# Introduction to Tools for IndoWordNet and Word Sense Disambiguation

**Arindam Chatterjee, Salil Rajeev Joshi, Mitesh M. Khapra, Pushpak Bhattacharyya**
**{ arindam, salilj, miteshk, pb }@cse.iitb.ac.in**
Department of Computer Science and Engineering,
Indian Institute of Technology Bombay, Mumbai

## Abstract

Lexically rich resources form the foundation to all NLP tasks. Maintaining the high quality of resources is thus a high priority issue.

In this paper we exhibit the tools developed at IIT Bombay, for the purpose of creation, enhancement and maintenance of the WordNets, as well as the ones used for NLP tasks that use WordNets directly, like Word Sense Disambiguation.

The paper presents online and offline tools for WordNet creation, synset categorization tool, sense marking tool, and the Hindi Morphology analyzer tool.

## 1. Introduction

Princeton English WordNet (C. Fellbaum, 1998) is an ontological, machine readable lexical database for English language developed at Princeton University. It graduated to become one of the most used and prized among language resources. When a language resource of quality as high as the English WordNet comes into being, several tools are developed to utilize, enhance and maintain the resource as best as possible.

Since the birth of the English WordNet, WordNets for many other languages have spawned. In case of Indian languages, Hindi WordNet (Dipak Narayan et al., 2002) was the first of its kind. The Hindi WordNet (`http://www.cfilt.iitb.ac.in/wordnet/webhwn/`) was developed at the Indian Institute of Technology, Bombay. Consequently, Marathi (`http://www.cfilt.iitb.ac.in/wordnet/webmwn/`) and Sanskrit (under development) WordNets have also been developed. Correspondingly, a number of tools were developed to provide better functionality and transparency to this impeccable language resource, which not only forms the heart of WordNets for all other Indian languages, but also of all NLP work in India.

Word Sense Disambiguation or WSD is the problem of computationally identifying the correct sense of a word in a context. Hence, tools developed for WordNets, which are characterized by senses become equally important for sense disambiguation purposes.

The purpose of these tools is four fold. First, they are an immense aid in faster completion of WordNet related work. Secondly, these tools bring about a uniform structure to the WordNets, which is of paramount importance to bring high quality to any ontological knowledge resource. The tools are mostly used by linguists, who may not be savvy, technically, to elude the technical details associated with the WordNet. Last but not the least, the tools provide transparency and clarity to the structure and intricate details of the resource as well provide security for the data embedded in the each WordNet.

## 2. Roadmap

The layout is as follows: Section 3 provides a summary of the tools used for the Hindi WordNet. Section 4 follows up with tools used for sense disambiguation purposes, and Section 5 deals with tools used in WordNet APIs.

The following diagram shows the tools and their dependencies:

