Part of Speech Tagging in Manipuri: A Rule-based Approach

Kh Raju Singha
Dept. of Computer Science
Assam University, Silchar

Bipul Syam Purkayastha
Dept. of Computer Science
Assam University, Silchar

Kh Dhiren Singha
Dept. of Linguistics
Assam University, Silchar

ABSTRACT
The process of assigning morpho-syntactic categories of each morpheme including punctuation marks in a given text document according to the context is called Part of Speech (POS) tagging. In this paper we represent the rule-based Part of Speech Tagger of Manipuri by applying a set of hand written linguistic rules of Manipuri language. Nevertheless, it is very difficult to classify the lexical categories of Manipuri, an agglutinating Tibeto-Burman language of Northeast India. So, in this tagger we are using the affix stripping technique to segment the affixes from the root. As Manipuri has limited POS tagged corpus, the tagged output of this tagger will be very helpful to analyze Manipuri Part of speech by using many statistical models.

General Terms
Part of Speech Tagging, Manipuri Language, Natural Language Processing, Algorithms, Morpho-syntactic categories.

Keywords
Tagset, tokenizer, lexicon, corpus, affix, stemmer, information retrieval.

1. INTRODUCTION
Part of Speech Tagger is one of the important components in the development of any serious application in different fields of Natural Language Processing (NLP) in the present world. Part of Speech tagging is a technique for automatic annotation of lexical categories. It assigns an appropriate tag for each word in a sentence of a language as corresponding to part of speech, based on its definition, as well as its context [6].

A POS tagger takes a sentence as input and assigns a unique part of speech tag to each lexical item of the sentence. POS tagging is used as an early stage of linguistic text analysis in many applications including subcategory acquisition, text to speech synthesis, and alignment of parallel corpora [6]. The POS tagger can be used in other areas of Natural Language Processing such as semantic analysis, information retrieval, shallow parsing, information extraction and machine translation etc.

POS taggers can be divided as supervised and unsupervised. Both the supervised and unsupervised taggers can be categorised as rule-based and statistic models. Rule based Part of Speech Tagging is the approach that uses hand written rules for tagging. Rule-based tagger depends on dictionary or lexicon to get possible tags for each word to be tagged. Hand-written rules are used to identify the correct tag when a word has more than one possible tag. Disambiguation is done by analyzing the linguistic features of the word, its preceding word, its following word and other aspects. For example, if the preceding word is adjective then the word in question must be adjective or noun in Manipuri language. This information is coded in the form of rules [18].

2. RELATED WORKS
Many works related to POS tagging has been done in languages like English, Chinese, German and Arabic etc. Several POS tagger of these languages are developed by using different algorithms. For instance, English language has developed POS tagger using rule based, statistical method, neural network and transformational based method etc [15]. In the year 1992 Eric Brill has been developed a rule based POS tagger with the accuracy rate of 95-99% [2]. POS tagging of some languages like Turkish [3], Czech [5] has been attempted using a combination of hand-crafted rules and statistical learning.

Similarly, Indian languages like Hindi, Bengali, Punjabi and Dravidian languages have many POS taggers. Manish Shrivastava and Pushpak Bhattacharyya proposed POS tagger for Hindi based on HMM in the year 2008 [5]. Adopting rule based approach a POS tagger for Marathi has been developed in 2006 using a technique called SRR (suffix replacement rule) by Sachin Burange et al. [10]. A Punjabi POS tagger is also developed by Singh Mandeep, Lehal Gurpeet and Sharma Shiv in 2008 with accuracy performance of 88.86% excluding unknown words [16].

As per the literature, there is a few works related to POS tagging in Manipuri and other Tibeto-Burman languages in the Indian Sub-continent. In the year 2004, Sirajul Islam Choudhury, Leihaorambam Sarbajit Singh, Samir Borgohain, P.K. Das have designed and implemented a morphological analyzer for Manipuri language [7]. Besides, D. S. Thoudam et al. developed morphology driven Manipuri POS tagger in 2008 [14]. Furthermore, Kh Raju Singha, Bipul Syam Purkayastha, Kh Dhiren Singha, Arindam Roy designed a tagset for Manipuri POS tagging consisting of 97 tags including generic attributes and language specific attribute values based on the ILPOST framework in 2011 [17].

