Morphology Driven Manipuri POS Tagger

Thoudam Doren Singh
Computer Science Department
St. Anthony's College
Shillong-793001, Meghalaya, India
thoudam_doren@rediffmail.com

Sivaji Bandyopadhyay
Computer Science & Engineering Department
Jadavpur University
Kolkata – 700 032, India
sivaji_cse_ju@yahoo.com

Abstract

A good POS tagger is a critical component of a machine translation system and other related NLP applications where an appropriate POS tag will be assigned to individual words in a collection of texts. There is not enough POS tagged corpus available in Manipuri language ruling out machine learning approaches for a POS tagger in the language. A morphology driven Manipuri POS tagger that uses three dictionaries containing root words, prefixes and suffixes has been designed and implemented using the affix information irrespective of the context of the words. We have tested the current POS tagger on 3784 sentences containing 10917 unique words. The POS tagger demonstrated an accuracy of 69%. Among the incorrectly tagged 31% words, 23% were unknown words (includes 9% named entities) and 8% known words were wrongly tagged.

1 Introduction

Manipuri (Meiteilon or Meiteiron) belongs to the Tibeto-Burman language family and is highly agglutinative in behavior, monosyllabic, influenced and enriched by the Indo-Aryan languages of Sanskrit origin and English. The affixes play the most important role in the structure of the language. A clear -cut demarcation between morphology and syntax is not possible. In Manipuri, words are formed in three processes called affixation, derivation and compounding (Thoudam, 2006). The majority of the roots found in the language are bound and the affixes are the determining factor of the class of the words in the language. Classification of words using the role of affix helps to implement the tagger for a resource poor language like Manipuri with high performance.

There are many POS taggers developed using different techniques for many major languages such as transformation-based error-driven learning (Brill, 1995), decision trees (Black et al., 1992), Markov model (Cutting et al., 1992), maximum entropy methods (Ratnaparkhi, 1996) etc for English. Decision trees are used to estimate marginal probabilities in a maximum entropy model for predicting the parts-of-speech of a word given the context in which it appears (Black et al., 1992). The rules in a rule-based system are usually difficult to construct and typically are not very robust (Brill, 1992). Large tables of statistics are not needed for the rule-based tagger. In a stochastic tagger, tens of thousands of lines of statistical information are needed to capture the contextual information (Brill, 1992). For a tagger to function as a practical component in a language processing system, a tagger must be robust, efficient, accurate, tunable and reusable (Cutting, 1992).

2 Previous work on Manipuri POS tagger

Morphology based POS tagging of some languages like Turkish (Oflazer and Kuruoz, 1994), Czech (Hajic, et al., 2001) has been tried out using a combination of hand-crafted rules and statistical learning. A Marathi rule based POS tagger used a technique called SRR (suffix replacement rule) (Burange et al., 2006) with considerable accuracy. A POS tagger for Hindi overcomes the handicap of annotated corpora scarcity by exploiting the rich morphology of the language (Singh et al., 2006). To the best of our knowledge, there is no record available of work done on a Manipuri POS tagger. A related work of word class and sentence type identification in a Manipuri Morphological Analyzer

Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 91–98,
Hyderabad, India, January 2008.

is found in (Thoudam and Bandyopadhyay, 2006) where the classification of few word categories and sentence type identification are discussed based on affix rules.

3 Manipuri Morphemes

There are free and bound roots in Manipuri. All the verb roots are bound roots. There are also a few bound noun roots, the interrogative and demonstrative pronoun roots. They cannot occur without some particle prefixed or suffixed to it. The bound root may form a compound by the addition of another root. The free roots are pure nouns, pronouns, time adverbials and some numerals. The bound roots are mostly verb roots although there are a few noun and other roots. The suffixes, which are attached to the nouns, derived nouns, to the adjectives in noun phrases including numerals, the case markers and the bound coordinators are the nominal suffixes. In Manipuri, the nominal suffixes are always attached to the numeral in a noun phrase and the noun cannot take the suffixes. Since numerals are considered as adjectives, the position occupied by the numerals in Manipuri may be regarded adjective position (Thoudam, 2006). There are a few prefixes in Manipuri. These prefixes are mostly attached to the verb roots. They can also be attached to the derived nouns and bound noun roots. There are also a few prefixes derived from the personal pronouns.

