HWN has been created by manually looking up the various listed meanings of words in different dictionaries, MWN has been created by expansion from HWN. That is, the synsets of HWN are adapted to MWN via addition or deletion of synonyms in the synset.
Figure 1 shows the creation of the synset for the word *peR* 'tree' in MWN *via* addition and deletion of synonyms from HWN. The synset in HWN for this word is {*peR, vriksh, paadap, drum, taru, viTap, ruuksh, ruukh, adhrip, taruvar*} *'tree'*. MWN deletes {*peR,viTap, ruuksh, ruukh, adhrip*} and adds {*jhaaR*} to it. Thus, the synset for tree in MWN is {*jhaaR, vriksh, taruvar, drum, taru, paadap*} *'tree'*. Hindi and Marathi being close members of the same language family, many Hindi words have the same meaning in Marathi. This is especially so for *tatsam* words, which are directly borrowed from Sanskrit. The semantic relations can be transferred directly, thus saving time and effort.

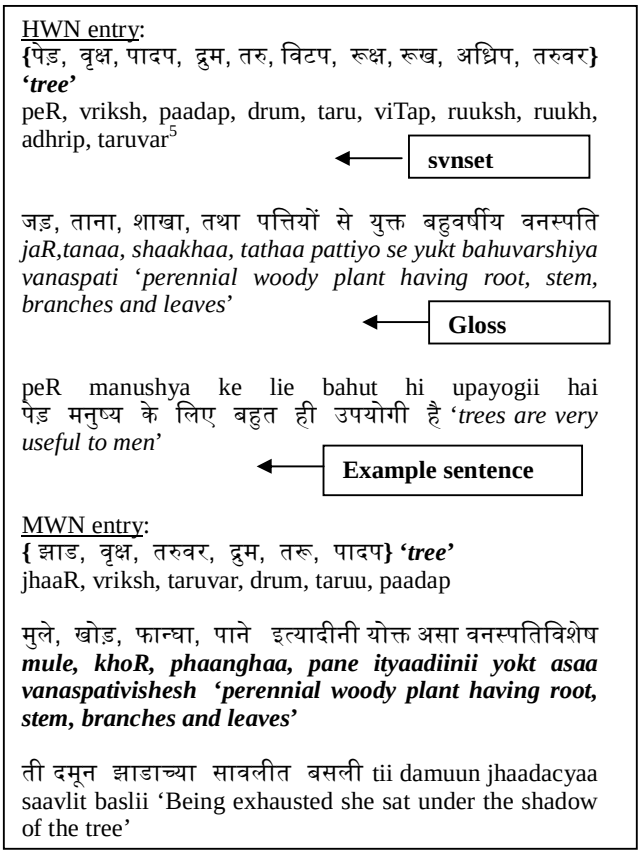

Figure 1: MWN synset creation from HWN

3.1 Synset making

The principles of *minimality*, *coverage* and *replaceability* govern the creation of the synsets:

(i) ***Minimality***: Only the minimal set that uniquely identifies the meaning is first used to create the sysnet, *e.g.*,

{ghar, kamaraa} (room)

ghar- which is ambiguous- is not by itself sufficient to denote the concept of a *room.* The addition of *kamaraa* to the synset brings out this meaning uniquely.

(ii) ***Coverage***: Next, the synset should contain all the words denoting a particular meaning. The words are listed in order of decreasing frequency of their occurrence in the corpus.

{ghar, kamaraa, kaksh} (room)

(iii) ***Replaceability*:** The words forming the synset should

[5] Henceforth we will use only Roman script for expressing Hindi text. This is to avoid any problem of font encoding and also for the readability of non-Hindi readers.

be mutually replaceable in a specific context. Consider,

{ स्वदेश, घर *}* (*motherland*)– अपना देश apanaa desh (*the country where one is born*)
svadesh, ghar

अमेरिका में दो साल बिताने के बाद श्याम स्वदेश /घर लौटा
amerikaa meN do saal bitaane ke baad shyaam svadesh/ghar lauTaa

America in two years stay after Shyam motherland returned

'*Shyam returned to his motherland after spending two years in America*'

Figure 2: Illustration of replaceability

Here *svadesh* and *ghar* can replace each other.

The replaceability criterion is observed with respect to synonymy (semantic properties) and not with respect to the syntactic properties (such as subcategorization) of a lexeme.