3. MANIPURI LANGUAGE
Manipuri (Meiteilon) is one of the oldest languages in the South-East Asia which has its own script (Meitei Mayek) and literature. At Present, Manipuri used to write in Bengali Script from 1709 A.D. onwards i.e.; during the reign of king Pamheiba. Manipuri is widely spoken in Manipur, Assam, Tripura, Bangladesh and Myanmar, which has been included in the eight schedule of Indian Constitution since 1992. Interestingly, it is the first Tibeto-Burman language which has obtained its due place and recognition in Indian Constitution [11]. Linguistically, it belongs to the Kuki-Chin group of the Tibeto-Burman family of Languages [1] influenced and

enriched by the Indo-Aryan languages of Sanskrit origin and English. The total number of people who return Manipuri as their mother tongue was 1,500,000 out of which 1,466,705 speakers reside in India (Census of India, 2001).

Manipuri is a tonal, agglutinating and verb final language. Like other OV languages adjectives may precede or follow the noun in NP constructions. As in many other Tibeto-Burman languages, adjective is not a distinct category of words in Manipuri as adjectives are derived from the intransitive verb particularly the stative verbs. Unlike other Indo-Aryan languages of the sub-continent, there is lack of relative pronoun; the relative clause is expressed by means of participle.

4. PROPOSED TAGSET

Many research institutions and individual has been developed various tagset worldwide in different languages. Some common examples of English tagset are Brown Corpus tagset, Penn Treebank tagset, C5 tagset and C7 tagset. Every language has different tagset as there are many linguistic variations among the languages. In this paper we have designed a 3-tier tagset for Manipuri based on ILPOST framework [13]. It has been customized for Manipuri to meet the morpho-syntactic requirements of the language and in accordance with language specific and orthography follows in Manipuri.

The proposed tagset consists of 97 tags including generic attributes and language specific attribute values. The tagset has two tables: table-1 contains 32 tags in which 31 tags of sub categories for 13 major categories and 1 tag for unknown category. There are 4 tags for noun, 6 tags for pronoun, 1 tag for verb, 2 tags for modifier, 1 tag for specifier, 1 tag for adverb, 2 tags for demonstrative, 3 tags for participle, 3 tags for particle, 1 tag for punctuation, 4 tags for numeral, 1 tag for reduplication, 2 tags for residual and 1 tag for unknown category. Table-2 contains the morpho-syntactic features or attributes of the sub categories. There are 32 attributes having 65 attribute value tags. The full tagset is given in the appendix. A partial graphical representation of the proposed tagset with noun as an example is given below:

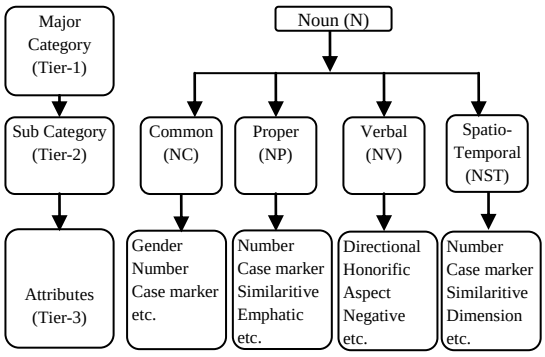


Fig 1: A partial graphical representation of the proposed tagset

5. PROPOSED ARCHITECTURE

The proposed system design for rule-based Manipuri POS tagger is given below:

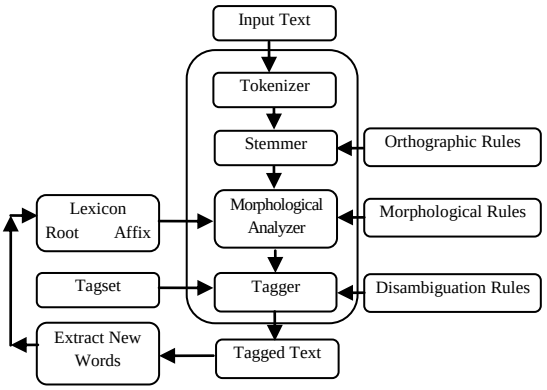


Fig 2: Architecture of Rule-Based Manipuri POS Tagger

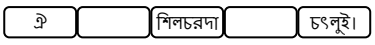
The different modules involved in this architecture are explained as following:

I. Tokenizer: Tokenization is the first step in part of speech tagging of any natural language. It separates the words including punctuation marks and the symbols of the input text into tokens by using the whitespace between consecutive words. Manipuri has its own writing convention of keeping a whitespace between the words but there is no whitespace between a word and punctuation mark or symbol. So for the proper tokenization of Manipuri text, we have developed a simple rule for such type of identification as follows:

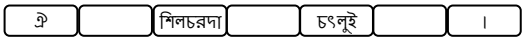
If (word.endswith ("punctuation mark" | |"symbol")) then insert a whitespace between the word and punctuation mark or symbol;

An example of tokenization of a simple Manipuri sentence is given, here the filled box is word and blank box is whitespace.

