Building Multilingual Lexical Resources Using Wordnets: Structure, Design and Implementation

Shikhar Kr. Sarma[1] *Dibyajyoti Sarmah*[1]
Biswajit Brahma[1] *Mayashree Mahanta*[1] *Himadri Bharali*[1] *Utpal Saikia*[1]
(1) Department of Information Technology,
Institute of Science & Technology, Gauhati University, Guwahati – 14 Assam, India
{sks001, dibyasarmah, bswjtbrahma, mayashreemahanta, himadri0001, utpal.sk}@gmail.com

Abstract

The present paper deals with the design and implementation of multilingual lexical resources of Assamese and Bodo Language with the help of Hindi Wordnet. Here, we present the multilingual dictionaries (for Hindi, Assamese and Bodo), synset based word search for Assamese-Hindi and Bodo-Hindi language. These words, of course, will have to go through some pre-processing before finally being uploaded to a database. The user-interface is being developed for specific language (Assamese, Bodo and Hindi language).

KEYWORDS: Lexical Resources, Concept Based Dictionary, Multilingual Dictionary Database, Web-based Interface

1 Introduction

In recent years, mono and multilingual lexical resources, Wordnet and other lexical resources are in high demand. Wordnet is a very recent and rich multilingual lexical resource which is being used in MT (Machine Translation), cross-lingual search, information extraction etc. Among the Indian language Wordnet, the Hindi Wordnet[1] was the first one to come into existence from 2000 onwards. It was inspired by the English Wordnet[2] which contains nouns, verbs, adjectives and adverbs organized into synonym sets, each representing one underlying lexical concept (Fellbaum, 1998). Different relations like hypernymy, hyponymy etc. link the synonym sets to each other. Soon, other Indian language Wordnet started getting created. The Wordnet for Assamese and Bodo have followed the Hindi Wordnet.

The present model tries to represent the lexical elements and their multilingual counterparts efficiently and economically. The present frameworks are derived inspiration from the Hindi Wordnet.

2 A case study: Introduction of Assamese language and Bodo language

Assamese language is the mainly spoken in the state of Assam. According to the VIII schedule of Indian Constitution, Assamese is recognized as the regional language. It becomes the official language of Assam. It is also used as a medium of communication in many north-eastern states specially Arunachal Pradesh and Nagaland and also in outside the north-eastern regions such as Bhutan and Bangladesh. Apart from these, a large number of Assamese speaking people settled in different parts of India and outside India like U.K. and U.S. due to various reasons. The

[1] http://www.cfilt.iitb.ac.in/wordnet/webhwn/
[2] http://wordnet.princeton.edu/

tentative number of Assamese speaker in the state of Assam and neighboring states of north-east India is 1.4 million and across India is approximately 14.3 million.

Bodo language became the scheduled language in the year 2003. It is spoken in the northern part of the Brahmaputra valley of Assam and also in the southern part of the valley. A small section of Bodo speakers are also found in the border areas like Meghalaya, Nagaland, North Bengal, Nepal and Bhutan adjoining Assam. According to the census 1991, there are approximately 11, 84, 569 Bodo speakers. However, the Bodo language has its written record from the last part of the 19th century. In the year 1963, it was introduced in the primary level of education in Assam and presently, it becomes the medium of instruction up to 10th standard in the state of Assam. The script of the Bodo is Devanagiri.

UNICODE compliant font sets, keyboard drivers, corpus, word-processors, spelling checkers, CLDR (Common Locale Data Repository) etc. are being developed with Government of India initiative very recently. Work has also started simultaneously for developing the Assamese and Bodo Wordnet as part of the North East Indo Wordnet development, which will ultimately be linked to the composite Indo Wordnet [Sarma, 2010].

