A Dictionary Based POS Tagger for Morphologically Rich Language

No Author Given

No Institute Given

Abstract. In this paper we present a dictionary based part of speech (POS) tagger for Assamese, an inflectional, relatively free word order Indic language. The main contribution of this paper is a POS tagger based on linguistic rules, that may work well morphologically rich languages with a strong case marking system. We have obtained an overall accuracy of 90%.

1 Introduction

Part of speech (POS) tagging is one of the main steps of any natural language processing task. It is a process of automatically assigning accurate part of speech tags to each word of a sentence. Though there are a number of methods to POS tagging, the set of POS tags themselves are language dependent as each language has its own distinct characteristics. Two factors determine the syntactic category of a word. The first is lexical information directly related to the category of the word and the other is the contextual information related to the environment of the word. [1] classified all POS tagging algorithms into three basic categories, viz., rule based, stochastic, and hybrid. Most taggers have originally been developed for English and later adapted to other languages.

Among Indo-Aryan languages, Sanskrit is a purely free word order language [2], but some other Indo-Aryan languages like Hindi, Bengali and Assamese have partially lost the free word ordering in the course of evolution. In fixed word order languages, position plays an important role in identifying the word category whereas this is not true for relatively free word order languages. Most Indian languages are morphologically rich, inflection being pre-dominant. In this report we use Assamese as the target language for all our experiments.

In the next section, we describe prior work in POS-tagging of morphologically rich languages. Section 3 describes some relevant linguistic characteristics of Assamese. We describe available POS tagsets for Assamese in Section 4. In sections 5 and 6, we describe our approach and experimental results, respectively. Section 7 discusses evaluation metrics for the tagger and Section 8 concludes our paper.

2 Literature Survey

Techniques for building POS taggers fall under two broad approaches, supervised and unsupervised. Both supervised and unsupervised tagging can be of three sub-types.

They are rule based, stochastic and neural network based. Each of these methods has its own pros and cons. During the last two decades, many different types of taggers have been developed, especially for corpus rich languages such as English and Turkish. In this paper, we are interested in dictionary based POS tagging. [3] developed a dictionary based morphology driven POS tagger for five morphologically rich languages Romanian, Czech, Estonian, Hungarian, and Slovene and concluded that an approach based on morphological dictionaries is a better choice for inflectionally rich languages. [4] reported a morphology driven rule based POS tagger for Turkish, using a combination of handcrafted rules and statistical learning. [5] reported a hybrid morphology based POS tagger for Persian where they combine the features of probabilistic and rule-based taggers to tag Persian unknown words.

Due to relative free word order, agglutinative nature, lack of resources and the general lateness in entering the computational linguistics field, reported tagger development work on Indian languages is relatively scanty. Among published works, Dandapat [6] developed a hybrid model of POS tagging by combining both supervised and unsupervised stochastic techniques. Avinesh and Karthik [7] used conditional random fields (CRF) and transformation based learning. The heart of the system developed by Singh et al. [8] for Hindi was the detailed linguistic analysis of morpho-syntactic phenomena. Saha et al. [9] developed a system for machine assisted POS tagging of Bangla corpora. Pammi and Prahllad [10] developed a POS tagger and chunker using decision forests. This work explored different methods for POS tagging of Indian languages using sub-words as units. [11] tried out a morphology driven POS tagger for Manipuri language with accuracy 65% for single tagged correct words. We have only one reported evidence of supervised part of speech tagging for Assamese [12] with accuracy nearly 87%.

3 Linguistic Characteristics of Assamese

Though Assamese is relatively free word order, predominant word order is SOV (subject-object-verb). In Assamese, secondary forms of words are formed through affixation (inflection and derivation), and compounding. Affixes play a very important role in word formation. Affixes are used in the formation of relational nouns and pronouns, and in the inflection of verbs with respect to number, person, tense, aspect and mood. For example, Table 1 shows how a relational noun দেউতা (*deutA*: father) is inflected depending on number and person.

There are 5 tenses in Assamese, namely, present, past, future, present perfect and past perfect tense [13]. Besides these, every root verb changes with case, tense, person. In Table 2 we present some possible forms of the root verb কৰ (*kr*: to do). The following paragraphs describe just a few of many characteristics of Assamese text that make the tagging task complex.

