Automatic Generation of Multilingual Lexicon by Using Wordnet

Nitin Verma and Pushpak Bhattacharyya

Department of Computer Science and Engineering, I.I.T. Bombay,
{nitinv,pb}@iitb.ac.in

Abstract. A lexicon is the heart of any language processing system. Accurate words with grammatical and semantic attributes are essential or highly desirable for any application- be it machine translation, information extraction, various forms of tagging or text mining. However, good quality lexicons are difficult to construct requiring enormous amount of time and manpower. In this paper, we present a method for automatically generating multilingual Universal Word (UW) dictionaries (for English, Hindi and Marathi) from an input document-making use of English, Hindi and Marathi WordNets. The dictionary entries are in the form of Universal Words (UWs) which are language words (primarily English) concatenated with disambiguation information. The entries are associated with syntactic and semantic properties- most of which too are generated automatically. In addition to the WordNet, the system uses a word sense disambiguator, an inferencer and the knowledge base (KB) of the Universal Networking Language which is a recently proposed interlingua. The lexicon so constructed is sufficiently accurate and reduces the manual labor substantially.

1 Introduction

Construction of good quality lexicons enriched with syntactic and semantic properties for the words is time consuming and manpower intensive. Also word sense disambiguation presents a challenge to any language processing application, which can be posed as the following question: *given a document **D** and a word **W** therein, which sense **S** of **W** should be picked up from the lexicon?*. It is, however, a redeeming observation that a particular **W** in a given **D** is mostly used in a single sense throughout the document. This motivates the following problem: *can the task of disambiguation be relegated to the background before the actual application starts? In particular, can one construct a **Document Specific Dictionary*** wherein single senses of the words are stored?

Such a problem is relevant, for example, in a machine translation context [1]. For the input document in the source language, if the *document specific dictionary* is available a-priory, the generation of the target language document reduces to essentially syntax planning and morphology processing for the pair of languages involved. The WSD problem has been solved before the MT process starts, by putting in place a lexicon with the document specific senses of the words.

Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 380–391.

In this paper we describe a methodology for automatic generation of *document specific UW dictionaries* (particularly for English, Hindi, and Marathi)- by making use of the *English, Hindi* and Marathi WordNets} [2, 5, 6, 8]. The methodology described in this paper for generating document specific English-UW dictionaries has an improved performance for adjectives and adverbs over [3].

Section 2 briefly describes the UNL system. The format of L-UW dictionary is described in section 3. Section 4 illustrates the method of *document-specific* English-UW dictionary generation. The method of generating Hindi-UW dictionary by using the *Hindi WordNet* is described in section 5. Section 6 gives the future directions for improving the performance of multilingual lexicon generation system.

2 Universal Networking Language

UNL [7] is an interlingua for machine translation [1] and is an attractive proposition for the multilingual context. In this scheme, a source language sentence is converted to the UNL form using a tool called the *EnConverter* [7]. Subsequently, the UNL representation is converted to the target language sentence by a tool called the *DeConverter* [7]. The sentential information in UNL is represented as a hyper-graph with concepts as nodes and relations as arcs. The UNL graph is a hyper-graph because the node itself can be a graph, in which case the node is called a *compound word* (CW). Figure 1 represents the sentence *John eats rice with a spoon.*

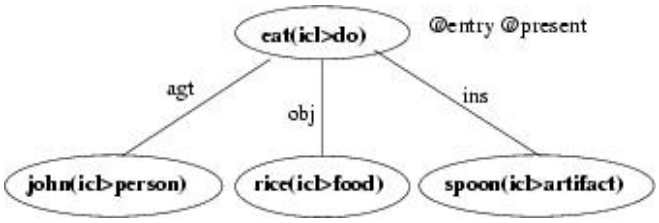

Fig. 1. UNL Graph of *John Eats Rice With A Spoon.*

In the above graph the arcs denoting *agt* (agent), *obj* (object) and *ins* (instrument) are the relation labels as defined in the UNL specification. This graph is represented as a set of directed binary relations between two concepts present in the sentence. The relation *agt* stands for *agent*, *obj* for *object* and *ins* for *instrument.* The binary relations are the basic building blocks of the UNL system, which are represented as strings of 3 characters or less each. There are 41 relations in the UNL system.

