A Wordnet for Bodo Language: Structure and Development

Dr. Shikhar Kr. Sarma
Department of Computer Science
Gauhati University
Guwahati 781014: Assam, India
sks001@gmail.com; sks@gauhati.ac.in

Biswajit Brahma
Department of Computer Science
Gauhati University
Guwahati 781014: Assam, India
bswjtbrahma@gmail.com

Moromi Gogoi
Department of Computer Science
Gauhati University
Guwahati 781014: Assam, India
moromi_gogoi@yahoo.com

Mane Bala Ramchiary
Department of Computer Science
Gauhati University
Guwahati 781014: Assam, India
maner@gmail.com

Abstract

This paper discusses the linguistics foundations for developing a Bodo Wordnet, describing the Bodo language characteristics and properties specific to the development of Wordnet. The characteristics of the Bodo language in terms of its morphological and syntactic structure are outlined. Important characteristics related to building of Wordnet are discussed with examples. As the Bodo Wordnet is being developed as an expansion of the Hindi-English Wordnet, the experience gathered during the initial startup works are very important in carrying out the whole work. Such experiences during the building of core 2000 synsets are discussed in the paper, alongwith the challenges faced during linking.

1 Introduction

Wordnet building in relatively newly developed languages has been a challenge for the linguists, researchers and computing professionals. This is mainly because of the technologically immature language scenario, as well as lack of properly structured linguistics resources. Bodo language is also falls in this category. The language is in its developing state, and in recent years only proper linguistics and literal emphasizes have been started. The Bodo language is a scheduled language of India, mainly spoken by the Bodo Community of the state of Assam. The first generation researchers in the field of Bodo Linguistics are just coming up, and technological developments have just started. The bodo language uses the Devnagiri scripts with additional symbols. UNICODE compliant font sets, keyboard drivers, corpus, word-processors, spelling checkers, CLDR etc are being developed with Government of India initiative very recently. Work has also started simultaneously for developing the Bodo Wordnet as part of the North East Indo Wordnet development, which will ultimately be linked to the composite Indo Wordnet.

1.1 Bodo Language

The Bodo language has its written record from the last part of the 19th century. This language was introduced in the primary level of education in Assam from the year 1963 and presently is the medium of instruction upto 10th standard in the state of Assam. It was recognized by the government of Assam as official language in the Kokrajhar district and Udalguri sub-division from the year 1984. The language also got Indian govt. recognition as scheduled language from 2003. According to the census of 1991 it has a total of 11, 84,569 speakers.

The Bodo population has basic concentration in the northern part of the Brahmaputra valley of Assam. They have also thin concentration in the southern part of the valley. Besides that they have also concentration in small number in the border areas of Meghalaya, Nagaland, North Bengal, Nepal, and Bhutan adjoining Assam.

1.2 Origin and History

The Bodo language belongs to the Tibeto-Burman branch of the Sino-Tibetan language family. It is a major language of the North-Eastern part of India and has very close resemblance with the Rabha, Garo, Dimasa, Kokborok, Tiwa, Hajong and other allied languages of N-E India. It is thought that this

language speakers have migrated through two different routes into Assam: one by the western route adjoining Himalayas and the other by the stream of the Brahmaputra river by eastern side of Assam. It is thought that the origin of this language is the headwaters of the Huang-Ho and Yang-Tsze-Kiang rivers in China. According to the scholars it is considered that this language presently has three distinct different dialect groups.

2 Characteristics of Bodo Language

This language has a total of 22 phonemes: 6 vowels and 16 consonants. Use of the high back unrounded vowel phoneme /w/ is very frequent in Bodo language. The Bodo language has different special characteristics such as: It has intonation pattern, juncture and two types of tones. The words in Bodo are highly mono-syllabic. It has agglutinative features also.

2.1 Morphological characteristics

a) The morphological feature of this language is discussed under two basic heads: primary and secondary grammatical categories. Primary consists of Noun, Pronoun, Verb, Adjective, Adverb, Conjunction and Interjection. Secondary consists of Number, Gender, Person, Case and Case-Endings, Numerals and Numeral Classifiers and Tense.
b) Noun has basic, derived as well as compound form composed of noun and verb, verb and noun as well as noun and noun.
c) Pronoun has five different categories.
d) Verb has simple, complex and compound as well as transitive, intransitive, causative, finite and infinite based on structure and function.
e) Adjective has basic and derived form and its basic foundation is verb.
f) Adverbs are basically derived from the adjectives by using derivational suffixes.
g) Numbers are two in this language and are inflected basically with nouns, pronouns also with adjectives.
h) It is basically a natural gender language having two genders i.e. masculine and feminine. Traditionally common and neuter are also used. It has three different phases of gender formation.
i) It has three persons: 1^{st}, 2^{nd}, 3^{rd} and is discussed with the personal pronouns.
j) It has seven cases including ablative and genitive.
k) Numerals have basic and derived forms. Classifiers are prefixed with the numerals.
l) Traditionally tense has three different forms: past, present and future, but are very difficult to completely differentiate in some cases.
m) It has two affixes: prefix and suffix. In comparison to suffix the number of prefix is relatively small. Suffixes are inflectional and derivational as well as class maintaining and changing.
n) Kinship terms are discussed only with the personal pronouns.

