Chapter 3

WordNet, IndoWordNet And Sense Marking Tool

WordNet is the sense repository for various WSD approaches. IndoWordNet is a linked WordNet consisting of eighteen of the scheduled languages of India. This chapter presents a discussion on WordNet , IndoWordNet and a Sense tagging tool for tagging the corpora of languages constituting IndoWordNet.

3.1 Introduction

Now-a-days Natural Language Processing (NLP) R & D in India consist , inter alia, of Cross Lingual Search, English to Indian Language Machine Translation and Indian Language to Indian Language Machine Translation. A multilingual WordNet-Synset-based dictionary forms the cornerstone of these large scale activities, with Word Sense disambiguation (WSD) forming a critical component of the system A novel and effective method of storage and usage of dictionary in a multilingual framework was proposed by[1] .Table 3.1 shows the structure of the multilingual dictionary.

Given a row, the first column is the key for *n* number of languages describing a concept. Each concept is assigned a unique ID. In Table 3.1 the columns2-4 show the appropriate words expressing the concepts in respective languages. To express the concept *'4265: a youthful male person'*, there are two lexical elements in English, which constitute a *synset.*

Table 3.1 Multi Dictionary Model

Concept	L1(English)	L2(Hindi)	L3(Bengali)
Concept id	W_1 , W_2 , W_3 , W_4	W_1 , W_2 , W_3 , W_4 , W_5 , W_6	W_1 , W_2 , W_3 , W_4
4265: a youthful male person	(male_child , boy)	(लड़का, बालक, बाल, बच्चा, छोकड़ा, छोरा, छोकरा, लौं□ा, वत्स)	(ছেলে , বালক)

There are nine words in Hindi which form the Hindi synset, and two words in Bengali which constitute the Bengali synset. The members of a particular synset are arranged in the order of their frequency. The model thus defines an *M X N* matrix as the multilingual dictionary, where each row is for a concept and each column for a particular language.

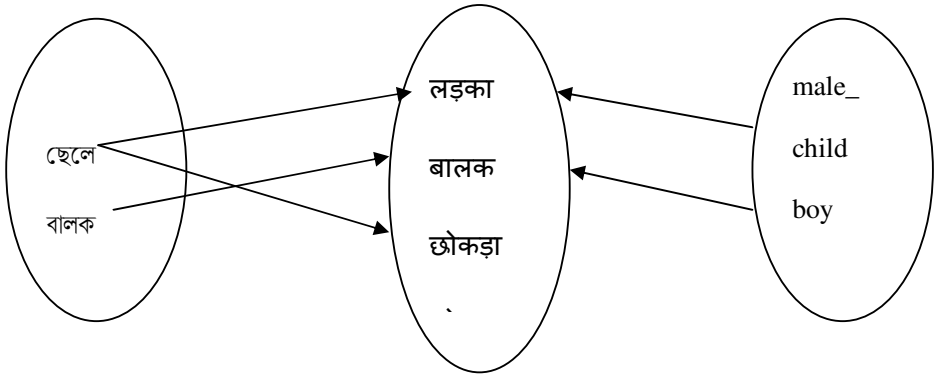

Figure3.1 Illustration of aligned synset members for the concept id 4265: a youthful male person

3.2WordNet

WordNet[2][9] has automated the Multi Lingual Dictionary Model. WordNet is a machine readable lexical database and it is a key resource that aids in several NLP tasks. The salient feature of WordNet is its emphasis on organizing lexical information in terms of word meanings/concepts in lieu of word forms [9]. A WordNet for a language is a linked structure of concept nodes represented by sets of synonymous words called synsets, which are connected through lexico-semantic relations. For open/content words, *i.e.,* Noun, Verb, Adjective and Adverb, synsets are created. In WordNet design, the stress is on concepts rather than on words. For

example, सागर (Sea), नदी (River) *etc.* are very common concepts . All the words representing particular concepts become members of the synset. [10].

English WordNet is the first WordNet developed at Princeton University for English. Then WordNets for European Languages followed suit. EuroWordNet[11]is a system of semantic networks for European languages, based on English WordNet[2][9].Similarly, Hindi WordNet[3] developed at the IIT Bombay was the first Indian language WordNet along the lines of the English WordNet following the Principle of expansion [3]. WordNets for other Indian languages were developed / are being developed following expansion approach along the lines of Hindi WordNet[3] and English WordNet[2][9] .

