Foundation and Structure of Developing an Assamese Wordnet

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

Rakesh Medhi
Department of Computer Science
Gauhati University
Guwahati 781014: Assam, India
rakesh.medhi@gmail.com

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

Utpal Saikia
Department of Computer Science
Gauhati University
Guwahati 781014: Assam, India
utpal.sk@gmail.com

Abstract

Development of Wordnets of regional languages has been of great concern in recent years. This is mainly due to the ever increasing demands and requirements of putting those languages as effective media of the digital world, including the internet. As the technologies for putting regional languages in the digital media are being developed, research and development works related to Wordnets in those languages are also starting. Efforts have been taken at different level, including academic researchers and at government level for developing language technologies for the Assamese language, a scheduled language in India mainly spoken by the people in the state of Assam. Basic technologies like,
UNICODE compliant fonts and keyboards, CLDR, Corpus, Spelling Checker etc. have been developed. As a part of the Government of India efforts on Technology Development of Indian Languages, creation of Assamese Wordnets has been started. This paper focuses on the foundations of the Assamese Wordnet, and describes the complete background concepts, language specifications, properties and characteristics, as well as the plan and challenges of creating such a Wordnet of Assamese language. The details of the structure, workflow are outlined. Considerable contents are included on the experience of handling developing the Assamese Wordnet with linkage to the Hindi and English Wordnets. The paper also focuses on the challenges faced on mapping to the Hindi and English Wordnets. The qualitative and quantitative achievements on structuring the core preliminary part of the Assamese Wordnet are also presented.

1 Introduction

Assamese language is the main spoken language of the state of Assam. This language is recognized as regional language in the eighth schedule of Indian constitution. This is one of the official languages of Assam. This language is also used as interstate communication language in many north eastern states specially Arunachal Pradesh and Nagaland. Apart from the states in Indian Territory, Assamese spoken people are found in Bhutan and Bangladesh also. There is a huge population of Assamese spoken people in different parts of India, who originated from Assam, but professionally settled in other states. There is also a considerable number of Assamese speakers in different other countries specially in UK and US. The tentative number of people speaking Assamese in the state of Assam and neighboring states of North East India, is 14 million, and the all India tentative counting is about 14.3 million.

While tracing the origin of Assamese language its relation is found to the Indo-Aryan language group and also a little bit to Sino-Tibetan language group. In 1870 linguist Ascoli divided Indo-European language in two main groups viz Satam and Centum. Satam group of language is again divided in four groups. One of which is Indo-Iranian group from where Indo-Aryan group is derived. Indo-Aryan language group is again divided into three parts[1], they are-

1. Old Indo Aryan (1500 BC to 600 BC)
2. Middle Indo Aryan(600 BC to 1000 AD)
3. New Indo Aryan(1000 AD to till now)

In Indo Aryan Languages, Assamese, Bangla, Oriya etc are derived from Modern Apabhransha. These languages are originally derived from the Magadhi-Prakrit. Presence of Assamese language dated back to the literatures of *Charyapadas*, written by Budhist scholars. The

Assamese language present in *charyapadas* reflects its evolutionary stages in initial state. Literatures with distinct Assamese language are found from the *Kavyas* of the pre Sankari era. This was in 13th century AD. From that time onwards pure Assamese language with its structured forms evolved.

Assamese script is derived from Brahmi script. It played a vital role in the evolution of the Indian script. The rock inscription and copper plate from 5th to 9th century showed the evolution of Assamese script. There are eight vowel phonemes in Assamese. There are twenty-one consonant and two semi-vowel phonemes in the Standard Colloquial Assamese.

2 Characteristics of the language

2.1 Morphological Characteristics:

The Assamese language has many special morphological characteristics [6]. Out of which few are outlined below:

i. Numbers are not grammatically marked in Assamese Language. There are two types of Numbers in Assamese Language, Singular and Plural.
ii. Gender is also not grammatically marked in Assamese. Linguistically, there are two types of Gender in this Language, Masculine and Feminine. But traditionally common and neuter genders are also used.
iii. Kinship nouns are inflected for personal pronominal possession.
iv. There are two types of affixes in Assamese Language, Prefix and Suffix. Both Prefix and Suffix are very commonly found in Assamese language. Suffix included derivational, Inflectional and conjugational forms.
v. There are six types of Cases in Assamese language, Nominative, Accusative, Instrumental, Dative, Ablative, and Locative.