- Suffixation of nouns is very extensive in Assamese. There are more than 100 suffixes for the Assamese noun. These are mostly placed singly, but sometimes in sequence after the root word.
- We need special care for honorific particles like ডাঙ্গৰীয়া *dAngrIyA*. Assamese and other Indian languages have a practice of adding particles such as দেউ *deu*,

Person	**Singular**	**Plural**
প্ৰথম 1st	মোৰ দেউতা My father	আমাৰ দেউতা Our father
মান্য মধ্যম 2nd	তোমাৰ দেউতাৰা Your father	তোমালোকৰ দেউতাৰা Your father
তুচ্ছ মধ্যম 2nd, Familiar	তোৰ দেউতাৰ Your father	তহঁতৰ দেউতাৰ Your father
তৃতীয় 3rd	তাইৰ দেউতাক Her father	সিহঁতৰ দেউতাক Their father

Table 1. Personal definitives are inflected on person and number

কৰ	প্ৰথম	তুচ্ছ মধ্যম	মান্য মধ্যম	তৃতীয়
Present	কৰোঁ *karo*	কৰ *kar*	কৰক *karaka*	কৰা *karA*
Past	কৰিলোঁ *karilo*	কৰিলি *karili*	কৰিলে *karile*	কৰিলা *karilA*
Future	কৰিম *karim*	কৰিবি *karibi*	কৰিব *kariba*	কৰিবা *karibA*
Present P.	কৰিছোঁ *karicho*	কৰিছ *karicha*	কৰিছে *kariche*	কৰিছা *karichA*
Past P.	কৰিছিলোঁ *karichilo*	কৰিছিলি *karichili*	কৰিছিল *karichil*	কৰিছিলা *karichilA*
Causative	——	কৰাবা *karAbA*	কৰোঁৱাওক *karowAok*	কৰোঁৱা *karowA*

Table 2. Verbs are conjugated/inflected on person and number

ডাঙৰীয়া *dAngrIyA*, মহোদয় *mahodaya*, মহোদয়া *mahodayA*, মহাশয় *mahAsay*, মহাশয়া *mahAsayA*, etc., after proper nouns or personal pronouns. They are added to indicate respect to the person being addressed.

- Use of foreign words is also common in Assamese. Often such words are used along with regular suffixes of Assamese. Such foreign words will be tagged as per the syntactic function of the word in the sentence.
- Some prepositions or particles are used as suffix if they occur after nouns, personal pronouns or verbs. For example, সিহে গৈছিল। TF: *Sihe goisil.*

  Actually হে (*he*) is a particle, but it is merged with the personal pronoun সি (*si*).
- An affix denoting number, gender or person, can be added to an adjective or other category word to create a noun word. For example,

  ধুনীয়াজনী হৈ আহিছা।

  TF : *DhuniyAjoni hoi aAhisA.*

  Here ধুনীয়া (*dhuniyA*) is an adjective, but after adding feminine definitive জনী the whole constituent becomes a noun word. Table 3 shows some other examples of formation of derived words in Assamese.

Prefix	Stem	Suffix	Category	Example
-	NN	আ	VB	আঙুলিয়া
-	VB	আ	VB	কৰা
-	VB	অন	NN	চলন
-	VB	উৰা	ADJ	কান্দুৰা
-	NN	অলীয়া	ADJ	গাওঁলীয়া
-	ADJ	জনী	NN	ডাঙৰজনী
-	ADJ	কৈ	ADV	ডাঙৰকৈ
ন	VB	-	VB	নহয়
ন	VB	এ	VB	নকৰে
ন	NN	-	NN	নকৰে
অপ	NN	-	NN	অপশক্তি

Table 3. Formation of derivational words in Assamese

– Even conjunctions can be used as other parts of speech.
হৰি আৰু যদু ভায়েক ককায়েক।
TF : *Hari aAru Jadu bhAyek kokAyek.*
ET : Hari and Jadu are brothers.

যোৱাকালিৰ ঘটনাটোৱে বিষয়টোক আৰু অধিক ৰহস্যজনক কৰি তুলিলে।
TF : *JowAkAlir ghotonAtowe bishoitok aAru adhik rahashyajanak kori tulile.*
ET : The incident last night has made the matter more mysterious.

