Development of Assamese WordNet

Iftikar Hussain, Navanath Saharia and Utpal Sharma

Department of Computer Science & Engineering
Tezpur University, Napaam, India-784028
iftikar.it@gmail.com, nava.nath@yahoo.in, utpal@tezu.ernet.in

Abstract. An important challenge in the task of semantic analysis is the fact that in natural languages several words may denote the same concept, and a single word form may denote different concepts. WordNet is a repository of information about such characteristics of words, in a readily accessible format. A WordNet may be developed for a language for which computational processing is attempted. It calls for efforts in the domain of linguistics as well as computer science. In this paper we describe our works towards the development of WordNet for Assamese language, an official language of India. We cover proposed formats for the WordNet database text files and also an Application Programming Interface (API) of the Assamese WordNet.

1 Introduction

One of the main challenges in Natural Language Processing (NLP) is determining the appropriate sense of each word that occurs in input expressions. Words in natural languages often have multiple senses, and often several distinct words denote the same sense. WordNet helps to overcome such challenges. WordNet is a database that consists of words or collocations. Words having similar senses are grouped together and the groups are interconnected through some lexical and semantic relations. Users or their applications can make queries on it and can find out appropriate senses of words distinctly.

Assamese is one of the 22 official languages in India, spoken by nearly 30 million people. WordNets are being built for about thirteen of these official languages at different institutions. Hindi WordNet, developed at IIT Bombay is the first WordNet developed for an Indian language. Assamese is a morphologically rich, free word order Indic language, where very little computational work is reported, viz. [1, 2, ?,?,?]. Our work reported in this paper is one of the first efforts towards building an Assamese WordNet. Though work on building Assamese WordNet has been taken up in Gauhati University[3] recently, results thereof are still awaited.

The proposed architecture of our Assamese WordNet comprises a database and a graphical user interface (GUI). The WordNet database consists of text files which include the synonym set (synset) of a word and the sense of the word. Synsets are interconnected with other synsets via a number of lexical and semantic relations. The database consist of three text files namely index file, data

file and ontology file. Assamese linguistic data are maintained in the database files in some pre-defined format. An Application Programming Interface (API) is developed through which other applications can access the WordNet database for their purposes. In course of our work, we had to deal with several issues related to the support of Assamese script in the computer, such as encoding, text typing etc.

In the next section we describe the basics of WordNet. Section 3 explores some existing work on WordNet and section 4 discusses issues faced during implementation. In section 5 we discuss our proposed architecture for Assamese WordNet and current status. Section 6 concludes this paper indicating future directions.

2 WordNet

WordNet is a repository of words of a language. The words are grouped together according to their similarity of meanings. For each word there is a synonym set called *synset*, representing one lexical relation. For each synset there is another element called *gloss* that describes the concept. Synsets in the WordNet are connected to other synsets via a number of lexical and semantic relations. Each entry of the wordNet[1] consists of following elements

1. **Synset :** Words in a synset are arranged according to the frequency of usage.
   Synset: { মৌ, মধু, মকৰন্দ};
   TF[2]: {*mou, modhu, makaranda*};
   EM[3]: Honey.
2. **Gloss :** It describes the concepts. It consists of two parts
   (a) **Text Definition:**It explains concepts denoted by the synset. Example-
   ফুলৰ মিঠা ৰস;
   TF: *phulor mithA ras*;
   EM: Flower's honey.
   (b) **Example Sentence:**It gives the usage of the words in the sentence. Example-
   মৌ মাখিয়ে ফুলৰ পৰা মৌ গোটাই;
   TF: *mou mAkhiye phulor porA mou gotAi* ;
   EM: Honey bee collects honey from flowers.;
3. **Position in Ontology :** An ontology is a hierarchical organization of concepts, more specifically, a categorization of entities and actions. For each syntactic category namely noun, verb, adjective and adverb, a separate

[1] Hindi WordNet Data and Associated Software License Agreement. The IIT-Bombay represented for the purpose of the signature of this agreement by the Dean of Research and Development, IIT Bombay or by his authorized representative Dr. Pushpak Bhattacharyya, Professor of the Department of Computer Science and Engineering, IIT Bombay.