There are six parts of speech (POS) in Assamese language [6]. They are :-

- Noun (Common, Proper, Collective, Material, Abstract.)
- Pronoun (Personal, Demonstrative, Inclusive, Relative, Indefinite, Interrogative, Reflexive)
- Verb (Transitive, Intransitive)
- Adverb (Manner, Place, Time)
- adjective (Nominal, Qualifying)
- Indeclinable (Conjunction, Interjection etc.)

2.2 Syntactic Characteristics

Few general syntactic characteristics of Assamese language are mentioned below [6]:

i. The general syntactic structure of Assamese language is Subject+Object+Verb (SOV).
ii. Syntactically Assamese sentence structure is mainly divided into three types - Simple, Complex and Compound.
iii. Assamese sentence structure is flexible. Depending on the context or mood of the speaker it might vary.
iv. Assamese sentence structure is of different kinds. Very short sentences are found frequently. Sometimes long expressions are made by adding indeclinables.
v. In Assamese, there are subject-verb agreements. The verbs in Assamese agree with the subjects in person. There is no agreement in number or gender like some other languages, English or Spanish etc.
vi. Verb-less sentences are also very frequent in Assamese language.
vii. Idiomatic expressions are also found in Assamese Language.

2.3 **Assamese Synonyms/ Antonyms**

There are large numbers of synonyms and antonyms used in Assamese language. Few examples are given below:

Assamese Synonyms

[English: Sun (सूर्य)]: সূৰ্য （*xurja*）, বেলি （*beli*）, ৰবি （*ravi*）, ভানু （*bhanu*）, তপন （*tapan*）, মিহিৰ （*mihir*）, দিবাকৰ （*dibakar*）, সূৰুয （*xuruj*）, সবিতা （*xabita*）, চাকা （*saka*）...

[English: Moon (चंन्द्र)]: চন্দ্ৰ （*sandra*）, জোন （*jon*）, জোনবাই （*jonbai*）, শশী （*sasi*）, চন্দ্ৰমা （*sandrama*）, শশাংক （*sasanka*）, মৃগাংক （*mrigangka*）, নিশাকৰ （*nisakar*）...

[English: Earth (पृथ्वी)]: পৃথিৱী （*prithiwi*）, জগত （*jagat*）, ধৰণী （*dharani*）, ধৰা （*dhara*）, ধৰিত্ৰী （*dharatri*）, মেদিনী （*medini*）, বিশ্ব （*biswa*）, বসুমতী （*baxumati*）, অৱনী （*awani*）, জাহান （*jahan*）...

[English: Sky (आकाश)]: আকাশ （*akash*）, গগন （*gagan*）, অম্বৰ （*ambar*）, আছমান （*asman*）, অভ্ৰ （*abhra*）, নীলিমা （*nilima*）, অন্তৰীক্ষ （*antariksha*）...

Assamese Antonyms

[English: Day-Night (दिन -रात)]: দিন-ৰাতি （*din–rati*）

[English : Black-White(काला -चाफ)]: ক'লা-বগা (*kola-baga*)

[English : Good-Bad (अच्छा-बुरा)]: ভাল- বেয়া (*bhal-beya*)

[English: Dark-Light (अँधेरा-उजाला)]: আন্ধাৰ-পোহৰ (*andhar-pohar*)

[English : Warm-Cold (गर्म-ठंढ)]: গৰম-ঠাণ্ডা (*garam-thanda*)

[English : Birth-Death (जन्म-मरण)]: জন্ম-মৃত্যু (*janma-mrityu*)

[English : Win-Defeat (जीत-हार)] : জয়-পৰাজয় (*jay-parajay*)

2.4 Hyponymy and Hypernyms

Hyponymy involves in the notion of inclusion. Hyponymy is the relationship which obtains between specific and general lexical items, such the former is 'included' in the latter (i.e. 'is a hyponym of' the latter). In each case, there is a superordinate term, sometimes called a 'hypernym', with reference to which the subordinate term can be defined.
For instance,
Belpat ; a leaf of a tree named *Bel*
=>*pat*; leaf
Here, *Belpat*; a leaf of a tree named *Bel* is a kind of *pat*; *pat* means leaf; *pat* is a hypernym and *belpat* is the hyponym. Similar case is with the example of *kolpat* also.

