ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

Some Challenges of Automated Annotation in A Multilingual Scenario

Arindam Roy[1], Sunita Sarkar[2], B. S. Purkayastha[3]

Department of Computer Science, Assam University, Silchar, Assam, India

ABSTRACT: A key ingredient of today's NLP scenario is annotation and this paper discusses challenges involved in one of the toughest annotation tasks which is sense marking. A large amount of data needs to be sense marked accurately by human annotators in order to train the machine to understand the spoken languages. The sense marked corpus for various languages facilitate the task of Word Sense Disambiguation (WSD) which is required for translation. For accurately sense marking voluminous data, a standard and definitive lexicon is required. In the work reported here, the corpus is taken from the newspaper domain and tourism domain. The Princeton WordNet (Version 2.1) is used as the sense repertoire for English text while the Hindi and Nepali WordNets have been used for Hindi and Nepali texts respectively. The corpus was independently tagged by different annotators and it was found that the agreement level on word sense disambiguation was about 85% across the three languages, i.e., English, Hindi and Nepali. Different senses of a particular word in WordNet are quite specific, yet there have been cases when the senses provided had limitations and posed challenges to the human sense markers.

KEYWORDS: Sense-marking, Synset, WordNet, Word sense disambiguation, Expansion approach.

I. INTRODUCTION

The famous Princeton English WordNet[1] is an ontological, machine readable lexical database for English language developed at Princeton University. It delineated the design for the nouns, verbs, adjectives and adverbs of a language to be grouped under sets of synonyms, or synsets . Apart from functioning as a dictionary and thesaurus combined into one, it is used greatly in various NLP applications. English wordnet[4], in course of time ,became one of the most used and valuable language resources. Over a period of time, wordnets in other languages got developed along the lines of English wordnet. In case of Indian languages, Hindi WordNet [2] was the first of its kind as far as Indian languages are concerned. Hindi Wordnet was developed at the IIT Bombay. Consequent to the development of Hindi WordNet, number of tools were developed to utilize this valuable language resource which not only forms the heart of all WordNets in India, but also of all NLP work in India.

The Nepali WordNet [9] has been developed at Assam University, Silchar as part of a Consortium Project 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. It is also a machine readable lexical database for the Nepali language along the lines of the famous English WordNet and the Hindi WordNet.

The roadmap of the paper is as follows: Section II describes, in brief, the features of Nepali Language. Section III provides a description of Nepali WordNet and the IndoWordNet Project of which Nepali WordNet is a part. Section IV describes the methodology for sense-marking, the sense-marker tool, a description of how it works, and also the screenshot of this tool. Section V and all its subsections describe the choices for sense-marking that have been considered along with examples to illustrate the point. Section VI and its subsections describes some of the challenges faced during the process of sense marking in a multi lingual set up. Section VII concludes the paper.

DOI: 10.15680/IJIRSET.2014.0312066

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

II. RELATED WORK

Sense marked corpus have been central to WSD system. Sense marked corpus for English and other languages have been developed in order to facilitate the task of word sense disambiguation. The first sense-tagged corpora produced for any language is English SemCor corpus, a sense-tagged corpus of English [14]. The corpus consists of a subset of the Brown Corpus containing 700,000 words. All the words are tagged by PoS in SemCor and more than 200,000 content words are also lemmatized and sense-tagged. Japanese SemCor (JSEMCOR) [11] has been developed using annotation transfer method. In annotation transfer method, sense tagged corpus in one language is translated into the target language and sense annotations are also projected to the target language. The sense projection is carried out using a WordNet in the target language which is aligned with the WordNet that was used to sense tag the source language text. The target language WordNet is Japanese WordNet. The final corpus consists of 14,169 sentences with 150,555 content words of which 58,265 are sense tagged. Bulgarian sense tagged corpus [13] has been constructed from Brown corpus of Bulgarian. The corpus for annotation consists of 500 excerpts of approximately 100 words each thus total of 63 440 words. The source corpus is lemmatised and tagged with PoS. The words in BulSemCor are assigned senses from the Bulgarian WordNet. Sense annotation of Bulgarian Brown corpus is conducted using a annotation tool Chooser. The sense-annotated corpus consists of 45562 semantically annotated single and multiwords from the Bulgarian WordNet (BulNet). An attempt for constructing multilingual sense annotated corpus for Indian language has been made [12].The languages are English, Hindi and Marathi. The source corpus for the sense annotation is constructed from tourism domain. The sense inventories are used for assigning the appropriate sense to the word English wordnet, Hindi wordnet and Marathi wordnet. A sense marking tool is given to different annotators and the sense annotators independently tagged the corpus with senses. It was found that the inter annotator agreement on word sense disambiguation was about 80 % across the three languages, *i.e.*, English, Hindi and Marathi.