3.2.1 Lexical entries

The basic idea of a WordNet can be explained by a lexical matrix. It is an abstract representation of the organization of lexical information. In the lexical matrix columns represents word forms and rows represents word meanings. Rows express synonymy while columns express polysemy as shown in Table 3.2. An entry in a cell of the matrix denotes that the word form in that column can be used in an appropriate context to express the meaning in that row. Thus, entry $V_{1,1}$ implies that word form W_1 can be used to express word meaning S_1. If there are more than one entry in a column, then the word form is polysemous; if there are more than one entry in a row, then word forms are synonyms. The concept of Lexical Matrix is illustrated in Table 3.2 [10] where W1 and W2 are synonyms, W2 is polysemous

Table 3.2 Lexical Matrix

Words	Word Forms					
Meaning	W_1	W_2	W_3	W_4	....	W_n
S_1	$V_{1,1}$	$V_{1.2}$				
S_2		$V_{2,2}$				
S_3			$V_{3.3}$			
.....					.....	
S_m						$V_{m, n}$

Mappings between word forms and meanings are many: many. This impliesthat some word forms can have several different meanings, and some meanings can be expressed by several different words.

3.3 Constituent Elements of WordNet

WordNet stores words and meanings like a dictionary but it is unlike traditional dictionaries in many ways. For example, words in WordNet are arranged semantically instead of alphabetically. Only content words i.e. nouns, verbs, adjectives, and adverbs are arranged in synsets. Prepositions and conjunctions do not find their place in any synset. Also, WordNet provides position of a word in ontology as an additional feature. Ontology is a hierarchical organization of concepts. Distinct ontological hierarchy exists for each category viz. noun, verb, adjective and adverb. Each synset is positioned in some place in the ontology.

3.3.1 Synset and Concept

In WordNet words are grouped together according to the similarity in their meanings. For each word, there is a synonym set, or 'synset', representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings. Synsets are the basic building blocks of WordNets. For e.g.eachsynset in the Hindi WordNet is linked with other synsets, through the well-known lexical and semantic relations. Semantic relations are between synsets and lexical relations are between words. These relations serve to organise the lexical knowledge base.

For e.g. let us take the word “house”. It has 12 senses as Noun and 2 senses as Verb in the English WordNet[2][9].

{house}---------- ambiguous, senses may range from a family, a legislative house, a social unit etc.

{ house , home } ---------- ---still ambiguous , senses may be family or a social unit

{ family , house , home } -------Completely unambiguous , has the sense of family and it is a synset.

Therefore a synset has the following properties:-

1. Minimal unit – Only the synonymous words that uniquely identifies a particular sense or concept should come within the synset.

2. Coverage – All the words representing a particular sense or concept should come within the synset. The words are listed in order of decreasing frequency of their usage by native speakers.
3. Replaceability - The first few words of the synset , which are the words commonly spoken by native speakers , should be mutually replaceable.

Different kinds of words found in WordNetare :-

- **Polysemous:** Words which have multiple senses are known as Polysemous. In WordNet, each word occurs in as many synsets as it has senses. For example, the word अर्थ (Artha) has several senses so it occurs in several noun synsets
- **Monosemous:** Words which can have only one sense are known as monosemous words. For example, the wordस्वाधीनता(Swadhinata) has only one sense and hence it appears in only one synset.
- **Compound Words:** Words compose of two or more words but are treated as single words are known as compound words. Compound words also sometimes appear in WordNetsynsets.

3.3.2 Relations in WordNet

The basic relations in WordNet are Semantic relations and Lexical relations which are explained below.

3.3.2.1 Semantic Relations

The relations between two synsets are known as semantic relations. Semantic relations are reciprocated *i.e.,* if '*R*' is a semantic relation between meaning { *a, a′, . .*} and meaning {*b, b′, . .}*, then there also exists a relation '*R′* between { *b, b′, . . .* } and { *a, a′, . .* }. Table 3.3 shows different kinds of semantic relations present between two synsets.