3.2 Lexical relations

HWN incorporates several commonly used semantic and lexical relationships along with a few new ones.

a) Antonymy is a lexical relation indicating 'opposites'. For instance, {**moTaa**, sthuulkaay}'*fat*' → {**patlaa**, dublaa} '*thin*'
patlaa (*thin*) is the antonym of moTaa (*fat*) and vice versa.

Criterion	Examples	Gloss
Size	(chhoTaa-badzaa, moTaa -patlaa)	big-small, thick-thin
Quality	(achchhaa-buraa, pyaar-ghriNaa)	good-bad, love-hatred
Gender	(betaa-beTii, maataa-pitaa)	son-daughter, father-mother
State	(shuruu-ant)	beginning-end
Personality	(raam-raavaN)	Rama-Ravana
Direction	(puurv-pashchim, aage-piichhe)	eat-west, front-behind
Action	(lenaa-denaa, khariid-bikrii)	take- give, buy-sell
Amount	(kam-jyaadaa, halkaa-bhaarii)	little-much, light-heavy
Place	(duur-paas)	far-near
Time	(din-raat, subaha-shaam)	Day-night,morning-evening

Table 1: Criteria for Antonymy

HWN indicates the criterion under which the antonomy holds. In the above example, the antonymy criterion is *size*. Other criteria are given in Table 1.

b) Gradation is a lexical relation that represents possible intermediate states between two antonyms. For example, figure 3 shows the gradation relation among time words.

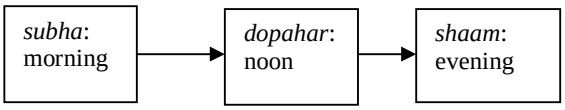

Figure 3: Gradation

c) Hypernymy and Hyponymy encode lexical relations between a more general term and specific instances of it.

{belpatra, belpattii, bilvapatra} '*leaf of a tree named bela*'
→ {pattaa, paat, parN, patra, dal} '*leaf*'

Here, *belpatra* (*a leaf of the bel tree*) is a kind of *pattaa* (*leaf*). *pattaa* is the hypernym of *belpatra* and *belpatra* is a hyponym of *pattaa*.

d) Meronymy and Holonymy express the *part-of relationship* and its inverse.

{*jaR, muul, sor*} '*root*' → {*peR, vriksh, paadap, drum*} '*tree*'
Here, *jaR* (*root*) is a part of *peR* (*tree*) and therefore, *jaR* is the meronym of *peR* and *peR* (*tree*) is the holonym of *jaR* (*root*).

e) Entailment is a semantic relationship between two verbs. A verb *A* entails a verb *B*, if the meaning of *B* follows logically and is strictly included in the meaning of *A*. This relation is unidirectional. For instance, *snoring* entails *sleeping*, but *sleeping* does not entail *snoring*.

{*kharraaTaa lenaa, naak bajaanaa*} '*snore*'→ {*sonaa*} '*sleep*'

f) Troponymy is a semantic relation between two verbs when one is a specific 'manner' elaboration of another. For instance,

{dahaaRanaa} '*to roar*' is the troponym of {bolanaa} '*to speak*'

g) HWN also cross-links synsets across different parts of speech. Cross-links between 'nouns' and 'verbs' include the following:

i. Ability link specifies the features *intrinsic to* a nominal. For example,
{*machlii, macchii, matsya, miin, maahii*} '*fish*'→ {*tairnaa, pairnaa, paunrnaa*} '*swim*'

ii. Capability link specifies the features that may be *acquired* by a nominal. For example,
{*vyakti, maanas*} '*person*' → {*tairnaa, pairnaa, paunrnaa*} '*swim*'

iii. Function link specifies function(s) associated with a nominal. For example,

{*adhyaapak, shikshak*}'*teacher*' → {*paRhanaa, shikshaa denaa*} '*teach*'

Cross-links between 'nouns' and 'adjectives' are used to indicate typical properties of a noun. For example, {*sher*} '*tiger*' → {*maansaahaarii*} '*carnivorous*'. Links between morphologically derived forms mark the root form from which a particular word is derived by affixation. For example, {*bhaaratiiyataa*} '*indianness*' is derived from {*bhaaratiiya*} '*Indian*' and is linked to it.