III. FEATURES OF NEPALI LANGUAGE

The hereditary structure of Nepali language is:- Indo European>Indo Iranian> Indo Aryan>North Western> Khasa Prakrit>Pahadi Language>Eastern Pahadi(Nepali). Nepali, like Hindi and its ancestor Sanskrit , unlike English, is a *Subject Object Verb* (SOV) language, i.e., in Nepali, the subject, object, and verb of a sentence usually appear in that order.For example:-

Sentence: उसले मेरो केरा खायो ।
Transliteration: *usle mero keraa khaayo.*
Gloss: he my banana ate.
Parts: Subject Object Verb
Translation: He ate my banana.

Nepali is written in Devnagari script. Nepali is a Head-right language i.e. in every phrase the head is on the right. The typical order of a VP is NP-VP. The typical order of a NP is ADJ-NP. The typical order of ADJP is ADV-ADJP. The typical order of PP is NP-PP i.e. the language is postpositional. It is written phonetically, that is, the sounds correspond almost exactly to the written letters. Nepali has many loanwords from Arabic and Persian languages, as well as some Hindi and English borrowings. There are some deviating features of Nepali from the other Indo- Aryan languages. These are:-

A. Unknown Past Aspect

Let us take the sentence '*kukhuro marechha*' (Nepali). A loose translation in English is '*chicken is dead*'. But actually it should be '*chicken was found to be dead*' (at the time the speaker came to know of this fact). '*chicken is dead*' has an equivalent Nepali translation '*kukhuro maryo*'.But as the *death* has occured in some unknown past, the Nepali speakers tend to say '*kukhuro marechha*'.

B. Gender

Human genders are treated as masculine or feminine. Apart from humans, all other nouns are treated as masculine. For e.g. '*Ram aayo*'(*Ram came*), '*Sita aai*'(*Sita came*). '*Goru aayo*'(*Ox came*), '*Gaii aayo*'*(cow came).*

DOI: 10.15680/IJIRSET.2014.0312066

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

C. Number

There are two numbers in Nepali-*eka bachan* and *bahu bachan*. *'Haru'* is the plural marker. But it is not imperative to use the plural marker for all nouns. For e.g. *keto aayo(boy came,*singular), *Ketaa aaye(boys came,*plural). Here *'o'* ending nouns change into *'aa'* ending for plural sense and verb *'aayo'* becomes *'aaye'* for plural sense. As a result, sentences can be made plural without using plural marker.

IV. NEPALI WORDNET

Nepali Wordnet is a system for bringing together different lexical and semantic relations between the Nepali words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Nepali WordNet is based on the principle of "expansion" from the Hindi Wordnet and English Wordnet. This principle was first proposed within the EuroWordNet project [5]. Thereafter it has been used by a number of WordNet development teams for the creation of new WordNets. Examples include the WordNets for Spanish, French , Hungarian language etc. In the Expansion Approach, synsets of a preexisting WordNet are understood by the lexicographer and the corresponding target language synsets expressing the same sense are created

A. Features of Nepali Wordnet

In Nepali Wordnet, the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set, or synset, in the Nepali WordNet, 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 WordNet. The Nepali WordNet deals with the content words, or open class category of words. Thus, the Nepali WordNet contains the following category of words- Noun, Verb, Adjective and Adverb.

Each entry in the Nepali Synset consist of the following elements [10]:-
ID: The synset identifier.
CAT: The syntactic category of the sense.
CONCEPT: It explains the concept represented by the synset. For example, "यस्तो कुरा वा काम जसले कसैको मान वा प्रतिष्ठा कम गराउँछ" (*yastokuraa waa kaam jasle kasaiko maan waa pratishTha kam garaaũcha*) explains the concept of insult as some saying or deed which diminishes somebody's reputation.
EXAMPLE: It gives the usage of the words of the synsets in the sentence. In general, the words in a synset are replaceable in the sentence. For example: "हामीले कसैलाई पिन अपमान गनुहुँदैन (*haameele kasailaaee pani apmaan garnuhũdain*) gives the usage for the words in the synset of 'अपमान', '*apmaan*' representing insult as something that should not be done to anybody.