2.2 Syntactic structure

a) Structurally syntax has three forms: simple, complex and compound.
b) General syntactic structure is of S-O-V pattern.
c) It has no concord relation.
d) Its word order is flexible and is based on the context and mood of the speaker.
e) It has idiomatic and non-idiomatic use of sentences.
f) It has the use of verb and verb less sentences.

2.3 Bodo Synonyms and Antonyms

Few Examples of Bodo Synsets:

[World, English]: [पृथ्वी,हिन्दी]: बुहुम, मुलुग, भुम, संसार, हालुर, बैसोमाथा, बिलाथलाथा [Bodo]।

[Jungle, English]: [जंगल,हिन्दी]: हाग्रामा, अरन, हाग्रा, हाग्राबारि, जाहार, आरंगा [Bodo] ।

[Body, English]: [शरीर,हिन्दी] : देहा, मोदोम, सोलेर, सावस्रि [Bodo] ।

[God, English]: [भगवान,हिन्दी]: इसोर, गसाइ, आनान_गसाइ, अबंलावरि, अबं [Bodo] ।

Few Examples of Bodo Antonyms are:

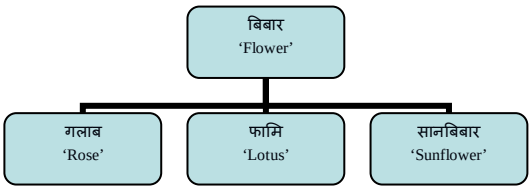
[Big-small, English]: [बड़ा-छोटा, हिन्दी]: गिदिर-फिसा [Bodo]

[Good-Bad, English]: [अच्छा-बुरा, हिन्दी] : मोजां - गाज्रि [Bodo]

[High-Low, English]:[ऊँचा-नाटा, हिन्दी] : गोजौ – गाहाय [Bodo]

2.4 Bodo Hyponymy and Hypernyms

Examples:



Here, गलाब, 'rose', फामि 'lotus', सानबिबार 'sunflower' are hyponyms of the superordinate term or hypernym of बिबार 'flower'.

Another Example:

Thaijwubilai; a leaf of a tree named *Thaijwu*
=>*bilai; leaf*

Here, *Thaijwubilai*, a leaf of a tree named *Thaijwu* is a kind of *bilai*; *bilai* means leaf; *bilai* is a hypernym and *Thaijwubilai* is the hyponym. Similar case with the example of *phakhribilai* also.

2.5 Meronymy and Holonymy

आसि : Finger

=>आखाय : Hand

बिथराय : Petal

=>बिबार : Flower

Here, आसि 'finger' and बिथराय 'petal' are meronymy and आखाय 'hand' and 'बिबार'flower' are holonymy.

2.6 Entailment

हंखदनाय : To snore

=>उन्दु : To sleep

2.7 Troponymy:

मिनिखैरो : smiling

=> *मिनि : laughing*

2.8 Antonymy

gwjwu ;High
=>*gahai*; Low
geder ;Big
=>*undwi*; Small

2.9 Inter-language Element

There are inter-language elements in Bodo and these elements are from the same language family. Few of these are shown below-

Bodo	Garo	Rabha	Dimasa	Kok-Borok	Hindi
न(house)	नक	नक	नक	न	घर
हा(land)	हा	हा	हा	हा	जमीन
लामा(road)	---	----	लामा	लामा	रास्ता

3 The Plan for Building Bodo Wordnet

Bodo Wordnet building is part of the North East Indian language Wordnet. This is in turn a sub part of the composite Indo Wordnet. As a policy, the Wordnet of Bodo language is planned as an expansion of the existing Hindi Wordnet. The work has been started with building of a prototype wordnet with 2000 core synset taken out of the Hindi Wordnet. The synset entries are done against concepts of those core Hindi synsets. A common interface has been used for data entry. Entries are started based on the corresponding Hindi Synset ID. The interface provides facilities for displaying already existing synsets with ID, Concept, Gloss, and the complete synset. For new entry, it provides the blank spaces and facilitates saving into the existing text file updating it on every saving attempt. The interface also provides display of corresponding English synset with examples and concepts, in a popup screen.