In this agglutinative language the numbers of verbal suffixes are more than that of the nominal suffixes (Singh, 2000). New words are easily formed in Manipuri using morphological rules. Inflectional morphology is more productive than derivative morphology (Chelliah, 1997). There are 8 inflectional (INFL) suffixes and 23 enclitics (ENC). There are 5 derivational prefixes out of which 2 are category changing and 3 are non-category changing. There are 31 non-category changing derivational suffixes and 2 category changing suffixes. The non-category changing derivational suffixes may be divided into first level derivatives (1st LD) of 8 suffixes, second level derivatives (2nd LD) of 16 suffixes and third level derivatives (3rd LD) of 7 suffixes. Enclitics in Manipuri fall in six categories: determiners, case markers, the copula, mood markers, inclusive/exclusive and pragmatic peak markers and attitude markers. The categories are determined on the basis of position in the word (category 1 occurs before category 2, category 2 occurs before category 3 and so on).

4 Dictionaries

Three different dictionaries namely prefix which contains prefix information, suffix which contains suffix information and root containing 2051 entries are used for the system. The format of root is <root><category>.

A bilingual dictionary consisting of Manipuri word and its corresponding pronunciation, POS, 1st English (Eng1) word meaning, 2nd English (Eng2) word meaning (if any), 3rd English (Eng3) word meaning (if any), a Manipuri sentence or phrase using the word and corresponding English meaning has been developed based on the work of Manipuri to English Dictionary (Imoba, 2004). The bilingual parallel dictionary is used for testing POS tagger and later on will be used for EBMT system. The Manipuri sentences/phrases using a particular word are used as the input to the POS tagger thus enabling to sort out words with multiple meaning.

5 Morphological analysis of Major Lexical categories

The lexical categories in Manipuri can be of two types – major and minor (Chelliah, 1997). Major lexical categories can be of two types, namely "actual" and "potential". The lexicon of actual lexical categories i.e., actual lexicon consists of an unordered list of roots and affixes and lexicalized forms. Each lexical entry in the actual lexicon consists of what lexical category it belongs to and what its meaning is. On the other hand, the output of the potential lexicon consists of words created through productive morphological processes. In the actual lexicon, roots may be bound or free. Nouns and verbs from the actual lexicon can be distinguished on formal grounds in that bound roots are verbs and free roots are nouns. In the potential lexicon, adjectives, adverbs and nominal forms can be derived from verb roots and stative verbs can be derived from noun roots. There are several instances where the words belonging to some class or category plays the role of some other category sometimes based on its position in the sentences (P.C. Thoudam,

2006) Some of the generalized handcrafted rules to identify the lexical are given as below.

5.1 Nouns

Nouns can be distinguished from other lexical categories on morphological grounds. Unlike verbs, nouns can be suffixed by gender, number or case markers. Proper nouns and common nouns are free standing forms.

The following is the list of word structure rules for nouns (Chelliah, 1997)

N → Root INFL (ENC)
Root → Root (2nd LD)
Root → Root (1st LD)
Root → (prefix) root (root)

Figure 1 shows the general form of noun morphology in Manipuri. Examples of some singular/plural noun forms are listed in Table 1.