[2] TF: Transliterated Assamese Form.

[3] EM: English Meaning (Concept).

ontological hierarchy is present. Each synset is mapped into some place in the ontology. A synset may have multiple parents. The ontology for the synset representing the concept *school* is shown in figure 1. Example-

Synset = { স্কুল, বিদ্যালয়, পাঠশালা };
TF: {*skul, bidyAlay, pAthchAlA*};
EM: School

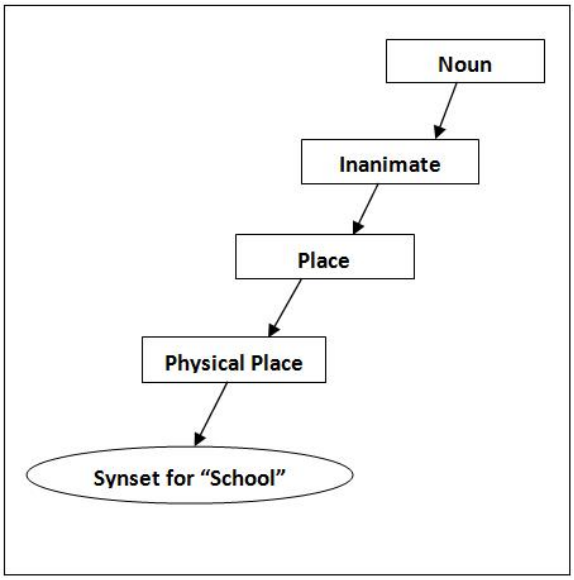

Fig. 1. Ontology for Synset for "School "

3 Related Work

The Princeton English WordNet[4] was developed by Professor G. A. Miller at the Cognitive Sciences Laboratory at Princeton University. A variety of lexical and semantic relations are used to represent the organization of lexical knowledge. Inputs provided through text files written by lexicographers are converted to database files. Two kinds of building blocks are distinguished in the source files: word forms and word meanings. There are separate files corresponding to each syntactic category such as noun, verb, adjective and adverb. All of the synsets in a lexicographer file are in the same syntactic category. For each syntactic category, two files are needed to represent the contents of the WordNet database - index.pos and data.pos, where pos is noun, verb, adj and adv. Each index file is an alphabetized list of all the words found in WordNet in the corresponding part

[4] http://wordnet.princeton.edu

of speech. A data file for a syntactic category contains information corresponding to the synsets that were specified in the lexicographer files.

The creation of WordNet is a time consuming and manpower intensive exercise. The effort can be reduced to some extent by using text repositories such as the web and certain corpora, and also by translating an existing WordNet into another language. But results of such attempt are often far from ideal, in the sense that the WordNet so produced contains synsets that have outlier words and/or missing words. Additionally, semantic relations may be inappropriately set up or may be missing altogether. [4] reported an automatic method of WordNet evaluation for the first time. They focused on verifying synonymy within non-singleton synsets and also on hypernymy between synsets. They made some rule based algorithm to validate the synonyms and hypernyms. The synonym validation was tested on the Princeton WordNet (v2.1) noun synsets. Out of the 81426 noun synsets, 39840 are synsets with more than one word, and only these were given as input to the validator. The result gave 70% accuracy where all words in synsets were validated, approximately 90% where half were validated and about 9% where no words were validated. The Hypernym validation algorithm was able to validate 56203 out of 79297 noun hypernymy relation pairs in the Princeton WordNet, giving a validation percentage of 70.88%. The validation algorithm is available only for Princeton WordNet. However, the approach should broadly be applicable to other language WordNets as well.

Hindi WordNet is the first WordNet developed for an Indian language. It is developed at CFILT, IIT Bombay. Among other Indian languages WordNets for, Marathi, Bengali, Nepali, Oriya, Telugu, Malayalam, Konkani, Kashmiri, Manipuri etc. are being developed at different Indian institutions. The NE WordNet[3] covering Assamese and Bodo languages is being developed at Gauhati University.

