Tools for IndoWordNet Development

Shilpa Desai
Dept. of Computer Science
& Tech., Goa University
sndesai@gmail.com

Ramdas Karmali
Dept. of Computer Science
& Tech., Goa University
rnk@unigoa.ac.in

Sushant Naik
Indradhanush WordNet
Goa University
susha410@gmail.com

Shantaram Walawalikar
Consultant, Indradhanush WordNet
Goa University
goembab@yahoo.co.in

Damodar Ghanekar
ILCI - Konkani Team
Goa University
damodar.ghanekar@rediffmail.com

Abstract

WordNet is an essential resource in Natural language processing. A lexical database like WordNet has a variety of practical applications like machine translation, information retrieval and many more. The creation of a comprehensive WordNet requires many hands and minds working collaboratively. In addition WordNet creation and maintenance demands creation of a wide range of software tools. All this is possible with help from some funding agency. Department of Information Technology, Ministry of Communications and Information Technology, Government of India, New Delhi is one such funding agency who has been a driving force behind the ongoing research work of WordNet development in India. Currently the IndoWordNet group in India is working towards the development of a multilingual WordNet which includes 16 Indian languages. Two more have been added recently. Many tools have been developed to assist in the WordNet development process. Based on the multilingual WordNet work being carried out by the IndoWordNet family need is felt for more tools for WordNet. This paper discusses the various existing WordNet tools from different dimensions. We also propose a tool to augment the building of IndoWordNet. The key features for this new tool being proposed are also covered.

Keywords: WordNet, IndoWordNet, Software tools

1 Introduction

WordNet is a lexical database which comprises of synonym sets, gloss, position in ontology and relations. A synonym set in a WordNet represents some lexical concept (Miller, 1993). For example the English synonym set {family, household, house, home, ménage} represents the concept of “A social unit that lives together”. The gloss gives definition of the underlying lexical concept and an example sentence to illustrate the concept. For each syntactic category namely noun, verb, adjective and adverb, a separate ontological hierarchy is present. Each synset is mapped into some place in the ontology. The WordNet also maintains semantic and lexical relations. Semantic relations are between synsets and lexical relations are between words. Semantic relations are Hyponymy, Hypernymy, Meronymy, Holonymy etc. Lexical relations are antonomy and such (Bhattacharyya, 2010). Thus we can say a WordNet is a dictionary plus a thesaurus and much more. Building a functional WordNet for a language is no easy task. It requires lexicographers and computer scientists to work jointly to create such an online lexical resource. This gives rise to a need to develop a wide range of software tools to create, maintain and make potential use of the WordNet.

This paper is organized as follows. Section 2 discusses the approaches used to build the WordNet and types of WordNet. Software tools for WordNet are covered in section 3. Section 4 and section 5 deal with tools for stand-alone and multilingual WordNets respectively. A comparison between the tools is given in section 6. In section 7 we propose an augmentation to the expansion approach framework and the tools required thereof and section 8 concludes the paper.

2 Approaches used to build WordNet and types of WordNet

WordNets can be built using the *merge approach* or the *expansion approach* (Vossen, 1998). The merge approach is also referred to as WordNet construction from *first principles* (Bhattacharyya, 2010). This approach uses an exhaustive dictionary or a collection of dictionaries of the language for which the WordNet is being created as a base. A dictionary lists the words in a language and the different senses in which the word is used. The lexicographers then construct synsets for each sense of the word by following the three principles of minimality, coverage and replacebility (Bhattacharyya, 2010). In this approach the lexicographers enter many details such as synonym set, gloss, position in ontology and the relations.

The other approach towards WordNet creation is the expansion approach. This approach makes use of an already existing functional WordNet of language as the source. Each synset entry from the source WordNet is carefully studied by the lexicographer to understand the underlying concept. The lexicographer then gathers the corresponding words for that concept in the target language for which the WordNet is being developed. Here the lexicographer just needs to add the synonym set and the gloss; the position in ontology and the semantic relations are directly borrowed from the source WordNet.