Pronominal prefix	Root	gender	number	Quantifier	Case

Figure 1. General form of Noun Morphology

Singular Form	Plural Form
উচেক -Uchek (bird)	উচেকশিং -Ucheksing(birds)
ম -Ma (He/She)	মখোয় -Makhoy (they)
মী -Mi (man)	মীয়াম -Mi-yaam (men)

Table 1: Singular/Plural forms

Although case markers are functionally inflectional, they exhibit the clitic like characteristic of docking at the edge of a phrase. The word structure of rules of verbs and nouns are identical except for the category of the word level node, the possible terminal elements of the derivational and inflectional categories and the lack of the third level nominal derivation. Two examples to demonstrate the noun morphology are given below:-

মচানুপীশিংনা *(mə-ca-nu-pi-siŋ-nə)* 'by his/her daughters'

মচানুপাশিংনা (*mə-ca-nu-pa-siŋ- nə*) 'by his/her sons'

The ম *-mə* 'his/her' is the pronominal suffix and চা *-ca* 'child' is the noun root. The নু *-nu* 'human' is suffixed by পী *-pi* to indicate a female human and পা*–pa* to indicate a male human. শিং *–siŋ* or খোই *-khoy* or য়াম*–yaam* can be used to indicate plurality. *-siŋ* cannot be used with pronouns or proper nouns and *-khoy* cannot be used with nonhuman nouns. না *-nə* meaning 'by the' is the instrumental case marker.

5.2 Pronouns

The singular personal pronouns are ঐ *-əy* 'I', নং *-nəŋ* 'you' and মা *-ma* 'he/she'. Possessive pronouns are formed through the suffixation of কি *-ki* 'genitive' on these personal pronouns. Indefinite pronouns are also lexicalized forms that consists of a question word which may be followed by সু *-su* 'also' or the sequence কুম্ব *-kumbə* composed of কুম *–kum*, 'like', 'kind of' and ব *–bə* 'nominalizer'. The strategy for creating relative clause in Manipuri is to place the relativized noun directly after a normalized clause; there is no relative pronoun to mark the relative clause. The determiner may occur either as an independent pronoun or encliticized on the noun phrase with no difference in meaning. The determiners সি *–si* 'proximate' and তু *–tu* 'distal' are stems that function as enclitics. সি *–si* indicated that the object or person being spoken of is near or currently seen or known to be near., even if not viewable by the speaker, or is currently the topic of conversation; তু *-tu* signifies something or someone not present at the time of speech or newly introduced in the conversation. Possessive pronominal prefix may be affixed to the root শা *sa* 'body' to form pronouns emphasizing that the subject of the verb is a particular person or thing and no one or nothing else: ইশানা *isanə* 'by myself' নশানা *nəsanə* 'by yourself' and মশানা *məsanə* 'by him/her/itself/. The set of Manipuri Pronominal prefixes differ for different persons (ই {I} for 1st person, ন {Na} for 2nd person and ম {Ma} for 3rd person) while the set of pronominal suffixes differ only on gender (পা –pa for masculine gender, পী -Pi for feminine gender).

5.3 Verbs

Verbs roots are in the actual lexicon and are bound forms. A verb may be free standing word if it is minimally suffixed by an inflectional marker. The verb root may also be followed by one of the enclitics. Three derivational categories may optionally precede the final inflectional suffix. The 1st LD suffixes signal adverbial meanings, the 2nd LD suffixes indicate evidentiality, the deitic reference of

a verb, or the number of persons performing the action and the 3rd LD suffixes signal aspect and mood. Verb roots may also be used to form verbal nouns, adjectives and adverbs. Verbal nouns are formed through the suffixation of the nominalizer পা –pə to the verb root.

The following is the list of word structure rules for verbs (Chelliah, 1997)

a. Verb → Root INFL
b. Root → Root (3rd LD)
c. Root → Root (2nd LD)
d. Root → Root (1st LD)
e. Root → root (root)
f. 3rd LD → (mood1)(mood2)(aspect)
g. 2nd LD → (2nd LD1),(2nd LD2),(2nd LD3)..
h. 1st LD → 1st LD

Derivational Prefixation	Root	1st Level derivation	2nd level derivation	3rd level derivation	Inflection

Figure 2. General form of Verb Morphology

There are 3 categories (mood1, mood2, and aspect) belonging to the third level derivational (3rd LD) markers. The general form of verb morphology is shown in figure 2.