3 The Multilingual Lexical Resources

A lexical resource (LR) is a database consisting of one or several dictionaries. Depending on the type of languages that are addressed, the LR may be qualified as monolingual, bilingual or multilingual. For bilingual and multilingual LRs, the words may be connected or not connected, from a language to another. When connected, the equivalence from a language to another, is performed through a bilingual link (for bilingual LRs) or through multilingual notations (for Multilingual LRs).

Following is the linked synset in Assamese and Bodo Wordnet

Assamese Linked Synset	Bodo Linked Synset
14958	15785

TABLE 1 – Synset of Assamese and Bodo Wordnet

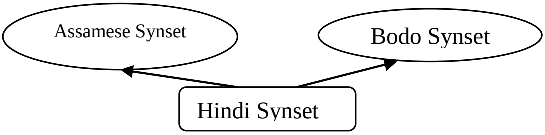


FIGURE 1– Relation between Assamese and Bodo synset with Hindi

Here we define the source language to target language flow diagram. For creating the target language synset, we derive help from Hindi Wordnet. For building the Multilingual (Assamese, Bodo and Hindi) lexical resources we used root Wordnet Hindi for Assamese and Bodo language and mapping words by compare with Hindi.

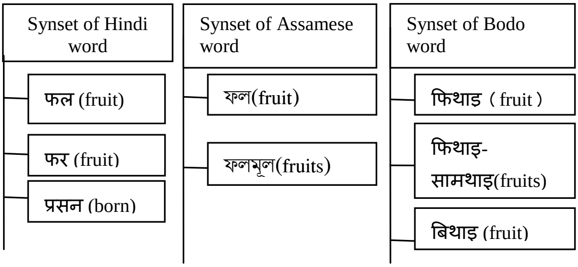


FIGURE 2 –Synset of English 'fruit' word sense in different

There are three words (फल, फर, प्रसून) in Hindi which form the Hindi synset, two words (ফল, ফলমূল) in Assamese from Assamese synset for the same concept and another three words (फिथाइ, फिथाइ-सामथाइ, बिथाइ) in Bodo from Bodo synset, as illustrated in FIGURE 2.

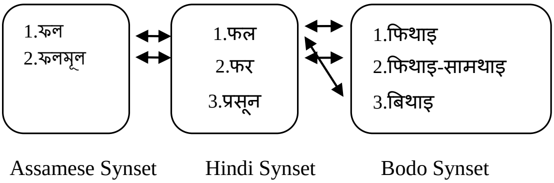


FIGURE 3 –Mapping with root synset (Hindi Synset)

In FIGURE 3 we show the mapping with Assamese and Bodo with root synset Hindi. Here the फल (fruit) word is mapping with Assamese word ফল (phal:fruit) and Bodo word फिथाइ (fithai:Fruit). In the same way फर (phar:fruit) word is related with ফলমূল (phalmul:fruits) (Assamese synset) and फिथाइ-सामथाइ (fithai-samthai:fruits) (Bodo synset). But there is no equivalent Assamese word for Hindi प्रसून word. So, we cannot map this प्रसून with Assamese synset.

4 Challenges in Lexical Resources

Morphological Characteristics (Assamese Language)

Assamese is very rich in morphological features[3]. Some of them are outlined below

1. There is no inflection for number and gender in Assamese. There are two kinds of numbers, viz., singular and plural. Linguistically, Gender is of two types – Masculine and Feminine. But traditionally Common and Neuter gender are also used.
2. Relational nouns or kinship terms are inflected for person and case.
3. Derivation is done by various processes – prefixation, suffixation, zero modification, compounding and change of consonant and vowel phoneme.

[3] Golock C Goswami. 1983. Structure of Assamese, Gauhati University, Assam

4. There are two types of affixes in Assamese language – Prefix and suffix. But there is no infix found in the language.
5. Assamese language contains six types of case markings, Nominative, Accusative, Instrumental, Dative, Ablative and Locative.
6. In negation, the negative ‘n-’ is prefixed to the verb and morphophonemic changes are also common in the language.