Table 3.3: Semantic Relations in WordNet

Relation	Meaning
Synonymy	Similarity of meaning
Hypernymy/Hyponymy	Is-A (Kind-Of)
Entailment/Troponymy	Manner-Of (for verbs)
Meronymy/Holonymy	Has-A (Part-Whole)

Synonymy**:-**Synonymy means similarity of meaning. Words having similar meanings are represented using this relation. The relation is symmetric: if 'a' is similar to 'b', then 'b' is equally similar to 'a' [9]. Synsets for different synonymous words are shown in table 3.4

Table 3.4 Examples of Synonyms in Bengali WordNet

ফল	পরিণাম, অন্ত, ফলাফল, ফল, পরিণতি,
নদী	নদী, তটিনী, তরঙ্গিনী, কল্লোলিনী

Hypernymy/Hyponymy:-Hierarchical organization of synsets via super-class/sub-class relationship is referred to as hypernymy/hyponymy [9]. If synset A is a kind of synset B then synset A is the hyponym of B and synset B is the hypernym of A. Hyponymy / hypernymy are a semantic relation between word meanings.For example,{**ক্রিয়া ,কাজ**}(Work, Action) is a hypernym of {**ফল**} (result , outcome) whereas {**কর্মফল , পরীক্ষার_ফল** } (result_of_one's_deeds , result_of_exam) is a hyponym of{**ফল**}(result , outcome) . This relation is also called IS-A relation. A hypernym is the generic concept whereas hyponym is the specific concept. A hyponym possesses all the features of a more generic concept but has at least one feature that distinguishes it from its superordinate and from any other hyponyms of that superordinate. Examples depicting hypernymy and hyponymy relations are shown in Figure 3.2.

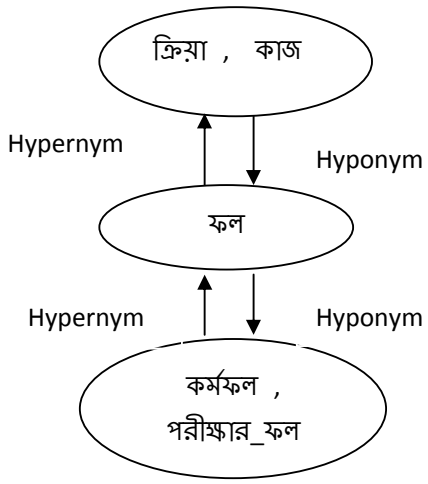

Figure 3.2 Showing hypernym and hyponym

Meronymy/Holonymy:- This is the part-whole (or HAS A) relation [9]. Synset A is a meronym of Synset B if A is a part of B. Conversely B is a holonym of A if B has A as a part. For example, চোখ(eyes), নাক (nose) and (head) are all parts of শরীর(body). The example depicting Meronymy and Holonymy relations are shown in Figure 3.3

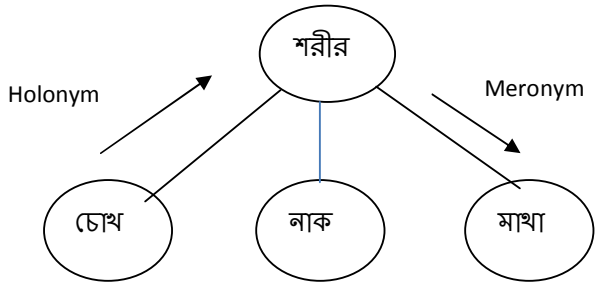

Figure 3.3: Meronymy/Holonymy relations

Troponymy:- Verbs are related through the relations hypernymy and troponymy. Synset A is the hypernym of B if B is one way to A ; B is then the troponym of A.Troponymy is a semantic relation between two verbs when one is a specific '*manner*' elaboration of another [10]. It shows manner of an action, i.e., X is a troponym of Y if to X is to Y in some manner. For e.g. verb synset {station, post, base, send, place} is a troponym of {move, displace} since to {station, post, base, send, place} is one way to { move, displace} [47].

Entailment: 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'*. Entailment is a semantic relation because it involves reference to the kinds of situations that A and B stand for. This use of entailment can be called lexical entailment. For eg. snoring entails sleeping. This relation is unidirectional [10].

Modifies Noun: Certain adjectives can only modify certain nouns. Such adjectives and nouns are linked in the Hindi WordNet by the relation Modifies Noun.