The sub-categorization frames of affixes will restrict that only nominal affixes occur with a noun and verbal affixes occur with a verb root. The derivational suffix order of the word চেকখাইরক্কনি is given below:-

চেক	খাই.	রক	ক	নি
cek	*–khay*	*-rək*	*-kə*	*-ni*
crack	*-totally affect (1st LD)*	*-distal (2nd LD)*	*-potential (3rd LD)*	*–copula*

The রক *-rək* has allomorph লক-*lək*. রক *-rək* occurs after vowels while লক-*lək* occurs after consonants. Such allomorph is an example of orthographic change and it is taken care by the system by making individual entries into the dictionary.

চারকএ *–ca-rək-y* (ate there and came here)
চাম্লকএ – *cam-lək-y* (washed there and came here)

The formation of verb can be of the form

Verb stem + aspect/mood → verb
থক *-thək* (drink) + লে *-le-* → থকলে *thəkle* (has drunk)

The verbal noun is formed with the rule as given as

Verb Stem + Nominalizer → Verbal noun
থোং *-thong* (cook)+ বা *-ba* → থোংবা *thongba* (to cook)

5.4 Adjectives

An adjective is derived through the affixation of the attributive, derivational prefix অ *ə-* to a verbal noun.
e.g.
অ *-ə* + *Verbal noun* → *Adjective*
অ *-ə* + সি *-si* (die) + বা *-ba*–>অসিবা *əsiba* (something dead)

Adjectives may appear before or after the nouns they modify. Possessive adjectives are formed through the suffixation of the genitive marker কি *–ki* to the possessor of a noun.

5.5 Adverbs

Manner adverbs are formed through suffixation of না *–nə* 'adverbial' to a verb root. e.g. লোয়না *loynə* 'completely, all' from *loy* 'complete,finish'. e.g.,

Stem + না - na → Adverb
কপ *-Kəp* (cry)+ না - *nə* –> কপনা *-kəp-nə* (cryingly)

Locative adverbs are derived through the prefixation of ম *mə* 'noun marker' to a noun or verb roots. e.g. মখা *məkha* 'below, underneath' *from* খা *kha 'south'*

6 Morphological analyses of some minor lexical categories

The three minor lexical categories of Manipuri are quantifiers, numerals and interjections. These are considered minor categories because these lexical items are closed sets which express meanings most often encoded by affixal morphology. The lexical items in interjection is defined on the semantic similarity of its members, all express strong emotion.

6.1 Quantifiers

Most quantifiers in Manipuri are lexicalized forms consisting of the unproductive prefix *kh*V- (where the vowel can be a, i, u). These are খরা *-khəra* 'some' which indicates an indeterminate amount; খিতং *-khitəŋ* 'ever so little, a particle' of some tangible material. These quantifiers can be combined as in

ঈশিং খরা খিতং পুরকউ.
Ishiŋ khəra khitəŋ purək-u
'Bring me just a little bit of water'.

6.2 Numerals

The numerals are nouns. Ordinal numerals are adjectives, derived through the affixation of the attributive prefix অ *–ə* and the nominalizer বা *–bə* to any numeral with শু *–su* 'also': thus অনিশুবা *ənisubə 'second one'*.

6.3 Interjections

The lexical items of this category which is defined on the semantic similarity of its members, all express strong emotion. Some of these are composite forms where one syllable is identifiable as the exasperative enclitic হে *–he* and the second syllable is not identifiable as a productive affix or stem.

7 Manipuri Tagset

The basic Manipuri POS tag set used in the POS tagger is listed below. কুক্রু কুক্রু *kukru kukru* (a pigeon's cry) is ideophone. তু *tu* 'that' is a determiner. হায়বশি *haybasi* is a determiner complementizer.

Sl. No.	Category name	Tag
1	adjective	ADJ
2	adverb	ADV
3	conjunction	CONJ
4	complementizer	CMP
5	determiner	DET
6	ideophone	IDEO
7	interjection	INTJ
8	noun	N
9	pronoun	PN
10	quantifier	QU
11	verb	VB
12	Verbal noun	VN
13	Unknown	UNK

Table 2. Manipuri POS tagset

8 Design of Manipuri POS tagger

In Manipuri, the basic POS tags are assigned to the words on the basis of morphological rules. Figure 3 shows the system diagram of Manipuri POS tagger.