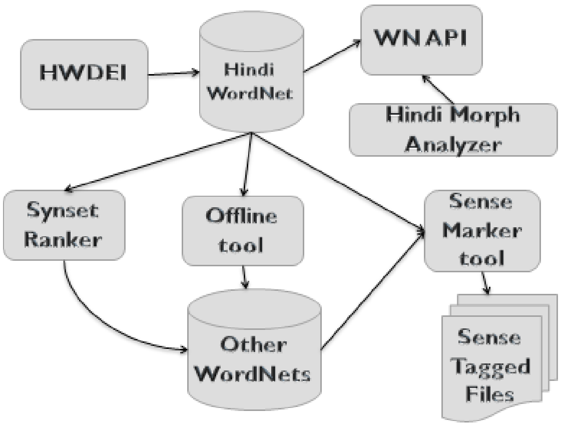
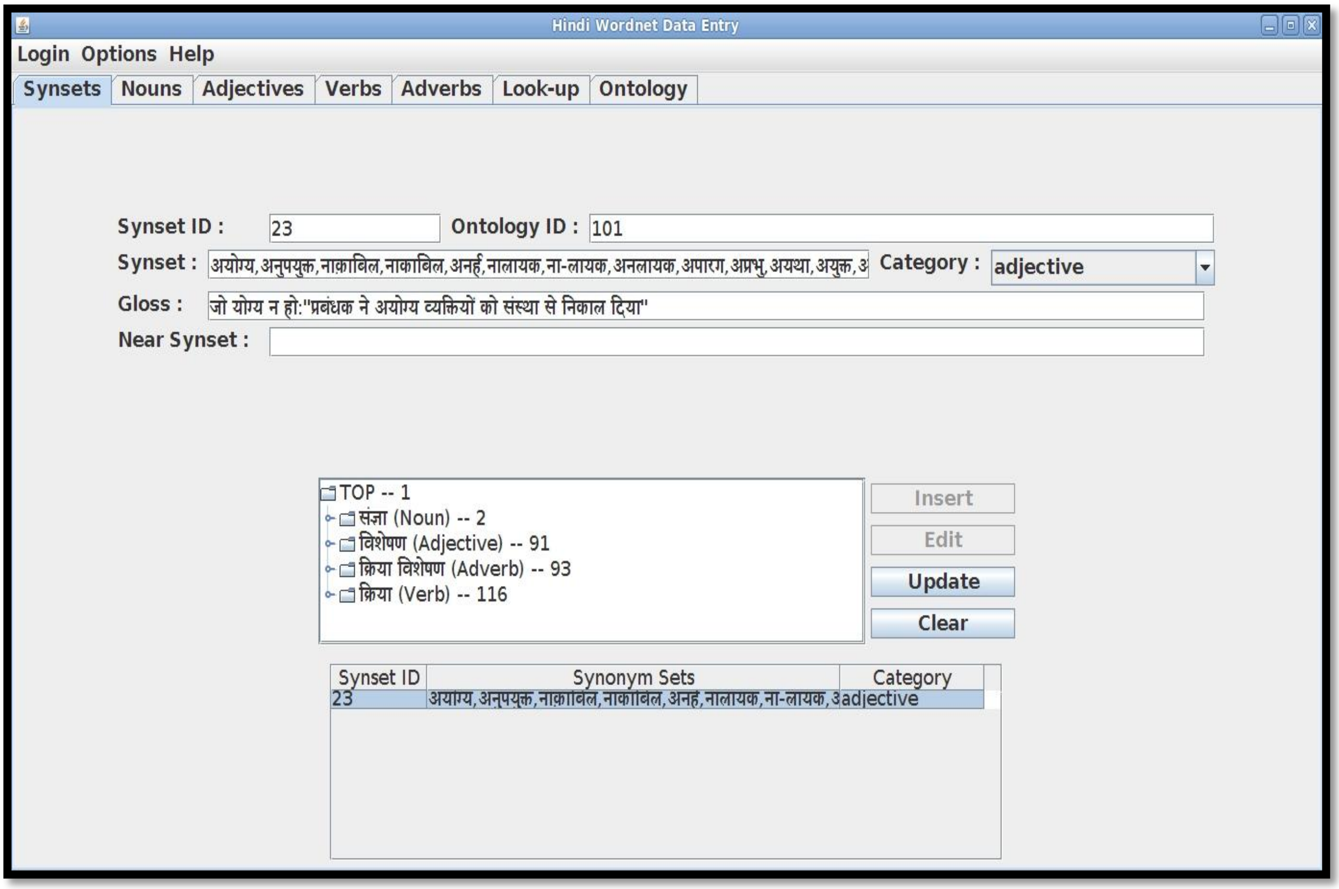


***Figure 1**: All the tools and their dependencies*



Synset ID : 23 Ontology ID : 101

Synset : अयोग्य,अनुपयुक्त,नाक़ाबिल,नाकाबिल,अनर्ह,नालायक,ना-लायक,अनलायक,अपारग,अप्रभु,अयथा,अयुक्त,अ Category : adjective

Gloss : जो योग्य न हो:"प्रबंधक ने अयोग्य व्यक्तियों को संस्था से निकाल दिया"

Near Synset :

TOP -- 1
संज्ञा (Noun) -- 2
विशेषण (Adjective) -- 91
क्रिया विशेषण (Adverb) -- 93
क्रिया (Verb) -- 116

| Synset ID | Synonym Sets | Category |
|---|---|---|
| 23 | अयोग्य,अनुपयुक्त,नाक़ाबिल,नाकाबिल,अनर्ह,नालायक,ना-लायक,अ | adjective |

*Figure 2: Hindi WordNet Data Entry Interface*

## 3. Tools for WordNet

In this section we will focus on the tools that were developed at IIT Bombay for the purpose of maintenance and enhancement of the Hindi WordNet.

