Building Manipuri-English Machine Readable Dictionary by Implementing Ontology

Shantikumar Ningombam*
Research Scholar, Department of Computer Science,
Assam University, Silchar-788011

Sagolsem Poireiton Meitei
Research Scholar, Department of Computer Science,
Assam University, Silchar-788011

Dr. Bipul Syam Purkayastha
Professor, Department of Computer Science,
Assam University, Silchar-788011

sk_ningomba@rediffmail.com[1], poireimeitei@gmail.com[2], bipul_sh@hotmail.com[3].

Abstract:

Any system that hopes to process natural languages as people do must have information about words, their meaning, concept, relative words in another language and meaningful sentences are composed of meaningful words. Traditionally information is provided through electronic dictionaries. But these dictionary entries evolved for the convenience of human readers, not for machines. So, machine readable electronic dictionary becomes the central resources for Natural Language applications. Dictionaries and other lexical resources are not yet widely available in electronic form for Manipuri language. And there is no Manipuri-English machine readable dictionary that can provide both of lexical resources and conceptual information. This paper describes the process for developing Manipuri-English dictionary by implementing ontology. This implementation should provide a more effective combination of traditional Manipuri-English bilingual lexicographic information and their conceptual information.

Keywords: Ontology, Lexicon, Meitei, Taxonomic.

1. INTRODUCTION

The ultimate goal of Natural Language Processing is to understand the language. The meaning of a sentence in natural language is derived from the meanings of its individual words together with its syntactic structure and surrounding context. So, to know the words is an extremely important part of knowing a language. Lexicons are storehouses of such information. But traditional bilingual lexicon just only provides the lexical resources of the words, not the conceptual meaning. Whereas, ontology can offer a way to address concepts, relational and meaning of words required for natural language processing. So, the combination of bilingual lexicon with the structure of ontology will produce not only the lexical resources but also their conceptual meaning (senses). As my knowledge, there is no ontology based Manipuri-English machine-readable bilingual dictionary in our country. To construct the Manipuri-English bilingual lexicon by implementing ontology that can provide both of lexical information and their concepts. So, this implementation will provide as a central promising resource for Manipuri-English language translation and further NLP research.

2. RELATED WORK

The main reasons to use ontology in machine translation (MT) are to enable source language analysers and target language generators to share knowledge, to store semantic constraints, and to resolve semantic ambiguities by making inferences with the concept network of the ontology [4]. In natural language, words have different meanings in different contexts are said to be ambiguous. To better facilitate Word Sense Disambiguation (WSD) for machine translation, ontology-based lexicon is required [3]. Segmenting sentences into words is a challenging task in Manipuri Language and needs to exploit stored word lists. So, the implementation of Manipuri-English electronic dictionary becomes an essential knowledge source for future Manipuri NLP researches such as WSD, Segmentation of words, Spell Checking, and other MT tasks, although it is very difficult and time consuming to define its concept and their relations.

3. MANIPURI LANGUAGE

The Manipuri Language belongs to Kuki-Chin group of Sino-Tibetan family of languages. As per the earliest written evidence on the history of Manipuri, the language dates back to the 11th century. Manipuri Language is originally known as Meitei. It is the official language of the south eastern Himalayan state of Manipur in India. There are about 1,500,000 people in the world which speak Manipuri. There are at least 33 different languages spoken in Manipur. It is also spoke in Assam, Tripura, Mizoram, Burma and Bangladesh. It is the Lingua Franca in the state. This is also the language used by the offices and government institutions in the state of Manipur. It has been recognized as a viii scheduled language by the Indian Union and is taught up to master and going on many esearch in various aspects of Manipuri for the degree of Ph.D. level in the Indian Universities. It is also the medium of education in Manipuri Schools.

3.1. Analysis of Manipuri Word

The analysis of Manipuri vocabularies, which are related to English vocabularies and how they relate each others are, perform in this phase. In Manipuri vocabularies, words can be polysemous (A word that has more than one sense in common) and synonymous (two or more words that share at least one sense in common). Here are some examples of polysemous and synonymous Manipuri words as shown in Table [1].