In the above figure the nodes such as *eat(icl>do), John(iof>person), rice(icl>food)* and *spoon(icl>artifact)* are the *Universal Words (UW)*. These are language words with *restrictions* in parentheses. *icl* stands for *inclusion* and *iof* stands for *instance* of. UWs can be annotated with attributes which provide further information about how the concept is being used in the specific sentence. Any of the three restriction labels, *viz., icl, iof* and *equ*, is attached to an UW for restricting its sense. For example, two senses of *state* will be represented in the UNL system in the following way:

- *state(icl>express)* to express something clearly and carefully.
- *state(icl>country)* a politically organized body of people under a single government.

A UW is created using the *specifications* of the *UNL Knowledge Base (KB)*. UNL KB organizes the UWs in a *hierarchy*. A *part* of the UW hierarchy for *nouns* in the UNL KB is shown in Figure 2 which is self-explanatory.

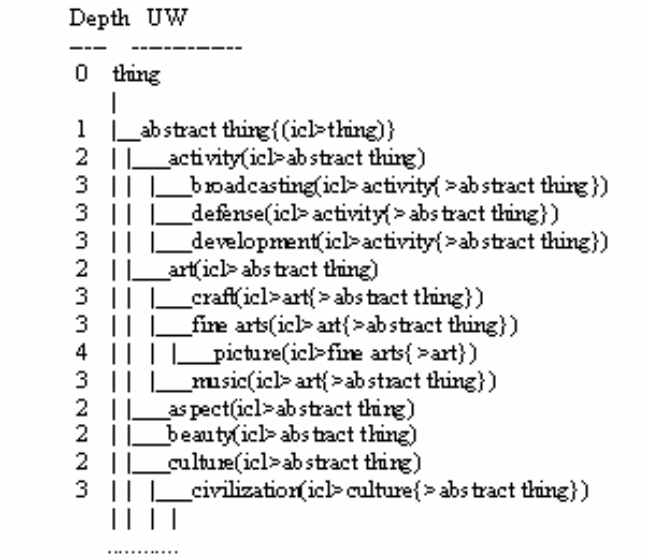

Fig. 2. Hierarchy of *noun* UW's in the UNL KB (a snapshot).

For verbs, the hierarchy is not so deep. All the verbs are organized under three categories, *viz*., *do*, *occur* and *be*. The first two are *aktionstat* verbs and the last one is the set of *stative verbs*. The adjective, adverb and preposition hierarchies too are quite shallow. The adjectives that are both *attributive* and *predicative* are given the restriction *(aoj>thing)*, where *aoj* is a semantic relation denoting *attribute* of the object and *thing* denotes a nominal concept. The adjectives which are only *predicative* are given the restriction *(mod>thing)* where *mod* is the *modifier* relation. The adverbs are uniformly expressed through *(icl>how)*.

3 L-UW Dictionary

The dictionary maps the *words* of a natural language to the *universal words* of the UNL system [9]. For example:

- [kuttaa]"dog(icl>mammal)" (**...** attributes **...**)
- [bhaukaa]"bark(icl>do)" (**...** attributes **...**)

- are the entries in a Hindi-UW dictionary [10]. Similarly:
- [dog]"dog(icl>mammal)" (**...** attributes **...**)
- [bark]"bark(icl>do)" (**...** attributes **...**)

are the entries is an English-UW dictionary. When the sentence *The dog barks* is given to an UNL-based English-Hindi MT system, the Uws *dog(icl>mammal)* and *bark(icl>do)* are picked up. These are disambiguated concepts different from other senses of *dog* and *bark*, for example the *pursue* sense of *dog* (*dog(icl>do)*) and the *skin of the tree* sense of *bark* (*bark(icl>skin)*). *If the L-UW dictionary contains only document specific UWs, the analyser and the generator systems do not commit error on account of WSD.*

The *attributes* attached to each entry in the L-UW dictionary are the *lexical, grammatical,* and *semantic* properties of the language specific words (*NOT of the UWs*). The syntactic attributes include the word category- *noun, verb, adjectives, adverb etc*. and attributes like *person* and *number* for nouns and *tense* for verbs. The *Semantic Attributes* are derived from an *ontology*. Figure 3 shows a part of the *ontology* used for obtaining semantic attributes [9].