### 3.1. Hindi WordNet Data Entry Interface Tool

To facilitate a simple GUI based synset insertion point for linguists working at IITB, this data entry interface was developed. The interface allows the linguists to insert or modify the Hindi synsets easily.

This tool is designed for creating language specific synsets, and was originally created for Hindi language only. The tool is Database based tool, and the data entered using this tool is directly updated in Hindi WordNet database maintained at IIT Bombay. For this reason, the tool is used only from within IIT Bombay by the linguistic team working on generation of Hindi WordNet.

A similar interface was later on designed for Marathi WordNet as well, as the linguistic team for Marathi language is also part of IIT Bombay NLP group.

For other languages, a generic, offline tool was later on created, and the details are presented in Section 3.2.

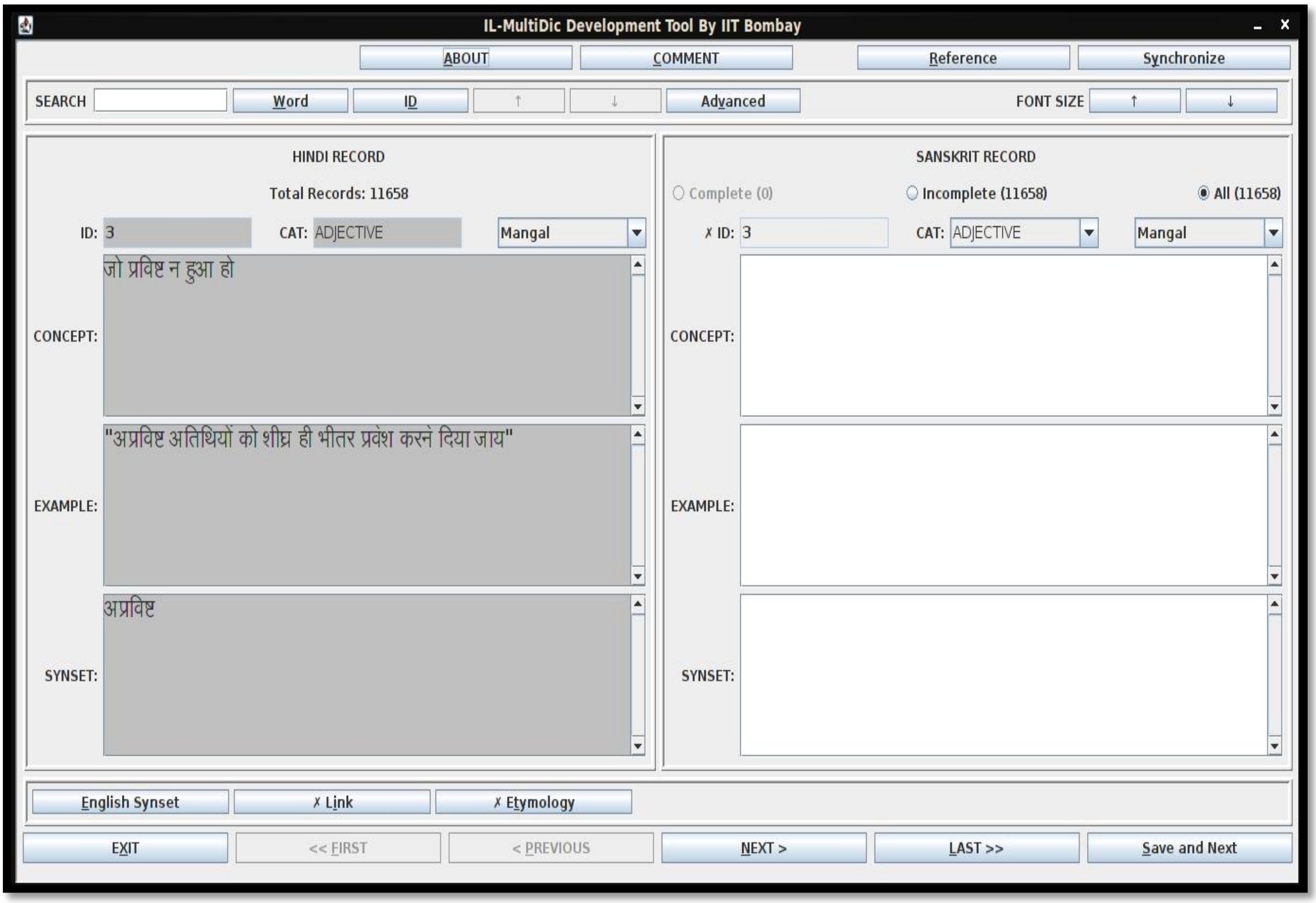
The tool has facilities for faster lookup of category (nouns, adjectives, verbs etc.) specific search and includes options for specifying relations between synsets.