B. *IndoWordnet Project*

The Nepali Wordnet is part of the Indo Wordnet[3] Project which is a linked wordnet of major Indian languages, viz, Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu and Urdu. These Wordnets have been created using the expansion approach from Hindi Wordnet and English.
Each entry in the IndoWordnet consist of the following elements:-
i. Synonymy
ii. Gloss
iii. Example Sentence

As an example let us see an entry in the IndoWordnet hosted at the TDIL website. The entry is the word आम and how the parallel synsets for this word is organized in Hindi and Nepali language.

Hindi synset for one of the senses of आम
Synset id : 3462 POS: Noun
Synonyms : आम, रसाल, आम्र, अंब, अम्ब, च्यूत, प्रियांबु, प्रियाम्बु, केशवायुध, कामायुध, कामशर
Gloss : एक फल जो खाया या चूसा जाता है

DOI: 10.15680/IJIRSET.2014.0312066

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

Example : "तोता पेड़ पर बैठकर आम खा रहा है / शास्त्रों ने आम को इंद्रासनी फल की संज्ञा दी है"
Gloss in English : large oval tropical fruit having smooth skin, juicy aromatic pulp, and a large hairy seed.
Nepali synset for one of the senses of आम
Synset id : 3462 POS: Noun
Synoyms : आँप , आम
Gloss : एउटा फल जुनचाहिँ खाने र चुस्ने गरिन्छ
Example : सुगा रूखमा बसेर आँप खाँदै छ/ शास्त्रले आँपलाई इन्द्रासनको फलको सज्ञा दिएको छ
Hindi synset for another sense of आम
Synset id : 3468 POS: adjective
Synonyms : सामान्य, आम, साधारण, कामचलाऊ, मामूली, औसत, अविशिष्ट, अविशेष, अदिव्य
Gloss : जिसमें कोई विशेषता न हो या अच्छे से कुछ हल्के दरजे क़ा
Example : "यह सामान्य साड़ी है / यह कामचालाऊ सरकार
अधिक दिन तक नहीं टिकने वाली है / खिलाड़ियों के औसत प्रदरशन से दरशक अप्रसन्न थे"
Gloss in English : Not exceptional in any way especially in quality or ability or size or degree; "ordinary everyday objects"; "ordinary decency"; "an ordinary day"; "an ordinary wine"
Nepali synset for another sense of आम
Synset id : 3468 POS: adjective
Synonyms : सामान्य, आम, साधारण, कामचलाऊ, अविशिष्ट
Gloss : जसमा कुनै विशेषता हुँदैनया राम्रो केही हल्का तहको
Example : "यो सामान्य साडी हो/ यो कामचलाउ सरकार धेरै दिनसम्म टिक्नेवाला छैन"
The above examples clearly expound the fact that a synset corresponds to one and only one sense of a word. The Indo Wordnet Project has linked the synsets of one Indian language to another. The synsets have a synset id which is uniform across languages and this synset id gets tagged to a word when it is sense marked as will be shown in the succeeding sections. As the words in the corpus gets tagged with an unique id so sense marking is a great facilitator in the process of translation. A very important task for which wordnet is used is the resolution of word sense ambiguity. Any Machine Translation system from English to an Indian language would require word sense disambiguation[8]. In a given text, the occurrence of a particular word will signify only one sense and the words in the neighbourhood of the target word to be tagged provides the clue for the appropriate sense of the target word which a sense marker is required to identify.