Before Tokenization:



After Tokenization:



II. Stemmer: It separates the affixes i.e.; prefixes and suffixes from the stem or root word. Stemmer plays an important role to identify the affixes and stem in this architecture because affixation is one of the word formation methods of Manipuri language. Stemmer separates the suffixes one at a time starting at the end of the word and working towards the beginning by using the iterative affix stripping algorithm. Figure 3 shows a diagrammatic view of stemming of a simple Manipuri word:

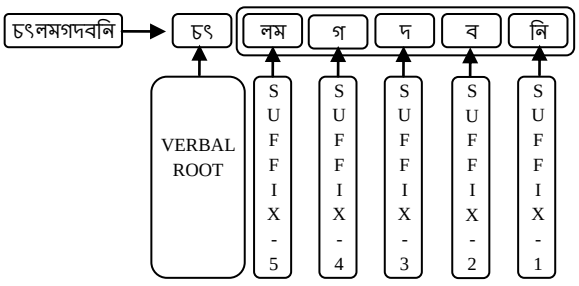


Fig 3: A diagrammatic view of stemming of a simple Manipuri word

III. Morphological Analyzer: Morphological analyzer analyzes the Manipuri words according to the morphological rules and identifies the morphosyntactic categories of root and affixes including the relations between the morphemes.

IV. Lexicon: Lexicon consists of the list of roots and affixes with their corresponding part of speech tags which are in the tagset. Initially a tagged lexicon is developed manually by collecting limited words from Manipuri newspapers, books and dictionaries.

V. Extraction of new words: In this architecture, the tagger first tag the words which are in the lexicon and the words which are not available in the lexicon are also tagged by applying rules. If the rules failed to tag such words then the new words are given a specific tag as "UNK" i.e.; unknown which can be extracted from the tagged output text. The new words are then entered in to the lexicon and new rules are created for the new words.

VI. Tagger: Tagger is used to resolve the ambiguity issues in tagging and tag the proper tag to a token. In some cases, a token has more than one tag but the tagger tagged only one tag to a token by applying hand written rules for those specific cases.

VII. Rules: We have applied three types of rules for Manipuri in implementing this rule-based tagger. Different rules are formulated with example as shown below:

a. ***Orthographic Rules:*** There are orthographic variations in the spelling system of Manipuri having difficulties in formulating orthographic rules consistently. However, attempt has been made to formulate the rules of orthography in Manipuri experimentally. A simple example of such kind of rule for Manipuri is given below:

If any stem getting after stemming the suffixes and ends with "ব" or "প" then "ব" or "প" will be replaced by "বা" or "পা" respectively like অঙাংবগী→অঙাংবা/MJ গী/GEN and চৎপগী→ চৎপা/NV গী/GEN

b. ***Morphological Rules****:* The word formation in Manipuri is employed by three morphological processes called affixation, derivation and compounding. Some verb roots can be formed noun, verb, adjective and adverb by affixation as shown below with examples in the underlying representation.

Prefix +bound root + suffix → Adjective

অ/Prefix + চেন/VR + বা/NMZ → অচেনবা / MJ

অ/Prefix + চৎ/VR + পা/NMZ → অচৎপা / MJ

If a word starts with "অ" and ends with "বা" or "পা" then tag the word as Adjective (MJ)

Apart from the above, some adjectives in Manipuri particularly colour terms are free morphemes.

c. ***Disambiguation Rules:*** Any natural language has the ambiguity issues as the single word has different tags or categories. To overcome the ambiguity issues and assigning a proper tag to a word, disambiguation rules are required. In Manipuri there are plenty of words which have multiple tags. Consider the following examples in this regard.

First example: ঐ চা থকলি।

Second example: ঐ কমলা চারি।

The word "চা" may be Common Noun (NC) in the first example or Verb Root (VR) in the second example. Now the rule of disambiguation is as

Given Input: "চা"

If (+1 is VR/MJ/NC) /* *if next word is verb root, adjective or*

*Common noun */*

Then assign NC tag

Else assign VR tag

5.1 Algorithm for POS tagging

Algorithm used for this tagging is as follows:

Step 1: Input the Manipuri text.

