Semi-Automatic Extension of Sanskrit Wordnet using Bilingual Dictionary

Sudha Bhingardive
Center for Indian Language Technology,
Indian Institute of Technology Bombay
sudha@cse.iitb.ac.in

Tanuja Ajotikar
Department of Humanities and Social Sciences,
Indian Institute of Technology Bombay
gtanu30@gmail.com

Irawati Kulkarni
Center for Indian Language Technology,
Indian Institute of Technology Bombay
irawatikulkarni@gmail.com

Malhar Kulkarni
Department of Humanities and Social Sciences,
Indian Institute Technology Bombay
malhar@iitb.ac.in

Pushpak Bhattacharyya
Center for Indian Language Technology,
Indian Institute Technology Bombay
pb@cse.iitb.ac.in

Abstract

In this paper, we report our methods and results of using, for the first time, semi-automatic approach to enhance an Indian language Wordnet. We apply our methods to enhancing an already existing Sanskrit Wordnet created from Hindi Wordnet (which is created from Princeton Wordnet) using expansion approach. We base our experiment on an existing bilingual Sanskrit English Dictionary and show how lemma in this dictionary can be mapped to Princeton Wordnet through which corresponding Sanskrit synsets can be populated by Sanskrit lexemes. This our method will also show how absence of resources of a pair of languages need not be an obstacle, if another resource of one of them is available. Sanskrit being historically related to languages of Indo-European family, we believe that this semi-automatic approach will help enhance Wordnets of other Indian languages of the same family.

1 Introduction

Wordnet is a lexical semantic network, widely used in various applications of natural language processing. Princeton wordnet (PWN) is the mother of all Wordnets (Fellbaum, 1988). It was created at the Cognitive Science Laboratory of Princeton University. EuroWordNet (Vossen, 1998; Vossen, 2000), CoreNet (Choi, 2004), IndoWordNet (Bhattacharyya, 2010), HowNet (Zhendong, 2000), MultiWordNet (Bentivogli, 2000; Bentivogli and Pianta 2000), BabelNet (Navigli, 2012) and so many other Nets are also some of the most commonly used semantic networks.

PWN is manually created using the knowledge from various dictionaries. Several Wordnets are created semi-automatically using the expansion approach from PWN. Many of them use bilingual dictionaries or Wikipedia. This type of creation saves enormous manual efforts and time. However, it demands high quality machine-readable resources in the respective languages.

Sanskrit wordnet (SWN) (Kulkarni *et.al*, 2010) is manually created using the expansion approach from Hindi wordnet (HWN), which in turn was created from the Princeton Wordnet. The current status of Sanskrit wordnet is stated in Table 1.

Total synsets: 22912 **Total unique words**: 44950

POS	Noun	Verb	Adverb	Adjective
synset counts	17413	1246	263	3990

Table 1: Sanskrit wordnet current status

2 Motivation

In this work, we aim to report our experiences to populate SWN by a semi-automated approach. Currently, manual approach is used which is time consuming and tedious. Following are the reasons that make manual approach time consuming.

2.1 Large number of synonyms for a Sanskrit word

In the available lexical literature of Sanskrit (given below in Section 3), normal range for number of words in any synset varies between 1–20 *e.g., līlā* [a game (6 synset members)], *vṛddhaḥ* [an old man (20 synset members)], *bhakṣaṇam* [an act of eating (20 synset members)]. Synsets with only one word are common in the cases of coined words, instrument names and kinship relations. However, some synsets exceed this limit and have huge number of words as its members. We note below some of the prominent phenomena.

- Synsets expressing concepts in the domain of mythology, culture, religion and philosophy contain large number of words *e.g., viṣṇuḥ* [Hindu deity (127 synset members)], *somaḥ* [a God (120 sysnet members)], *yuddha* [a war (97 synset members)], *sūryaḥ* [the Sun (85 synset members)], *samudraḥ* [an ocean (synset members)].
- Synsets of noun/adjective category containing words with features of derivational morphology tend to have large number of words *e.g., dyutimat* [bright (246 synset members)], *Shikhin* [one who possesses antenna (40 synset members)].
- The process of compound formation in Sanskrit allows creation of multiple synonyms and therefore synsets containing such compounds tend to have large number of words *e.g., devaalaya* [house of gods = temple (50 synset members)], *alpamati* [one who possesses little intellect (40 synset members)].