V. PROCEDURE FOR SENSE MARKING

The sense marker tool has been developed to automate the process of sense marking. It is a software tool developed to provide the lexicographers with an easy and efficient way of sense tagging the words. It has been developed keeping in mind the larger goal of word sense disambiguation (WSD) required for speedy translation. A sense annotated corpus using wordnet would, in a large measure, facilitate the task of WSD
The tool supports three languages viz English, Nepali and Hindi The Steps to sense mark a document are:
i.The Language i.e. English, Hindi or Nepali has to be chosen from the drop down menu box for which Sense Tagging is to be done.
ii. The file containing the corpus that is required to be tagged has to be opened by first clicking on the Open MenuItem of the File Menu.
iii. A File Chooser menu gets opened. The user has to select the file for tagging and press the open button of the same or double click on the file.
iv. The file/document gets opened in the Tagging Window.
v. For sense tagging the particular word needs to be single clicked and for a compound word the user has to drag to select.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

vi. When the word to be tagged is highlighted, the synset corresponding to that word appears in the synset window.
vii. A word may be monosemous or polysemous . All the senses related to the word are displayed in the synset window and the lexicographer has to use his/her judgement to select the most appropriate sense in the given context of the word by clicking on the respective synset.
viii. The synset id corresponding to the synset gets tagged to the word. In this way all the words in the file may be tagged and the file is saved by clicking on the Save Menu Item of the File Menu.

Screen shot of the Sense Marker Interface using a corpus from the tourism Domain is shown in Figure 1 below.

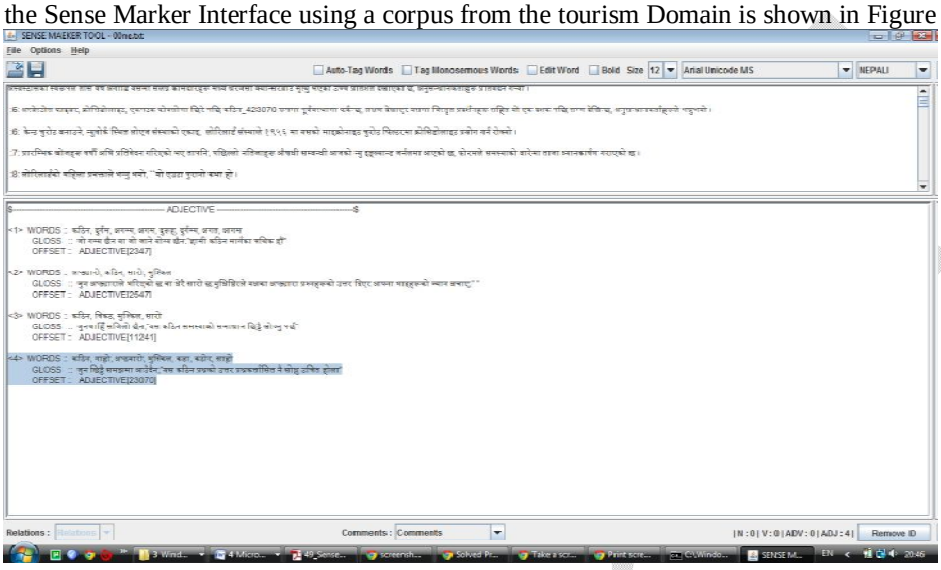

Fig1:Sense Marker Interface for a Corpus from Tourism Domain

As can be seen from the screen shots of the Sense Marker interface, it is composed of two broad windows. The top window is the place holder for the file containing the corpus. This file containing the corpus is actually the corpus which the annotator has decided to annotate. The bottom window is the place holder for the synsets related to a particular word which the annotator chooses to sense mark or annotate. As the word is chosen all the senses related to the word from the particular wordnet would appear in the bottom window. The annotator has to choose which of these senses is the most appropriate for the word in the given context i.e. the annotator or lexicographer has to choose the appropriate synset. The synset id corresponding to the particular synset in the wordnet would be tagged to the word and the file is saved. This is how an entire corpus can be sense tagged automatically using the wordnet. Of course assigning a particular sense to a word from a collection of senses is a subjective decision of the annotator and depends on his/her knowledge of the language and his/her understanding of the context in which the word appears.

But there have been situations when this task of sense marking 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. We now discuss these issues.