4 Assamese WordNet

Assamese WordNet is a database of Assamese word forms (words and collocations) which are grouped together in the form of synsets. The synsets are interconnected to other synsets via a number of lexical and semantic relations such as hypernym and hyponym (the **is-a** relation), meronym and holonym (the **part-of** relation), antonyms etc. The lexical relationships hold between semantically related forms of words and the semantic relationships hold between related word definitions. The subgraph of a WordNet holding different relationships are shown in figure 2 Relations between the synsets in the Assamese WordNet are described below.

– *Hyponym and hypernym (is a kind of)*: Hypernymy is a semantic relation between two synsets to capture super-set hood. Similarly, hyponymy is a semantic relation between two synsets to capture sub-set hood. Example:
{ গেন্ধাই-ফুল, নার্জি-ফুল, গেন্দা-ফুল } $=>^{hr}$ { ফুল, পুষ্প, কুসুম };
TF: { *gendhAi_phul, nArji_phul, gendA_phul* } $=>^{hr}$ {*phul, puspa,kusum* };

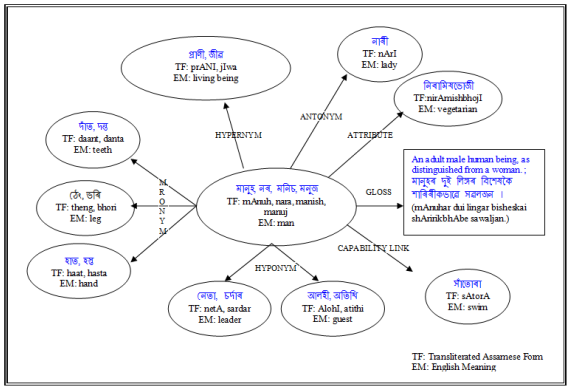

Fig. 2. Subgraph of a WordNet

EM: (name of a flower) $=>^{hr}$ (flower).
ফুল *(phul)* is Hypernym of গেন্ধাই_ফুল *(gendhai_phul)*.
গেন্ধাই_ফুল *(gendhai_phul)* is a Hyponym of ফুল *(phul)*.

– *Meronym and Holonym (part-whole relation)*: It is a semantic relation between two synsets. If the concepts A and B are related in such a manner that A is one of the constituent of B, then A is the meronym of B and B is the holonym of A. The meronymy relation is transitive and asymmetrical. Holonymy is the reverse of meronymy. It is used to construct a part-of hierarchy. Example:
{ কোঠালি, কোঠা, কক্ষ} $=>^{mh}$ { ঘৰ, বাস-ভৱন, গৃহ };
TF: {*kothAli, kothA, kakkhya*} $=>^{mh}$ { *ghor, bas-bhawan, griha*};
EM: (room) $=>^{mh}$ (house).
Here কোঠালি (*kothali*) is a part of ঘৰ (*ghor*). Therefore, কোঠালি (*kothali*) is a Meronym of ঘৰ(*ghor*) and ঘৰ is a Holonym of কোঠালি.

– *Entailment:* Entailment refers to a relationship between two verbs. Any verb A entails B, if the truth of B follows logically from the truth of A. The relation of entailment is a one way relation. Example:
{ সূর্য্য_উদয়_হোৱা, কুকুৰাই_ডাক_দিয়া } $=>^{en}$ { পোহৰ_হোৱা};
TF: { *surjya_uday_howA, kukurAi_dAk_diyA*} $=>^{en}$ { *pohar_howA* };
EM: (sunrise) $=>^{en}$ (morning).
The concept of { সূর্য্য_উদয়_হোৱা, কুকুৰাই_ডাক_দিয়া } logically means { পোহৰ_হোৱা}.

– *Troponymy:* Troponym denotes a specific manner elaboration of another verb. It shows manner of an action, i.e., X is a troponym of Y if to X is to Y in some manner. Example:
{ বক্তৃতা_দিয়া, ভাষণ_দিয়া } $=>^{tp}$ { কথা_কোৱা, বিৱৰণ_দিয়া, উক্তি_দিয়া };
TF: { *baktritA_diyA, bhAsan_diyA*} $=>^{tp}$ {*kathAkowA, bibwaran_diyA, ukti_diyA*};
EM: (speech) $=>^{tp}$ (explain).