The tool allows the users to search for existing synsets using either the Synset IDs or Synset words. The tool also keeps track of the ontology ID, category and other fields which are related to the synset.

The tool provides a facility of finding all synsets which contain a specific word or pattern. This is useful in cases where there are many synsets corresponding to a word or pattern, and the exact Synset ID is to be found out.

The tool also includes user friendly options for changing the font-size, feel and look etc.

The tool is developed in java in order to make it platform independent, and works for Indic languages on platforms which support Unicode.



HINDI RECORD

Total Records: 11658

ID: 3 CAT: ADJECTIVE Mangal

CONCEPT: जो प्रविष्ट न हुआ हो

EXAMPLE: "अप्रविष्ट अतिथियों को शीघ्र ही भीतर प्रवेश करने दिया जाय"

SYNSET: अप्रविष्ट

SANSKRIT RECORD

○ Complete (0) ○ Incomplete (11658) ● All (11658)

✗ ID: 3 CAT: ADJECTIVE Mangal

CONCEPT:

EXAMPLE:

SYNSET:

***Figure 3***: *Offline tool*

### 3.2. Offline Tool

The offline tool is a java-based open-source tool created for faster creation of Indic WordNets using Hindi WordNet as a pivot language. The tool provides a very similar feel to that of the online interface available for the Hindi WordNet and is publicly downloadable. The tool was created at IIT Bombay.

Juxtaposed to HWDEI tool explained in Section 3.1, the offline tool was created as a file based tool for creation of WordNets. The tool uses Hindi WordNet as a pivot, and provides a rapid way of constructing Synsets for any language for the linguists.

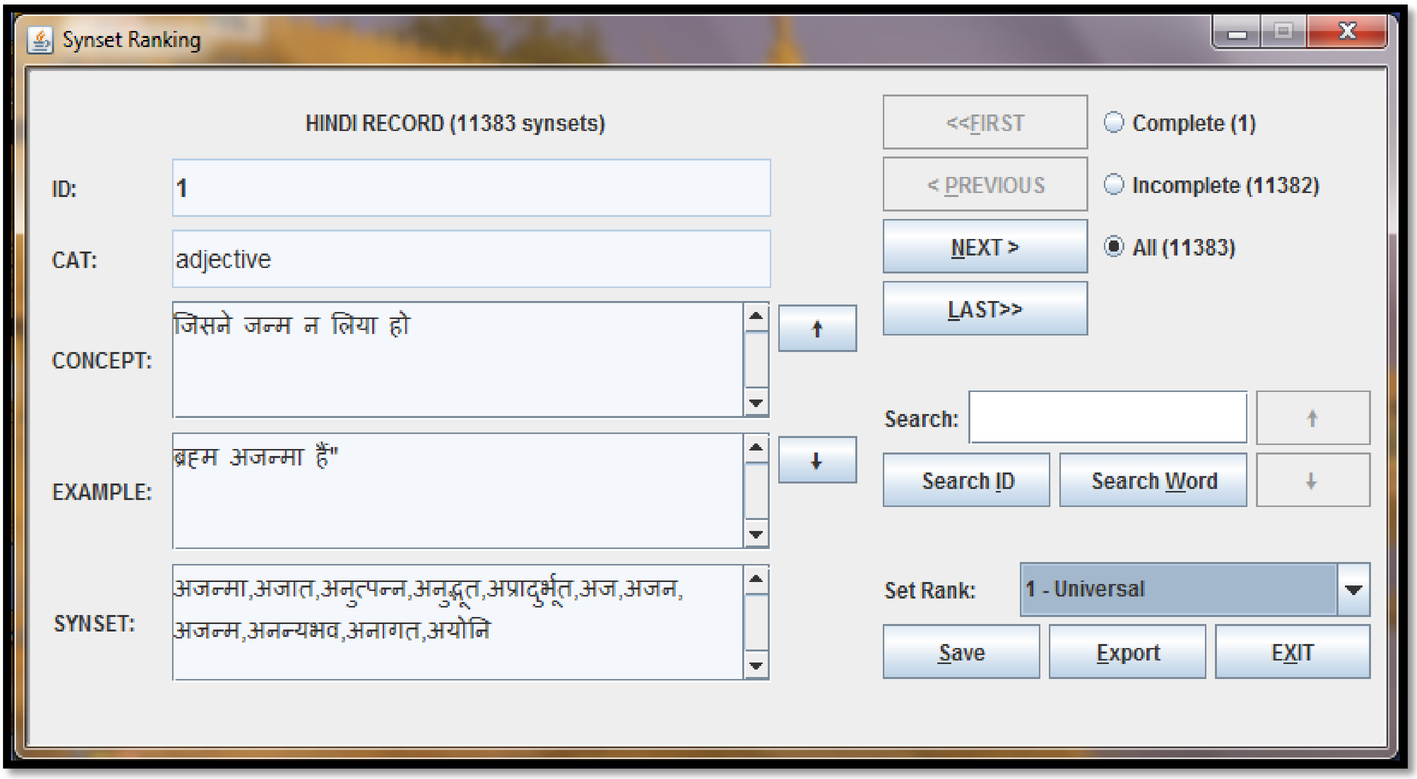
As shown in the screenshot, the left pane of the tool is where the Hindi Synsets get loaded. The Right side pane is to be filled by the user in order to create synset in the target language (the screenshot is showing Sanskrit as the target language).