Both the merge and expansion approaches have their advantages and limitations (Bhattacharyya, 2010). The lexicographer using the merge approach needs to be well versed with only the language for which the WordNet is being created. He does not have to deal with problems such as culture specific concepts of the source language. This becomes a limitation to the lexicographer in expansion approach. The lexicographer using the expansion approach should have enough knowledge of both source language and target language.

We can classify the WordNets as *stand-alone* WordNet or *multilingual* WordNets. A stand-alone WordNet is a single WordNet such as Princeton University's English WordNet or Hindi WordNet (S. Jha et. al., 2001) of Indian Institute of Technology Bombay (IIT-B). Such WordNets are not dependent directly on any other WordNet. They may be linked explicitly to some other WordNet. Such WordNets show more language specific characteristics and are usually built using the merge approach. They are a complete entity by themselves and cover most of the lexical concepts present in the concerned language.

Multilingual WordNets such as EuroWordNet and IndoWordNet (Bhattacharyya, 2010) is a collection of WordNets interlinked together. The links between the WordNets could be either implicit or explicit. Multilingual WordNets can be built either by linking existing WordNets or by simultaneous construction of WordNets using expansion approach. When the multilingual WordNets are created by expansion approach the target WordNet is implicitly dependent on the source WordNet as it borrows semantic relations from the source WordNet (Bhattacharyya et. al., 2010).

3 Software tools for WordNet

Tools for WordNet are required to assist in the overall development of a functional WordNet. Tools can be broadly classified depending on the WordNet type, purpose and WordNet creation approach. Figure 1 shows the classification of the different tools for a WordNet.

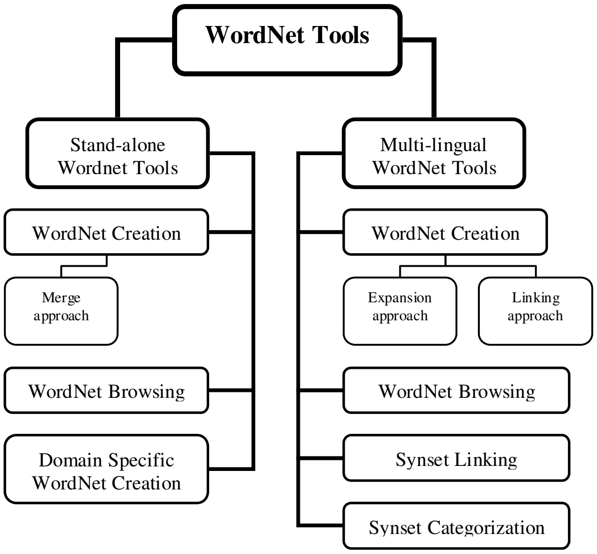

Figure 1: Categorization of Software tools for WordNet

The tools listed under stand-alone are tools which work on a stand-alone WordNet. These tools can work with a concerned language only and need not refer to or connect to a WordNet of any other language. The tools listed under multilingual are tools which work on more than one language WordNet simultaneously.

4 Tools for stand-alone WordNet

Different custom built tools are being used by the various WordNet building teams. In addition more tools need to be developed. The tools working with only one WordNet at a time are referred to as stand-alone WordNet tools. These tools are used for various purposes such as creating a WordNet, browsing a WordNet, Domain specific Sub-WordNet.

4.1 WordNet Creation

A stand-alone WordNet during creation is not linked to some other WordNet hence such a WordNet is created using the merge approach. The Hindi WordNet was developed using this approach. The tool for the same was created by IIT-B, Mumbai. The tool used for the same allows the lexicographer to make an entry of a synset (concept) in the WordNet.