For creating above mentioned synsets, lexicographers gathered information from various resources, *e.g.,* while creating a concept of *yuddha* (a war), 97 words were collected from various lexical resources given below: Spoken Sanskrit Dictionary[1] (7 words), Apate's Sanskrit-English Dictionar[2] (7 words), Monier William's English–Sanskrit Dictionary[3] (57 words) and Shabdakalpadrum (80 words).

After collecting the words, duplicate words were eliminated. Words representing proper meanings are entered in the synset. This process is monetarily expensive and time consuming. Automatic approach can help populate such synsets using bilingual dictionaries. In the process there will be over-generation which will have to be controlled by manual approach.

2.2 Appropriate selection of words for creating synsets

While creating the synsets, appropriate selection of words is required to express the precise meaning. In Hindu texts, which are mainly in Sanskrit there are various names for a single deity *e.g., Viṣṇu* (Hindu deity) has 132 names, *Kṛṣṇa* has 132 names and *Rāma* has 67 names. For creating synsets of these deities one must be very careful as *Kṛṣṇa* and *Rāma* are incarnations of *Viṣṇu* and can easily get interchanged and thereby affecting the intended meaning.

The road-map of the paper is as follows. Section 3 presents the related work. Section 4 explains the methodology used for extension of SWN. Section 5 illustrates results. Outcomes are presented in Section 6. Section 7 includes conclusion and future work.

3 Related Work

Most of the Wordnets are created by expansion approach using PWN. Several Wordnets have tried to increase their coverage using various automatic or semi-automatic approaches. Some of them are listed below. CoreNet (Choi, 2004) is an automatically constructed Wordnet, which uses a Japanese–Korean electronic dictionary. Korean words are programmatically generated during translation from Japanese. BabelNet (Navigli, 2012) is a very large, wide-coverage, multilingual semantic network. This resource is created by mapping a multilingual encyclopedic knowledge repository (Wikipedia) and a computational lexicon of English (PWN). The integration is performed via an automatic mapping and by filling in lexical gaps in resource-poor languages with the

[1] http://spokensanskrit.de/

[2] http://www.aa.tufs.ac.jp/~tjun/sktdic/

[3] http://www.sanskrit-lexicon.uni-koeln.de/monier/mwauthorities/mwauth_SktDevaUnicode.html#record_Lalit_

aid of Machine Translation. This provides concepts and named entities, lexicalized in many languages and connected with large amounts of semantic relations. Chinese Wordnet (Renjie Xu, 2008) is developed in an automatic manner by translating English words to Chinese using Chinese–English dictionary. Czech wordnet (Karel Pala, 2008) is automatically extended from PWN using machine-readable bilingual dictionary. Polish WordNet (M. Derwojedowa, 2008) is designed semi-automatically by extracting lexical relations from the large Polish corpora. Lexicographers are used for mapping these relations with PWN.

3.1 Why was Monier William's Sanskrit–English dictionary used for extending SWN?

We have used the publicly available Monier William's Sanskrit–English dictionary for SWN semi-automatic extension. The list of all the texts used by Monier Williams is publicly available. This dictionary includes over 1, 80, 000 words and definitions. All entries are organized according to the root of a word, the *dhatu*, which offers better understanding of the meaning of the word. It includes special references to cognate Indo-European languages as well as literary citations. It provides precise meanings for the words in the Vedic literature, which is useful for studying the scriptures. This is one of the most comprehensive and useful Sanskrit–English dictionaries. The other reason for using this dictionary for the present purpose, fortunately, is availability. Out of all the lexical resources mentioned above, only this is available in program readable format which makes this resource singularly important from the point of view of present research. One of the outputs of the use of this resource is extraction of proper nouns. We have automatically extracted them and added to SWN without linking them to PWN. This method is explained in Section 4.2.

4 Methodologies used for extending SWN

SWN is created by expansion approach from HWN, which was in turn created by PWN.

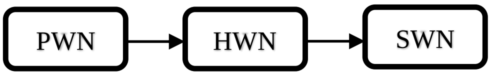


Figure 1: SWN manual creation

Our selected resource is in Sanskrit and English. Therefore, in order to utilize it for the present purpose we have to link PWN directly to SWN.