VI. CHOICES FOR SENSE MARKING

For sense marking, the annotators or the sense markers had the following choices:

A. Marking the word with the exact sense

This is the ideal and most desirable situation. It is the task of the sense-marker to assign senses to as many words as possible. When the word is available in the sense repository with its complete and correct set of senses, the sense marker essentially has to apply her/his knowledge of language and the understanding of context to assign the sense accurately.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

B. Marking with the exact sense even if the sense does not mention the particular word as a synset member
If the exact sense was found in some other existing synset of the wordnet, then the word was made a part of that synset. It was decided that such words should be enclosed with a hash mark followed by the ID of that synset. Quite a few words fell in this category. For example:

i. 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. The tagging looked like this - #Ganga#_9153625 [6].
ii. Pulao:- The concept was same as that of pilaf, pilaff, pilau, pilaw (rice cooked in well-seasoned broth with onions or celery and usually poultry or game or shellfish and sometimes tomatoes), and so it was tagged with this ID.
C. Creating a new sense
This option was adopted on occasions where a word appears in the document, the sense of which is either present in the wordnet but is not appropriate in the context or is completely absent from the wordnet. This is clearly obvious in cases of culture-specific word entities. It was decided that a new sense should be created for them and stored in the local copy of the wordnet. All such words were enclosed between an opening # and a closing # symbol. A script to parse the words between these symbols would be written and new synsets for these words would be created. The synset IDs will start from 200000.

VII. CATEGORIES FOR NEW SENSE

There are a few categories which have been identified for creating a new sense for a word.
A. A sense is present in Hindi but not in Nepali:
To establish a sense in such a case the following steps are to be followed:-
i. Transliteration
ii. Use of multiword expression (short phrases)
iii. Coining of new words
The steps should be used in the given order of priority.
B. A sense is present in Nepali but not in Hindi:
Such a sense may be termed as a Nepali specific sense. The relations for such a sense in Nepali have to be established manually. Some example are:

{ पेवा [pewaa, *a portion of the property of family owned by a female member*]}

{ईस्कुस [iskus, *a kind of vegetable*]}

{ पुनिउँ [punion, *a flat spoon for laying out rice*]}

{कुराउनी [kuraoni, *milk boiled down till nearly solid*)}
C. A sense is present in both Nepali and Hindi but not in English
Such a sense may be termed as a Nepali/Hindi specific
sense. Some examples are:

{जेठाजु [jethaju, *husband's elder brother*]}

{जेठान [jethaan, *elder brother's wife*]}

{सम्धी [samdhi, son's father-in-law, daughter's father-inlaw]}

{सम्धिनी [samdhini, son's mother-in-law, daughter's mother-in-law]}

In such cases the decision taken was to communicate to the English Wordnet developers to earmark a range of ids for such senses which are present in Nepali and Hindi but not in English and to transliterate the words representing such senses.

D. Multi words in the corpus
There are two kinds of multiword expressions (MWE): one which can have compositional interpretation and the other conveying the non-compositional. Machine cannot infer non compositional multiword expressions, so they have to be

DOI: 10.15680/IJIRSET.2014.0312066

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

stored in the sense repositories[7]. For example, in a sports corpus ,the English expression red card conveys the sense of a card that is used to send out an earring player from the soccer field. This meaning obviously does not come from the individual components red and card and hence has to be specifically stored in a sense repository including the wordnet. An example of this in Nepali is हाती चले बजार तो कुता भुके हजार , the gloss of which is Elephant going market then dog bark thousand which is when an elephant goes to market, thousand dogs bark but that does not convey the actual meaning that normally people tend to envy the successful.

Another source of multi-word expressions is when a sense in English or Hindi can be expressed by a single word but the corresponding sense in Nepali cannot be expressed by a single word and vice-versa. Some examples of this from newspaper domain and tourism domain are:

{Agrarian, कृषि या भूमि सम्बन्धि (krishi ya bhumi sambandhi)
Gloss-agriculture or land related}

{ambush, आक्रमणका लागि झाड़ीमा फौजरूको लुक्ने ठाउं, gloss-attack for bush troop concealment}

{cluster, गुच्छामा जम्मा गरियको, gloss- group formed into}

{paragon, उत्तमताको नमूना gloss-excellence of example}

{coastline, समुद्रि किनार, gloss-sea of near}

{afforestation (English), जंगल जमाउने काम (jangal jamaune kaam) , gloss-forest in sum do work}

While sense marking ,the sense markers have also come across senses in Nepali or Hindi which can be represented by single words but these senses can be represented in English only through a multi-word expression (Adj+Noun).Some examples are:

{ खेतिकमाई (khetikamai), agricultural income }

{ खोरिया(khoriya), uncultivated land }

{लघुकथा (laghukatha), short story }

These multi words are frequently found as translation candidates but the problem is they are also frequently not present in the wordnet. In the context of WSD, this is a cause for major concern.

