Introduction to Synskarta: An Online Interface for Synset Creation with Special Reference to Sanskrit

Hanumant Redkar, Jai Paranjape, Nilesh Joshi,
Irawati Kulkarni, Malhar Kulkarni, Pushpak Bhattacharyya
Center for Indian Language Technology,
Indian Institute of Technology Bombay, India.
hanumantredkar@gmail.com, jai.para20@gmail.com, joshinilesh60@gmail.com,
irawatikulkarni@gmail.com, malhar@iitb.ac.in, pb@cse.iitb.ac.in

Abstract

WordNet is a large lexical resource expressing distinct concepts in a language. Synset is a basic building block of the WordNet. In this paper, we introduce a web based lexicographer's interface 'Synskarta' which is developed to create synsets from source language to target language with special reference to Sanskrit WordNet. We focus on introduction and implementation of Synskarta and how it can help to overcome the limitations of the existing system. Further, we highlight the features, advantages, limitations and user evaluations of the same. Finally, we mention the scope and enhancements to the Synskarta and its usefulness in the entire IndoWordNet community.

1 Introduction

WordNet is a lexical resource 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, etc. IndoWordNet[1] is a linked structure of WordNets of major Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan families (Bhattacharyya, 2010). WordNets are constructed by following the merge approach or the expansion approach (Vossen, 1998). IndoWordNet is constructed using expansion approach wherein Hindi is used as the source language; however, the Hindi WordNet[2] is constructed using merge approach (Narayan et al., 2002).

In this paper, we have taken reference of Sanskrit WordNet[3]. Sanskrit is an Indo-Aryan language and is one of the ancient languages. It has vast literature and a rich tradition of creating léxica. The roots of all languages in the Indo European family in India can be traced to Sanskrit (Kulkarni et al., 2010). Sanskrit WordNet is constructed using expansion approach where Hindi WordNet is used as a source (Kulkarni et al., 2010).

While developing Sanskrit WordNet, lexicographers create Sanskrit synsets by referring to Hindi synsets and by following the three principles of synset creation (Bhattacharyya, 2010). Since Sanskrit came into existence much before Hindi, it has many words which are not present in Hindi WordNet. These words are frequently used in the Sanskrit texts; hence, there was a need to create new Sanskrit synsets. In this case, the lexicographer creates new Sanskrit synsets by referring to electronic lexical resources such as Monier-Williams Dictionary, Apte's Dictionary, Spoken Sanskrit Dictionary, etc. and linguistic resources such as *Shabdakalpadruma and Vaacaspatyam*, etc. (Kulkarni et al., 2010).

The IndoWordNet community uses the IL-Multidict tool to create synsets. This tool is designed and developed by researchers at IIT Bombay (Bhattacharyya, 2010; Kulkarni et al., 2010). Though this existing lexicographer's interface is popular and widely used, it has major limitations. Some of them are – it uses flat files, has chances of data redundancy, inconsistency, etc. To overcome these limitations, we developed a new web based synset creation tool – 'Synskarta'. The features, advantages, limitations and user evaluations of Synskarta are detailed in this paper.

[1] http://www.cfilt.iitb.ac.in/indowordnet/
[2] http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php
[3] http://www.cfilt.iitb.ac.in/wordnet/webswn/wn.php

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 95–100,
Goa, India. December 2014.

The rest of the paper is organized as follows: Section 2 describes the existing system – its advantages and disadvantages, section 3 describes Synskarta – its features, advantages, limitations and the user evaluations. Subsequently, the conclusion, scope and enhancements to the tool are presented.

2 Existing System

2.1 IL Multidic Development Tool

IL Multidict (Indian Language Multidict development tool) or the Offline Synset Creation tool is developed using Java and works with flat files. This tool, popularly known as the 'Lexicographer's Interface' is an offline tool which helps in creating synsets using the expansion approach.

The interface is vertically divided into the source language panel and the target language panel. At any given time, only the current source synset is displayed in the source language panel and its corresponding target synset is displayed in the target language panel. The source panel displays the details of the current synset of the source language such as number of records in source file, current synset id, its part-of-speech (POS) category, gloss or the concept definition, example(s) and synonym(s).

Similarly, the target panel displays the target synset details such as total number of synsets in the target file, number of complete synsets, number of incomplete synsets and the current synset id (which is the same as source synset id). These fields are non-editable. There are also editable fields which allow editing of target synset details such as gloss, example(s) and synonym(s). The lexicographer translates the source language synset into the target language synset, while the validator uses same tool to validate these translated synsets.