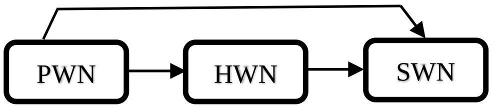
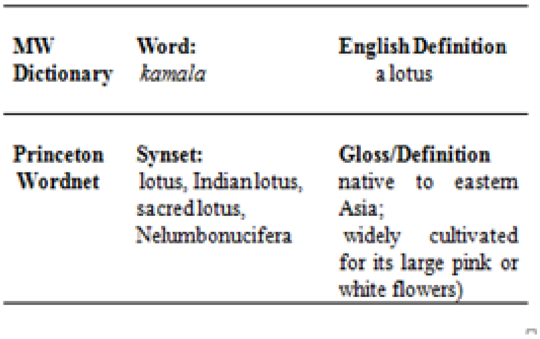


Figure 2: SWN semi-automatic creation

We link Sanskrit–English dictionary to PWN by using a heuristic. This will be automatic approach. These linkages are validated by lexicographers. This will be manual approach. Thus we will populate SWN by semi-automatic approach using this resource.

4.1 Heuristics used for linking William's dictionary to PWN

William's dictionary contains Sanskrit words along with its English description. The description is concise for most of the Sanskrit words, *e.g.*, *kamala* (lotus) has the description 'a lotus'. In comparison, PWN glosses are descriptive as shown in Figure 3.



MW Dictionary	Word: *kamala*	English Definition a lotus
Princeton Wordnet	Synset: lotus, Indian lotus, sacred lotus, Nelumbonucifera	Gloss/Definition native to eastern Asia; widely cultivated for its large pink or white flowers)

Figure 3: Dictionary and PWN entry for *kamala* (a lotus)

Finding the maximum overlap between the description words in dictionary and PWN gloss words is not efficient as we get several possible mappings. It is monetarily expensive and time consuming to generate and validate these mappings. Therefore, this type of heuristic is not suitable for linking dictionary to PWN.

William's dictionary is a very rich resource in Sanskrit language, which is useful for extending the SWN. Hence, we linked dictionary to PWN using a heuristic, which finds the maximum overlap between description words in dictionary and words in PWN synsets. Using this heuristic, the dictionary entries are linked to PWN. We got 14653 single and 55059 multiple possible mappings. Lexicographers are in the process of validating these mappings. The architecture diagram of the process is shown in Figure 4. Following are the steps for the procedure to link dictionary to PWN.

- For a Sanskrit word S_w, from dictionary, its equivalent English description is taken and its maximum overlap with words in the PWN synsets is found.
- S_w is directly mapped to the synset if the word in the description is found to be monosemous in PWN.
- The mapping is evaluated manually if the word in the description is found to be polysemous in PWN.

After successful mapping, all Sanskrit words are added in SWN.

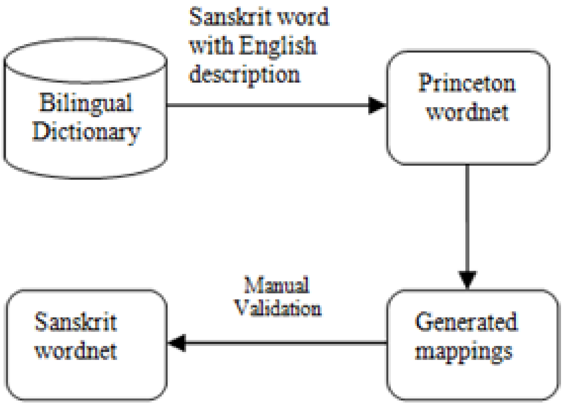


Figure 4: Architecture diagram

This task can be explained with the help of an example, for the word '*kartr*' (spinner), we found three possible mappings in PWN. For validating these multiple possible mappings, we designed an interface as shown in Figure 5. It provides various functionalities on mappings *viz.*, display, search, validate and delete. A lexicographer will select an appropriate mapping with the synset in PWN of correct sense.

After manual validation, all dictionary entries with valid mappings are inserted into the Sanskrit wordnet. Adding of all the dictionary entries maually requires excessive efforts. Thus, a semi-automatic approach will save these excessive manual efforts.

