A Suffix-based Noun and Verb Classifier for an Inflectional Language

Navanath Saharia
Department of CSE
Tezpur University
Assam, India 784028
nava.nath@yahoo.in

Utpal Sharma
Department of CSE
Tezpur University
Assam, India 784028
utpal@tezu.ernet.in

Jugal Kalita
Department of CS
University of Colorado
Colorado Springs, USA 80918
kalita@eas.ucsc.edu

Abstract—Nouns and verbs pose the major challenge in part-of-speech tagging exercises. In this paper we present a suffix based noun and verb classifier for Assamese, an inflectional, relatively free word order Indic language. We used a tiny dictionary of frequent words to increase the accuracy. We obtained F-score of around 85%.

Keywords—Assamese, Noun, Verb, POS tagging.

I. INTRODUCTION

In any language, nouns and verbs are the most crucial parts of a sentence. The noun class is always an open lexical category; its members can occur as the head word in the subject of a clause, the object of a verb, or the object of a preposition. In this paper we discuss our research towards part-of-speech (POS) tagging, an important step in any natural language processing task, where the goal is to automatically assign lexical category to each lexical object occurring in a given text. For any thorough study in computational linguistics, we need to POS tag each and every word of a sentence. Two factors determine the syntactic category of a word. The first is lexical information that is directly related to the category of the word, and other is contextual information related to the environment of the word. In this paper we mainly focus on the properties of a lexicon that may help in POS tagging. We choose Assamese, a morphologically rich inflectional Indic language, for the experiments. Spoken by about 30 million people, Assamese is the lingua-franca of north-eastern region of India, and the official language of the state of Assam.

This paper is organized as follows. In section II we give a brief survey of literature related to this work, the state-of-the-art and some linguistic characteristics of Assamese. In section III we present our methodology and a brief description of the corpora used. In section IV we report our experimental results and a discussion thereof. Section V concludes the paper with hints of future work.

II. LITERATURE SURVEY

A. Existing work

In [1] all POS tagging algorithms are categorised into three basic categories- rule based, stochastic, and hybrid.

Grammatical and morphological rules are defined for rule based POS tagger. Most taggers, either rule based, stochastic or hybrid, are initially developed for English, and afterwards adapted to other languages. Brill's tagger [2], is a widely discussed linguistically motivated rule based POS tagger for English. In the two stage architecture of Brill's tagger, in the first stage the input tokens are initially tagged with their most likely tags; after that lexical rules are employed to assign tags to unknown tokens. On the other hand, TnT [3], a widely discussed statistical POS tagger based on a second order Markov model, was developed for English and German. It calculates the lexical probabilities of unknown words based on their suffixes. Comparison between statistical and linguistic rule based taggers shows that for the same amount of remaining ambiguity, the error rate of a statistical tagger is one order of magnitude greater than that of the rule based one [4]. The taggers described above are specifically designed for relatively fixed word order languages, where position of the word plays an important role. For relatively free word order languages, Dincer et. al [5] described a suffix based POS tagging approach for Turkish. They use the well-known Hidden Markov Model with a closed lexicon that consists of a fixed number of letters from word endings, and obtained accuracy 90.2%.

Indian languages are highly inflectional, morphologically rich and relatively free word order. Morphological richness and free word order nature make morphological analysis a crucial task in tagging of Indian language texts. While in some Indic languages such as Assamese, most case markers occur as suffixes, in others such as Hindi they occur as separate words, leading to local word grouping. Beyond that Indic languages have similar degree of free word order. Table I shows some of the reported POS taggers for Indian languages.

B. Assamese noun and verb morphology

As Assamese nouns and verbs are open lexical categories, if we can tag words in these classes correctly, tagging the remaining words in a text will be facilitated. In this work, we consider only the morpho-syntactic properties of Assamese words. Assamese words can be categorized into inflected classes (noun, pronoun, adjective and verb) and un-inflected classes (adverb and particle). Among inflected classes two
main types of inflection are noun inflection and verb inflection.

C. Noun inflection

The inflection model of the noun in Assamese is depicted in Figure 1. Noun inflection represents gender, number and case in Assamese. For example, nouns ending with -ja (jan) and -ja (jan) are identified as masculine and feminine nouns, respectively. All rules applied to noun inflection can be applied to pronouns, adjectives and even numerals. Table II and Table III show formation of compound and derivational words, respectively. Most compounds in Assamese are noun; although other forms are also not rare. Derivation takes place for suffixes, prefixes, or a combinations of both. The base for derivation can be a simple word or a compound word. Only suffixes can change the word category, prefixes do not change the category of a word [cf. Table III].