There is a navigation panel which allows a lexicographer to navigate between synsets. The button 'Save & Next' saves the current synset and moves to the next synset. The source and target synset data is extracted from source and target synset files respectively. These files are in Dictionary Standard Format (DSF) with extension '.syns'. Some of the features of the existing system are: search by synset id, search by word, generate synset count, generate word count, reference to quotations, commenting on a current synset and linkage to the corresponding English synset, etc.

2.2 Advantages and Disadvantages of the Existing System

Some of the major advantages of the existing tool are: Firstly, it is a standalone tool. Hence it can be installed easily. Secondly, it is portable, i.e. the tool can be installed and used over different operating systems.

Though the existing Lexicographer's Interface has its own advantages and it is widely accepted by the IndoWordNet community, we can find several limitations with the system. Some of them are mentioned below:

- Tool works with flat files hence there is a high possibility of data redundancy, data inconsistency, etc.
- As the number of synsets increases, the processing time to perform various operations like searching, counting, synchronization, etc. increases.
- Being a standalone tool, the installation and configuration time increases with increase in number of machines.
- Merging data from different systems may lead to data loss, data redundancy as well as data inconsistency.
- Synset data is not rendered properly in the interface if there is any formatting mistake in the source or target file. Also, if any special character is added in a file then the synsets are not loaded in the system.

3 Developed System

3.1 Synskarta

The developed system, 'Synskarta' is an online interface for creating synsets by following the expansion approach. This web based tool is developed using PHP and MySQL which uses relational database management system to store and maintain the synset and related data. The IndoWordNet database structure (Prabhu et al., 2012) is used for storing and maintaining the synset data while

IndoWordNet APIs (Prabhugaonkar et al., 2012) are used for accessing and manipulating this data.

Synskarta overcomes the limitations of the standalone offline tool. The look and feel of the interface is kept similar to that of the existing system for ease of user adaptability. Most of the basic features of the existing system are incorporated in this developed system. Figure 1 shows the Lexicographer's Interface of the developed system.

3.2 Features of Synskarta

3.2.1 Features of Synskarta incorporated from the Existing System

The features of the existing system which are implemented with some improvements in Synskarta are as follows –

- *User Registration Module* - This module allows the system administrator to create user profiles and provide necessary access privileges to user. The user can login using the access privileges provided to him and accordingly the user interface is displayed to that particular user.
- *Configuration Module* – This module sets all the necessary parameters such as source language, target language and enables or disables certain features such as Source, Domain, Linking, Comment, References, etc.
- *Main Module* – This module allows the user to enter data in the target language panel by referring to data in the source language panel. The source panel and the target panel vertically divide the main module into two equal sized panels. Following are the major components of this module:
  - Source language panel – This panel is placed on the left of the screen which has fields for synset id, POS category, gloss, example(s) and synonyms of the source language synset.
  - Target language panel – This panel is placed on the right of the screen. This panel has non-editable fields such as synset id, POS category and editable fields such as gloss, example(s) and synonym(s) of the target language synset. The user is expected to enter the data in these editable fields.
  - Search – User can search a synset either by entering 'synset id' or a 'word' in a synset.
  - Advanced Search – Here user is allowed to search synset data by entering various parameters such as POS category, words appearing in gloss or in example, etc.
  - Comment – User can comment on a particular synset being translated.
  - English Synset – User can check the corresponding English synset for better clarity in translation process.
  - Navigation Panel – This panel allows the user to navigate between synsets. Button 'Save & Next' allows inserting or updating a current target synset and the data is directly stored in the IndoWordNet database.

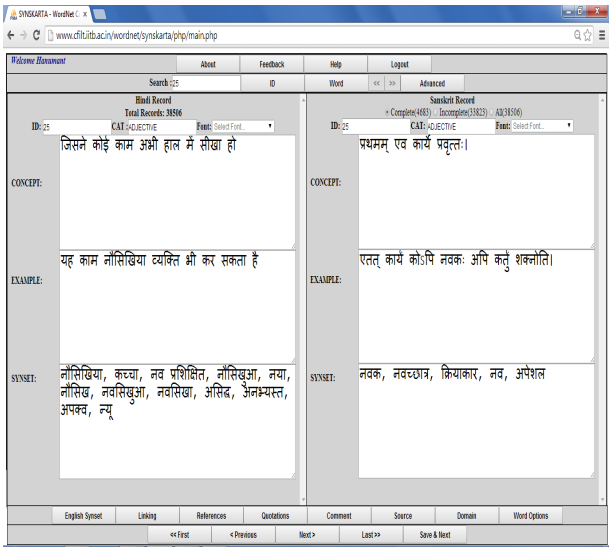

Figure 1. The Developed System – Synskarta

3.2.2 Features of Synskarta specific to Sanskrit Language

Apart from the features of the existing system, there are features which are specific to the Sanskrit language. Some of these features can also be applicable to other languages.