The word আৰু (*aAru*) shows ambiguity in these two sentences. In the first, it is used as conjunction and in the second, it is used as adjective of adjective.

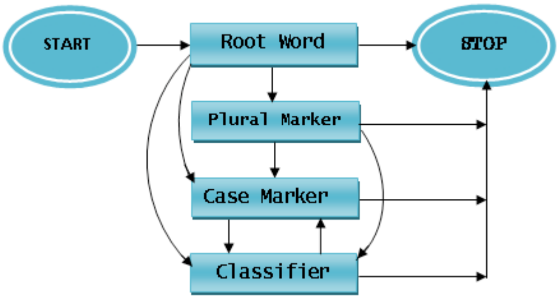


Fig. 1. Assamese noun inflection model

4 Assamese POS Tagset

Xobdo[1], an Assamese online dictionary project had developed a POS tagset[2] for all Northeast Indian languages. This tagset contains only 15 tags (Table 4). It groups all case endings, prefixes and suffixes into adposition. Though Assamese has a rich system of particles, Xobdo excludes particles other than the ones for interjection and conjunction. Another tagset[3] developed at Tezpur University solely for Assamese includes 172 tags. But this tagset is too large and has separate tags for general case markers as well as very specific noun case markers. For example the tag NCM is used for nominative case marker and CN1 and CNS1 are used for nominative singular common noun and nominative plural common noun, respectively. In this work, we follow the POS guidelines of the AnnCora (Bharati et al. 2006)[14], Penn treebank tagset[4] and MSRI-JNU Sanskrit tagset[5]. We use the same tags as in the Penn treebank when possible, so that they arel easily understandable to all annotators. The tags designed during this project for Assamese are shown in Table 5.

	Major POS	Minor POS
1	Noun	Common Noun
2		Proper Noun
3		Material Noun
4		Verbal Noun
5		Abstract Noun
6	Pronoun	-
7	Adjective	Proper Adjective
8		Verbal Adjective
9		Adjective of Adjective
10		Adverb
11	Verb	Transitive Verb
12		Intransitive Verb
13	Others	Ad-position
14		Interjection
15		Conjunction

Table 4. Xobdo's tagset

Our POS tagset covers the following lexical items:

1. Single word tokens: These are the common words in the vocabulary, e.g., nouns, verbs, adjectives, etc.

[1] http://xobdo.org

[2] http://xobdo.org/dic/help/pos.php

[3] http://tezu.ernet.in/¾nlp/posf.pdf

[4] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQP-HTMLDemo/PennTreebankTS.html

[5] http://sanskrit.jnu.ac.in/corpora/MSR-JNU-Sanskrit-Guidelines.htm

	Symbol	1st level	2nd level
1	NN	Noun	Common Proper Material Abstract Verbal Time Indicative Verb Indicative
2	PN	Pronoun	Personal Reflexive Reciprocal
3	VB	Verb	Main Auxiliary Causative
4	RB	Adverb	Time Location Manner
5	NOM	Nominal Modifier	Adjective Demonstrative Quantifier Pre-nominal
6	PAR	Particle	Conjuction Disjunction Exclamatory Vocative Particle
7	PSP	Post-position	Case Marker Classifier Plural Marker
8	QH	Question word	Interrogative Pronoun Interrogative Particle
9	RDP	Reduplication	Reduplicative Onomatopoetic Echo Word
10	NUM	Number	Cardinal Ordinal Date Time
11	SPS	Special symbol	
12	PUN	Punctuation	
13	UNK	Unknown word	

Table 5. Assamese hierarchical tagset

2. Named entities of various types: Names of people, topological items, titles of films, company names, scientific names, formulas, etc. Sometimes such lexical material is surrounded by inverted comma or brackets.

3. Compound word tokens .
4. Abbreviations.
5. Punctuations.

(1) and (2) above include items that belong to common dictionaries. (3) and (4) contain items that refer to real world entities and (5) contains text formatting items. The design of our annotation scheme does not rely only on linguistic assumptions of traditional Assamese grammar, but also on the output needed for further linguistic processing of data.