The tool provides easy configuration options for setting the source and target languages. By default the source language is Hindi (as it is the pivot language), but this can be changed to reflect any language of user's choice.

The configuration also allows the user to provide the English Synset file (for reference) if it exists. Once the Concept, Example, Synset (words), Link and Etymology fields are filled up for the target language, the synset is considered to be Complete. This count helps the end user to gather data statistics and also to easily navigate through the incomplete synsets.

The latest version (v2.1) of the tool provides option for Secure Shell (SSH) Synchronization. This allows multiple users from same linguistic team to work on parallel for the same target language, and then enables them to merge their work on a server through which the communication is done using SSH.

The tool provides standard options of changing the font size, navigation options, options for Synset level comments etc.



Synset Ranking

HINDI RECORD (11383 synsets)

ID: 1

CAT: adjective

CONCEPT: जिसने जन्म न लिया हो

EXAMPLE: ब्रहम अजन्मा हैं"

SYNSET: अजन्मा,अजात,अनुत्पन्न,अनुद्भूत,अप्रादुर्भूत,अज,अजन,
अजन्म,अनन्यभव,अनागत,अयोनि

○ Complete (1)
○ Incomplete (11382)
● All (11383)

Search:

Set Rank: 1 - Universal

***Figure 4***: *Synset Categorizer tool*

### 3.3. Synset Categorizer Tool

As per the new distribution of synsets, decided in last IndoWordNet workshop, there are six categories (ranks) of synsets for Indian languages:

*Universal* - Synset concept available in all languages in the world (e.g. “Sun”).

*Pan Indian* - Natural or indigenous lexeme available for the Synset concept across all Indic languages, but might be missing in other languages. (e.g. “Papad” – a crispy food item).

*Family-Specific* - Lexeme available in languages which originated from the same ancestor. For e.g. - Indo-Aryan, Sino-Tibetan etc. (e.g. “Bhatijaa”- brother's son; a naturally occurring expression in Indo Aryan family of languages).

*Language-Specific* - Lexeme available only in that particular language.(e.g.–“Bihu” - a festival in Assam).

*Rare* - Concept is available in only a few languages which apparently do not have a common ancestry.

*Synthesized* - The concept is taken as it is from some foreign language concept.