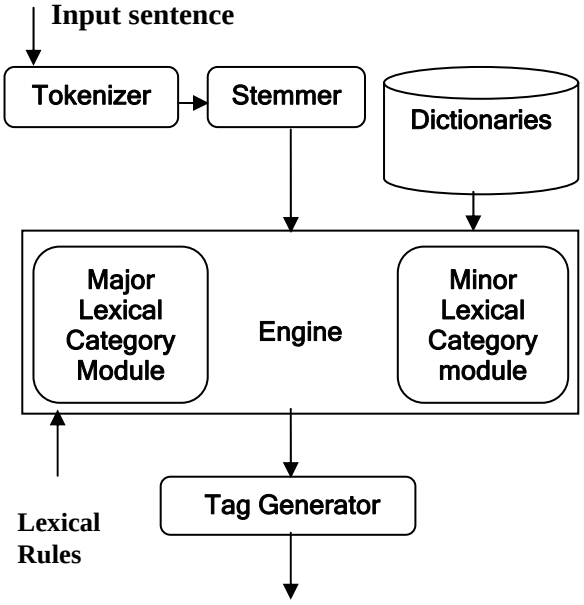

Figure 3. System Diagram

The different parts involved in the system are:-

a. **Tokenizer:** Words are separated based on the space given between consecutive words.
b. **Stemmer:** It separates the prefixes and suffixes from the words.
c. **Engine:** Different analysis and treatment of different words are performed based on the category.
d. **Tag Generator:** Tags are assigned to the words in the sentence input based on the tagset and morphology rules.
e. **Dictionaries:** Prefix, suffix and word dictionary along with sentences using the words are maintained.

8.1 Algorithm of POS tagging

Algorithm used for tagging is as follows:-

1. Input the Manipuri input texts to the Tokenizer.
2. Repeat steps 3 to 6 until the end of the texts for each token.
3. Feed the tokens to the stemmer.
4. Check the patterns and order of the different morphemes by looking at the stem category.
5. Apply the handcrafted morphological rules for identifying the category using the engine.
6. Generate the POS tags using Tag generator.
7. End.

The Visual C++, MsAccess and GIST SDK are used to develop the system. The Manipuri words are entered into the dictionary using Bengali script (BN1 TTBidisha font).

9 Evaluation

In Manipuri, word category is not so distinct except Noun. The verbs are also under bound category. Another problem is to classify basic root forms according to word class although the distinction between the noun class and verb classes is relatively clear, the distinction between nouns and adjectives is often vague. Distinction between a noun and an adverb becomes unclear because structurally a word may be a noun but contextually it is adverb. Thus, the assumption made for word categories are depending upon the root category and affix information available from the dictionaries. At the moment, we use a sequential search of a stem from the root dictionary in alphabetical order. It is found to be suitable for small size dictionary. Further a part of root may also be a prefix which leads to wrong tagging. The verb morphology is more complex than that of noun. A comparative study on the number of words tagged by the system and manually tagged had been carried out. The inputs of 3784 Manipuri sentences of 10917 unique words as input to the tagger engine. Sometimes two words get fused to form a complete word. Handling such collocations is difficult. Conjuncts require a separate dealing using a table. Verbs, nouns and noun phrases, subordinate sentences, and root sentences can be affixed by enclitics. Table 4 shows the percentage statistics of tagging output based on the actual and correctly tagged words. The accuracy of tagging can be further improved by populating more root morphemes to the root dictionary.

$$\text{Accuracy percentage} = \frac{\text{No. of single correct tags}}{\text{Total no. of tokens}} \times 100$$

Group Types	**Percentage**
Single tagged correct words	65%
Multiple tagged correct words	4%
Unknown words	23% (9% Named Entities)
Wrong tagged words	8%

Table 4. Tagger output statistics

The unknown words are the words which could not be tagged based on the linguistic rules and unavailability of entries mainly in root dictionary. In the process of word formation, only affixation: prefixing, suffixing or compounding takes the role of formation of new words in this language. Due to the fact that new words are easily formed in Manipuri, thus the number of unknown words (out of vocabulary) is relatively large (Sirajul et al., 2004).