3.3.2.2 Lexical Relations

Lexical relations are the relations between members of two different synsets. The difference between lexical and semantic relations is that former are relations between members of two different synsets, whereas the latter are relations between two whole synsets.

Antonymy: Antonymy is a relation that holds between two words that (in a given context) express opposite meanings. It is a lexical relation as it holds between two words and not the entire synset.Example**:** मोटा,स्थूलकाय (moTAA, sthuulkaay; fat) ==>पतला, दबुला, दबुला-पतला, छरहरा (patlaa, dublaa, dublaa-patlaa, charharaa; thin)

Gradation: Thislexical relation provides possible intermediate state between two antonyms. For example, to show gradation relation among time words we have, 'दुपूर' (noon) between 'सकाल' (*morning)* and 'सन्ध्या' (*evening*) .

3.4 IndoWordnet

India is a multilingual country where machine translation and cross lingual search are highly relevant problems. These problems require large resources- WordNets and lexicons- of high quality and coverage.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 . The IndoWordnet Project was built by a consortium of Indian Universities headed by IIT, Bombay with a generous grant from Technology Development of Indian Language Programme, Department Of Information Technology, Ministry of Communications and Information Technology, India. IndoWordNet is a linked lexical knowledge base of WordNet of 18 of the scheduled languages of India, viz., Assamese, Bangla, Bodo,

Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu and Urdu. Figure 3.4 shows some languages which constitute a part of the Indo WordNet project.

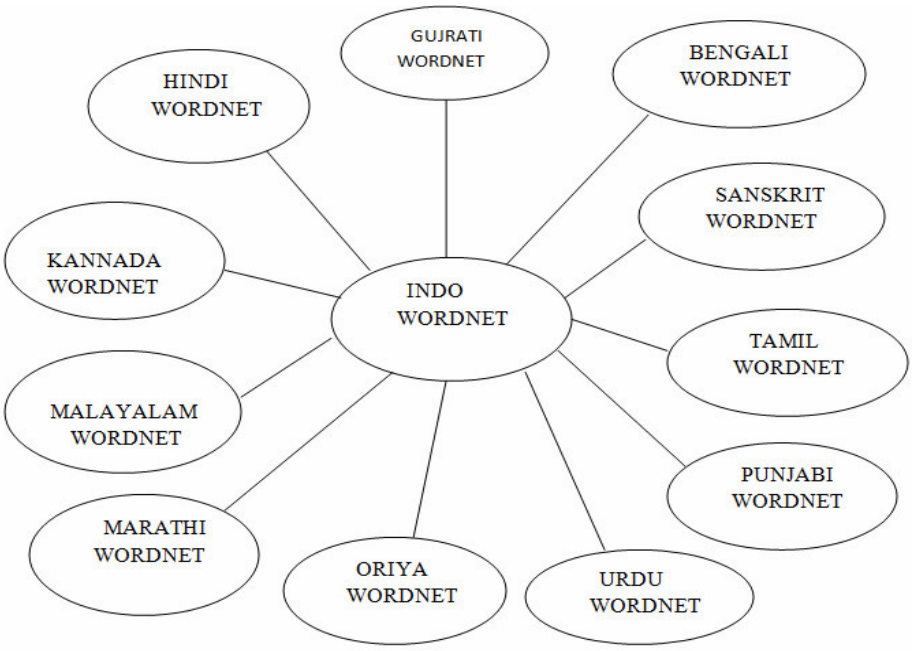

Figure 3.4: Linked Structure of IndoWordnet

Each entry in the IndoWordnet consists of the following elements :-

- Synset id of the Synset. For e.g. the synsetशाम (evening) would be written as शाम_8164
- Part Of Speech Tag (POS)
- Synonymous words constituting the synset
- Gloss which explicates the meaning of the synset
- Example sentence to show the usage.

A remarkable feature of the IndoWordnet is that a synset of any of the 18 scheduled languages of India forming a part of IndoWordnet is aligned with the corresponding synsets of the other languages and also to English following the principle of Multi Lingual cross linked Dictionary as discussed earlier. This design principle was adopted to give a fillip to Machine Translation and Cross Lingual Search in Indian

languages. All the relations – Semantic and Lexical- discussed earlier are available in the IndoWordnet. The Fig 3.5- 3.6 show snapshots of the Home page and search results of two concepts in Hindi and Bengali.