Syntactic Characteristics

a) The basic sentence structure in Assamese language is Subject + Object + Verb (SOV). But it may vary according to the context or mood of the speaker
b) Depending on the form, the sentence in the language is of three kinds – Simple, Complex and Compound.
c) Semantically, sentences in Assamese are classed into – Declarative, Interrogative, Exclamatory, Imperative. In fact, Intonation plays a significant role in determining the sentence type.

Bodo Morpho-syntactic features

a) Sentence pattern of the Bodo is Subject + Object + Verb (SOV) pattern.
b) The language does not follow the concord relation which is the agreement of verb and person.
c) There is no change of verb according to the person and number. In each sentence the verb does not possess change of its character regarding person and number where it is singular or plural form in the sentence.

5 Challenges of Lexical Resources

The linkage task has to do a fine balance between maintaining accuracy and providing maximum linkages. While trying to do this for the linkage between the Hindi, Assamese and Bodo Wordnet, several challenges were encountered. The specific such problems were faced are the synset denoting the following:

a) It is often the case that a concept is expressed through a synthetic expression in one language, but through a single word expression in the other language.eg. For Bodo language a single word express a whole sentence.
For example,

[4]HC: एक प्रकार के छोटे जंतु जिनके मुँह में, विशेषकर कुतरने में सहायक, छोटे और पैने दाँत होते हैं

[5]ET: Relatively small gnawing animals having a single pair of constantly growing incisor teeth specialized for gnawing.

[6]HS: कृंतक जन्तु (rodent, gnawer, gnawing_animal)

[7]BS: गोफार_हाथाय_गोनां_जुनार (gwfar-hathai-gwnang-junar: sharped teeth animal)

In this example Hindi Synset कृंतक जन्तु word meaning is like as गोफार_हाथाय_गोनां_जुनार in Bodo Wordnet. This word is a combination of four parts.

[4] HC-Hindi Concept
[5] ET-English Translation
[6]HS-Hindi Synset
[7] BS-Bodo Synset

b) Sometime there is no equivalent concept in target language. For example, the Hindi concept like साधु बन जाना (to become a monk) is not found any equivalent term in the target language Assamese.

Some cultural terms may be missed out from the target languages as these are not available in the Hindi Wordnet. It prevents the true representation of the target language in digital world. For example: the terms relating to festival like विहु (Bihu) in Assamese and बैसागु (Boisagu) in Bodo are not found in the Hindi Wordnet.

In source language and target language, we have found words with same structure with different meanings in different time. For instance, धुरन्धर (dhurandhar) in Hindi means ‘renowned one’, but in Assamese धुৰन्धৰ (dhurandhar) refers to ‘a scoundrel’.

6 Multilingual Lexical Database for Computational Framework

Design of multilingual database by help of root Wordnet (Hindi Wordnet) is shown in below. First we create our target language synset from Hindi Wordnet by using multilingual tool. After creating our own language we put that file in our database. In FIGURE 4 we show the DFD (Data Flow Diagram) of multilingual lexical resources.

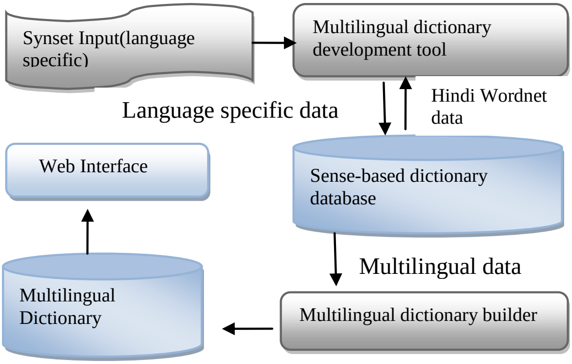

Synset Input(language specific)
Multilingual dictionary development tool
Language specific data
Hindi Wordnet data
Web Interface
Sense-based dictionary database
Multilingual data
Multilingual Dictionary
Multilingual dictionary builder