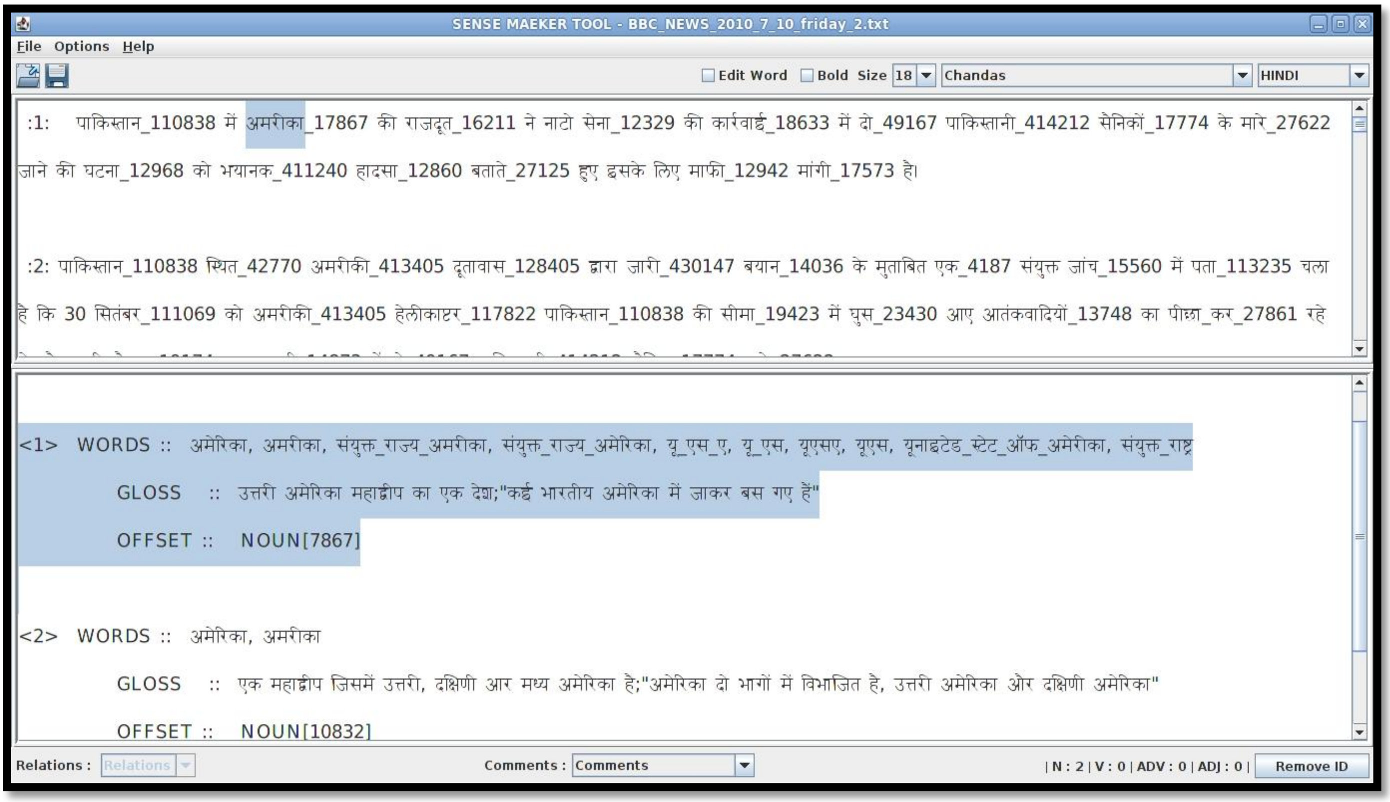
The synset Categorizer or Ranker tool helps linguists categorize the synsets into one of the six categories. The Synset Categorizer tool is a very significant tool for IndoWordNet because, synset making happens in a prioritized way in the following order: Universal → Pan-Indian → Family-specific → Language-specific → Rare → Synthesized.

Besides, it automatically imposes a preliminary ontological structure on the synsets and if synsets are completed in this way, natural lexemes of the language get covered early, and the disambiguity, that might have cropped up later is reduced at the initial stages.

The design for the tool is rather simple, compared to tools discussed in Sections 3.1, 3.2 and 3.4. The tool shows the language synset in the left side pane. The linguist’s job is limited to deciding the category for this Synset from the aforementioned categories.