Table1. Examples of Polysemous and Synonymous Manipuri Words

	Manipuri Word	POS	Concept (Class)	English Def:
Polysemous	Ləi	Noun	Human/Animal	Tongue
	Ləi	Noun	Plant	Flower
Synonymous	Lem	Verb	Quality	Handsome/Decent
	P^{h}ə.jə	Verb	Quality	Handsome/Beauty

Words in Manipuri language can be divided into: (1) Simple Words, and (2) Compound Words. Simple words consisting of one or two root and the combination of two or more roots become Compound word. Ontology can support the information to identify the given word is simple or compound and if it is compound word, which types of Part of Speech are being composed to form this word. For example, the Manipuri Word, "***potpumən***→ pot-pubə-məmən" is the type of Compound Word and it is composed of "free root + bound root + free root" combination. So, these providing features can be partially solved Word Sense Disambiguation problem in Word Segmentation of Manipuri Sentence.

4. CONSTRUCTION OF THE ONTOLOGY

By using OWL (Web Ontology Language) and Protégé 3.2 beta software which is the editor of OWL can store the words. Other NLP applications (such as Spelling Checker, Machine Translation) can also be used for ontology as storage of word list. Unicode format of Manipuri word can store using "BN-TTDurga, BN-TTBedisha and AS-TTBidisha". The construction of the ontology involves implementing Manipuri-English lexicon.

4.1. Taxonomic Structure of the Ontology

Regardless of the language, the knowledge in the discourse universe is conventionally divided in two classes: conceptual and linguistic. Terms and sentences refer to concepts, but they have particular structural and morphological features in each language. In Manipuri language, each vocabulary can also be represented as two types (semantically and syntactically). It can define the subclasses for each POS semantically or conceptually. And it also defines the subclasses for Word Structure, to provide syntactic or morphological information for each POS (Noun and Verb).

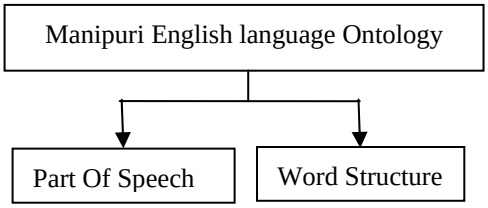

Figure1:- Top-level annotation

4.1.1 Part of Speech (POS)

The class "POS" is used to represent all Manipuri vocabularies. It contains two main classes as shown in Figure [2].

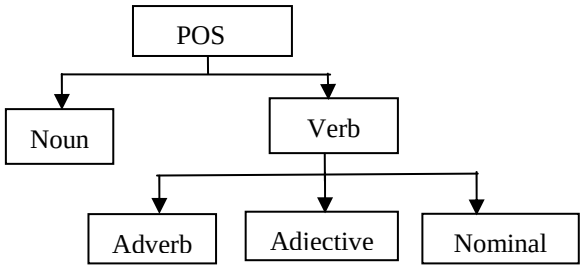

Figure2:- Annotation of POS Class

To developed sub-classes for Noun more semantically, the taxonomic structure of Noun class is represented by the psychological concepts. For example, one of the "Noun" class "***ləi.sa.bi*** → Matured _ Feminine _Human" is derived from the taxonomic structure like this; Matured _Feminine _Human →Vertebrate →Human →Living Things →Physical _Entity →Entity →Noun. Where, → represent "is-a" relation.

"Verb" class contains three subclasses:

(1) Stative _Verb
(2) Action _Verb
(3) Process _Verb

"Adverb" contains five subclasses:

(1) Conditional _Adverb
(2) Interrogative _Adverb
(3) Manner _Adverb
(4) Time _Adverb
(5) Quantitative_ Adverb

"Adjective" In Manipuri an adjective derived through the affixation of the attribute prefix or formative particle (FP) **ə-** to a verbal noun. There are some adjectives which are without the attribute **ə-** .Adjective can form by adding the nominalizer ***pə- bə*** to the verb in progressive and perfect. In Manipuri language adjective can be derived from any verbal root. Each POS is related to each other under the relation, Synonyms and Antonyms.