Figure 3.5 Home Page of IndoWordnet

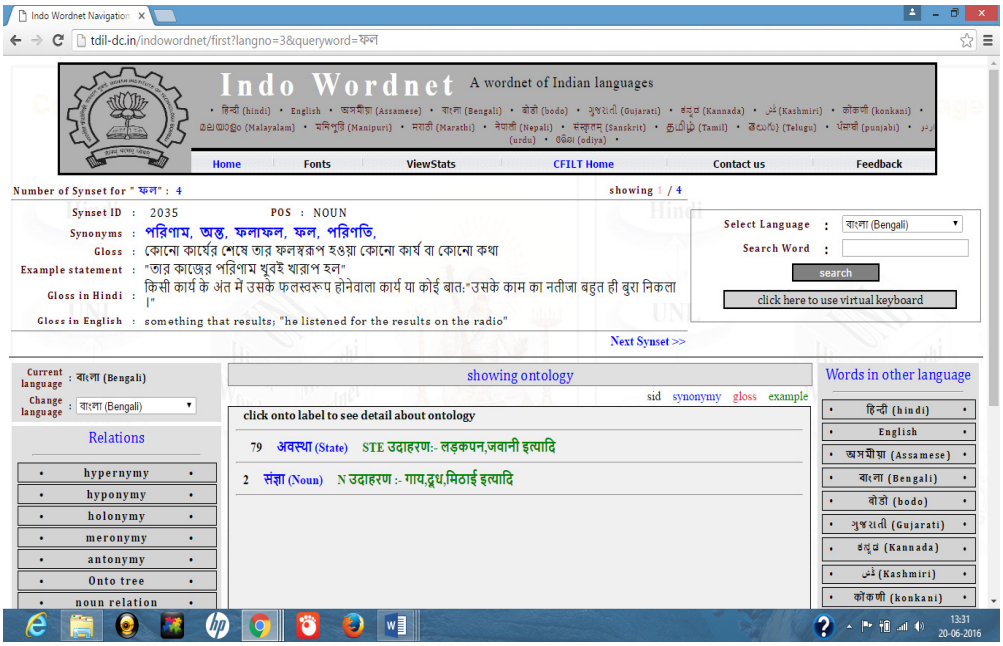

Figure 3.6(a) : A search page showing the result of search of the concept 'फल'

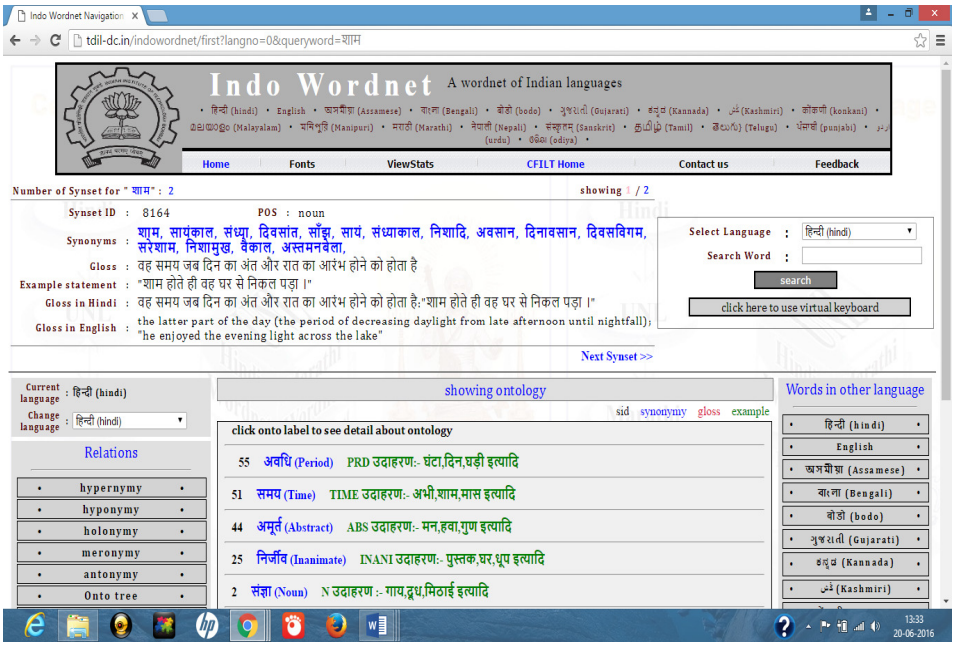

Figure 3.6(b): A search page showing the result of search of the concept '**शाम**'