– *Antonymy :*Antonymy is a relation that holds between two words that (in a given context) express opposite meanings. It is a lexical relation as it holds between two words and not the entire synset. Example:
{ সচা, সত্য, প্রকৃত, যথার্থ} $=>^{an}$ { মিছা, অসত্য, অপ্রকৃত, নিস্ফল,অর্থহীন, অসাৰ };
TF: {*sochA, satya, prakrita, jathArtha*} $=>^{an}$ { *misA, asatya, aprakrita, nisfal, arthaheen, asAr*};
EM: (truth) $=>^{an}$ (false).
– *Gradation:* Gradation is a lexical relation that represents the intermediate concept between two opposite concepts. Example:
ল'ৰালি [TF: *lorAli*; EM: (childhood)] and বৃদ্ধা [TF: *briddha*; EM: (old age)] are two opposite concepts. Here যৌৱন [TF: *jouwan*; EM: (young age)] will be the intermediate concept between these two.
– *Causative:* In some languages (Hindi/Assamese) there is a convention of forming causation by making morphological change in the base verb. The Causative relation links the causative verbs and the base verbs and show interdependency between them. Example-
কন্দা $=>^{ca}$ কন্দুওৱা ;
TF: {*kandA*} $=>^{ca}$ { *kanduowA* };
EM: (cry) $=>^{ca}$ (to make someone to cry).
কন্দুওৱা (*kanduowA*) is a causative verb of কন্দা (*kandA*).

4.1 Relations between Synsets of different Part-Of-Speech

Relations may hold between synsets of different parts-of-speech, too, as listed below. Except the relation '*Modifies Noun*', which is a lexical relation, the rest are semantic relations.

Nominal and Verbal Concept :

1. *Ability Link :* This link specifies the inherited features of a nominal concept.
Example: { মাছ, মৎস্য, মীন } $=>^{al}$ { সাঁতোৰা, সন্তৰন_কৰা};
TF: { *mAsh, matsya, mIn*} $=>^{al}$ {*sAtorA, santaran_kara*};
EM: (fish) $=>^{al}$ (swim).
2. *Capability Link :* This link specifies the acquired features of a nominal concept.
Example: { মানুহ, মনিচ, মণুষ্য} $=>^{cl}$ { সাঁতোৰা, সন্তৰন_কৰা};
TF: { *mAnuh, manish, manusya* } $=>^{cl}$ {*sAtorA, santaran_karA*};
EM: (man) $=>^{cl}$ (swim).
3. *Function Link :*This link specifies the function of a nominal concept.
Example: { শিক্ষক, অধ্যাপক, আচার্য্য, গুৰু, মাষ্টৰ} $=>^{fl}$ { পঢ়ুওৱা, শিক্ষা_দিয়া};
TF: { *sikshak, adhyApak, Acharya, guru, mAstar* } $=>^{fl}$ { *parhuowA, sikshA_diyA* };
EM: (Teacher) $=>^{fl}$ (Teaching).

Nominal and Adjectival Concept :

1. *Attribute :*This denotes the properties of noun. It is a linkage between noun and an adjective.

Example: { গৰু, গো} =>^{at} { নোমাল};
TF: {*garu, go*} =>^{at} {*nomAl*};
EM: (cow) =>^{at} (hursute).

2. *Modifies Noun :*Certain adjectives can only modify certain nouns. Such adjectives and nouns are linked in the Hindi WordNet by the relation Modifies Noun.
Example:{ দয়ালু, মৰমিয়াল} =>^{mn} { মানুহ, জন, মনুষ্য, মনিচ };
TF: { *dayAlu, maramiAl*} =>^{mn} {*mAnuh, jana, manushya, manish*};
EM: (kind) =>^{mn} (man).

4.2 Issues in Assamese WordNet Development

There are some issues in implementing a WordNet for Assamese language.