10 Challenges for future work

The noun group words handling are not incorporated. For example অখাক অরাও (pronounced as *əkhak əraw*) meaning *thunderbolt,* অঙম অরাই খঙদবা (pronounced as *əŋam əray khəŋdəba*) meaning *wanton* are noun group words and are not tagged by the POS tagger correctly. The Noun-Adjective ambiguity disambiguation scheme is required as a separate module and implementations are to be included in the future work. The Manipuri tagging is very much dependent on the morphological analysis and lexical rules of each category. There is a cleaning process of all word and morphemes specially the spelling to ensure that the lexical rules are implemented. This has not yet been implemented. Collocations handling and more disambiguation rules will be developed in further phases of the work. The output of the POS tagger will be used in a Manipuri-English machine translation system.

References

E. Black, F. Jelinek, J. Lafferty, R. Mercer and S. Roukos. 1992. Decision tree models applied to labeling of texts

with parts of speech. In *DARPA Workshop on Speech and Natural Language*. San Mateo, CA, 1992, Morgan Kaufman.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings Third Conference on Applied Natural Language Processing, ACL*, Trento, Italy.

Eric Brill. 1995. Transformation-Based Error Driven Learning and Natural Language Processing: A case study in Parts-Of-Speech tagging. *Computational Linguistics* 21(94): pp 543-566.

Sachin Burange, Sushant Devlakar, Pushpak Bhattacharyya. 2006. Rule Governed Marathi POS Tagging. In *Proceeding of MSPIL*, IIT Bombay, pp 69-78.

Shobhana L. Chelliah. 1997. *A Grammar of Meithei*. Mouton de Gruyter, Berlin, pp 77-92.

Sirajul Islam Choudhury, Leihaorambam Sarbajit Singh, Samir Borgohain, P.K. Das. 2004. Morphological Analyzer for Manipuri: Design and Implementation. In *Proceedings of AACC*, Kathmandu, Nepal, pp 123-129.

D. Cutting. 1992. A practical part-of-speech tagger. In *Proceeding of third conference on Applied Natural Language Processing. ACL*, 1992. pp 133-140.

J. Hajic, P. Krbec, P. Kveton, K. Oliva, V.Petkevic, 2001. A Case Study in Czech Tagging. In *proceedings of the 39th Annual Meeting of the ACL.*

S. Imoba. 2004. *Manipuri to English Dictionary*. S. Ibetombi Devi, Imphal.

K. Oflazer, I Kuruoz. 1994. Tagging and morphological disambiguation of Turkish text. In *Proceedings of 4th ACL conference on Applied Natural Language Processing Conference.*

A. Ratnaparakhi. 1996. A maximum entropy Parts-Of-Speech Tagger. In *Proceedings EMNLP-ACL*. pp 133-142.

Smriti Singh, Kuhoo Gupta, Manish Shrivastava, Pushpak Bhattacharya. 2006. Morphological Richness offsets Resource Demand – Experiences in constructing a POS tagger for Hindi. In *Proceedings of COLING-ACL*, Sydney, Australia.

Ch. Yashawanta Singh. 2000. *Manipuri Grammar*. Rajesh Publications, New Delhi.

P.C. Thoudam. 2006. *Problems in the Analysis of Manipuri Language*. www.ciil-ebooks.net, CIIL, Mysore.

D. S. Thoudam and S. Bandyopadhyay. 2006. Word Class and Sentence Type Identification in Manipuri Morphological Analyzer. In *Proceedings of MSPIL*, IIT Bombay, pp 11-17.