4 Automatic Generation of English-UW Dictionary

For generating the document specific English-UW dictionary we use the *English WordNet, a WSD System*, the *UNL KB* and an *inferencer*. The approach is *Knowledge Based* [12]. The UNL KB as shown in figure 2 is stored as a *mysql* database. The table *UNL-KB-table* in figure 4 shows a part of this storage structure for nouns.

The *word sense disambiguator* (integrated with our lexicon generation system) uses a method called as *Soft Word Sense Disambiguation* [4]. In soft word sense disambiguation method, the *sense disambiguation system* does not commit to a *particular sense* but it gives us a *set of senses* which are not necessarily orthogonal or mutually exclusive. The senses are expressed by the WordNet synsets and are arranged according to their relevance in the given context. A detailed description of *soft word sense disambiguation* method is given in [4].

Soft word sense disambiguation system attaches a *confidence value* (relevance score or probability) with every relevant sense of a word present in the document. In the final English-UW dictionary the entries with the low confidence value of their sense are disabled by placing a semicolon at their beginning. Everything after a semicolon (in a particular line) is ignored by the EnConverter automatically by the lexicon generation system and the one with the highest score is kept enabled. This method still keeps the dictionary *document-specific* and gives a *flexibility* to the lexicographers to enable an appropriate sense in the dictionary generated.

The steps involved in the generation of *document specific English-UW dictionary* are as follows.

4.1 POS Tagging and Sense Disambiguation

The document is passed to the word sense disambiguator [4], which gives us a *part of speech* and *sense tagged* document. The output of this step is a list of entries in the