- The creation of a WordNet is a time consuming and manpower intensive exercise. Manpower such as both linguist and computer persons are required.
- Adapting WordNet to a domain[5]. Usually WordNet contains a large amount of data covering many domains. Sometimes this may be create difficulties an application to efficiently access such a large database.
- Updating the WordNet database is very complicated and requires experts.
- Identification of different relations between the synsets is difficult.
- Handling the morphology of Assamese words is non-trivial.
- Different corpora used for such an exercise vary in the encoding scheme used. Appropriate software is required to convert these into a common encoding scheme such as Unicode [5]
- Use of Unicode for Assamese text needs appropriate system set-up including availability of Assamese fonts. Moreover, implementation of a convenient user interface requires programming environment set-up.
- Typing Assamese text for user queries and for database entries requires appropriate skill. We have developed a virtual keyboard for Assamese for the purpose (see figure 3).
- Data collected for WordNet database requires certain pre-processing. For instance, we have used some programs to extract tagged words from electronic dictionaries.

A large amount of effort in this work goes into addressing the above issues, some of which are not apparently specific to WordNet.

5 Architecture and Implementation

Our implementation of Assamese WordNet is based on the idea of construction of Hindi WordNet[6]. Most of the Indian language WordNets follow the expansion principle, where Hindi WordNet database is borrowed or expanded to get the

[5] http://www.tezu.ernet.in/∼nlp/r2u.htm
[6] http://www.cfilt.iitb.ac.in/

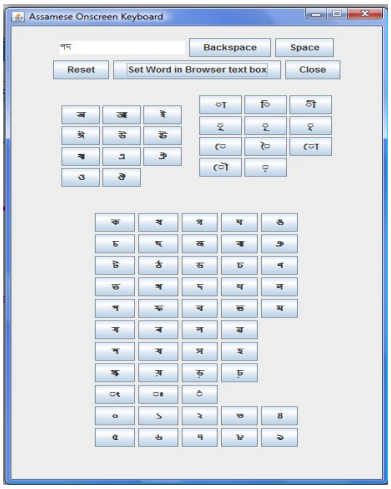

Fig. 3. Assamese Keyboard Interface

database of its own language. We have created the database by entering data word by word. For this we have implemented a simple data entry interface, so that a normal user can insert data into the database conveniently. We have also created a browser through which users can retrieve different information about a word from the database. An Application Programming Interface (API) is also developed through which other applications can use the WordNet database. The architecture of the Assamese WordNet (Figure 4) with detailed description of each unit is given below.

5.1 WordNet Database

The core of a WordNet is the database of words. The Assamese WordNet database is of text files. There are mainly three files in the database.

– **Index File**
It contains detailed information of every word in the WordNet database. Each line of the file is of the format given below.
word pos r_cnt r_type [r_type] s_cnt synset_offset [synset_offset]
Where,
- word - Unicode text of word or collocation.
- r_cnt - number of relations in the synsets containing the word.
- r_type - Type of the relations in the synsets containing the word. There are r_cnt numbers of relations.
- s_cnt - Number of synsets containing this word i.e., number of different senses of the word.
- synset_offset - Byte offset of the synset in the data file containing the word. Each index file entry contains *s_cnt* number of synset offsets.

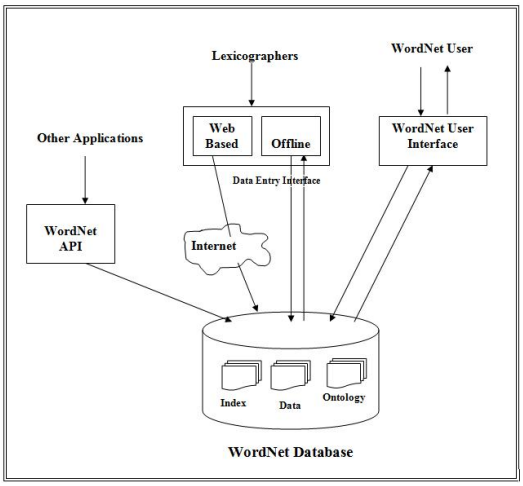

Fig. 4. Architecture of Assamese WordNet

For example: ফুল **01 01 12 03 00000007 00000008 00000009**
TF: phul 01 01 12 03 00000007 00000008 00000009
EM: (Flower).
Here, ফুল is the word, 01 means the type of the word i.e., noun. Next 01 indicates that it has only one relation with other synsets and the synset containing it. *12* indicates the type of the relation i.e., Hyponym. *03* indicates that the word is in three synsets in the data file. 00000007, 00000008 and 00000009 are the three synset offsets of the synset in the data file containing that word.
Index file is used for fast retrieval of data.