5 Our Approach

The declension of an Assamese noun is given in Table 1. Assamese verbs and nouns are open class word categories. Assamese pronoun, particle, adjective and adverb classes are small and closed. So we can easily tag them. Assamese nouns are inflected with case markers (CM), plural markers (PM) and classifiers (CL). See examples below.

1. মানুহৰ = মানুহ [root] + ৰ [CM]
2. মানুহবোৰ = মানুহ [root] + বোৰ [PM]
3. মানুহটো = মানুহ [root] + টো [CL]
4. মানুহকলৈ = মানুহ [root] + ক [CM] + লৈ [CM]
5. মানুহৰহে = মানুহ [root] + ৰ [CM] + হে [CL]
6. মানুহবোৰৰ = মানুহ [root] + বোৰ [PM] + ৰ [CM]
7. মানুহবোৰৰহে = মানুহ [root] + বোৰ [PM] + ৰ [CM] + হে [CL]

5.1 Corpus

We use a part of the EMILLE Assamese text corpus of nearly 2.6M words jointly developed by Lancaster University and CIIL-Mysore. Though the texts are in Unicode format, it required a lot of preprocessing. Here are some examples of errors we had to correct in the Unicodified EMILLE corpus.

1. Bengali *ra* occurs in the corpus instead of Assamese ৰ.
2. য় occurs where য should and vice versa.
3. An unrecognized character occurs where Assamese ৱ should.
4. If second character of a conjunct is ব *ba*, then it disappears.
5. There are many patternless spelling mistakes.

We corrected as many errors as possible errors in the texts programatically and by extensive manual checking.

5.2 Pre-processing

In this phase, we tokenized our corpus. For tokenizing we consider white space as word separator and and a punctuation symbol (।,?,!) as sentence terminator. Some problems we face during preparation of the corpus are listed below.

1. Detecting boundaries for some words such as নতুনপাৰা and নতুন পাৰা. Many place names are written in two ways, sometime with space and sometime without space.
2. Irregularities in placing hyphens with reduplicative words: Reduplication is a special phenomenon in most Indian languages. Here, either the same word is written twice for indicating emphasis, deriving a category from another category (for example, কেৰ-কেৰ); or some nonsense word is used after a regular lexical word indicating the sense ‘etc’ (for example, মাছ-তাছ); or some tonally similar lexical word is used after a regular lexical word (for example, কেৰ-মেৰ). Sometimes hyphens are used between the two tokens and sometimes, not.
3. Sometimes adverbs such as লাহে-লাহে are written as লাহে লাহে, that is, without hyphens.
4. Phrases such as গাই বাই ছা are sometime written as গাই-বাই-ছা. The complete phrase is considered as a single token, if it is written as গাই-বাই-ছা.

5.3 Dictionary

In our dictionary file, we store words, corresponding tags and whether they are inflected. A word is not inflected means the word is a root word. The simple way is to search the corpus for a dictionary word and its all possible combinations of affixes. Here we assume that all inflected dictionary words are 3 or more characters long. We can get all possible suffix sequence information for noun from Figure 1. Similarly, suffix sequence for verbs can also be determined. A Java module searches all possible suffix sequences of a dictionary word and tag them.

Words	Number of entries
Prefix	25
Pronoun	89
Suffix	110
Particle	162
Adverb	392
Verb	881
Adjective	3942
Noun	4855
Total	10456

Table 6. Dictionary Information. 20 Assamese prefixes originate from Sanskrit, and other 5 prefixes are of native origin [15]

We store 102 suffixes for noun and 27 suffixes for verb. Our dictionary file contains only 10456 entries with tags. The dictionary is used primarily for reducing ambiguity. In Assamese, most words less than 4 characters long have more than one meaning [16]. Our dictionary stores all root words and their corresponding tags, and all words which show ambiguity at word level and their corresponding tags. We maintain another file of suffixes that inflect nouns and verbs.

Algorithm 1: Algorithm for dictionary based POS tagging.

```
Input: A dictionary file, a suffix file and a corpus crps
Output: Tagged Corpus
Read dictionary and suffix file and store it in separate array
for Each token in the corpus crps do
    if the token is in dictionary file then
        Tag it with corresponding tag against the dictionary element.
    end
    else if The token ends with any element of suffix file then
        Tag the token with corresponding tag against the suffix.
    end
    else if The token starts with any element of prefix file then
        Tag the token with corresponding tag against the prefix.
    end
    Check whether tagged token satisfies the handcrafted rules or not.
end
```

Our tagging algorithm is described as Algorithm 1. We used Java to implement this algorithm. The results obtained are shown in Table 7. To resolve ambiguities such as noun-adjective ambiguity, noun-verb ambiguity, and adjective-adverb ambiguity, we used a simple rule base. Some of the rules we use are listed below.