4.1.2. Word Structure

There are two main classes for "Word_ Structure":

(1) Simple Word
(2) Compound Word

Table [2] represents the subclasses for Simple Word classes. "Compound Word" may be composed of the alternative combination of Noun, Verb and Adjective. There are six combinations for "Com-pound Noun" and two combinations for "Compound Verb". These combinations are the instances of those classes respectively. Table [3] and Table [4] will represent these combinations and some examples of Manipuri Compound Noun and Verb.

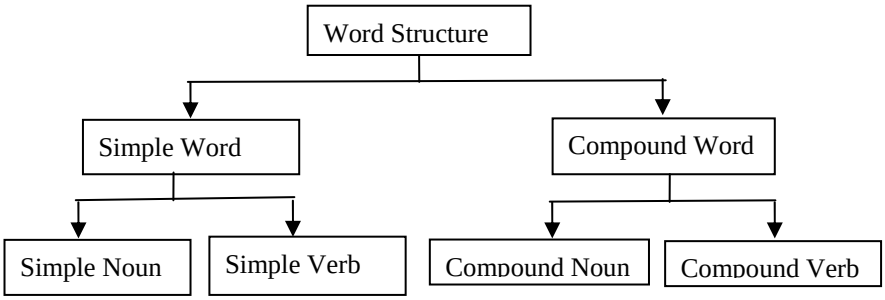

Figure3. Annotation of Word Structure Class

Table2. Some examples of Simple Noun and Verb

Simple Noun		**Examples**
1	Original Noun	***mi*** 'man', ***tha*** 'moon', ***nupi*** 'woman', ***layrik*** 'book'
2	Derived Noun	***məpok*** 'birth', ***pabə*** 'reading', ***thəkpə*** 'drinking'
Simple Verb		***i-*** 'write',***ca-*** 'eat', ***lai-*** 'be easy/ simple'

Table3.Six combinations of Compound Noun

No	**Combinations of Compound Noun**	**Examples**
1	Noun+Noun	***məyphəm*** (məy + phəm) 'fire place'
2	Noun + verbal -Noun	***usubə*** (u+ su-bə) 'carpenter'
3	Adjective +Noun	***ləmpak***(əpakpə+ləm) 'playing ground'
4	Noun +A-ObRoot +Noun	***mithuŋsəŋ***(mi+thuŋ+ səŋ) 'accomodation'
5	Noun+Adjective	***yathək***(ya+thək) 'upper teeth'
6	Noun+A-O bRoot	***ləmkoy***(ləm+koy) 'travel'

Table4. Two combinations of Compound Verb

No	**Combinations of Compound verb**	**Examples**
1	Noun +Verb	***sindəm*** (sin + təmbə), ***ləllon***- (ləl + onbə)
2	Verb +Verb	***satkai***-(sat- +kai-) 'be in (full)bloom', ***koidanə***-(koi- +ta-)'be round (especially of face)'

4.2. The Lexical Entries (Data Type Properties)

Each lexical entry are required to represents a distinct sense of word and contains: (1) a word form in Manipuri (2) Part of Speech (3) A particular concept of the word (4) Pronunciation of Manipuri word (5) Equivalent definition list in English and (6) Example usage.

4.3. The Relation (Object Properties)

Relations of lexical entries can be specifies to link various concepts and instances. This ontology contain "is-a" relations for all part of speech because concepts are organized into taxonomic structure. To create a link between two synonymous lexical entries by applying the relations "hasSynonym _Noun", "has-Synonym _Verb", "hasSynonym _Adj" and "has-Synonym _Adv" for Nouns, Verbs, Adjectives and Adverbs respectively. And it also create the relation "has Antonym _Noun", "hasAntonym _Verb", "has Antonym _Adj" or "has Antonym _Adv" between two opposite lexical entries. For example, Manipuri verb '***p^hə.jə***' (instance of the "Verb" class) and another Manipuri Adj '***t^hibə***'; (which is also instance of "Adjective" class) have "as Antonymy" relation. Word-structure information for the given term in the lexicon can provided through "hasNoun -StructureType" and "hasVerbStructure _Type" relations, in which, the link "Noun" to "SimpleNoun" or "CompoundNoun" class and the link between "Verb" to "SimpleVerb" or "CompoundVerb" class. For example, Manipuri Noun, ***'məhəytəmphəmsəŋ'*** is composed of 'məhəy' (Noun), 'təm' (Verb), 'phəm'(Noun) and 'səŋ' (Noun). So, this instance has a link to the "CmpNoun _NounVerbNounNoun", which is one of instance of the class "CompoundNoun". The representation of lexical entries, relations and taxonomic structure for Manipuri Noun "***məhəytəmphəmsəŋ***" is shown in Figure [4].