4. The Process of IndoWordnet Creation

Seeing the enormous potential of wordnet, 16 out of 22 official languages of India, have started making their wordnets under the leadership of IIT Bombay. These languages are: (1) Hindi [6] (already discussed), (2) Marathi[7], (3) Konkani[8], (4) Sanskrit[9], (5) Nepali[10], (6) Kashimiri[11], (7), Assamese[12], (8) Tamil[13], (9) Malyalam[14], (10) Telugu[15], (11) Kannad [16], (12) Manipuri[17] and (13) Bodo,[18] (14) Bangla[19], (15) Punjabi[20] and (16) Gujarati[21]. These languages cover the length and breadth of India and are used by about 900 million people. Table 2 shows the wordnets and the corresponding institutes developing them.

The experiences of various language groups of building these wordnets have been presented in the 5th International Conference of Global Wordnet (GWC2010) (Bhattacharyya et. al., 2010).

Wordnet –Language	Institute(s)
Assamese	Guahati University, Assam
Bengali	Indian Statistical Institute Kolkata, IIT Kharagpur and Jadavpur University
Bodo	Guahati University, Assam
Gugarati	DDU, Nadiad, Gujarat
Hindi	IIT Bombay
Kannad	Amrita University, Koimbatore
Kashmiri	Kashmir University, Srinagar
Malayalam	Amrita University, Koimbatore
Manipuri	Manipur University, Imphal, Manipur

[6] Hindi/Khadi boli belongs to the Indo-Aryan language sub-group of Indo-European language family. It is a dialect continuum of the Indic language family in the northern plains of India. 2001 census of India noted 422,048,642 speakers of this language. It is spoken in the Indian states and union territories of Bihar, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttarakhand.

[7] Marathi is an Indo-Aryan language spoken by the Marathi people of south western India and is the official language of the state of Maharashtra. 2001 census of India noted 71,936,894 speakers of this language.

[8] Konkani is an Indo-Aryan language belonging to the Indo-European family of languages spoken in the Konkan coast of India. It has approximately 7.6 million speakers of its two individual languages, Konkani and Goan Konkani.

[9] Sanskrit is a historical Indo-Aryan language and as per the 2001 census of India, there are 6,106 speakers of this language.

[10] Nepali is a language of the Indo-Aryan branch of the Indo-European language family. 2001 census of India records 13,168,484 speakers of this language.

[11] The origin of Kashmiri language is uncertain. According to one view it belongs to the Dardic languages which form a sub-group of the Indo-Aryan languages whereas others believe that it belongs to the Iranian languages. It is spoken in eastern Afghanistan, northern Pakistan, and in the Indian region of Jammu and Kashmir. 2001 census of India recorded 5,527,698 speakers of this language.

[12] Assamese is the easternmost Indo-Aryan language. According to the 2001 census of India there are 13,168,484 speakers of this particular language.

[13] Tamil is the only surviving Classical language in the world and is a Dravidian language. According to the 2001 census of India there are 60,793,814 speakers of this particular language.

[14] Malayalam is one of the four major Dravidian languages of South India. According to the 2001 census of India there are 33,066,392speakers of this particular language

[15] Telugu is a Dravidian language mostly spoken in the Indian state of Andhra Pradesh. According to the 2001 census of India there are 74,002,856 speakers of this particular language.

[16] Kannada is one of the major Dravidian languages of India, spoken predominantly in the state of Karnataka. 2001 census of India recorded 37,924,011 speakers of this language.

[17] Manipuri belongs to the Sino-Tibetan family of languages. It is the official language of south-eastern Himalayan state of Manipur, in north-eastern India. According to 2001 census, 1,466,705 speakers of this language are found in India.

[18] Bodo is a Tibeto-Burman language. 2001 census of India records 1,350,478 speakers of this language.

[19] Bangla is one of the major members of the Indo-Aryan family of languages, spoken by 215 million people. It is the national language of the country of Bangladesh and also the state language of the province of West Bengal in India.

[20] Punjabi is a member of the Indo-Aryan family spoken by about 88 million people. It is the state language of the province of Punjab in Western India.

[21] Gujarati is a member of the Indo-Aryan family spoken by about 46 million people. It is the state language of the province of Gujarat in Western India.