– **Data File**
It contains the synsets and glosses along with different relationships. Each entry is of the form as follows.
synset_offset pos w_cnt word[:word] r_cnt relations [relations] | gloss
Where,

- synset_offset - Byte offset of the synset in the data file. It can be considered as synset ID also.
- pos - Part Of Speech of the synset. 01 for noun, 02 for Adjective, 03 for Verb and 04 for Adverb.
- w_cnt- Number of words present in the synset.
- word - words or collocation in the synset.
- r_cnt - Two digit decimal number indicating the number of relations from this synset to other synsets.
- relations - relations are of the form
  r_type [words_loc] target_offset
  Where,

    - r_type - Two digit decimal number indicating the type of relation.
    - words_loc - It's a four digits decimal number where the first 2 digits indicate the source word location in the synset and the last two digit represents the target word location of that synset. This is optional and used for lexical relations such as Antonyms only.
    - target_offset- Byte offset of the synset with which the current synset is related.
  - gloss - Describes the senses of the words. It is multilingual i.e., the senses are described in Assamese as well as English.

  For example: **00000006 01 04** মৌ:মধু:মকৰন্দ **01 22 00000007** | ফুলৰ মিঠা ৰস; এজাতি মাখিয়ে বাহত গোটাই ৰখা ফুলৰ ৰস;
  TF: 00000006 01 04 mou:madhu:makaranda 01 22 00000007 | phular mithA ras ; ejAti makhiye bAhat gotAi rakhA phular ras ;
  Here,
  00000006 is the synset offset that distinguishes the synset. 01 indicates that the synset is of type noun.*04* indicates that the synset contains four words. *01* indicates that there is only one relation associated with this synset. *22* indicates the relation type and 00000007 is the synset offset of the relation to which this synset is related. The symbol "|" divides the gloss and the other things. "ফুলৰ মিঠা ৰস; এজাতি মাখয়িে বাহত গোতাই ৰখা ফুলৰ ৰস;" is called the gloss that describe the meaning or concept of the synset.

– **Ontology**

  Ontology is a hierarchical organization of concepts. The file format is as follows.

  offset level up_offset | category/subcategory example

  Where,

  - offset - Byte offset or ID of the category/subcategory
  - level - 0000 is the top level and 0001 is the lower level.
  - up_offset - offset of the upper level category or super category i.e., it is the subcategory of the up_offset.
  - category/subcategory - These are written in Unicode text. For each category there are subcategories. For each subcategory there are sub-subcategories and so on.
  - example - an example of words of that category.

  A sample of the entries of the file is mentioned here.
  00000001 0000 | TOP {Top Level Node}
  00000002 0001 00000001 | বিশেষ্য (Noun) {N উদাহৰন :- গৰু, গাখীৰ, কল ইত্যাদি }
  00000003 0001 00000002 | ব্যক্তিবাচক বিশেষ্য (Proper Noun) {PROP উদাহৰন :- তেজপুৰ, হিমালয়, গংগা ইত্যাদি}
  TF: 00000001 0000 | TOP {Top Level Node}
  00000002 0001 00000001 | biseshya (Noun) {N udAharan:- garu, gAkhIr kal ityAdi}
  00000003 0001 00000002 | byaktibAsak biseshya (Proper Noun) {PROP udAharan:- tezpur, himAlay, gangA ityAdi}

Here the first field is the offset that works as a unique ID, 00000001 is the top level of the hierarchy. The third field indicates which level belongs to which in the hierarchy. After the symbol '|' the first field indicates the class of the level and then there are some example words of that level.

5.2 WordNet User Interface

With the help of the WordNet User Interface a normal user can access the WordNet database. The interface contains one text field where a user can put their query word or collocation. There is a button named 'search' which may be clicked to search different senses of the word. Different synsets are retrieved from the database and these are displayed in another text area of the interface. Then a user can search for different relationships associated with the synsets containing the word. A user can also save the retrieved information into files for their use. The user interface also provides an Assamese keyboard (as shown in figure 3) through which a user can type Assamese words into the text box for searching.