As far as Sanskrit is concerned, some of the features can be specific to a particular word in the synset. These are given in table 1. To capture these word specific features, the 'Word Options' button is provided. This window has various features which users can set or unset for a selected word.

Feature	Details
Word Types	Word can have *vaidika* or *laukika* word types. *vaidika* words are from vedic literature and *laukika* words are from post-vedic literature.
Accent	Accent is an appropriate tone for utterance of any particular syllable.
Class	There are 10 classes of verbs in Sanskrit. They are *bhvādi*, *adādi*, *juhotyādi*, *divādi*, *svādi*, *tudādi*, *rudhādi*, *tanādi*, *kryādi* and *curādi*.
Etymology	It is the study of the origin of a word and historical development of its meaning.
Pada	There are *padas* for suffixes of a verb such as *parasmaipada*, *ātmanepada* or *ubhayapada*.
Ittva	Verb can be *aniṭ*, *seṭ* or *veṭ*.

Table 1. Special Features of Synskarta specific to Sanskrit

3.2.2.1 Noun Specific Features

Noun specific features are displayed only if the POS category of the synset is NOUN. Figure 2 shows a screenshot of the 'Word Options' window for noun synsets. Following are some of the noun specific features of a word in Sanskrit language –

- Indication of Word Type (शब्द प्रकार, *śabda prakāra*) – In Sanskrit, words can have वैदिक (*vaidika*) or लौकिक (*laukika*) word types.
- Indication of Accent (स्वर, *swara*) – In Sanskrit, if a word has *vaidika* word type then it can have accents such as उदात्त (*udaatta*), अनुदात्त (*anudaatta*) or स्वरित (*svarita*). Again, *udaata* has sub-accents such as आद्युदात्त (*aadyudaatta*), मध्योदात्त (*madhyodātta*) or अन्तोदात्त (*antodaatta*). It is needed particularly in Sanskrit because the meaning of a word changes according to the place of accent.
- Identification of Gender (लिङ्ग, *liṅga*) – In Sanskrit, a gender can be masculine (पुंल्लिङ्ग, *puṃlliṅga*), feminine (स्त्रीलिङ्ग, *strīliṅga*) or neutral (नपुंसकलिङ्ग, *napuṃsakaliṅga*). For example, a word तट *(taṭa*) has two genders, masculine for तटः (*taṭaḥ*) and neuter for तटम् (*taṭam*). Hence, there is a need to store gender information for such type of words.