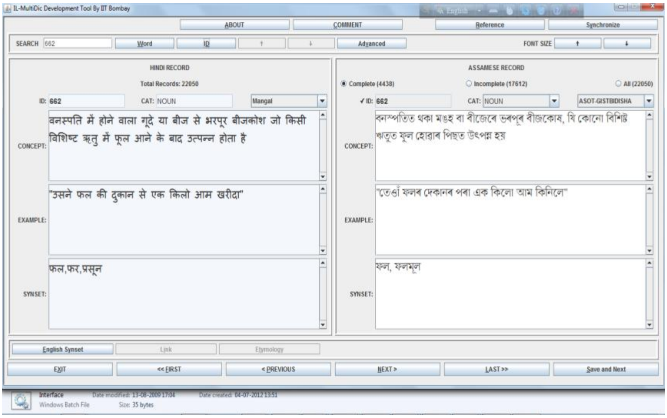
FIGURE 4 – DFD of Multilingual Dictionary creation

7 Multilingual Lexical Database for Computational Framework

The Multilingual tool, used by lexicographers for manually linking the two Wordnet, was developed at CFILT, IIT Bombay.

The offline multilingual tool takes as input a source file containing the number of query synset N, where N stands for total number of synset that are to be linked and N lines in following format:

- Synset ID
- POS category
- Concept
- EXAMPLE
- SYNSET



SEARCH 662

HINDI RECORD

Total Records: 22050

ID: 662 CAT: NOUN Mangal

CONCEPT: वनस्पति में होने वाला गूदे या बीज से भरपूर बीजकोश जो किसी विशिष्ट ऋतु में फूल आने के बाद उत्पन्न होता है

EXAMPLE: "उसने फल की दुकान से एक किलो आम खरीदा"

SYNSET: फल,फर,प्रसून

ASSAMESE RECORD

● Complete (4438) ○ Incomplete (17612) ○ All (22050)

✓ID: 662 CAT: NOUN ASOT-GISTBIDISHA

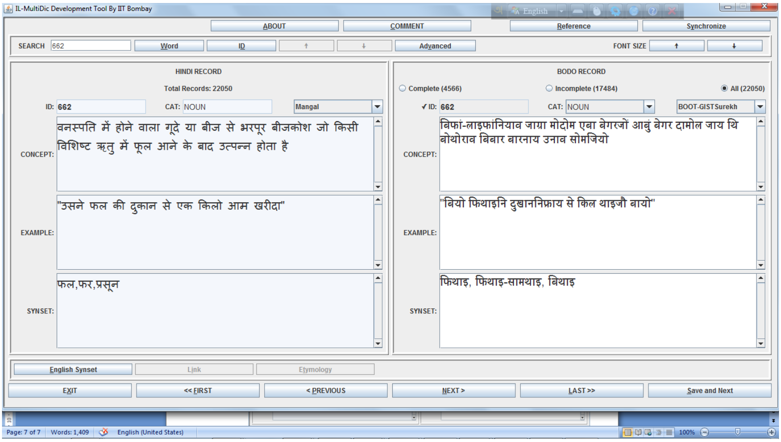
CONCEPT: বনস্পতিত থকা মঙহ বা বীজেৰে ভৰপূৰ বীজকোষ, যি কোনো বিশিষ্ট ঋতুত ফুল হোৱাৰ পিছত উৎপন্ন হয়

EXAMPLE: "তেওঁ ফলৰ দেকানৰ পৰা এক কিলো আম কিনিলে"

SYNSET: ফল, ফলমূল

FIGURE 5 –Multilingual Tool (for Assamese language)

In this tool, the synset (synset ID, POS category, Concept, example and synonyms) is displayed in the source synset panel at the top of the tool. Similar information is displayed in the candidate synset panel below it, for each of the N candidate synset. The candidates are displayed in decreasing order of their confidence score. Facility for searching synset in both source and target languages with respect to a word or synset ID is also provided in the tool.