Marathi	IIT Bombay
Nepali	Assam University, Silchar, Assam
Oriya	University of Hyderabad
Punjabi	Thapar Institute and Punjabi University, Patiala
Sanskrit	IIT Bombay
Tamil	Tamil University, Thanjavur and Amrita University
Telugu	University of Hyderabad and Dravidian University, Kuppam
Urdu	University or Hyderabad and International Institute of Information Technology, Allahabad

Table 2: Wordnets of different languages and institutes developing them

Wordnets creation for languages other than Hindi is going on using the Expansion Approach. Figure 4 below shows the big picture of the IndoWordnet.

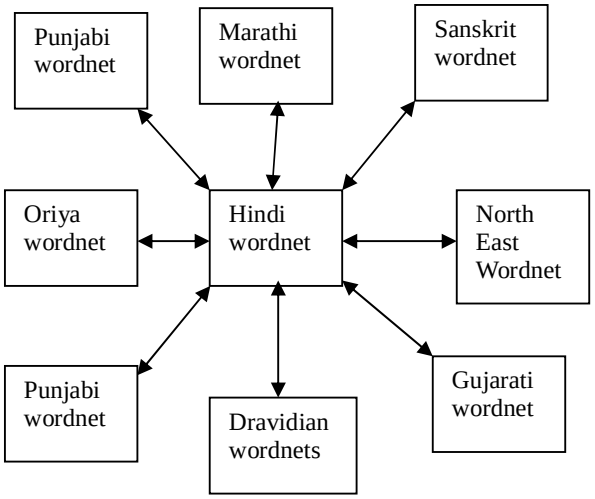

Figure 4: Linked IndoWordnet structure

4.1 Selection of core and common synsets

In the process of IndoWordnet creation, the first phase is the construction of synsets for most common concepts which are universal across languages. It was decided that all IndoWordnet members will first link to the core synsets.

To select the most common concepts from approximately 32000 synsets of HWN, the following steps were taken:

1. Initially, 32K synsets were distributed among 6 people. Each one classified them into 4 categories, *viz., (i) Common, (ii) Uncommon, (iii) Common in Indian languages* and *(iv) Region and language specific*. This was done with the help of a specially designed tool for synset classification. By this process 16K synsets were filtered.

2. These 16K synsets were again ranked by voting. 11K synsets were selected as common synsets.

3. An online interface was provided to rank these 11K synsets by the NLP group at IIT Bombay.

compiled by D. N. Narwane was used to create a set of core concepts necessary for everyday living and communication. 2000 synsets were selected as core synsets and distributed to other language groups.

Wordnet –Language	**#synsets/unique-words**
Assamese	3530/19609
Bengali	8679/ 18563
Bodo	3837/13357
Gugarati	970/2125
Hindi	33900/82000
Kannad	5920/7344
Kashmiri	6569/8674
Malayalam	6154/8622
Manipuri	2744/5231
Marathi	9739/21223
Nepali	5802/10278
Oriya	To start
Punjabi	To start
Sanskrit	3340/17820
Tamil	4750/9821
Telugu	10639/18250
Urdu	6123/9641

Table 3: Statistics of total synsets and the unique words for wordnets of various languages

Rest of the common synsets was also distributed, but these were linked only after finishing the 2000 core concepts. Table 3 shows the status of core and common synset linkage and the number of unique words covered thereby as on March 1, 2010.

The task of **linking the synsets of Hindi with those of English** has also been going on. As of now 13693 synsets of Hindi have been linked with English. Efforts are also on to automatise this process of linkage.

4.2 Lexicographers' interface for wordnet building

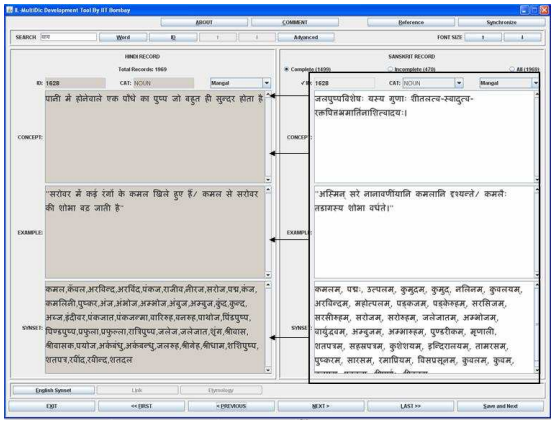

Figure 5: Lexicographer's interface.