1. Adverbs always precede verbs and adjectives precede nouns.
2. Words ending with কৈ are generally adverbs.
3. Words ending with plural markers or definitives are always noun.
4. Except single constituent sentences, particles do not occur in the initial position of a sentence.

The strength of our approach is based on affix information regarding words and categories of root words. We resolve ambiguity at the context level also. Suppose we get more than one tag for a specific token. In such a case, we check the previous token t_1 and using a handcrafted rule we mark the token t_2 and check the next token t_3. We backtrack to t_2 to determine whether it is correct considering the tag on t_3. To some extent, this simple procedure covers contextual information also. However, a problem will arise if t_3 has also more than one tag.

6 Results

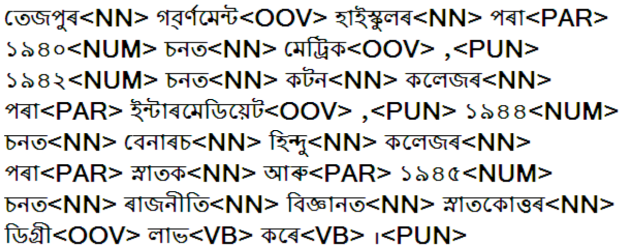
The results obtained are given in Table 7. In our corpus file there are 190897 numbers of sentence and total 1560677 words.

6.1 Handling OOV

The information of morphological features and contextual features are used to resolve the OOV. Morphological features like affix informations and other context rule determined the the tag against the word.

POS Tag	Number	Correct	Accuracy
NN	148829	128670	86.45
PN	14662	14654	99.94
NOM	26916	26184	97.28
RB	3813	3802	99.71
VB	10585	9554	90.25
NUM	1549	1549	100
PAR	17711	17676	99.80
PUN	37954	37954	1
PSP	27	27	100
Other	54	54	100
Total Sentences		190897	
Total Words		1560677	
OOV		205385	

Table 7. Obtained results



```
তেজপুৰ<NN> গৱৰ্ণমেন্ট<OOV> হাইস্কুলৰ<NN> পৰা<PAR>
১৯৪০<NUM> চনত<NN> মেট্ৰিক<OOV> ,<PUN>
১৯৪২<NUM> চনত<NN> কটন<NN> কলেজৰ<NN>
পৰা<PAR> ইন্টাৰমেডিয়েট<OOV> ,<PUN> ১৯৪৪<NUM>
চনত<NN> বেনাৰচ<NN> হিন্দু<NN> কলেজৰ<NN>
পৰা<PAR> স্নাতক<NN> আৰু<PAR> ১৯৪৫<NUM>
চনত<NN> ৰাজনীতি<NN> বিজ্ঞানত<NN> স্নাতকোত্তৰ<NN>
ডিগ্ৰী<OOV> লাভ<VB> কৰে<VB> ।<PUN>
```

Fig. 2. Example Output

Let us consider the output text in Figure 2. Here four words, viz., গৰ্ৱণমেন্ট (*government*), মেট্ৰিক (*metric*), ইন্টাৰমেডিয়েট (*intermediate*) and ডিগ্ৰী (*degree*) are marked as OOV. All the four words are English words written in Assamese script without a morphological marker. Therefore the algorithm cannot detect the category and mark them as unknown words. In the same figure, the word কলেজৰ is also a English word written in Assamese script and marked as noun because the English word কলেজ (*college*) is associated with the genitive case marker ৰ. A manual verification of the tagged text has been carried out after the tagging process is over. Table **??** summarizes the results obtained and verified corrected result is given in Table 7.