3.5 Sense Marking Tool

Words in a language may have multiple meanings. The meaning of a word is context sensitive. For example in the sentence, *the house is at the foot of the mountain,* foot refers to sense bottom part of the mountains instead of other possibilities like bottom part of the leg. The identification of the meaning of a word in a particular context, also called as lexical disambiguation or word sense disambiguation, is hardly a problem for humans except in some rare cases. But unlike a human being a machine cannot identify which sense of a word is appropriate in a particular framework. Given a word and its possible senses, as defined in a knowledge base, sense tagging is the process of assigning the most appropriate senses to the words in the corpus within a given context.Huge amount of data needs to be sense tagged accurately by human annotators in order to train the machine to understand the spoken languages. Sense tagging is the task of identification and tagging a particular sense to the word in the given context out of many senses available for that word. Sense tagging is one of the toughest annotation tasks. The Sense Tagger Tool developed to

sense tag the corpus of all the languages incorporated in IndoWordnet is described below:-

The steps to sense tagged a document are:-

- First we open the file containing the corpus that is required to be tagged in the corpus view window.
- When we select the word to be tagged, the definitions of the senses related to that word appears in the right upper window. The annotator uses his/her judgement to select the appropriate sense from the list of possible senses
- When the most appropriate sense in the given context is selected all the information available in the WordNet for that word appear in the WordNet view window as shown in figure 3.7
- By clicking on the synset in the right lower window, synset id corresponding to the synset gets tagged to the word as shown in figure 3.8.

In this way all the words in the file may be tagged.