Existing features of the tool: It provides interfaces to the lexicographer, which enables him/her to enter / set the following

- Word category such as noun, verb, adjective, adverb.
- Head word and synonym set
- Gloss or Concept
- Examples relevant to the concept
- Semantic and lexical relations
- Ontology details
- The reference and etymology if any.

The existing features of the tool serve the purpose of WordNet creation well. The tool could be made more sophisticated by adding a few features.

Enhancements to the tool: The tool to create a WordNet can have the following additional features

- An automatic example sentence selector module could be included. This makes the job of the lexicographer easy, as this module will provide example sentences which are extracted from a corpus. The lexicographer will have to choose the example sentence which has the appropriate sense.
- A semantic relationship suggestion module based on gloss, example sentences and synset entries already made. For example for the synset {herb, herbaceous_plant} the gloss is "a plant lacking a permanent woody stem, many are flowering garden plants or potherbs some having medicinal properties" from this gloss a relation between herb and plant could be suggested by the advance tool based on the fact that plant appears twice in the gloss.
- Using a tagged corpus of the language/ examples or gloss entered in the current WordNet an automatic head word suggestion list of words which do not already appear in the WordNet could be generated. This will be particularly helpful for languages where there is corpus data available but no proper exhaustive dictionary available in the language.

The above stated features can be considered as a refining cum validating tool to further improve the quality of the WordNet so as to expand the WordNet entries.

4.2 Tools to browse stand-alone WordNets

Such a tool for Hindi WordNet is available in both web based and java based interfaces developed by IIT-B, Mumbai. The salient features of this tool are as listed below

- It allows user to search a word in the WordNet.
- It displays the detailed listing of the word, in different categories such as noun, verb and also the different senses the word can assume.
- It provides a detailed list of relations the synset could have.
- It also gives the ontological position for the synset.
- It provides a category wise list of words present in the WordNet like nouns, verbs, adjectives and adverbs, from which one can choose a word and see the corresponding WordNet entry.
- The search accepts words in the root form as well as morphological variant and maps it to the relevant entries of the WordNet. Example फूलता and फूलना both map to फूलना
- It also provides a script specific keyboard for the concerned language to facilitate data entry for searches
- In case if the word being searched is wrongly spelt or not found the tool suggests possible close matches.

4.3 Domain-Specific Sub-WordNet creation

One of the applications of a WordNet is to aid translation. Translation of text could be domain specific at times, such as translation of medical documents, or tourism related data and such. Domain is to be identified at WordNet creation level. Thus only a subset of the WordNet having concepts related to the specific domain under consideration will be used in this case. Further a general WordNet for a language may not incorporate concepts for a specific domain. Thus tools to derive domain specific concepts and expand the WordNet (derive a domain specific Sub-WordNet) may be required.

5 Tools for multilingual WordNets

Multilingual WordNet tools work with multiple language WordNets simultaneously. The main categories of tools found are creation tools, browsing tools and Synset Categorization tool. We also propose the Concept Merging Tool and Interactive Synset Linking Tool to aid the development of a more effective IndoWordNet in the context of the Indian cultural scenario.

5.1 Multilingual WordNets Creation

As mentioned earlier multilingual WordNets are more than one WordNet whose synsets are linked together using common identification number. Such multilingual WordNets can be built using expansion approach tool or WordNet linking tool.

a. Expansion Approach Tool:
Currently the IndoWordNet which is a multilingual WordNet is being developed using the expansion approach. A tool to create WordNet using expansion approach named *MultiDict* (Bhattacharyya, 2010) has been developed by Indian Institute of Technology, Bombay. This tool provides an easier and faster means for the lexicographers to create WordNet entries. Here the source language WordNet is already prepared and it acts as a guide for the lexicographer to make corresponding entries for the target language. The WordNets for many Indian languages are being developed using this approach, as part of the IndoWordNet by using Hindi WordNet as the source (Narayan D et. al., 2010).

Existing features of the tool: The tool provides interfaces which allows us to enter the following

- Head word and synonym set
- Gloss or Concept
- Examples relevant to the concept
- Link corresponding words in synsets
- Check corresponding English synset
- The reference and etymology if any.