SEARCH 662

HINDI RECORD

Total Records: 22050

ID: 662 CAT: NOUN Mangal

CONCEPT: वनस्पति में होने वाला गूदे या बीज से भरपूर बीजकोश जो किसी विशिष्ट ऋतु में फूल आने के बाद उत्पन्न होता है

EXAMPLE: "उसने फल की दुकान से एक किलो आम खरीदा"

SYNSET: फल,फर,प्रसून

BODO RECORD

○ Complete (4566) ○ Incomplete (17484) ● All (22050)

✓ID: 662 CAT: NOUN BOOT-GISTSurekh

CONCEPT: बिफां-लाइफांनियाव जाग्रा मोदोम एबा बेगरजों आबुं बेगर दामोल जाय थि बोथोराव बिबार बारनाय उनाव सोमजियो

EXAMPLE: "बियो फिथाइनि दुखाननिफ्राय से किल थाइजौ बायो"

SYNSET: फिथाइ, फिथाइ-सामथाइ, बिथाइ

FIGURE 6 –Multilingual Tool (for Bodo language)

We have taken help from the Indo Wordnet website, when we did not find equivalent concept in our target languages. For example, Synset ID, POS (Part Of Speech), Concept, Example, Synset, Hyponymy etc. for respective languages.

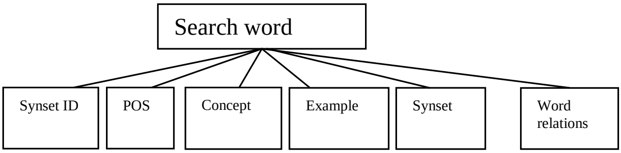
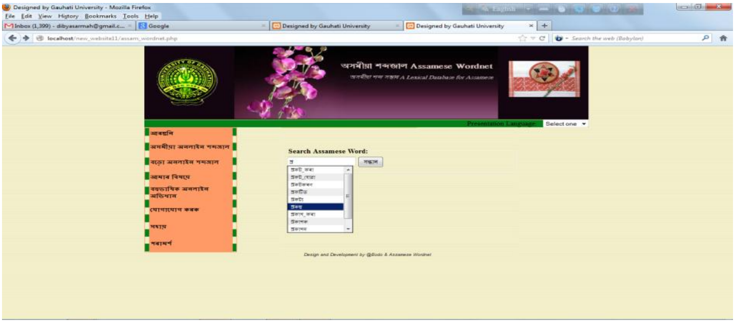


FIGURE 7– Word Structure of Hierarchical order

8 User Interface of lexical resources

A. Bilingual link (Assamese-Hindi synset based translation).
B. Bilingual link (Bodo- Hindi synset based translation).
C. Multilingual dictionary (Assamese-Bodo-Hindi).

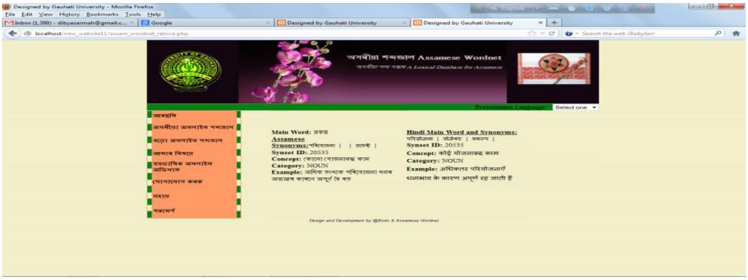

Designed by Gauhati University - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Inbox (1,399) - dibyasarmah@gmail.c...
Google
Designed by Gauhati University
Designed by Gauhati University
localhost/new_website11/assam_wordnet.php
Search the web (Babylon)
অসমীয়া শব্দজাল Assamese Wordnet
অসমীয়া শব্দ সম্ভাৰ A Lexical Database for Assamese
Presentation Language
Select one
আৰম্ভণি
অসমীয়া অনলাইন শব্দজাল
বড়ো অনলাইন শব্দজাল
আমাৰ বিষয়ে
বহুভাষিক অনলাইন অভিধান
যোগাযোগ কৰক
সহায়
পৰামৰ্শ
সন্ধান
Design and Development by @Bodo & Assamese Wordnet