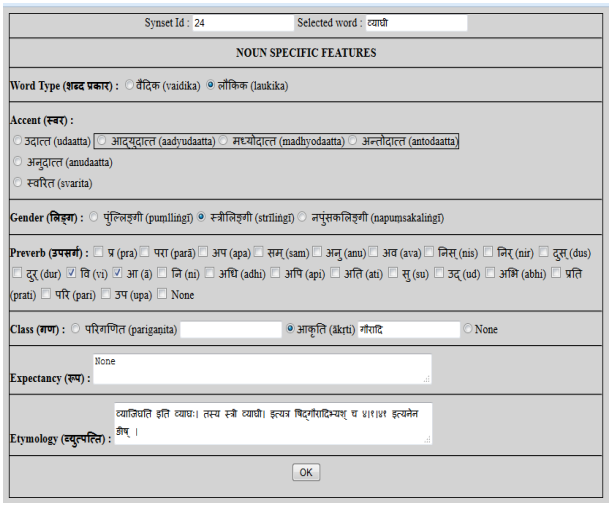

Figure 2. Word Options - Noun Specific Features

- Indication of Preverbs (उपसर्ग, *upasarga*) – In Sanskrit, there are 22 preverbs. Some of them are प्र (*pra*), परा (*parā*), अप (*apa*), etc. (Papke, 2005; Ajotikar et al. 2012). For example, for a word गन्ध (*gandha*), if we add preverbs सु (*su*), दुस् (*dus*) and उप (*upa*) we get words सुगन्ध (*sugandha*), दुर्गन्ध (*durgandha*) and उपगन्ध (*upagandha*) respectively.
- Indication of Class (गण, *gaṇa*) – Certain words in Sanskrit language belong to परिगणित (*parigaṇita*) or आकृति (*ākṛti*) class type. Each of this class type has class names. For example, a word शिव (*śiva*) belongs to *parigaṇita* class type having class name शिवादि (*śivādi*) and a word शौण्ड (*śauṇḍa*) belongs to *ākṛti* class type having class name शौण्डादि (*śauṇḍādi*).
- Expectancy (रूप, *rūpa*) – Certain words expect its related word to be in specific case(s). For example, a word अलम् (*alam*, 'enough') expects its related word to be in तृतीया (tṛtīyā, 'instrumental case') as अलं रोदनेन। (*alaṁ rōdanēna*, 'enough of crying').
- Etymology (व्युत्पत्ति, *vyutpatti*) – Most of the words in Sanskrit have etymology. For exam-

ple, a word कूलङ्कषः (*kūlaṅkaṣaḥ,* 'the sea'), कूलं कषति इति। (*kūlaṁ kaṣati iti,* 'one who cuts the shore').

3.2.2.2 Verb Specific Features

Verb specific feature list is displayed only if the POS category of the synset is VERB. Figure 3 shows the 'Word Options' window for verb synsets. Following are some of the verb specific features of a word in Sanskrit language –

- Indication of Word Type (शब्द प्रकार, *śabda prakāra*) – Verbs can have वैदिक (*vaidika*) or लौकिक (*laukika*) word types.
- Indication of Accent (स्वर, *swara*).
- Indication of Transitivity (कर्मकत्व,*karmakatva*) – A verb can be सकर्मक (*sakarmaka,* 'active') or अकर्मक (*akarmaka,* 'passive').
- Indication of Iṭtva (इट्त्व, *iṭtva*) - A verb can be अनिट् (*aniṭ*), सेट् (*seṭ*) or वेट् (*veṭ*).
- Indication of Class (गण, *gaṇa*) – In Sanskrit, there are 10 classes of verbs. Some of them are भ्वादि (*bhvādi*), अदादि (*adādi*), जुहोत्यादि (*juhotyādi*), etc.
- Indication of Pada (पद, *pada*) – In Sanskrit, there are padas for suffixes of a verb such as परस्मैपद (*parasmaipada*), आत्मनेपद (*ātmanepada*) or उभयपद (*ubhayapada*).
- Indication of Preverbs (उपसर्ग, *upasarga,*) – For verbs also there are 22 preverbs in Sanskrit. For example, when preverbs are attached to a root गम् (*gam,* 'to go'), we get verb forms such as आ√गम् (*ā√gam,* 'to come'), अनु√गम् (*anu√gam,* 'to follow'), निर्√गम् (*nir√gam,* 'to go out'), वि√गम् (*vi√gam,* 'to go away').
- Indication of Verbal Root Types (धातु प्रकार, *dhātu prakāra*) – A verbal root can have types such as धातु (*dhātu*), साधितधातु (*sādhitadhātu*), वैदिकधातु (*vaidikadhātu*) and सौत्रधातु (*sautradhātu*).
- Expectancy (रूप, *rūpa*) – Certain verbs expect its related word to be in specific case(s). For example, a root भी (*bhī,* 'to fear') expects its related word in पञ्चमी (*pañcamī,* 'ablative case') such as व्याघ्रात् भीतः (*vyāghrāt bhītaḥ,* 'feared of tiger').

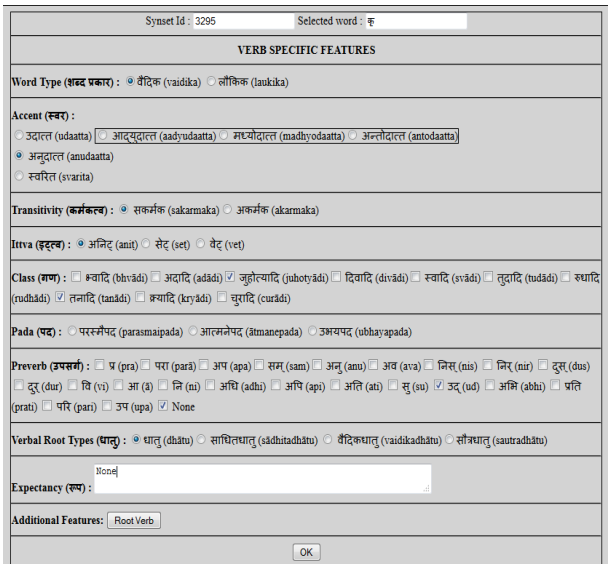

Figure 3. Word Options – Verb Specific Features

3.2.3 General Features not specific to any Language

Some of the general features which are not specific to any particular language are also provided in Synskarta, they are as follows:

- Vindication – This feature allows the user to record the special feature of a particular word in a current synset.
- Source – This feature allows the user to record the information about source of the synset.
- Domain – This feature allows the user to record the information about domain of the synset.
- Linking – Many-to-Many linking of words is supported in this feature.
- Quotations – The feature to add quotations as additional examples are supported.
- Root Verb – This feature allows the user to enter the root verb of a given word.
- Feedback – Feedback related to the tool and its features are captured here.