To create a lexical resource like wordnet, one needs a user friendly tool. Use is made of the *MultiDict* tool developed at the Center for Indian Language Technology, Computer Science Department, IIT Bombay (Figure 5 above). The tool provides an interface for linking the synsets that express the same meaning in different language. In this figure a synset of Sanskrit meaning *lotus* is linked with the corresponding synset for Hindi. The left panel shows the Hindi synset and the right panel is for the synset of the target languages.

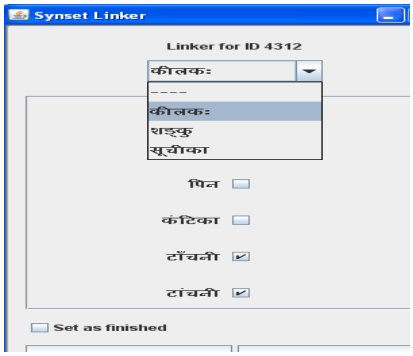

Figure 6: The Linker tool

The linker tool (Figure 6) is integrated in the interface for cross-linkage between the literals of source and target synsets. It allows a lexicographer to link a literal of the source language to one or more literals in the corresponding target language synset. The particular example shown in figure 6 depicts the word level linkage for words of Sanskrit and Hindi meaning *nail.*

5. Some experiences of lexicographers of representative Indian languages

Here we record the IndoWordnet experience of three languages, by describing the challenges faced by lexicographers of these languages. We have chosen Sanskrit as a heritage language, Bodo as a member of the Tibeto Burman family and Telugu as a member of the Dravidian family of languages.

5.1 Sanskrit

Sanskrit is the oldest member of the Indo-Aryan language family, a sub branch of Indo-Iranian, which in turn is a branch of Indo European language family. The wordnet for Sanskrit (Kulkarni *et. al*., 2010) is being created from Hindi wordnet.

There is a traditional fourfold division of lexical units of Indian languages into:

1. *tatsama*[22] - words having their origin in Sanskrit and accepted in the modern Indo-Aryan languages without any change in their phonology.
2. *tadbhava*[23] - words which have their origin in Sanskrit but their phonological forms are changed as per the rules of the modern Indo-Aryan languages.
3. *deshī* -words which are the native words of the particular language and
4. *videshī* - words borrowed from foreign languages.

The links to *tatsama* and *tadbhava* words, in particular, will be a great pan-Indian linguistic resource for computational purposes.

The challenge faced in creating the synsets of Sanskrit wordnet in consonance with those of Hindi is mainly in **finding equivalent words.** For example; the word { } chaaya (*tea*) is a frequently used word. The concept of *tea* is explained as follows in the Hindi wordnet:

(1) ,

cAya ke paudhe kI pattiyon ko pAnI mein DAlkar cinI dUdha Adi milAkar banAyA huA peya padArtha
(A drink prepared by mixing the leaves of the Tea-plant with sugar, milk and water)

But Sanskrit does not have a word of its own for this concept. Monier Williams in his Sanskrit-English dictionary (MW hereafter) suggests that " " *cahA* (which is actually is a Marathi word) should be used as a borrowed word. In the dictionary of spoken Sanskrit we find two different regional words " " *cAya* and " " *cAyA* belonging to the languages of North and South India respectively. The gloss field in the synset of { , , } {*kaSAyapeyaM, cAyaH, cAyA, cahA*} in the Sanskrit wordnet is created as follows:

(2) -

cAyaH cahA evaMvidhaiH shabdaiH bhAratIya-bhASAsu prasiddhasya kSupasya shuSka-parNAnAM cUrNam uSNajale abhipacya tasmin drave sharkarA-dugdhAdIn saMmishrya nirmitam uSNapeyam
(A hot drink which is prepared by first mixing the leaves of the plant, which is famous by the names like etc. in the Indian languages, into hot water and then mixing it with sugar and milk)

5.2 Bodo

The Bodo language belongs to the Tibeto-Burman branch of the Sino-Tibetan language family. It is a major language of the North-Eastern part of India and has very close resemblance with the Rabha, Garo, Dimasa, Kokborok, Tiwa, Hajong and other allied languages of N-E India. Bodo is a developing language. The language does not have rich linguistic and literary resources. New

[22] Tatsama Shabda Kosha (Tatsama words dictionary) is published by Kendriya Hindi Nideshalaya, Shiksha Vibhaga, Manava Samsadhana Vikasa Mantralaya, Bharata Sarakara in 1988.