Search Assamese Word:

প্ৰ

প্ৰকট_কৰা
প্ৰকট_হোৱা
প্ৰকটিকৰণ
প্ৰকটিত
প্ৰকটা
প্ৰকল্প
প্ৰকাশ_কৰা
প্ৰকাশক
প্ৰকাশন

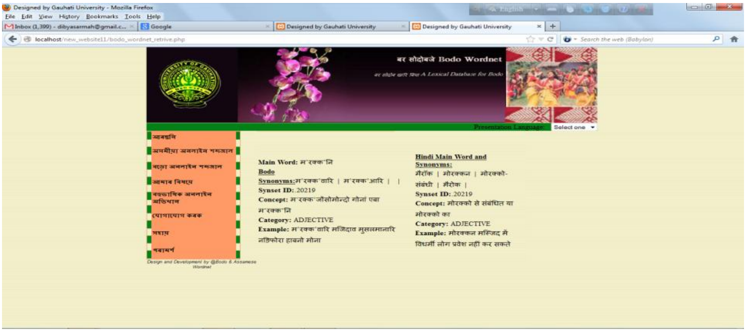
FIGURE 7–Searching a word (User Interface)


Designed by Gauhati University - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Inbox (1,399) - dibyasarmah@gmail.c...
Google
Designed by Gauhati University
Designed by Gauhati University
localhost/new_website11/assam_wordnet_retrive.php
Search the web (Babylon)
অসমীয়া শব্দজাল Assamese Wordnet
অসমীয়া শব্দ সম্ভাৰ A Lexical Database for Assamese
Presentation Language
Select one
আৰম্ভণি
অসমীয়া অনলাইন শব্দজাল
বড়ো অনলাইন শব্দজাল
আমাৰ বিষয়ে
বহুভাষিক অনলাইন অভিধান
যোগাযোগ কৰক
সহায়
পৰামৰ্শ
Design and Development by @Bodo & Assamese Wordnet

Main Word: প্ৰকল্প
Assamese Synonyms: পৰিযোজনা | | প্ৰজেক্ট |
Synset ID: 20535
Concept: কোনো যোজনাবদ্ধ কাম
Category: NOUN
Example: অধিক সংখ্যক পৰিযোজনা ধনৰ অভাৱৰ কাৰণে অপূৰ্ণ হৈ ৰয়

Hindi Main Word and Synonyms:
परियोजना | प्रोजेक्ट | प्रकल्प |
Synset ID: 20535
Concept: कोई योजनाबद्ध काम
Category: NOUN
Example: अधिकतर परियोजनाएँ धनाभाव के कारण अपूर्ण रह जाती हैं

FIGURE 8–Synset based word search (Assamese-Hindi)


Designed by Gauhati University - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Inbox (1,399) - dibyasarmah@gmail.c...
Google
Designed by Gauhati University
Designed by Gauhati University
localhost/new_website11/bodo_wordnet_retrive.php
Search the web (Babylon)
बर सोदोबजे Bodo Wordnet
A Lexical Database for Bodo
Presentation Language:
Select one
আৰম্ভণি
অসমীয়া অনলাইন শব্দজ্ঞান
বড়ো অনলাইন শব্দজ্ঞান
আমাৰ বিষয়ে
বহুভাষিক অনলাইন অভিধান
যোগাযোগ কৰক
সহায়
পৰামৰ্শ
Design and Development by @Bodo & Assamese Wordnet