Proposed enhancements to the tool: In addition the tool to create a WordNet using expansion approach can have the following additional features

- Show the ontological nodes to the lexicographer
- Show the borrowed relationships

b. WordNet Linking Tool:
If two or more WordNets are individually developed independent of each other using the merge approach then the two can be linked to form part of a multilingual WordNet using such tool.

Proposed features of the tool: This tool should have the following features

- Given a concept in language X, aid the search for a corresponding concept in language Y by the lexicographer, and link the two concepts. For example, if the concept is “A social unit that lives together” in English with the corresponding synonym set {family, household, house, home, ménage} then the lexicographer searches for परिवार (parivaar) in Hindi WordNet and links the two corresponding synsets.
- In synset {a,b,c,d} of language X and corresponding linked synset {p,q,r,s} of language Y , interface to link a to q if required.

5.2 Multilingual WordNets browsing tool

A tool to browse multilingual WordNet should incorporate all the features of a standalone WordNet browsing tool. These features will work the same for each individual WordNet of the multilingual WordNet group.

The IndoWordNet multilingual WordNet currently provides a web-based tool which allows us to see the synsets of the other member languages by choosing the corresponding synset of Hindi.

Proposed features of the tool: In addition to the features of stand-alone to exploit the true potential of a multilingual WordNet, a tool to browse the multilingual WordNet should incorporate the following features

- Enable search for word in WordNet of language X, say Konkani, and then for a chosen synset (concept) of language X the corresponding synsets in language Y, say Marathi, or all other languages of the multilingual WordNet. Here both languages X and Y are members of the multilingual WordNet.

- The replacebility or link if set between words in the synsets of languages X and Y should be shown
- The reference and etymology if any can also be shown.

5.3 Interactive synset linking tool

Since mostly multilingual WordNets are created using expansion approach, they use one language as a source language. For example IndoWordNet uses Hindi as the source language. For languages belonging to the same family like Dravidian languages Malayalam and Tamil will be linked to each other via Hindi synsets. This mapping may not be properly valid as there may be concepts in these two languages which do not appear in the same way in Hindi. For example consider Marathi and Konkani. The synset {roti, chappati} of Hindi is a more general concept while in both Marathi and Konkani we have chappati and Bhakri as more specific concepts. Hence a link between the synsets of these two languages is more appropriate for exchanges (translations) between these pairs of languages. Also the feature of linking word pairs in synsets from target language to Hindi will not capture the finer nuances when two target languages like Malayalam and Tamil are being linked together via Hindi. The tools can be used to specify such links between word pairs in synsets. Such a tool can take advantage of the fact that the synsets are built on concepts and linked. Also an automatic linking can be generated which can be interactively verified or changed by the lexicographers for the concerned language pair. This will also act as a check as to how well the expansion approach works to achieve mapping between the language pairs which do not belong to same family as the source WordNet.

5.4 Synset categorization tools

One problem encountered in linking WordNets in multilingual WordNets is that the lexicographer does not find a corresponding concept in the target language for a concept in the source language. This could happen due to cultural differences. To overcome such differences in IndoWordNet, a synset ranker tool was developed. This tool rates a concept with respect to a particular language, whether the concept is present in the language or not. Such a tool helps in rating the concepts in multilingual WordNets as universal concepts, PAN Indian or belonging to a certain family of languages. Such classification of concepts helps to concentrate on universal concepts first when building a WordNet using the expansion approach.

6 Comparison between tools for standalone WordNets and Multilingual WordNets

Tools for standalone WordNet and multilingual WordNets compared based on different aspects are shown in Table 1 and Table 2.
Table 1 shows the evaluation of the tool from lexicographer's point of view and table 2 shows the tool from computer programmer's view point.