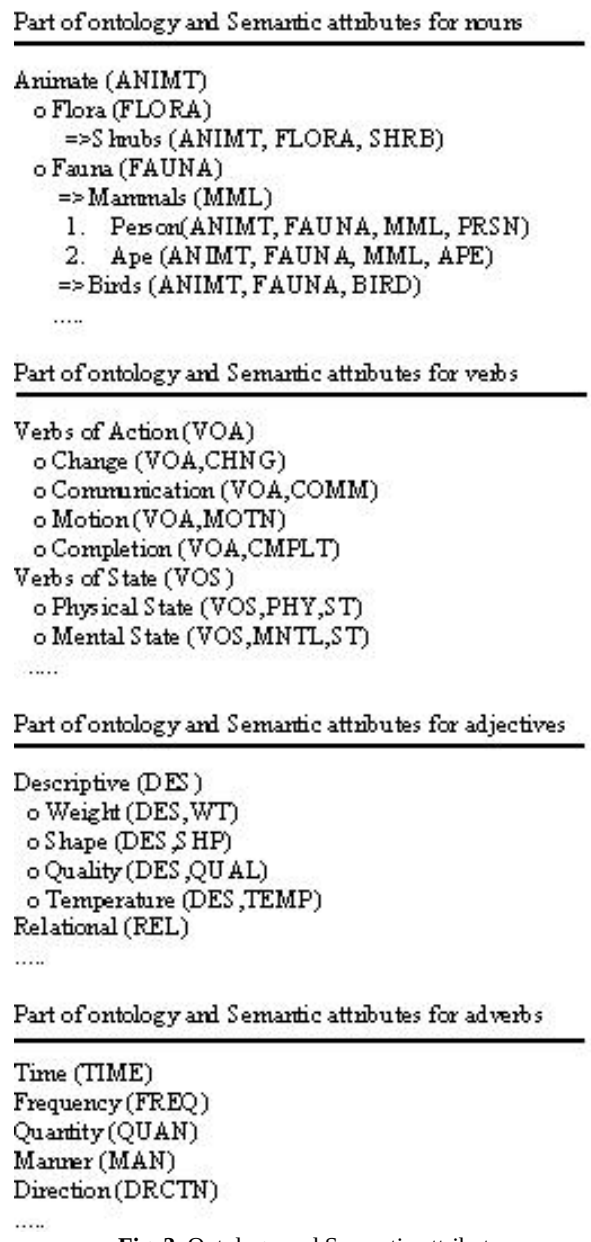

Fig. 3. Ontology and Semantic attributes

format **Word:POS:WSD**, where POS stands for *part of speech* and WSN indicates *WordNet sense number*. The *syntactic* attributes are obtained at this stage.

4.2 Generation of UW's

The WN and UNL KB are used to generate the restriction for the word. If the word is a noun, the WN is queried for the hypernymy for the marked sense. All the Hypernymy ancestors $\mathbf{H_1}$, $\mathbf{H_2}$, ..., $\mathbf{H_n}$ of **W** up-to the *unique beginner* are collected. If $W(icl{>}H_i)$ exists in the UNL KB, it is picked up and entered in the dictionary. If not, $W(icl{>}H_1)$ is asserted as the dictionary entry.

For example, for *crane* the *bird*-sense gives the hypernyms as *bird, fauna, animal, organism* and finally *living_thing*. *crane(icl>bird)* becomes the dictionary entry in this case. Figure 4 illustrates this process.

For verbs, the hypernymy ancestors are collected from the WN. If these include concepts like *be, hold, continue etc.*, then we generate the restriction *(icl>be)* (case of *be* verb). If not, the corresponding *nominal word* (for example, the nominal word for the verb *rain* is *rain* itself) of the verb is referred to in the WN. If the hypernyms of the nominal word include concepts like *phenomenon, natural_event etc.*, then we generate the restriction *(icl>occur)* signifying an *occur* verb. If both these conditions are not satisfied, then the restriction *(icl>do)* is generated.

For adjectives, use is made of the *is_a_value_of* semantic relation [8] in the WN. For example, for the adjective *heavy* the above relation links it to *weight*. If this relation is present then the restriction *(aoj>thing)* is generated. Else we generate *(mod>thing)* (please refer back to Section 2).

For adverbs, (icl>how) is by default generated, as per the specifications of the UNL system.

4.3 Creating the Semantic Attributes

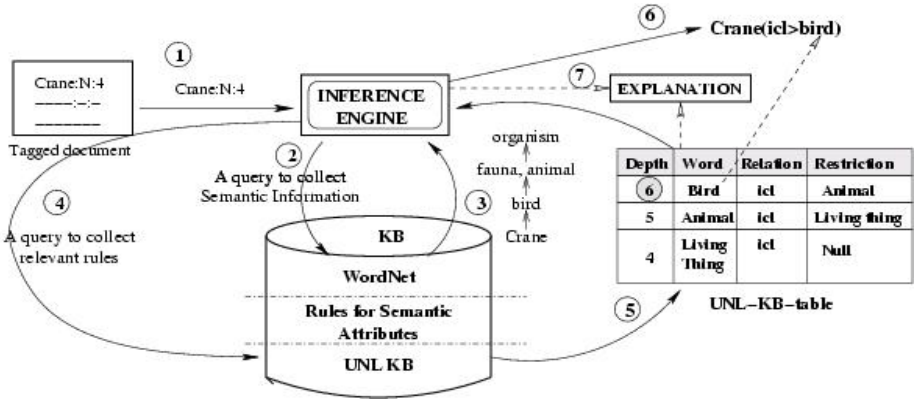

Fig. 4. Universal Word Creation: an example

The semantic attributes are generated from a *rule-base*, linking the lexico-semantic relations of the WN with the semantic properties of the word senses. To take an example, **if** the *hypernym* is *organism*, **then** the attribute *ANIMT* signifying *animate* is

generated. We have more than 5000 such rules in the rule base. The tables in the figure *rules* shows sample of such rules for all the POS words.