4.2 Other automatic application of William's dictionary to populate SWN

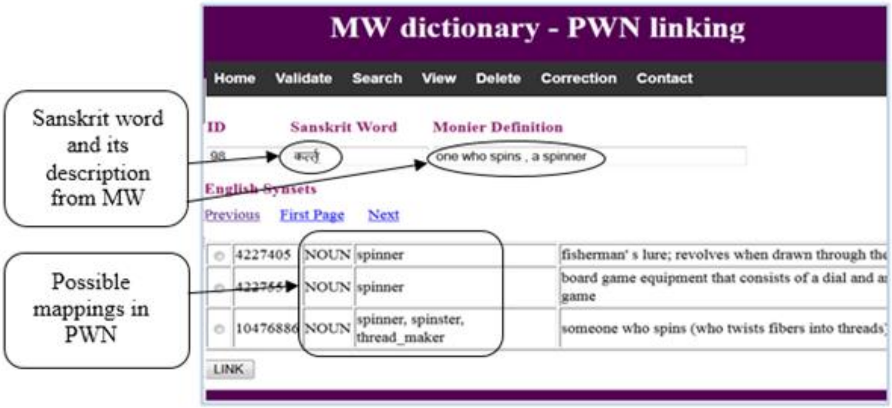
If the English description of the Sanskrit word began with the phrase 'Name of a', all such words can be considered as proper nouns. For example, the word '*Brahamhapuri'* has the description, 'Name of a location'. Currently all proper nouns are part of the Wordnet. However, it is yet to be decided whether these are maintained in a separate gazetteer (gazetteers are those which contain entities themselves that are proper nouns), which will in turn link to SWN. If it is decided that they are to be treated as a part of Wordnet then it would add 14,339 synsets to SWN.

Some of the extracted nouns are class names. For example, the word '*Ustika'* has the description 'Name of a kind of plant' and the word '*Bhaumadevalipi*' has the description 'Name of a kind writing'. Both these words are class names. All class names are not stored in a gazetteer. They are very much stored in the SWN. So far, fifty-five class names are extracted from the dictionary and stored in SWN.

5 Results

As discussed in Section 4.1 we are linking dictionary with PWN. There are 14, 653 Sanskrit words for which single mappings were found in PWN and 55, 059 words for which multiple mappings were found in PWN. The work of these mappings is still under validation process. We have extracted 14,339 proper nouns from dictionary, which are not covered by SWN.

These proper nouns must get inserted into SWN as these are most frequent occurrences in Sanskrit literature. Current synset coverage status of SWN is illustrated in Table 1. After adding dictionary entries, SWN coverage will increase considerably. With this semi-automatic approach, SWN will be a richer lexical resource in Sanskrit language.



ID	Sanskrit Word	Monier Definition
98	कर्तृ	one who spins , a spinner

English Synsets

○	4227405	NOUN	spinner	fisherman' s lure; revolves when drawn through the
○	[illegible]	NOUN	spinner	board game equipment that consists of a dial and a game
○	10476886	NOUN	spinner, spinster, thread_maker	someone who spins (who twists fibers into threads)

Figure 5: Interface for validating multiple possible mappings

6 Outcomes: Improving SWN-HWN-PWN linkages

6.1 SWN synsets can be corrected with the help of William's dictionary. For example, in SWN, one synset containing the word '*dīptiḥ*' is linked to the sense of 'luster' in PWN. However, in William's dictionary sense of '*dīptiḥ*' is {Brightness, Slight, splendor, beauty} which is different than this already linked to PWN sense (luster). As this dictionary is considered as an authentic lexical resource for Sanskrit we can remove the word '*dīptiḥ*' from the corresponding SWN synset.

6.2 Coverage of HWN will also improve with the help of dictionary. For example, dictionary provides the same English meaning 'moonless' for all the Sanskrit words namely '*acandra*', '*naṣṭacandra*', '*niḥsomaka*', and '*visoma*'. In the existing HWN, the concept of 'moonless' is not available. It is also not covered in SWN as it is created using expansion approach from HWN. The above mentioned words form a synset and can be added to SWN and then be further borrowed in HWN. In this way, we are also increasing the HWN coverage using dictionary and SWN as shown below.