Thus we see that the tool for standalone WordNet creation requires a lot of data entry per concept, hence the entire process becomes slow. It is easier to use the multilingual WordNet creation tool which uses the expansion approach. Since the categories and relations are borrowed from the source language the data entry required per concept is relatively less. One benefit of the standalone WordNet creation tool is that the category, relations are visible to the lexicographer hence the concept is more clear to him. On the other hand the lexicographer using multilingual WordNet creation tool (expansion approach) cannot see the corresponding relations, ontological nodes, neither can he set antonym relation.

Tool\Aspect	Ease of Use	Amount of Data entry required /time taken	Features provided
Stand-alone WordNet Tools			
WordNet Creation using merge approach	Relatively complex	High / more	Many
Browsing tool	easy	N.A.* / less	Many
Synset Ranker	Easy	Low / less	N.A.*
Multilingual WordNet Tools			
WordNet creation using expansion approach	Easy	Low/ less	Few
Multilingual WordNet Browsing	Moderate	N.A. / moderate	Many

* N.A. = Not Applicable

Table 1: Lexicographers view

Tool\Aspect	**Software Complexity**	**Features included**
Stand-alone WordNet Tools		

WordNet Creation using merge approach	Relatively easy	Many
Browsing tool	Moderately complex	Moderate
Synset Ranker	Relatively easy	Moderate
Multilingual WordNet Tools		
WordNet creation using expansion approach	Moderate	Moderate
Multilingual WordNet Browsing	Complex if morphology for all languages are incorporated	Moderate

Table 2: Computer programmer's view

Standalone WordNets browsing tools are easier to operate and simple, as they are not linked to other WordNets. The resources required to implement such a tool are limited to the specific language under consideration. A browsing tool for a multilingual WordNet is relatively complex to implement. Since the relations are not entered they need to be borrowed. Modules which search for morphological variants need to be implemented for all member languages to enable cross language searches.

7 Augmented framework for expansion Approach and the tools required

The rich and varied cultural scenario found in India results in many culture specific concepts in a language (Bhattacharyya et. al., 2010). The expansion approach used to develop IndoWordNet uses Hindi WordNet as the source WordNet. Two limitations which arise due to this approach are

a. ***Concept in Source language not found in Target language:*** In such a case there are three possibilities
   i. Concept not present in target language. The concept can be borrowed in the target language by transliteration (Bhattacharyya, 2010). For example : Baisakhi a Punjabi festival which appears in Hindi WordNet can be borrowed by Konkani
   ii. Concept present in target language as a more specific concepts. In such a case a more general concept may have to be coined in the target language as combination word. For example *roti* is a more general concept in Hindi which has specific concepts like chapatti, roti, bhakri, fulki, etc in Konkani.
   iii. Concept present in target language as a more generic concept. In this case again the target language can either borrow the specific concept or use combination word to illustrate the concept.

b. ***Concept found in target language but not found in source language:*** In this case the concept will have to be borrowed by the source language by means of transliteration. For example NavavArI in Konkani is a concept not found directly in Hindi (Walawalikar S et. al., 2010).

The IndoWordNet should include all concepts of all the member languages. We thus propose the *concept merging tool* to aid the development of a more effective IndoWordNet. Figure 2 shows this augmented framework for expansion approach

7.1 Language Specific Concept Collection

Each of the member languages of the Multilingual WordNet except for the source WordNet i.e. Source language will first have to list all their culture specific or language specific concepts which do not appear in the source WordNet. This list can be called *language-specific-concept-list*. This list can be prepared by using one of the two approaches

a. Dictionary approach: The target language will have to use a good dictionary of the language to identify the concepts that are present in the dictionary but not present in the source WordNet.
b. Corpus approach: The target language can use a sufficiently large corpus if available to identify the concepts that are present in the corpus but not present in the source WordNet.