Table 1: Rules for noun

HYPERNYM	ATTRIBUTE
organism	ANIMT
flora	FLORA
fauna	FAUNA
beast	FAUNA
bird	BIRD

Table 2: Rules for verb

HYPERNYM	ATTRIBUTE
change	VOA,CHNG
communicate	VOA,COMM
move	VOA,MOTN
complete	VOA,CMPLT
finish	VOA,CMPLT

Table 3.1: Rules for adjectives

IS_VALUE_OF	ATTRIBUTE
weight	DES,WT
strength	DES,STRNGTH
qual	DES,QUAL

Table 3.2: Rules for adjectives

SYNONYMY OR ANTONYMY	ATTRIBUTE
bright	DES,APPR
deep	DES,DPTH
shallow	DES,DPTH

Table 4: Rules for adverbs

SYNONYMY	ATTRIBUTE
backward	DRCTN
always	FREQ
frequent	FREQ
beautifully	MAN

Fig. 5. Rules For Generating Semantic Attributes

For nouns, Table 1 (Rules for noun) in figure 5 is used to generate semantic attributes. The first entry corresponds to the rule: IF hypernym = organism THEN generate ANIMT attribute. Semantic attributes for verbs are obtained in the same way by using Table 2 (Rules for verbs).

For adjectives, Tables 3.1 and 3.2 are used. The first entry in the table 3.1 corresponds to the rule: **IF** *input_word* IS_A_VALUE_OF(*weight*) **THEN** the attributes *DES,WT* signifying *weight* (classified in *descriptive* category) are generated. The first entry in the Table 3.2 is interpreted as: **IF** *input_word* is (SYNONYM_OF(*bright*) OR ANTONYM_OF(bright)) **THEN** the attributes *DES,APPR* (*descriptive, appearance*) are generated.

4.4 Experiments and Results

We have tested our system on documents from various domains like agriculture, science, arts, sports *etc*. each containing about 1000 words. We have *measured* the *performance* of this system by calculating its *precision* in every POS category. The precision is defined as:

$$precision = \frac{number\ of\ entries\ correctly\ generated}{total\ entries\ generated}.$$

Figure 6 shows the results.