Expansion approach is followed to build the Bodo Wordnet. As the Bodo Wordnet is ultimately a part of the Indo Wordnet, this facilitates the creation of the composite Wordnet. This has also drastically reduced the preliminary work of generating and compiling concepts. For a developing language like Bodo, it has significantly accelerated the entire workflow.

Few general principles adopted for constructing the Bodo Wordnet are:

- Concepts are borrowed from the Hindi Wordnet. All Hindi concepts will have corresponding Bodo concepts, to the maximum extent possible.
- Direct translation of the concepts are encouraged. Whenever direct translation is not possible, conceptual translation is done.
- Examples are also direct translation of the Hindi examples. When it does not provide the proper meaning, or if a more language specific example prevails, translation is avoided.
- Synsets are created language specific, not looking at the Hind synset, but looking at the just created Bodo concept.

- When the concept or example could not be completed with the existing vocabulary of Bodo language, transliteration is done.
- English concept and example are referred for removing ambiguities.
- Validating with experts before finalization.

4 Challenges in Expansion

Bodo is a developing language. It does not have a very strong linguistic resource. Also literature resource is very limited. The language does not have enough vocabulary, and new and new words are being discovered, coined and added. As a result, the development of Bodo Wordnet faces typical and frequent problems, and overcoming the problems to accommodate expansion of the Hindi Wordnet with one to one mapping has been a big challenge. Out of the 2000 core synsets mapped from Hindi to Bodo till now, in around 20 Hindi synsets, there could not be found Bodo sysnsets corresponding to the exact Hindi concept. A very interesting observation is that sometimes to represent a particular concept of Hindi, Bodo language requires a complete sentence or long phrase as the synset, rather than one single word or combination of two words. In such exceptional cases, the synset entries are done with underscore as the linking between words. Some of the frequently encountered challenges during the initial exercises are:

i. Lack of proper vocabulary/set of vocabulary to mean the concept, or the example.
ii. For certain concepts in Hindi/English, there is no such vocabulary as member of the synset
iii. In many cases, the synset is very small. Two members/three members synsets are very common
iv. Sometimes the synset contains members containing multiple words. In case of multiple words, they are joined by an underscore.
v. In many cases, the synset entry itself appears in the concept.

5 Conclusions

Building of Bodo Wordnet is aimed in developing a comprehensive and rich computational linguistics resources for the language. As the language itself is developing in terms of linguistics and literature, the Wordnet building of Bodo language is with lots of challenges. The demand and requirements of digital technologies in local languages have been felt among the users and policy makers, including at government and academics. As a result efforts have been initiated to developing tools, applications, and technologies for developing Bodo as an effective media of the digital world. The majority of the Bodo spoken population are in rural areas, and bringing digital technologies to the rural mass requires tools and technologies in the language. Wordnet development in Bodo language is an ambitious project, but we visualize its potential impact integrating with other digital technologies and applications. This will certainly add newer dimensions to the Bodo language in its journey towards its success, spreading, expansion, and in becoming an effective and efficient media of knowledge society.

Acknowledgements

The Bodo Wordnet Development Project is sponsored by the Department of Information Technology, Ministry of Communication and Information Technology, Government of India.

References

Chakrabarti Debasri, Narayan Dipak Kumar, Pandey Prabhakar, Bhattacharyya Pushpak. 2002. *Experiences in Building the Indo WordNet: A WordNet for Hindi*, Proceedings of the First Global WordNet Conference.

Dave Shachi and Bhattacharyya Pushpak. 2001. *Knowledge Extraction from Hindi Texts*, Journal of Institution of Electronic and Telecommunication Engineers, vol. 18, no. 4.

E. Pianta, L. Bentivogli, C. Girardi. 2002. *MultiWordNet: Developing an Aligned Multilingual Database*, Proceedings of the First International Conference on Global WordNet Mysore, India.

Fellbaum, C. (ed.), 1998. *WordNet: An Electronic Lexical Database*,. The MIT Press.

Miller,G., Beckwith,R., Fellbaum,C., Gross,D., and Miller,K, 1990. *Five Papers on WordNet.* CSL Report 43, Cognitive Science Laboratory, Princeton University, Princeton.

Mintu Narzary. 2009. *Standard Anglo Bodo Dictionary*, Nilima Prakashani, Baganpara, Nalbari, India

Promod Chandra Brahma. 2003. *Bodo-Engraji-Hindi Swdwb Bihung*, Bodo Sahitya Sabha, Kokrajhar, India

Ramchiary and Daimary. 2009. *Hindi-Bodo Swdwb Bakhri*, Bina Library, Guwahati, India.

Surendra Goyari. 2000. *Hindi-Bodo Swdwb Bihung*, Ganesh Prakashan, Guwahati, India.