7.2 Concept Merging Tool

The proposed *Concept Merge Tool* will merge the different *language-specific-concept-list* made available by the different target languages. This tool will finally create an *Assimilated-concept-list.* The tool should have the following features

- It provides an interface whereby the *language-specific-concept-list* which is in target language is converted to common language manually by the lexicographer. The common language will have to be decided by the members of the IndoWordNet group
- It merges the converted *language-specific-concept-list* given by each of the member languages into one common *Assimilated-concept-list* which will be in the common language. This process can use manual community based approach for this purpose.

- It removes any redundancy (duplication of the same concept) from the *Assimilated-concept-list*
- It generates a unique identification number (Id) for each concept in the *Assimilated-concept-list*
- It maintains a vector for each concept with an entry for each language. This vector can be used to identify in which all languages the concept under consideration is present. There may be some concepts which are present in more than one closely related language.

7.3 Expanding Source WordNet

Once the *Assimilated-concept-list* is ready, using the same, the source WordNet can be expanded to include the language specific concepts. All concepts in the expanded source WordNet are then included in the other target WordNet using expansion approach.

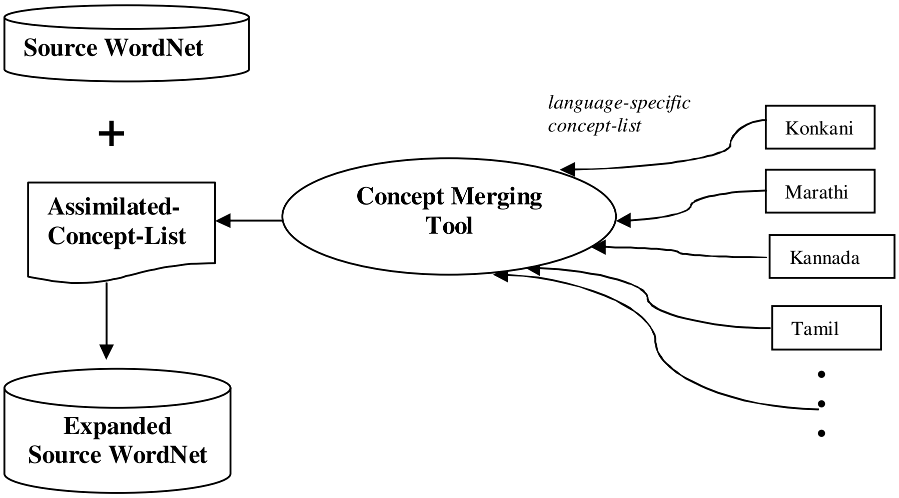

Figure 2: Augmented framework for expansion Approach

8 Conclusion

In this paper we have put forth our ideas about the WordNet tools. The existing features of the tools familiar to us have been summarized and possible enhancements to these tools have been proposed. We have also suggested an enhancement to the existing expansion approach IndoWordNet creation method.

The need for more new tools has been emphasized considering the importance of multilingual WordNet development in the current NLP era. The essential features for the new proposed tools have also been highlighted.

References

Bhattacharyya P., *IndoWordNet*, Lexical Resources Engineering Conference 2010 (**LREC 2010**), Malta, May, 2010.

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.

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande and P. Bhattacharyya, *An Experience in Building the Indo WordNet - a WordNet for Hindi*, First International Conference on Global WordNet, Mysore, India, January 2002.

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

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller, 1993, *Introduction to WordNet: An On-line Lexical Database*

S. Jha, D. Narayan, P. Pande, P. Bhattacharyya, *A WordNet for Hindi,* International Workshop on Lexical Resources in Natural Language Processing, Hyderabad, India, January 2001.

Shantaram Walawalikar, Shilpa Desai, Ramdas Karmali, Sushant Naik, Damodar Ghanekar, Chandralekha D'Souza and Jyoti Pawar, *Experiences In Building The Konkani WordNet Using The Expansion Approach*, Proceedings of the 5th Global WordNet Conference, Mumbai, Narosa Publishing House, India.

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