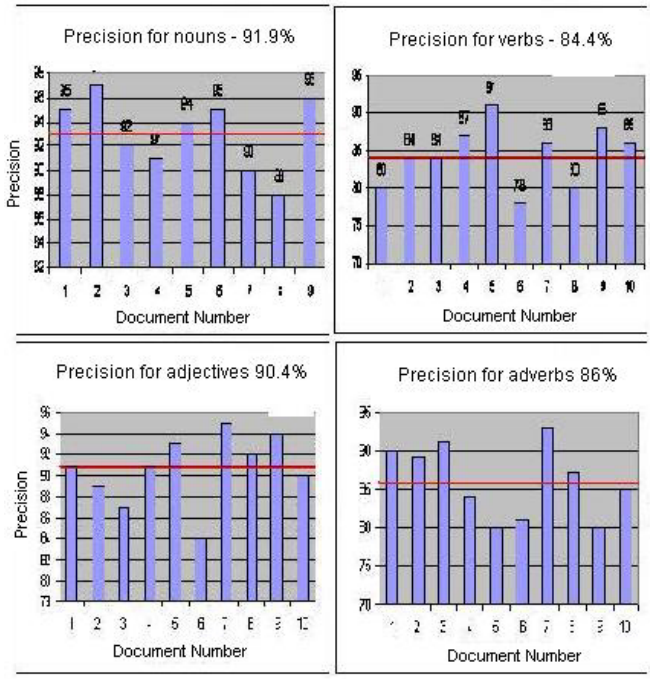

Fig. 6. Experiments And Results

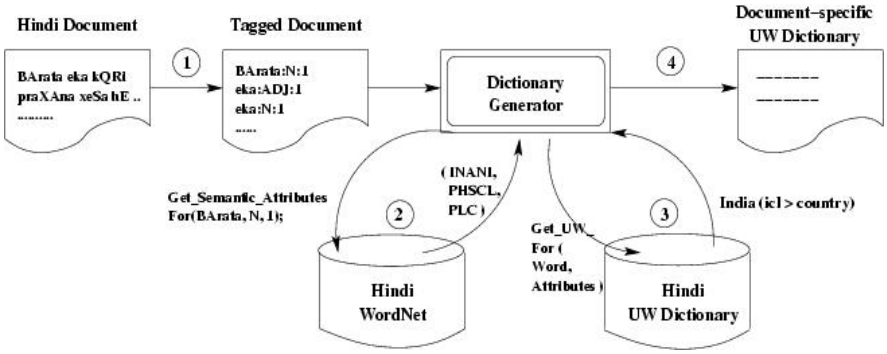

Fig. 7. Document-Specific Hindi-UW Dictionary Generation

The average precision for nouns is **93.9**%, for *verbs* **84.4**%, for *adjectives* **90.6**% and for *adverbs* **86**%. The system is being routinely used in our work on machine translation in a tri-language setting (*English*, *Hindi* and *Marathi*). It has reduced the burden of lexicography considerably. The incorrect entries- which are not many- are

corrected manually by the lexicographer. A snapshot of document specific English-UW dictionary generated (the entries with the low score value are disabled *automatically* by placing a semicolon at the beginning) after running our system on a document containing the following paragraph is shown below.

Modern agriculture depends heavily on engineering and technology and on the biological and physical sciences. Irrigation, drainage, conservation, and sanitary engineering- each of which is important in successful farming- are some of the fields requiring the specialized knowledge of agricultural engineers.

- [Modern] {} "modern(icl>character)" (N, INANI, PROP, ACT, COMM, ABS) <E,0,0>; SCORE=0.917893
- ; [Modern] {} "modern(icl>person)" (N, PROP, ANIMT, FAUNA, MML, PRSN, PHSCL) <E,0,0>; SCORE=0.901949
- [agriculture]{}"agriculture(icl>industry)"(N,INANI,EVENT,ABS)<E,0,0>; SCORE=0.931336
- ; [agriculture] {} "agriculture(icl>business)" (N, INANI, EVENT, ABS) <E,0,0>; SCORE=0.90433
- [depend] {} "depend(icl>be(aoj>thing{,obj>thing}))" (VRB, CONT, VOS-PHY-ST) <E,0,0>; SCORE=0.937532
- ; [depend] {} "depend(icl>trust{>be}(aoj>thing))" (VRB, VOA-COGN, VOA-POSS, VOS-MNT-ST) <E,0,0>; SCORE=0.923279
- [engineering] {} "engineering(icl>subject)" (N, INANI, PSYFTR, ABS) <E,0,0>; SCORE=0.924104
- ;[engineering] {} "engineering(icl>structure)" (N, INANI, OBJCT, ARTFCT, PHSCL) <E,0,0>; SCORE=0.90438
- [technology] {} "technology(icl>subject)" (N, INANI, PSYFTR, ABS) <E,0,0>; SCORE=0.924104
- ; [technology] {} "technology(icl>exercise)" (N, INANI, EVENT, ABS) <E,0,0>; SCORE=0.894572
- [biological] {} "biological(mod<thing)" (ADJ, REL) <E,0,0>; SCORE=0.924506
- [physical] {} "physical(mod<thing)" (ADJ, DES, QUAL) <E,0,0>; SCORE=0.924204
- [scienc] {} "science(icl>power)" (N, INANI, PSYFTR, ABS) <E,0,0>; SCORE=0.926118
- ; [scienc] {} "science(icl>subject)" (N, INANI, PSYFTR, ABS) <E,0,0>; SCORE=0.898305
- [Irrigation] {} "irrigation(icl>act)" (N, INANI, PROP, EVENT, ABS) <E,0,0>; SCORE=0.926247
- [conservation] {} "conservation(icl>protection)" (N, INANI, EVENT, ABS) <E,0,0>; SCORE=0.919366