7 Evaluation and Discussion

For calculating *precision* and *recall*, the formulas are as follows.

$$Precision = \frac{Number\ of\ tagged\ words}{Number\ of\ total\ words}$$

$$Recall = \frac{Number\ of\ correctly\ tagged\ word}{Number\ of\ tagged\ words}$$

To combine *precision* and *recall* into a single measure of over all performance, we can measure the *F-measure* is as follows-

$$F\text{-}measure = \frac{2 * P * R}{P + R}$$

Table **??** shows the *precision*, *recall* and *F-measure* values. As this is the first step towards developing a rule based approach for Assamese POS-tagging, we intend to investigate if modifying some of our rules or creating additional rules increases the performance of the tagger.

Author	Language	Accuracy
[4]	Turkish	98%
[11]	Manipuri	69%
[3]	Romanian, Czech, Hungarian Estonian, Slovene	94.23%
[17]	Polish	88%
Ours	Assamese	90%

Table 8. Compared result with other dictionary based works.

We compare our result with published results in other languages in Table 8. There were aproximately 24K words in the lexicon in the work of [4] whereas [11] use only 2.1K root words in the dictionary file. As mentioned above we use a lexicon of size 10.5K in the dictionary and obtain 90% accuracy. [3] repoted 3.72% error rate for Czech, 8.20% for Estonian, 5.64% for Hungarian, 5.04% for Romanian and 5.12% for Slovene.

8 Conclusion

From our experiments, we observe that dictionary based tagging gives promising results for a morphologically rich Indic language. The F-measure values obtained are nearly 90%. There is hardly any other reported work on POS tagging of Assamese. Hence our work assumes significance. We strongly feel that this approach combined with techniques such as HMM, ME, CRF, etc., will produce even better results.

References

1. Jurafsky, D., Martin, J.H.: SPEECH and LANGUAGE PROCESSING, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Education (2000)
2. Ray, P.R., V., H., Sarkar, S., Basu, A.: Part of speech taggging and local word grouping techniques for natural language parsing in Hindi
3. Hajič, J.: Morphological tagging: Data vs. dictionaries. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. (2000)

4. Oflazer, K., Kuruöz, I.: Tagging and morphological disambiguation of Turkish text. In: Proceedings of 4th Conference on ANLP. (1994)
5. Shamsfard, M., Fadaee, H.: A hybrid morphology-based pos tagger for Persian. In: Proceedings of the the International Conference on Language Resources and Evaluation. (2008)
6. Dandapat, S.: Part-of-speech tagging and chunking with maximum entropy model. In: Proceedings of Workshop on Shallow Parsing for South Asian Languages (SPSAL). (2007)
7. PVS, A., G, K.: Part-of-speech tagging and chunking using Conditional Random Field and Transformation based learning. In: Proceedings of IJCAI-07 workshop on Shallow Parsing for South Asian Languages (SPSAL). (2007)
8. Singh, S., Gupta, K., Shrivastava, M., Bhattacharyya, P.: Morphological richness offsets resource demand-experiences in constructing a POS tagger for Hindi. In: In Proceedings of the COLING/ACL on Main conference poster. (2006)
9. Saha, G.K., Saha, A.B., Debnath, S.: Computer assisted Bangla words POS tagging. In: Proc. International Symposium on Machine Translation NLP & TSS. (2004)
10. Pammi, S.C., Prahallad, K.: POS tagging and chunking using Decision Forests. In: Proceedings of Workshop on Shallow Parsing for South Asian Languages (SPSAL). (2007)
11. Singh, T.D., Bandyopadhyay, S.: Morphology driven Manipuri pos tagger. In: Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. (2008)
12. Saharia, N., Das, D., Sharma, U., Kalita, J.: Part-of-Speech Tagger for Assamese Text. In: Proceedings of ACL/IJCNLP 2009. (2009)
13. Goswami, G.C.: Asamiya Vyakaran Pravesh. Bina Library (2000)
14. Bharati, A., Chaitanya, V., Sangal, R.: Natural Language Processing: A Paninian Perspective. Prentice-Hall, India (1993)
15. Goswami, U.: Asamiya Bhashar Vyakaran. Mani Manik Prakash (2001)
16. Sharma, U.: Unsupervised Learning of Morphology of A Highly Inflectional Language. PhD thesis, Tezpur University (2007)
17. Galus, S.: Dictionary-based part-of-speech tagging of polish. In: Intelligent Information Processing and Web Mining, Springer Berlin / Heidelberg (2005)