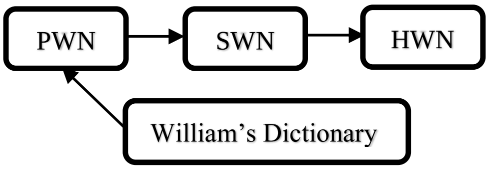


Figure 6: HWN enriched with SWN and William's dictionary

6.3 Some existing SWN synsets are not linked with PWN as SWN–PWN linking is via HWN. We are also improving these linkages using the dictionary. For example, an HWN synset corresponding to one of the synsets of *vilāsin* in SWN, is not linked with PWN. English description of *vilāsin* is given as '*coquettish*' in the William's dictionary. Both Sanskrit and English interpretation are under the same POS category of adjective. Thus, now we can link this SWN synset to PWN synset. In this way we are improving SWN–HWN–PWN linkages.

7 Conclusion and Future work

We have attempted to implement a semi-automatic approach for Sanskrit wordnet extension using Monier William's Sanskrit–English dictionary. Dictionary entries are automatically extracted and linked to PWN which need manual validation. For this purpose we have created a tool (Figure 5) which is language independent and therefore can be adopted by other similar language pairs. Post manual validation, all these entries will be inserted to SWN. Also, we have automatically extracted proper nouns from dictionary, which play an important role in Sanskrit literature. With the help of this approach we are correcting existing synset members of SWN and existing SWN–HWN–PWN linkages. HWN coverage can also be increased with the help of this approach. Following this approach, we will generate all semantic and lexical relations automatically from the same bilingual dictionary. This work can be extended using other resources like Böhtlingk and Roth's Sanskrit–German dictionary along with Monier William's dictionary for learning some useful patterns to make SWN a rich resource in Sanskrit language.

Acknowledgements

We thank to Dr. Peter Scharf, President, Sanskrit Library, United States of America, for providing the database. We acknowledge the work which provides an XML encoding of the Monier Williams Sanskrit-English Dictionary which was done in collaboration between the Institute of Indology and Tamil Studies (IITS) of the University of Cologne and Brown University's Sanskrit Department.

References

Luisa Bentivogli, Emanuele Pianta and Fabio Pianesi, 2000. Coping with lexical gaps when building aligned multilingual wordnets, In Proceedings of LREC2000, Athens, Greece.

Luisa Bentivogli, Emanuele Pianta, 2000. Looking for lexical gaps, Proceedings of Euralex, Stuttgart, Germany.

Pushpak Bhattacharyya, 2010. IndoWordNet, Lexical Resources Engineering Conference (LREC), Malta.

Key-Sun Choi and Hee-Sook Bae, 2004. Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy, GWC2004.

Magdalena Derwojedowa, Maciej Piasecki, Stanisaw Szpakowicz, Magdalena Zawisawska and Bartosz Broda, 2008. Words, concepts and relations in the construction of Polish WordNet, In Proceedings of the Global WordNet Conference, Seged, Hungary.

Dong Zhen Dong, 1988. Knowledge Description: What, How and Who? , In Proceedings of the International Symposium on Electronic Dictionaries, Tokyo, Japan.

Christiane Fellbaum, 1998. WordNet: An Electronic Database, MIT Press, Cambridge, MA.

Marti Hearst, 1992. Automatic Acquisition of Hyponyms from Large Text Corpora, Proc. of International Conference on Computational Linguistics, COLING1992.

Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda and Pushpak Bhattacharyya, 2010. Introducing Sanskrit Wordnet, 5th International Conference on Global Wordnet (GWC2010), Mumbai.

Malhar Kulkarni, 2008. Lexicographic traditions in India and Sanskrit, Journal of Language Technology, (1) pp. 160-165.

Roberto Navigli and Simone Ponzetto, 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network, Artificial Intelligence, Elsevier.

Karel Pala, Dana Hlavácková, and Vašek, 2008. Semiautomatic Linking of New Czech Synsets Using Princeton WordNet, Intelligent Information Systems.

Madhukar Mangesh Patkar, 1981. History of Sanskrit Lexicography, Munshiram Manoharlal Publishers, Delhi.

Ellen Riloff and Rosie Jones, 1999. Learning Dictionaries for Information Extraction using Multilevel Bootstrapping, Proc. of National Conference on Artificial Intelligence.

Piek Vossen, 2002. Euro WordNet General Document, University of Amsterdam.

Piek Vossen, 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer, Dordrecht, Netherlands.

Renjie Xu, Zhiqiang Gao, Yingji Pan, Yuzhong Qu, Zhisheng Huang, 2008. An Integrated Approach for Automatic Construction of Bilingual Chinese-English WordNet, ASWC 2008, LNCS 5367, 302–314.