2.5 Meronymy and Holonymy

anguli ;Finger
=>*hat*;Hand
pahi ; Petal
=> *phul*; Flower.

2.6 Entailment

usup ; To sob
=> *kand*; To cry, To weep

2.7 Troponymy:

misikiya; To smile
=>*hah*; to laugh
dour; To run
=>*za*; to go

2.8 Antonymy

suti ; Short
=>*dighol*; Long
din; Day
=>*rati;* Night

2.9 Inter-language element in Assamese synset/synonym

Assamese is rich in synonymous words. In fact, we can find many inter-language elements in Assamese synset. For instance, Hindi and Bangla elements are commonly found in Assamese synset, few of which are shown below:

Assamese	**Bengali**	**Hindi**
নদী (River)	নদী	नदी
বৰষুণ (Rain)	বৃষ্টি	वृष्टि
বিবাহ, বিয়া (Marriage)	বিয়া, বিবাহ	विवाह
স্বামী, পতি (Husband)	স্বামী, পতি	स्वामी, पति
কিতাপ, পুস্তক, গ্ৰন্থ (Book)	পুস্তক, গ্রন্থ	पुस्तक, ग्रन्थ

3 The developing plan

The Assamese Wordnet is being developed as a sub part of the North East Wordnet Development effort, which is a part of the Indo Wordnet. The basic principles adopted have been uniform for all the North Eastern Language Wordnet developments. A common interface for creation of Wordnet is being used. The interface has simultaneous multiple active spaces. One is for the Hindi Synset display. And the other is for the Assamese entry and display. The Hindi space includes the Hindi Synset ID, Category, Concept, gloss, and the synset. The corresponding Assamese space is on the right hand. The ID is same for the same concept. Part of Speech category could be changed. This is because, sometimes the corresponding concepts in Hindi and Assamese might have part of speech ambiguities. This space either displays the already entered contents, or facilitates the content entry. The interface has a third space, which is normally hidden, and could be made active on clicking button. This third space contains the corresponding English synsets.

As the whole idea is to create an integrated Indo Wordnet, we have adopted the expansion approach. The framework guidelines adopted for creating the Assamese Wordnet as an expansion of existing Hindi and English Wordnets are as follows:

- For Concepts:
  - Understanding using the Hindi one. For clarification, refer to the English one, if and only if it is required
  - Direct translation from Hindi in all possible cases
  - In case the translation is extremely difficult, not meaningful, or could not be represented with available set of vocabulary, create concept in local language style.
- For Examples:
  - Translation from Hindi Examples
  - If it is ambiguous/not fully meaningful in local language, create new one
- Building the synsets:
  - Standardisation
    - Use of official dictionaries
    - No new coining of Words
    - Rather Coining of proposals for new required words
  - Creating a reporting interface
    - Printing the entries in a tabular format
    - Distributing to the expert linguists and feedback
    - Putting in public domain and feedback
- Referring to the English entries, concepts and examples for any clarifications, or to resolving any ambiguities

4 Expansion from Hindi/English

The main challenge in expansion approach is in one to one mapping. Although initially in most of the cases, Hindi-Assamese mapping has shown one to one correspondence, but as we progressed, we started encountering many problems. The problem ranges from word meaning ambiguities to concept mismatch. Few important challenges faced during initial 2000 core synset creation are outlined below:

i. There is no equivalent concept in Assamese language against a Hindi concept. This has been found particularly in meaning corresponding to human relations
ii. There is no Assamese synset against a Hindi concept. The synset overlaps with another Hindi concept
iii. The Hindi concept itself sometimes differ from the English concept in meaning
iv. Direct translation of Hindi examples result in awkward/less-meaningful sentences in Assamese
v. In many cases in the Hindi Wordnet, the Synset entries are found in the Concept itself. This result in similar overlapping in the Assamese entries. This could have been resolved provided the flexibility to modifying the concept would have been there. But as a principle of Indo Wordnet, at present this flexibility is not allowed.
vi. In few cases, the Hindi and English concepts conflict within them.