Like offline tool, this is a file-based tool, and portable as it is written in Java. Since the tool follows the same syntax for the input file as that of the offline tool, the file generated by offline tool can be directly used for categorizing the synsets of that particular language.

As of now, for the task of identifying Universal and Pan-Indian synsets, the tool was modified to decide whether a particular Hindi Synset is available in particular language, and the tool was distributed to all linguistic groups under Indo WordNet family to make a decision.


SENSE MAEKER TOOL - BBC_NEWS_2010_7_10_friday_2.txt
File Options Help
□ Edit Word □ Bold Size 18 Chandas HINDI
Relations : Relations
Comments : Comments
| N : 2 | V : 0 | ADV : 0 | ADJ : 0 |
Remove ID

:1: पाकिस्तान_110838 में अमरीका_17867 की राजदूत_16211 ने नाटो सेना_12329 की कार्रवाई_18633 में दो_49167 पाकिस्तानी_414212 सैनिकों_17774 के मारे_27622 जाने की घटना_12968 को भयानक_411240 हादसा_12860 बताते_27125 हुए इसके लिए माफ़ी_12942 मांगी_17573 है।

:2: पाकिस्तान_110838 स्थित_42770 अमरीकी_413405 दूतावास_128405 द्वारा जारी_430147 बयान_14036 के मुताबित एक_4187 संयुक्त जांच_15560 में पता_113235 चला हे कि 30 सितंबर_111069 को अमरीकी_413405 हेलीकाप्टर_117822 पाकिस्तान_110838 की सीमा_19423 में घुस_23430 आए आतंकवादियों_13748 का पीछा_कर_27861 रहे

<1> WORDS :: अमेरिका, अमरीका, संयुक्त_राज्य_अमरीका, संयुक्त_राज्य_अमेरिका, यू_एस_ए, यू_एस, यूएसए, यूएस, यूनाइटेड_स्टेट_ऑफ_अमेरीका, संयुक्त_राष्ट्र

GLOSS :: उत्तरी अमेरिका महाद्वीप का एक देश;"कई भारतीय अमेरिका में जाकर बस गए हैं"

OFFSET :: NOUN[7867]

<2> WORDS :: अमेरिका, अमरीका

GLOSS :: एक महाद्वीप जिसमें उत्तरी, दक्षिणी आर मध्य अमेरिका हे;"अमेरिका दो भागों में विभाजित है, उत्तरी अमेरिका और दक्षिणी अमेरिका"

OFFSET :: NOUN[10832]

***Figure 5****: Sense Marking Tool*

## 4. Tools for Word Sense Disambiguation

Among all tools developed at IIT Bombay, the one that serves as a backbone for WSD is the Sense Marker Tool.

### 4.1. The Sense Marker Tool

Sense marking is the task of marking each word in the sentence with the correct sense of the word.

To train the machine to understand the written language, a huge amount of data needs to be sense-marked accurately by humans. A word may have a number of senses and to identify and mark which particular sense has been used in the given context is known as sense marking.

At IIT Bombay, this work is being done in 3 languages – English, Hindi and Marathi. The corpus used so far have been taken from Tourism, Health, Environment and Travel review domains and the Princeton WordNet is used as the sense inventory for English text while the Hindi and Marathi WordNets have been used for Hindi and Marathi texts respectively.

The sense-marker tool developed by IITB supports 9 languages (English, Hindi, Marathi, Tamil, Telugu, Kannada, Malayalam, Bengali and Punjabi).

The Sense Marker tool is a Graphical User Interface based tool developed using Java which facilitates the task of manual sense marking. This tool displays the senses of the word (as available in the Marathi, Hindi and Princeton (English) WordNets) and allows the user to select the correct sense of the word from the candidate senses.

The table shown alongside is the statistics of sense-marking done for Tourism and Health files (for two languages: Marathi & Hindi):

| Domain | Total Documents | Total Sentences | Tagged Words |
|---|---|---|---|
| Tourism Hindi | 152 | 15200 | 1,80,525 |
| Tourism Marathi | 152 | 15200 | 1,25,387 |
| Health Hindi | 89 | 8900 | 94,209 |
| Health Marathi | 72 | 7200 | 51,415 |
| Tourism English | 152 | 15200 | 181964 |
| Health English | 140 | 14000 | 149259 |
| Total | 757 | 75700 | 782759 |

***Table 1:*** *Statistics of sense marking*

## 5. Tools for WordNet APIs

The main tool whose output is fed into the Hindi WordNet API is the Morph analyzer tool.