3.3 Advantages and Limitations of Synskarta

Major advantages of Synskarta are as follows:

- Centralized system
- No data redundancy and inconsistency.

- Online access from anywhere in the world
- No text files to maintain data
- Faster processing and updating
- Multiple users can work at the same time
- Can be used by all the language WordNets

Synskarta has a few limitations. The major limitation is that, it is heavily dependent on internet access or networking. Another limitation is that currently there are rendering issues depending on the device being used.

3.4 User Evaluation

The beta version of Synskarta was released to the Sanskrit WordNet creation team to capture users' experiences and the feedback which are as follows:

- Searching of synset from source as well as target language is faster.
- User does not have to maintain files for source and target language.
- Separate synchronization is not required. Everything is handled by the tool itself.
- The system always shows the updated synsets.
- User has to depend on the internet while using the system.

4 Conclusion

It has been observed that the existing system for synset creation activity is not up to the expectations of the user, as it uses text files for storage and processing. This may lead in creation of redundant synset data and hence, data maintenance becomes difficult. A web based tool 'Synskarta' is developed to overcome the limitations of the existing system. This is a centralized system which uses relational database to store and maintain data. The difficulty of maintaining data in flat files is taken care of. User feedback is found to be positive and this tool may prove essential for the IndoWordNet community.

5 Future Scope and Enhancements

In future, the tool can be expanded to link to the other Indian languages. This tool can have additional features such as capturing appropriate pronunciation of the synset, features to capture video, images or documents related to the synset. Features such as automatic translation, transliteration, transcription can also be implemented over the time. Further, there can be a feature to link IndoWordNet synsets to the foreign language WordNets. A feature to generate produced words can also be implemented. Link to FrameNet of Sanskrit verbs is an important work that will be undertaken in future. This will be useful in the light of development of dependency tree banks.

Acknowledgements

We sincerely thank the members of the CFILT lab – IIT Bombay and IndoWordNet community.

References

Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda and Pushpak Bhattacharyya. 2010. *Introducing Sanskrit Wordnet.* In Principles, Construction and Application of Multilingual WordNets, Proceedings of the 5th GWC, edited by Pushpak Bhattacharyya, Christiane Fellbaum and Piek Vossen, Narosa Publishing House, New Delhi, 2010, pp 257 – 294.

Malhar Kulkarni, Irawati Kulkarni, Chaitali Dangarikar and Pushpak Bhattacharyya. 2010. *Gloss in Sanskrit Wordnet.* In Proceedings of Sanskrit Computational Linguistics. Jha. G. Berlin: Springer-Verlag / Heidelberg. pp 190-197.

Tanuja Ajotikar Malhar Kulkarni, and Pushpak Bhattacharyya. 2012. *Verbs in Sanskrit Wordnet.* Proceeding of 6th Global WordNet Conference, Matsue, Japan. pp 30 – 35.

Narayan D., Chakrabarty D., Pande P. and Bhattacharyya P. 2002. *An Experience in Building the Indo WordNet - a WordNet for Hindi*. 1st Global WordNet Conference, Mysore, India.

Neha R Prabhugaonkar, Apurva S Nagvenkar, Ramdas N Karmali. 2012. *IndoWordNet Application Programming Interfaces.* COLING2012, Mumbai, India.

Papke Julia. 2005. *Order and Meaning in Sanskrit Preverbs.* 17th International Conference on Historical Linguistics, Madison, Wisconsin.

Pushpak Bhattacharyya. 2010. *IndoWordNet*. In the Proceedings of Lexical Resources Engineering Conference (LREC), Malta.

Venkatesh Prabhu, Shilpa Desai, Hanumant Redkar, Neha Prabhugaonkar, Apurva Nagvenkar, Ramdas, Karmali. 2012. *An Efficient Database Design for IndoWordNet Development Using Hybrid Approach.* COLING 2012, Mumbai, India. p 229.

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