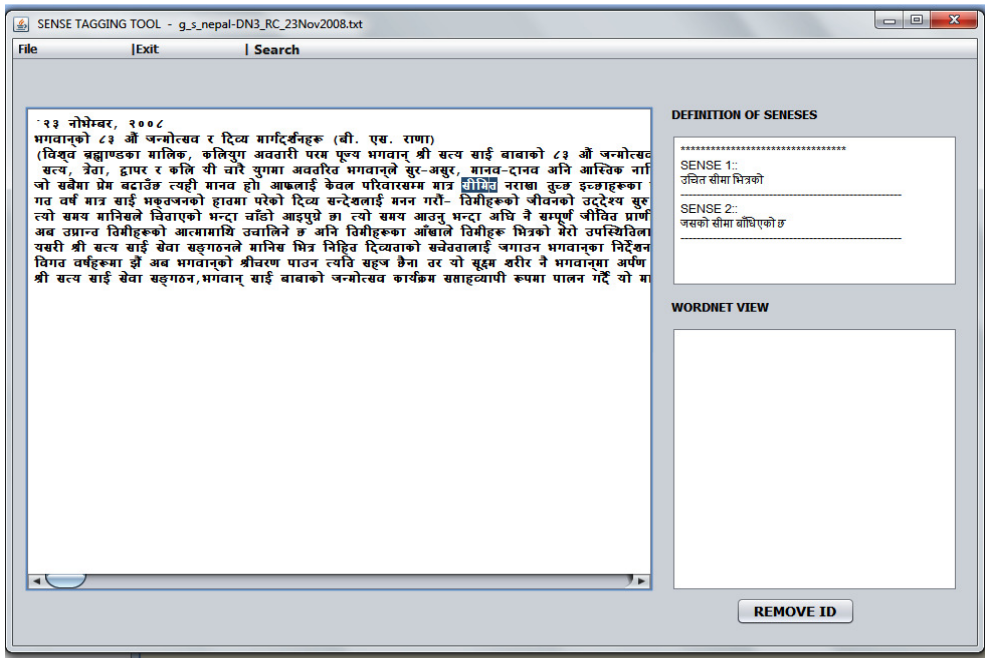

Figure 3.7 Layout of Sense Marking Tool

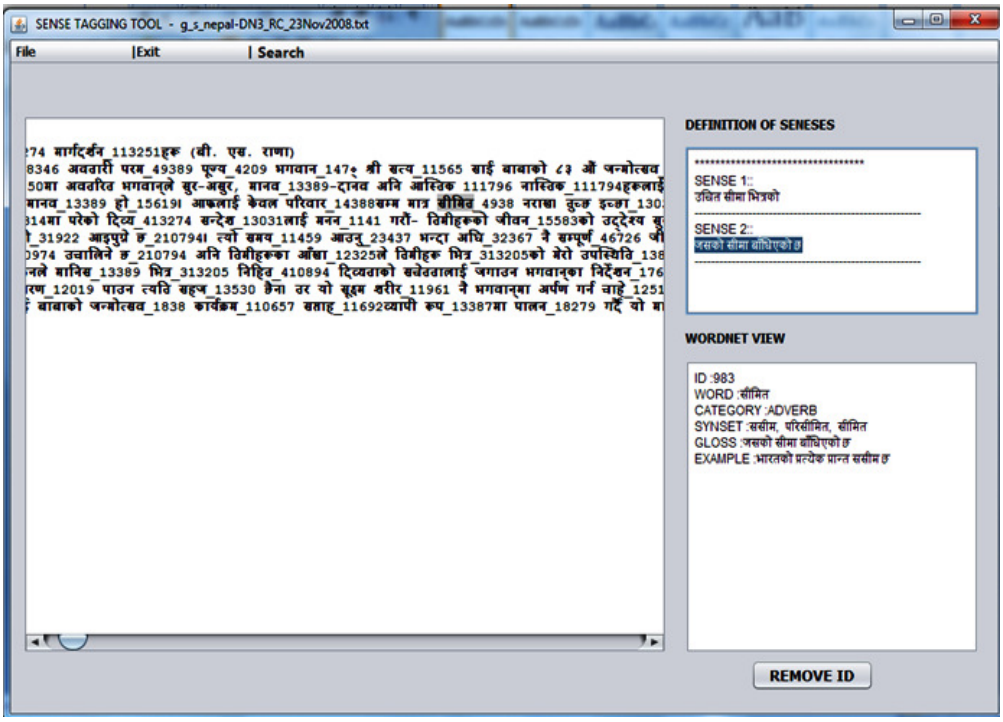

Figure 3.8 Sense Tagged Document

During the course of sense tagging there have been situations when the task of sense tagging got complicated because the sense was either entirely missing from the Wordnet or the existing sense was only a proximate one or the compound expressions could not be separately disambiguated to provide the correct sense. These issues are discussed below:

- The most ideal and desirable situation is when the correct set of senses for the word to be tagged is available in the WordNet. When words are present in the WordNet with exact senses, the sense marker essentially has to assign the sense accurately by applying her/his knowledge of language and the understanding of context.
- A particular word is absent in the sense repository but the exact sense is detected in some other existing synset of the WordNet, then the word is added to the synset and tagged with the appropriate synset id. . A number of words happened to be in this category. For example , the word “Ganga”. This word was tagged to the concept (an Asian river; rises in the Himalayas and flows east into the Bay of Bengal; a sacred river of the Hindus) where the existing

synset was Ganges, Ganges River. The ID of this synset was given to the word Ganga as, definitely, it should have been a member of this synset. [7].

- In the event of the exact sense of a word in the document being either absent or although available in the WordNet but not suitable in the context then a new sense for that word should be created. This is apparent in cases of MultiWords, or word entities pertaining to specific cultures or languages. For e.g. the word "লক্ষণ" whose Gloss in Bengali is "রাজা দশরথের পুত্র িতিনি সুমিত্রার গর্ভেজন্ম নিয়েছিলেন" (the son of King Dasaratha who was born of the womb of Sumitra) does not have a corresponding sense in English and thus a new sense id for the word "লক্ষণ" in the sense mentioned above was created and given the id 8117 . The Gloss for লক্ষণ_8117 given in English was mythical_being.

3.6 Chapter Summary

In this chapterwe have discussed, at length, about WordNet, IndoWordnet and a Sense marker tool for sense marking the corpora of the languages constituting Indo WordNet. We have also discussed the constituent elements of WordNet and made a detailed study about the relations available in a WordNet as these relations can provide key features for a good WSD algorithm.