5.3 Data Entry Interface

The data entry interface is used to insert data into the WordNet database. This interface is also called as lexicographers' interface as it is usually used by lexicographers to insert data. There are two types of interfaces -

1. Web based Data Entry Form
2. Offline Data Entry Form

With the help of the web based data entry interface, a remote user can suggest or insert new synsets which are again modified at the server side and then inserted into the WordNet database. Through offline data entry interface the synsets are inserted to the database directly. This interface also facilitates modification of the existing entries of the database.

5.4 WordNet API

An Application Programming Interface (API) is a set of routines, protocols and tools for building software applications that can directly access or use the WordNet database.

5.5 Current Status of the Assamese WordNet

We have collected data from different sources to build the Assamese WordNet-

– Online Dictionary Projects [7] [8]

[7] http://www.xobdo.net
[8] http://dsal.uchicago.edu/dictionaries/candrakanta

– Asomiya Jatia Abhidhan[9] - A unilingual, comprehensive, scientific and encyclopedic Assamese national dictionary.
– Samartha Sabdakosh[6] - An Assamese Thesaurus.
– CIIL-EMILLE Corpus
– Assamese Pratidin Corpus[7]

The dictionaries listed above have helped identifying some relationships between synsets. Presently, our Assamese WordNet contains about 2000 unique Assamese words grouped in more than 300 synsets. With the entry of new synsets by lexicographers it can be enriched in near future. A web based Data entry form has been provided via the Internet through which a user can suggest synsets for the database.

6 Conclusion and Future Work

A WordNet is a significant resource for NLP. Building a WordNet is takes a lot of effort. Our efforts have mainly been in creation of the technology and filling in the initial data. At present our Assamese WordNet includes about 2000 distinct words grouped into about 300 synsets. With the set-up that we have been able to put in place, the Assamese WordNet can grow further with inputs. Followings are some directions to carry on the work in future.

– No attempt is made for handling the problem of entering a sub-synset or super-synset of a synset which is already present in the database.
– Ontology hierarchies for the synsets are not included in this project as no data are inserted into the ontology file. But the file format of the ontology file is defined. So, according to that file format data can be inserted on it in future.
– The Assamese WordNet technology we have created can be considered as a general technique and this can be utilized to develop WordNets for other languages, particularly those languages for which WordNet work has not yet been started. This can facilitate the creation of multilingual IndoWordNet among all Indian languages.
– Assamese language is a morphology rich language. The Assamese WordNet requires handling different morphological analysis. One can work for handling morphology which will give better options to the WordNet users.

Our work is a very significant step for the cause of Assamese language in particular, and likely to be useful in the IndoWordNet initiative in future.

References

1. Das, M., Borgohain, S., Gogoi, J., Nair, S.B.: Design and Implementation of a Spell Checker for Assamese. In: Language Engineering Conference (LEC'02). (2002)

[9] http://www.nationaldictionary.org

2. Sharma, U., Kalita, J., Das, R.: Unsupervised learning of morphology for building lexicon for a highly inflectional language. In: Proceedings of the ACL-02 workshop on Morphological and phonological learning. (2002)
3. Sarma, S.K., Gogoi, M., Medhi, R., Saikia, U.: Foundation and structure of developing an assamese wordnet. In: Proceedings of 5th International Conference of the Global WordNet Association (GWC-2010), Department of Computer Science, Gauhati University (2010)
4. Nadig, R., Ramanand, J., Bhattacharyya, P.: Automatic Evaluation of Wordnet Synonyms and Hypernyms. In: International Conference on NLP (ICON08), Pune, India. (2008)
5. Jing, H.: Usage of WordNet in Natural Language Generation. In: Proceedings of the Joint 17th International Conference on Computational Linguistics 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98) workshop on Usage of WordNet in Natural Language Processing Systems, Universite de Montreal, Quebec, Canada, Department of Computer Science; Columbia University (1998)
6. Kotoky, P.: Samartha Sabdakosh. Jyoti Prakashan, 2nd Ed. (2007)
7. Sharma, U., Kalita, J.K., Das, R.K.: Acquisition of morphology of an Indic language from text corpus. ACM Transactions on Asian Language Information Processing (TALIP) (2008)