### 5.1. Hindi Morphology Analyzer

(Kuhoo Gupta et al. 2006)

A morphology analyzer is a tool that takes a word (morphed) as input and outputs the set of roots along with its features.

The *root* of the word is its base form, on which the suffixes are applied.

For e.g. - लड़का (*ladakaa*) in लड़को (*ladakon*)

The *features* detected and output by the Hindi morph analyzer are as follows:

1. *Class* – It is the paradigm to which the word belongs.
2. *Category* – It is the POS of the word.
3. *Gender* – Masculine or feminine.
4. *Person* - 1st, 2nd or 3rd.
5. *Number* – Singular or Plural.
6. *Aspec*t – Perfective, Habitual, Progressive or Completive.
7. *Mode* – Honorary, Intimate etc.
8. *Tense* – Past, Present or Future

The current Hindi morph analyzer has a rich lexicon and the modular code that makes it very efficient. Currently, this is a command line tool, with no option for GUI.

#### 5.1.1. The Algorithm for Hindi Morph analyzer

**a. For Nouns:**

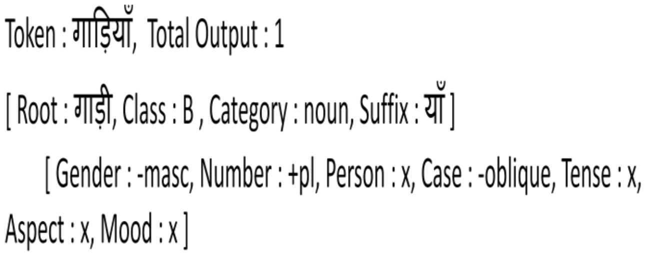
The Hindi morph analyzer operates using the following algorithm for nouns:

i. After the input word is detected as a noun, it goes into the *stemmer* or the *tokenizer*, where the root and the suffix are separated, using re-adjustment rules. For e.g. the stemmer output for the word गाड़ियाँ *(gadiyaan)* is गाड़ी + याँ *(gadi+ yaan)* as root and suffix.

ii. The lexicon contains entries for the Class and Category features, along with each root word. Once separated, the root is searched in the lexicon and these features are obtained.

iii. The suffix is then analyzed, and the other features are generated using rules for suffixes.

The following shows a sample output generated by the Hindi morph analyzer for the word 'गाड़ियाँ':



```
Token : गाड़ियाँ,  Total Output : 1

[ Root : गाड़ी, Class : B , Category : noun, Suffix : याँ ]

     [ Gender : -masc, Number : +pl, Person : x, Case : -oblique, Tense : x,
Aspect : x, Mood : x ]
```

***Figure 6:*** *Sample Output of the Hindi Morph Analyzer*

**b. For Verbs:**

In case of verbs the algorithm is as follows:

i. In case of verbs the input word goes through the stemmer and the root and suffix pair is detected.

ii. The root is searched in the lexicon.

iii. The suffix is then analyzed to get the corresponding features of verb morphology like aspect, tense, modality etc.

**c. For other POSs:**

i. In case of other POSs, the word is searched in the lexicon and only the category feature is output.

## 6. Conclusion

In this paper we have discussed some tools that facilitate better functioning of IndoWordNet both online and offline. The tools have been developed keeping the linguists and lexicographers in mind.

The paper also emphasizes the importance of tools in order to keep up the high quality of lexical resources like WordNets.

**Acknowledgement**

We gratefully acknowledge the support from Department of Technology, Ministry of Communication and Information Technology. We are highly indebted to all the students of IIT Bombay who have worked for the development of these tools.

## References

[1] C. Fellbaum, “WordNet: An Electronic Lexical Database.”, MIT Press, 1998.

[2] Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande and P. Bhattacharyya “An Experience in Building the Indo WordNet- a WordNet for Hindi”, International Conference on Global WordNet (GWC 02), Mysore, India, January, 2002.

[3] Kuhoo Gupta, Manish Shrivastava, Smriti Singh and Pushpak Bhattacharyya, Morphological Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi, COLING/ACL-2006, Sydney, Australia, July, 2006.