5 Semi-Automatic Generation of Hindi-UW Dictionary

The prime resources we use for generating document specific Hindi-UW dictionaries are *Hindi-WordNet* [5] [6] and a Hindi UW dictionary which contains about 80,000 entries. The difficulty in automatic generation of document specific Hindi-UW dictionary is the absence of a *part-of-speech tagger* and a *Word Sense Disambiguator*. In our method we generate dictionary entries for all possible parts of speech and for all possible senses of the input word (present in the Hindi WordNet). Once the dictionary is generated, the irrelevant entries are disabled by the lexicographer by placing a semi-colon at the beginning of the entry. The document-specific dictionary generation in this case is not fully automatic (because of absence of POS tagger and WSD system for Hindi) like English-UW dictionary generation, but it has reduced the manual efforts required for Hindi lexicon generation substantially. The methodology of generating document specific Hindi-UW dictionary is described in the sub-section below, and the process is also shown in the Figure 7.

5.1 Methodology Used for Dictionary Generation

1. For every input word we use *Hindi-WordNet API* to obtain all possible *parts of speech* and all possible senses (present in the Hindi WordNet) for that word. In this step, an intermediate *tagged document* is generated in which the entries are in the format- *Word:POS:SenseNo* (shown in figure 7).
2. In Hindi WordNet every *synset* is linked to an *ontology node* (figure 3, which makes it easy for us to derive *semantic attributes* for a word (present in the Hindi WordNet) in its given POS and sense number. Hindi WordNet design has special support for linking *synsets* with the *ontology nodes* [5] [6]. For every *Word:POS:SenseNo* pair in the *tagged document*, Hindi WordNet is queried (by using Hindi WordNet API) to obtain the *semantic attributes*.
3. For generating an appropriate UW, we use a Hindi UW dictionary which contains about 80,000 entries. For the efficient retrieval of the UWs, we have stored the Hindi UW dictionary in a MySQL database table having the structure (*Hindi Word, UW, POS, attributes*). After obtaining the *semantic attributes* from the Hindi WordNet database, the Hindi UW dictionary database is queried to obtain an appropriate UW.
4. After collecting appropriate Semantic Attributes in step 2 and obtaining UWs from step 3, the *document-specific* Hindi-UW dictionary is generated. In this step the irrelevant entries (entries with irrelevant POS and Sense) are disabled and the incorrect ones are corrected manually by the lexicographer. This process reduces the burden of lexicography considerably. A snapshot of Document specific Hindi UW dictionary generated for a document containing the following paragraph on Indian Agriculture (Written in phonetic font format) is shown below. The incorrect entries are marked by a *.

"*bhaarat eka krishi pradhaan desh hai. yahaan kii aadhe se adhika janasankhyaa gaavon mai nivaas karatii hai. jinakaa mukhya vyavasaaya krishi hai. swatantrata ke baad bhaarat ne krishi ke kshetra mai bahut vikaas kiyaa hai*".

– [bhaarat] {} "India(icl>country)" (N, INANI, PHSCL, PLC) <H,0,0>;
– [eka] {} "a(icl>number)" (ADJ, DES, NUM) <H,0,0>;
– * [eka] {} "unity(scn>mathematics)" (N, INANI, ABS, MATHS) <H,0,0>;
– [krishi] {} "agriculture(icl>farming)" (N, INANI, ABS, ACT, PHSCLACT) <H,0,0>;
– [pradhaan] {} "cardinal" (ADJ, DES, QUAL) <H,0,0>;
– [hai] {} "have(icl>be)" (V, VE) <H,0,0>;
– * [se] {} "since" (ADV) <H,0,0>;
– [adhika] {} "full(equ>most)" (ADJ, DES, QUAN) <H,0,0>;
– * [janasankhyaa] {} "population(fld>biology)" (N, INANI, GRP) <H,0,0>;
– [nivaas] {} "lodging(icl>accommodation)" (N, INANI, PHSCL, PLC) <H,0,0>;
– [mukhya] {} "arch(icl>chief)" (ADJ, DES, QUAL) <H,0,0>;
– [vyavasaay] {} "firm(icl>shop)" (N, INANI, ABS, ACT, OCP) <H,0,0>;
– [krishi] {} "agriculture(icl>farming)" (N, INANI, ABS, ACT, PHSCLACT) <H,0,0>;
– [hai] {} "have(icl>be)" (V, VE) <H,0,0>;
– [baad] {} "later(icl>afterwards)" (ADV) <H,0,0>;
– [bhaarat] {} "India(icl>country)" (N, INANI, PHSCL, PLC) <H,0,0>;
– [krishi] {} "agriculture(icl>farming)" (N, INANI, ABS, ACT, PHSCLACT) <H,0,0>;
– [bahut] {} "abundant(icl>lot)" (ADJ, DES, QUAN) <H,0,0>;
– [vikaas] {} "advance(icl>development)" (N, INANI, ABS, ACT) <H,0,0>;