E. Fine grained sense in Nepali but not in Hindi and English

Another very serious problem faced by the sense markers was that senses which have very specific meanings in Nepali do not have corresponding senses in Hindi and English Wordnets. Some examples are:

{ खुँडा (khunda), a kind of sword, a kind of scmitar }

{ पोते (pote), glass bead strands plain or with design worn
by married women in Nepal }

{ हकु छइला (haku chhoila), smoked buffalo meat }

{ बडा (bara) , deep fried black lentil patties }

{ सगुण (sagun), a traditional plate consisting of boiled
egg, smoked fish, bara, haku chhoila and ends with
yoghurt }

{ खुत्रक्क(khutrukka), sound made by little particles when dropped }

For all such cases the decision taken was to convey to the Hindi and English Wordnet developers an exhaustive list of such cases so that they can take appropriate measures for e.g. coining short multi-word phrases for such senses.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

VIII. CONCLUSION

In this paper we have discussed problems faced in annotating newspaper and tourism domain corpora in three languages using the sense repositories of English, Hindi and Nepali wordnets. In general we faced challenges in assigning senses to

i.When a sense is borrowed from Hindi Wordnet to Nepali Wordnet and vice versa

ii. When a sense is present in both Hindi and Nepali but not in English

iii.Multi-word expressions

iv.Words having specific senses in Nepali but no such sense exist in both English and Hindi Wordnets.

REFERENCES

[1] Felbaum .C,"Wordnet : An Electronic Lexical Database" MIT Press,1998

[2] Narayan D., Chakrabarty D., Pande P. And Bhattacharyya P. "An Experience In Building The Indo Wordnet-A Wordnet For Hindi" International Conference On Global Wordnet (GWC 02), Mysore, India, January, 2002.

[3] Sinha M., Reddy M., Bhattacharyya P. "An Approach towards Constructionand Application of Multilingual Indo- WordNet", 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea, January, 2006

[4] George A. Miller, "English WordNet -A Lexical Database for English", Communications of the association for Computing Machinery, 38(11):39-41. 1995

[5] Bhattacharyya P., Fellbaum C. and Vossen P. (eds), "Principles, Construction and Application of Multlingual Wordnets", Proceedings of the 5th Global WordNet Conference, Mumbai, Narosa Publishing House, India, Mumbai,2010

[6] Saraswati J., Shukla R., Pathade S., Solanki T. and Bhattacharyya P., "Challenges in Multilingual Domain-Specific Sense-Marking", Proceedings of the 5th Global WordNet Conference, Mumbai, Narosa Publishing House, India,Mumbai, Jan, 2010

[7] Khapra M. M., Sohoney S., Pande P. and Bhattacharyya P., "Value for Money:Balancing annotation effort, Lexicon Building and Accuracy for Multingual WSD",Computational Linguistics Conference(COLING 2010), Beijing,China,August 2010

[8] Chatterjee A., Joshi S.R., Khapra M.M. and Bhattacharyya P., " Introduction to Tools for Indo Wordnet and WSD" , Proceedings of the 5th Global WordNet Conference, Mumbai,Narosa Publishing House , India,Mumbai, Jan, 2010.

[9] Chakraborty A., Purkayastha B. S., Roy A., "Experiences in building the Nepali Wordnet" Proceedings of the 5th Global WordNet Conference, Mumbai, Narosa Publishing House, India, Mumbai 2010

[10] Roy A., Sarkar S., Purkayastha B. S., "A Proposed Nepali Synset Entry and Extraction Tool", Proceedings of the 6th Global WordNet Conference, librix.eu, Matsue,Japan Jan 2012.

[11] F. Bond, T. Baldwin, R. Fothergill & K. Uchimoto, "Japanese SemCor: A sense-tagged corpus of Japanese", Proceedings of the 6th International Conference of the *Global WordNet Association* (GWC), 2012.

[12] Saraswati J., Shukla R., Solanki T. and Bhattacharyya P., "Challenges in Multilingual Domain-Specific Sense-marking " GWC 2010

[13] Koeva S., Lesseva S., Todorov M., Bulgarian Sense Tagged Corpus, SALTMIL Workshop on Minority Languages, 2006 , Genoa, Italy.

[14] S. Landes, C. Leacock and R. I. Tengi, "Building semantic concordances", *WordNet: An electronic lexical database*, 1998: 199-216.