[23] See Hindi ki Tadbhava Shabdavali[**Error! Reference source not found.**]

words and phrases are being discovered, coined and added to the language. Few examples of synsets from Bodo wordnet (Sarma *et. al.*, 2010) follow:

[World, English]: [पृथ्वी,हिन्दी]: बुहुम, मुलुग, भुम, संसार, हालुर, बैसोमाथा, बिलाथलाथा [Bodo]।
[Jungle, English]: [जंगल,हिन्दी]: हाग्रामा, अरन, हाग्रा, हाग्राबारि, जाहार, आरंगा [Bodo] ।
[Body, English]: [शरीर,हिन्दी] : देहा, मोदोम, सोलेर, सावस्रि [Bodo] ।
[God, English]: [भगवान,हिन्दी]: इसोर, गसाइ, आनान_गसाइ, अबंलावरि, अबं [Bodo] ।

Some of the frequently encountered challenges faced by Bodo wordnet construction are:

i. Lack of proper vocabulary to mean the concept, or the example as given by the Hindi synset.
ii. Small size the Bodo synset. Two/three members synsets are very common.
iii. Multiwords in synsets which are often coined.

As is apparent, these challenges result from Bodo's relative newness as a language.

5.3 Telugu

Wordnet for Telugu (Selvaraj A., 2010) is being constructed expanding from Hindi wordnet. Telugu belongs to the South Central Dravidian subgroup of the Dravidian family of languages.

The vocabulary of Telugu is highly Sanskritized in addition to the Persian-Arabic borrowings కబురు /kaburu/ `story', జవాబు /javaabu/ `answer'; Urdu తరాజు /taraaju/ `balance'. It does have cognates in other Dravidan languages such as పులి /puli/ `tiger', ఊరు /uuru/ `village'; తల /tala/ `head'.

Words in Dravidian languages, especially in Telugu are long and complex. This is because of highly rich morphology and poly-agglutination.

In Telugu (as well as in other languages of the Dravidian family: *Tamil, Malayalam and Kannad*), the lexicographers typically consult the English wordnet, even though they are expanding from the Hindi wordnet. The reason is the relatively lower level of proficiency with Hindi and much wider usage of English in the Southern part of India. This is an interesting situation where both English and Hindi wordnets are consulted. IIT Bombay has provided both Hindi and English synsets for facilitating the expansion.

This gives rise to one of the main challenges in the construction of the Telugu wordnet. Sometimes the Hindi and English synsets do not completely agree in the finer nuances of the meaning, and the lexicographers is left with the tough task of which language's meaning to choose.

The other challenge is that *kinship terms* differ between Hindi and Telugu. For example, for the Hindi synset with id #7379 containing the word भतीजा (*bhatiijaa* meaning *brother's son*), Telugu has finer distinction between *elder brother's son* and *younger brother's son.* Currently Telugu wordnet expands this particular Hindi synset into one with both these terms included, instead of the one with just {సోదరుని కుమారుడు} (soudaaruni kumaarudu; *brother's son* which is ambiguous in Telugu).

6. Some guideline for IndoWordnet construction

In developing the IndoWordNet the following considerations have been kept in mind:

1. Wordnet's central concern is to express a concept unambiguously. To express concepts with a set of word (s) we can follow these options: (a) dictionary words, b) transliteration (c) short phrase and (d) coined word.

2. Dictionary words are included in the wordnet according to the frequency of their use. Options (b), (c) and (d) are typically needed in expanding from a culture or region specific concept. However, these options should be used with discretion, respecting the native speakers' sensitivities.

3. Same synset ID has to be maintained across languages.

4. As for including newly coined words, it is felt that *Standardization* may be a problem. Coining of new words should be avoided till the method of coining and the procedure of standardization are decided. Some ways of standardization are proposed but there is no consensus among the language groups. One possibility is to validate the words by keeping them on the web and asking for opinions.

5. Regarding region specific and culture specific words, the general policy adopted is to assign a specific **ID range** for such concepts. However, this needs synchronization among lexicographers. It was decided that IDs ranging from 0-50000 will be reserved for common concepts across languages. After that ID ranges will be allotted for specific languages, *e.g.*, 50000-60000 for Marathi, 60000-70000 for Konkani and so on.