6 Conclusion and Future Work

In the machine translation process using UNL as an *interlingua*, the burden of lexicography has been reduced considerably by using the *multilingual lexicon generation system*. The system is being routinely used in our work on machine translation in a trilanguage setting (English, Hindi, and Marathi). The incorrect entries- which are not many are corrected manually.

Efforts are also on to implement the automatic lexicon generation system for *Marathi* language. The architecture of *Marathi WordNet* is same as that of *Hindi WordNet*. Like Hindi WordNet- every *Marathi synset* is linked to an *ontology node* (shown in Figure 3). The method of generating *semantic attributes* for Marathi words is same as that of Hindi (described in Section 5.1). At present we are making efforts to prepare a UNL KB dedicated to Marathi language which will enable us to automatically generate Universal Words for Marathi-UW dictionary.

The presence of *part of speech tagger* and *word sense disambiguator* for Hindi and Marathi will improve the performance of multi-lingual lexicon generation by many folds. Our future work will also be directed towards the implementation of *part of speech* tagger and *word sense disambiguator* for Hindi and Marathi languages.

References

1. W. John Hutchins and Harold L. Somers. “An Introduction to Machine Translation”. *Academic Press*, 1992.
2. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. “Five papers on WordNet”. Available at URL: http://clarity.princeton.edu:80/~wn/ , 1993.
3. N. Verma and P. Bhattacharyya. “Automatic lexicon generation through WordNet.” *Global WordNet Conference*, Jan 2004.
4. G. Ramakrishnan, Prithviraj, A. Deepa, P. Bhattacharyya and S. Chakrabarti. “Soft Word Sense Disambiguation”. *Global WordNet Conference*, Jan 2004. Available at URL: www.cse.iitb.ac.in/~nitinv/softWSD.ps.
5. S.Jha, D. Narayan, P. Pande and P. Bhattacharyya. “A WordNet for Hindi”. *Workshop on Lexical Resources in Natural Language Processing*, Hyderabad, India, January, 2001.
6. 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 (Global Wordnet 2002)*, Mysore, India, January, 2002.
7. “The Universal Networking Language (UNL) Specifications”, *United Nations University*. Available at URL: http://www.unl.ias.unu.edu/unlsys/ , July 2003.
8. Christiane Fellbaum. WordNet: “An Electronic Lexical Database”. *The MIT Press*, 1998.
9. P. Bhattacharyya. “Multilingual information processing using UNL”. in *Indo UK workshop* on Language Engineering for South Asian Languages LESAI, 2001.
10. Shachi Dave, Jignashu Parikh and Pushpak Bhattacharyya, “Interlingua Based English Hindi Machine Translation and Language Divergence”, *Journal of Machine Translation*, Volume 17, September, 2002. (to appear).
11. Hiroshi Uchida and Meiying Zhu. “The Universal Networking Language beyond Machine Translation”. *UNDL Foundation*, September 2001.
12. Adrian A. Hopgood. “Knowledge-Based Systems for Engineers and Scientists”. *CRC Press LLC*, 1992.