Step 2: Repeat the Step 3 to Step 7 till the end of the input.

Step 3: Tokenize the input text.

Step 4: If the word is in the form affixation, derivation and compounding then feed the word to stemmer for splitting.

Step 5: Morphological Analyzer checks the word with the lexicon for a match.

Step 6: If match is found the word is properly tagged by the tagger.

Step 7: If multiple tags exists for a single word then tagger tagged the word by using rules.

Step 8: Returned the tagged output text.

Step 9: Extract those unknown new words from the tagged output.

Step 10: Make the new entry for the unknown new word to the lexicon.

Step 11: Add the new rules for newly entered words.

5.2 Rule-based Manipuri POS tagger tool

A Graphical User Interface tool has been developed by using NetBeans IDE 7.1, JDK 6 and JRE 6. The front-end of the tool has been implemented in java and its interface is connected with a text file of Manipuri lexical items called "lexicon" as the back-end. The selection of textual database is for simplicity and to extend support for multiple platforms without the need of the installation of any DBMS server like MYSQL etc. by the end user. Each lexical item entry in "lexicon" file has two fields:

ITEM: Manipuri lexical item like "লাইরিক" or "Lairik" i.e.; "book".

CATEGORY: The morpho-syntactic category of the lexical item like NC (Common Noun).

The Manipuri words are entered into the lexicon using Kalpurush font of Bengali-Script. A screenshot view of the tool is shown as below.

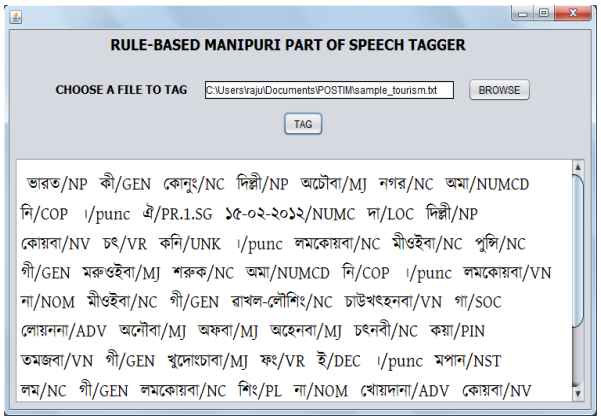


Fig 4: Rule-based Manipuri POS Tagger GUI tool

6. TEST AND RESULTS

A lexicon of limited words and few rules are present in this tagger. We have tested the tagger by using a lexicon with different number of words and rules. A summary of test and result of the tagger is given below:

Table1: Result Table

No. of words in the lexicon	No. of rules applied	Accuracy
100	7	50%
500	15	77%
1000	25	85%

The above table shows that the accuracy level is increasing with increase the no. of words in the lexicon and the no. of rules applied.

7. CONCLUSION AND FUTURE WORK

In this paper a rule-based Manipuri Part of Speech Tagger is implemented as part of the larger goal of computational analysis of Manipuri language. The tagged output of the tagger can be used as corpus in analysis of Manipuri language by applying many statistical methods like unigram, bigram and HMM model etc. This tagger gives a good accuracy rate of tagging Manipuri lexical items excluding MWE and named entity.

The future work would be to design a tagging model by hybridization of rule-based and statistical method to attain better accuracy rate. The design should be enabled to detect MWE and named entity recognition in tagging Manipuri text.

8. REFERENCES

[1] G.A. Grierson's Linguistic Survey of India. Vol. III, Pt. III, 1976.

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

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

[4] Ch. Yashawanta Singh "*Manipuri Grammar*." Rajesh Publications, New Delhi, 2000.

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

[6] Sandipan Dandapat, Sudeshna Sarkar and Anupam Basu "A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali", Transactions on Engineering, Computing and Technology V1 December, 2004.

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

[8] S. Imoba. "Manipuri to English Dictionary". S. Ibetombi Devi, Imphal, 2004.

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

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

[11] Kh. Dhiren Singha, "Loan Words in Manipuri", Bilingualism and North-East India, an Assam University Publication, 2008.

[12] Manish Shrivastava and Pushpak Bhattacharyya, Hindi POS Tagger Using Naïve Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge, International Conference on NLP (ICON08), Pune, India, 2008.