6. It was emphasized again and again to the lexicographers **never to translate the words in the Hindi synset, but to understand the meaning expressed by the synset and its attached gloss and example sentence and then to put down the words in frequency order the words of the language**.

7. Differences of IndoWordnet (IWN) from EuroWordnet (EWN)

The expansion approach of wordnet creation adopted and elaborated in EuroWordnet (Vossen, 1998) is also the principal methodology for Indwordnet construction. In EWN, English provided the Interlingual Index (ILI). In IWN, the same is provided by Hindi.

There are, however, some crucial differences between IWN and EWN:

(i) Right from the beginning, IWN insisted on storing lexical links expressing relationship of derivational morphology. Indian languages are rich in morphology. In Sanskrit wordnet, for example, the theory that all words are derived from verbal roots- *dhaatus*- is being seriously examined for its use as a fundamental guiding principle for storing and linking word.

(ii) *Causative verb forms* are a typically occurring phenomenon in Indian languages. For example, *khaanaa (to eat), khilaana (to feed)* and *khilwaanaa (to cause to feed)* are forms derived from the same root *khaanaa*. It has been decided to take special care to store causative forms in IWN and link them to their basic roots.

(iii) *Complex predicates* (CPs) (also known as *complex verbs*) abound in South Asian languages. They occur in the form of *nominal+verb* combinations (called *conjunct verbs*) and *verb+verb* combinations (called *compound verbs*). Key questions on complex predicates are:

A. *Given a N(oun)+V(erb) combination, is the noun incorporated into the verb complex or is it an overt argument of the verb?*
B. *Given a V(erb)+V(erb) combination, is the second verb an aspectual/modal or is it the polar (intensifier) in a polar-vector combination?*

IWN is drawing heavily on the research on complex predicates (Chakrabarty *et. al.*, 2007) and is devising means for storing them and linking them with their constituent N/V and Vs.

(iv) IWN has from the start taken cross part of speech linkages very seriously, especially between nouns and verbs. *Ability* and *capability* links discussed in section 2- between nouns and verbs- are being incorporated exhaustively.

(v) IWN has finer categories for antonymy and gradation relations compared to EWN.

8. Conclusion

In this paper we have described the India wide effort of creating the linked structure of Indian language wordnets- collectively called the IndoWordnet. Members of three language families spanning the length and breadth of the country are creating and linking lexical knowledge. Hindi is serving as the pivot language in this endeavour.

A key observation in this effort- which is by no means new, but important nonetheless- is that culture and region specific words do form a large component of the lexicography work and provisions have to be made for their storage and linking. Language specific ID ranges are an important step towards addressing this problem.

Our future work consists in putting in place a common background ontology for IndoWordnet. This will form the backdrop against which the synsets can be analysed for their quality.

9. Acknowledgements

The support of the Department of Information Technology (DIT), Ministry of Communication and Information Technology, Government of India and also of Ministry of Human Resource Development is gratefully acknowledged.

10. References

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

Chakrabarty D., Sarma V. and Bhattacharyya P. (2007), *Complex Predicates in Indian Language Wordnets*, Lexical Resources and Evaluation Journal, 40 (3-4).

Cruse D.A. (1986), *Lexical Semantics*, Cambridge University Press.

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*, MIT Press.

Kulkarni M., Dangarikar C., Kulkarni I., Nanda A. and Bhattacharyya P. (2010), *Introducing Sanskrit Wordnet*, Global Wordnet Conference (GWC10), Mumbai, India.

Mohanty R., Bhattacharyya P., Pande P., Kalele S., Khapra M. and Sharma A. (2008), *Synset Based Multilingual Dictionary: Insights, Applications and Challenges*, Global Wordnet Conference (GWC08), Szeged, Hungary.

Narayan D., Chakrabarty D., Pande P. and Bhattacharyya P., *An Experience in Building the Indo WordNet- a WordNet for Hindi*, 1st International Conference on Global WordNet (GWC 02), Mysore, India.

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

Selvaraj A., (2010), *Telugu Wordnet*, Global Wordnet Conference (GWC10), Mumbai, India.

Vossen P. (ed.). 1998 *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, Dordrecht.