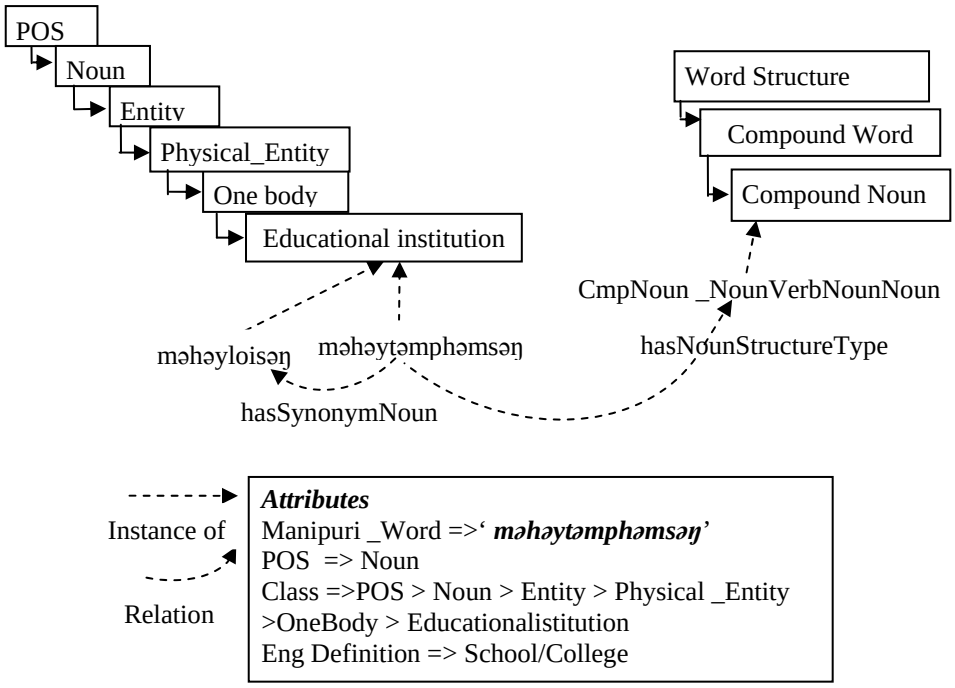

Figure4. Example representation of a Noun

5. IMPLEMENTATION OF MANIPURI-ENGLISH DICTIONARY

To developed ontology based Manipuri-English lexicon as a central resource for building user readable Manipuri-English electronic dictionary. It can contain the two main phases; (1) Look Up and (2) Data Entry. "Look Up" portion is any user enable to access. This portion is very useful as an electronic dictionary to search for the information of Manipuri vocabularies. The objective is to exploit the lexical resources from the ontology. Retrieving appropriate English definition list and lexical information such as pronunciation, synonym and antonym words and component structure for the given Manipuri word. The other portion, "Data Entry" can only be used by Administrator who has linguistic knowledge and can enter data and save by entering the correct password. All the necessary information such as the word represented in Manipuri language, its pronunciation, part-of-speech, concept hierarchy, related definition or definition list in English and its example usage. Moreover, their object properties, relation, can be done in this portion. By using Data Entry insert the vocabularies into the ontology. To create Synonym/ Antonym relation, can be done when two instances explicitly which have already exist. For developing the interface we used Java and Jess (Java Expert System Shell), which can map Java and OWL. Detail architecture of the system implementation is described in the following algorithms.

ALGORITHM: *LOOK UP*

(1) – Input the Manipuri Word, Wm, you would like to search.
(2) – Find Wm in the ontology
- if (found)
-choose specific class name, C
- retrieve and show all the information of the instance, Im, from the ontology, where its' class name is C
- if (Im has the "Synonym" relation with Is), retrieve and show Ws of Is.