Main Word: म'रक्क'नि
Bodo Synonyms:म'रक्क'वारि | म'रक्क'आरि | |
Synset ID: 20219
Concept: म'रक्क'जौसोमोन्दो गौनां एबा म'रक्क'नि
Category: ADJECTIVE
Example: म'रक्क'वारि मजिदाव मुसलमानारि नङिफोरा हाबनो मोना

Hindi Main Word and Synonyms:
मैरॉक | मोरक्कन | मोरक्को-संबंधी | मैरोक |
Synset ID: 20219
Concept: मोरक्को से संबंधित या मोरक्को का
Category: ADJECTIVE
Example: मोरक्कन मस्जिद में विधर्मी लोग प्रवेश नहीं कर सकते

FIGURE 9–Synset based Word search(Bodo-Hindi)

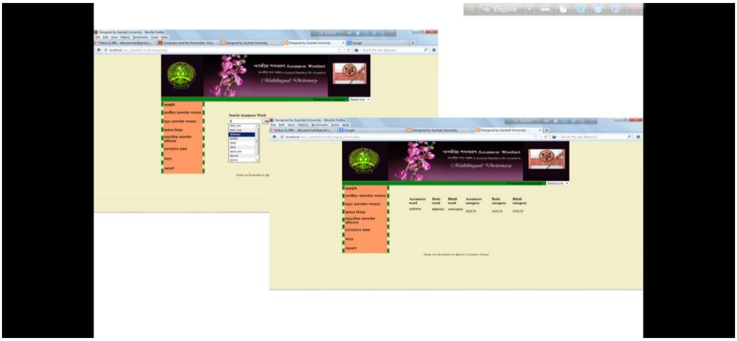


FIGURE 10–Interface of Multilingual Dictionary (Assamese-Bodo-Hindi)

9 Conclusions

In this paper, we present a discussion on the structure, design and implementation of the multilingual lexical resources for Assamese and Bodo Wordnet which is done by mapping with the Hindi Wordnet. Besides, the present paper also highlights the challenges faced in creating the Wordnet in Assamese as well as in Bodo such as script issue, cultural terms, similar structure but different meaning etc.

In future, attempts should be taken to create Wordnet for other north-eastern languages as well as other Indic languages which would not only preserve the language but also standardize the language in digital world. This kind of research would help the user for easy browsing of any language data in digital format.

Acknowledgment

The works have been done during the NE Wordnet development project. The authors acknowledge support from DIT, Ministry of Communication & IT, Govt. of India, & Indwordnet team, IIT Bombay.

References

Awasthi, S. and (Smt.) I. Awasthi. 2000. Chambers English-Hindi Dictionary (ed.). Allied Publisher Limited, New Delhi, India.

Fellbaum, C. 1998. Wordnet: An Electronic Lexical Database. The MIT Press.

Kamil, Bulke. 1997. An English-Hindi Dictionary (ed.). S. Chand & Co, New Delhi, India.

Khapra, Mitesh, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya. 2009. Projecting Parameters for Multilingual Word Sense Disambiguation. Empirical Methods in Natural Language Processing (EMNLP09), Singapore.

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

Ramanand J.. Akshay Ukey, Brahm Kiran Singh, Pushpak Bhattacharyya. 2007. Mapping and Structural Analysis of Multi-lingual Wordnets. IEEE Data Engineering Bulletin, 30(1).

Sarma, Shikhar Kr., Moromi Gogoi, Rakesh Medhi and Utpal Saikia, 2010. Foundation and Structure of Developing an Assamese Wordnet, Global Wordnet Conference, IIT Bombay.

Sarma, Shikhar Kr.,Moromi Gogoi, Biswajit Brahma, Mane Bala Ramchiary,2010. A Wordnet for Bodo Language: Structure and Development.

Sinha Manish,Mahesh Kumar Reddy and Pushpak Bhattacharyya.2006.An Approach towards Construction and Application of Multilingual Indo-WordNet, 3rd Global WordNet Conference (GWC 06), Jeju Island, Korea.