[13] S. Baskaran et al." Designing a Common POS-Tagset Framework for Indian Languages" The 6th Workshop on Asian Languae Resources, 2008.

[14] Thoudam Doren Singh & Sivaji Bandyopadhyay "Morphology Driven Manipuri POS Tagger", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 91–98, Hyderabad, India, January 2008.

[15] D. Jurafsky, and J. H. Martin, "Speech and Language Processing", Second edition, Published by Pearson Education, 2009.

[16] Dinesh Kumar and Gurpreet Singh Josan, "Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey", International Journal of Computer Applications (0975 – 8887) Volume6–No.5, September, 2010.

[17] Kh Raju Singha, Bipul Syam Purkayastha, Kh Dhiren Singha, Arindam Roy "Developing a Tagset for Manipuri Part of Speech Tagging" Journal of Computer Science and Engineering, Volume-5-issue-1-january 2011.http://sites.google.com/site./jcseuk/volume-5-issue 1-january-2011.

[18] http://language.worldofcomputing.net/pos-tagging/rule based-pos-tagging.html.2012.

APPENDIX

Manipuri Tagset Table: 1

Major Category	Tag	Description
Noun(N)	NC	Common noun
	NP	Proper noun
	NV	Verbal noun
	NST	Spatio-Temporal
Pronoun(P)	PPN	Personal pronoun
	PPS	Possessive pronoun
	PDM	Demonstrative pronoun
	PRF	Reflexive pronoun
	PRC	Reciprocal pronoun
	PIN	Interrogative pronoun
Verb (V)	VR	Verb Root
Modifier(M)	MJ	Adjective
	MQ	Quantifier
Specifier(SPEC)	SPEC	Specifier
Demonstrative(D)	DAB	Absolute demonstrative
	DWH	Wh-demonstrative
Adverb(ADV)	ADV	Adverb
Participle(PL)	PLRL	Relative participle
	PLV	Verbal participle
	PLC	Conditional participle
Particle(C)	CCD	Co-ordinating particle
	CSB	Subordinating particle
	CINT	Interjection particle
Residual(RD)	RDF	Foreign word residual
	RDS	Symbol residual
Punctuation(PUN)	PUN	Punctuation
Numeral(NUM)	NUMR	Real Numeral
	NUMC	Calendric Numeral
	NUMCD	Cardinal Numeral
	NUMO	Ordinal Numeral
Reduplication (RDP)	RDP	Reduplication
Unknown(UNK)	UNK	Unknown

Manipuri Tagset Table: 2

Attributes	Tag	Description
Gender	MAS	Masculine gender
	FEM	Feminine gender
Number	SG	Singular number
	DU	Dual number
	PL	Plural number
Case Marker	1	First person
	2	Second person
	3	Third person
	ERG	Ergative
	NOM	Nominative
	ACC	Accusative
	INS	Instrumental
	DAT	Dative
	GEN	Genitive
	ABL	Ablative
	SOC	Sociative
	LOC	Locative
	NMZ	Nominalizer
Allative (Towards)	ALL	Allative
Approximate	APP	Approximate
Simmilaritive	SIM	Simmilaritive
Aspect	PRG	Progressive
	PRF	Perfective
Prospective	PROS	Prospective
Inceptive	INC	Inceptive
Habitual	HAB	Habitual
Mood	DEC	Declarative
	SUP	Suplicative
	PROH	Prohibitive

	IMP	Imperative
	PERM	Permissive
	OPT	Optative
	INT	Interrogative
	POT	Potential
	NPOT	Non-potential
Modality	OBL	Obligation
	VOL	Volition
Evidential	EVI	Evidential
Certainty	CERT	Certainty
Directional	UP	Upward
	DOWN	Downward
	IN	Inward
	OUT	Outward
Causative	CAUS	Causative
Reflexive	REFL	Reflexive
Reciprocal	RECI	Reciprocal
Purposive	PURP	Purposive
Commutative	COMM	Commutative
Copula	COP	Copula
Evaluative	SURP	Surprise
	DUB	Dubitative
	CONF	Confirmation
	EXAS	Exasperation
	PERSU	Persuation
	CMPL	Complaint
	INSIS	Insistent
Distributive	DTRB	Distributive
Definiteness	DEF	Definiteness
Emphatic	EMPH	Emphatic
Negative	NEG	Negative

Dimension	PRX	Proximal
	DST	Distal
Inclusive/ Exclusive	INL	Inclusive
	EXL	Exclusive
Honorificity	HON	Honorificity