- if (Im has the "Antonym" relation with Ia), retrieve and show Wa of Ia.
-else return "Not Found!" information to user.

Where,
Wm =Manipuri Word that you want to search
C =the specific class name
Im =Instance name for Wm

ALGORITHM: *DATA ENTRY*

(1) -Enter the required lexical information for Wm
(2) -Creating an instance Im
-if (!the mandatory information are contained) then, go to stage (1)
- if(the synonym fields contain Ws)
- find the word Ws in the ontology
- if (found)
-retrieve the instance name, Is, of Ws.
-Saving the data and object properties of Im, where, "Synonym" relation (object property) of Im is Is
- else
-enter the required information to create the new instance, Is, for word Ws
-Saving the data and object properties of Is in the ontology
-Saving the data and object properties of Im, where, "Synonym" relation (object property) of Im is Is.
-Saving the "Synonym" relation for Is to Im.
- else
-Saving the data and object properties of Im in the ontology.
Where,
Wm = Manipuri Word (main), data type property of Im
Im = Instance (main)
Ws = Manipuri Word (synonym), data type property of Is
Is = Instance (synonym)

ALGORITHM: *CREATE SYN/ANT RELATION*

(1) -Enter W*1*, W2, POS and Class name as Input
(2) -Find W*1* and W2 in the ontology
- if (found)
-Saving the "Synonym" relation from I*1* to I2
-Saving the "Synonym" relation from I2 to I*1*
-else Display "Not Found" message
Where,
W*1* = Manipuri Word *1*
W2 = Manipuri Word 2
I*1* = Instance *1*
I2 = Instance 2

6. CONCLUSIONS

Development of electronic dictionaries would facilitate many NLP tasks such as Spelling checker, Machine translation, Information Ex-traction and Information Retrieval. Linguistic ontologies and other lexical resources are very scarce in Manipuri. In this paper describe how to build a Manipuri-English machine-readable electronic dictionary by implementing ontology. It provides both of lexical information and their concepts, it may become the central resources of further Manipuri NLP research although developing lexical resources and defining semantic concepts for Manipuri vocabularies are very tedious and time consuming task. Moreover, it also provides the user interface to exploit these resources. So, it can serve as User-readable electronic dictionary as well as Machine-readable and will play crucial role in further Manipuri NLP research and applications. Since the concepts in the Ontology can optionally be classified and organized, the system can possible to hold additional concepts or properties or both. And the system can be modified to specific domain, particularly for technical domain, medical domain, agricultural domain, etc. Properties can then be added to any class according to the chosen domain. Since words in different languages have a common equivalent sense or concept, it can extend the existing ontology based Manipuri-English dictionary as Multi-lingual dictionary.

References

[1] Christiane Fellbaum, (1998) editor. Wordnet: An Electronic Lexical Database. The MIT Press,.
[2] DNS Bhat & M.S Ningomba (1997) “Manipuri Grammar”, Munchen- New castle.
[3] Grigoris Antoniou[1] and Frank van Harmelen[2] “Web Ontology Language: OWL” 1. Department of Computer Science, University of Crete ga@csd.uoc.gr 2. Department of AI, Vrije Universiteit Amsterdam, Harmelen@cs.vu.nl
[4] Lian-Tze Lim and Tang Enya Kong, "Building an Ontology-based Multilingual Lexicon for Word Sense Disambiguation in Machine Translation", Unit Terjemahan Melalui Komputer, Universiti Sains Malaysia.
[5] Mahesh, K. (1996). Ontology Development for Machine Translation: Ideology and Methodology, Technical Report MCCS 96-292, Computing Re-search Laboratory, New Mexico State University, Las Cruces, NM.
[6] M.S. Ningomba (2010). “A Dictionary of Manipuri Verbs”, Bir Computer Printing Works, 85-PDA Complex, Lamphepaat, Imphaal
[7] Sharma.H. Surmangol (2006). “Learners’ Manipuri Dictionary”, Sangam Book Store, Paona Bazar, Imphal.
[8] Singh. Ch. Yashawanta (2000). “Manipuri Grammar”, Rajesh Publication, New Delhi-110002.