5 System Architecture

The computational infrastructure for creating the Assamese Wordnet has been built in a Client Server Architecture. Although the standalone tool for creating the Wordnet is used at desktop node, the backup, replication and updating has been automated in the client-server configuration. The default text file where the entries are stored has been used to structure a database of the Wordnet entities against fields. A Web based interface has been developed integrating the main database, feedback forms, feedback database, and the corrected database of different levels. The block diagram of the integrated system is depicted below:

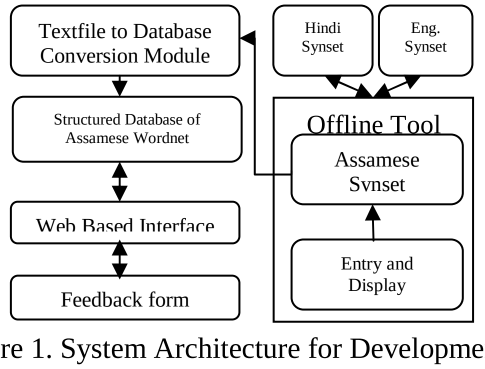

Figure 1. System Architecture for Development of Assamese Wordnet

6 Conclusions and Future work

Development of Wordnet is a collaborative work. Equal participation of academicians, researchers in fields like linguistics and computations are important to the building of a successful Wordnet. Such a structured and well-built Wordnet in a particular language carries the language to new heights. Attempt has been given in the present work to look at this angle so that a piece of well structured and valued finished product could be delivered. Language technology development in Assamese language signifies a lot, as Assamese is the most spoken language in the state of Assam, and also this is the official language of majority of the state population. Technology development in the language is the backbone for effective utilization and spreading of digital tools, appliances and applications to the mass people. An Assamese Wordnet will definitely openup the scope and opportunities for the language as an effective media for digital processing and communications. The work has been started with an integrated team comprising of lexicographers, computer professionals, and overseen by experienced linguistics, and language technologists. Although this is a part of the composite Indo Wordnet, but linguistically its a complete new work. The work has been integrated with fresh computational facilities like web-based feedback system, structured database, and also an automated rectification and correction system, with provision of different levels of modified and indexed databases. The target is to develop the complete Wordnet with 100% mapping of the present Hindi Wordnet with 12282 synsets, and also to add new concepts in the language specific space. The Assamese Wordnet so developed will definitely be a piece of technological importance which will lead to an extremely rich and useful lexical base facilitating automatic bi-lingual dictionary construction, machine translation between the Assamese and English, Hindi, and other Indian languages, Cross lingual information retrieval etc. The work will also produce user manuals and software in modules, which will ultimately have impact on social empowerment by IT, economic benefits, language learning, crossing of language barrier etc. An important secondary output of the whole exercise is in regards of manpower training and generation. Through this Wordnet development exercise a new breed of researchers in language technologies will be trained for proper skills and knowledge sets. As in Assamese, the linguistic and literature studies in formal education are with 0% computational linkage, and with no training/exposure for interlinking of linguistics and computing, the work will facilitate in developing a team of interdisciplinary researchers.

Acknowledgements

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

References

1. Banikanta Kakati. 2008. *Assamese: Its Formation and Development*, Lawyers Book Stall, Guwahati, Assam.
2. 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.
3. Dave Shachi and Bhattacharyya Pushpak. 2001. *Knowledge Extraction from Hindi Texts*, Journal of Institution of Electronic and Telecommunication Engineers, vol. 18, no. 4.
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.
5. Fellbaum, C. (ed.), 1998. *WordNet: An Electronic Lexical Database*,. The MIT Press.
6. Golock C Goswami. 1983. *Structure of Assamese*, Gauhati University, Assam.
7. 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.