![Assamese noun inflection model](image)

**Fig. 1. Assamese noun inflection model**

D. Verb inflection

Assamese verbs are inflected with tense, aspect and modality (TAM). Traditionally, Assamese verbs are categorised as either finite or non-finite. Verb roots are in non-finite form for which tense, person or grammatical markers are added. In comparison to nouns, Assamese verb inflection is complex. [13], [14] reported 520 inflectional forms for root verb ड़ (bh : to sit). Table IV shows some inflectional forms of verb ड़ (kr : to do). An Assamese verb conjugator is available in [http://www.tezu.ernet.in/~nlp/res.htm](http://www.tezu.ernet.in/~nlp/res.htm). Table IV summarizes of some inflectional form of verb ड़ (kr : to do) and Table V shows some suffixes and their categories with example.

### III. OUR APPROACH

We use a part of the EMILLE Assamese text corpus1 (3500 sentences), jointly developed by Lancaster University and CIIL-Mysore. We tokenized our corpus as far as possible considering white space as word separator and punctuations (.,!!) as sentence terminator. Some examples of the difficulties in this task are given below.

1) Quite frequently the same place name is written in two ways, such as नतुपड़ा and नन पपाद. पा (paA) and नन (bArI) are among most popular Assamese suffixes placed after village or neighbourhood names. In नन पपाद, नन is considered separate adjective word, which qualifies पा whereas actually नतुपड़ा कर when used as a single word is a noun. Thus irregularities in placing white space and hyphen make tokenizing process a complex job.

2) Foreign words are very commonly used in Assamese, especially in news reports and scientific or technical writing. Foreign words take Assamese suffix. Based on the category of suffix we can identify the foreign words. For Example- U.G.C.’ or U.G.C.- (of U.G.C). This makes the number of OOV words high.

In our method, we follow the following three basic steps to tagged tokenized text.

1) **Brute-force determination of suffix sequences:** In this step, we obtain all possible sequences of noun suffixes following our model shown in Figure 1. Assamese nouns and pronouns take more than one suffix in a sequence, though not all suffix sequences are grammatically

---

1[http://www.emille.lancs.ac.uk/](http://www.emille.lancs.ac.uk/)
Aek-iCelAw, -iCil, -ib, person relational noun etc. person. For example in
In this step, we identify the noun + infix + mAnuhr vAeyk 2 etc. (etc.
+ etc.
jnI mAnuhebAr + eh Suffix stripping: mAnuhrjn etc. [15] reported that Assamese has the highest number of kinship
in -ebAr, -ht, -mKA, -esApA -jn, -jnI, -kN, -Kn + kkA + kiriCelAw (16x45) I krAbA (16x45) kiribecAn (16x45)
krk (16x45) kiriClA (16x45) Person (Familiar) kirbAecAn – (16x45) mAnuh kr : to do
is tagged as kerAwWA (16x45) Person (Respect) kirel
mAnuhebArr ‘(Nominative case kirb.
1 kirC kiriCelAw Person (kirelAw
VERB WITH RESPECT TO TENSE AND PERSON
noun among Indo-Aryan languages. Table VI shows some inflected form of
kkAeyk 3 rd 2
All relational nouns in Assamese are added with inflected form to form sequences. From this model we obtain rules of
suffix sequences. From this model we obtain rules of suffix sequences attached to a noun.
3) Suffix stripping: In this step, we identify the noun
and verb roots based on the single suffix that occurs immediately after the root. For example, if we found the
word mAnuhw, it will first identify the suffix sequence nAitnIeykekjnImAeneh etc. (16x45) krk 
kiriCl
Person (Respect)
krA)
: plural marker + genitive marker), which
kiriCl
(kiriCA
YA
gA
correct. For example
Noun+ inflected form of kinship noun and infixed feminine marker + plural marker + nominative case marker + particle.
We can obtain all possible sequences of noun suffixes from Figure 1. Some suffixes are always used with words
from a small class of roots. For example the suffix -Ah (aKhi) is always placed after kal : banana) that is
due. So in the next step, i.e., in sequence pruning we try to minimize the search space.
2) Suffix sequence pruning: In this step we filter out the non-valid suffix sequences from among all the sequences
obtained in Step 1. Though a number of suffix sequences can possibly occur after a root word, we usually
find only 3 suffix sequence in the corpus we studied, though there are exceptions. All suffix sequences are not
valid. So if we list all the valid suffix sequences beforehand using our linguistic knowledge, we need not
go through all possible combinations of the suffixes. A Java module is employed to prune the suffix sequence.
Figure 1 depicts a model to order legal noun suffixes to form sequences. From this model we obtain rules of
suffix sequences attached to a noun.

TABLE IV
EXAMPLE OF SUFFIXES WITH CATEGORIES IN ASSAMESE

| Case Marker | - Nb, - Nt, - Nt, - Nt, etc. | मानकाय |
| Plural suffix | -DAl, -pAt -eTA, -TA |
| Classifiers | -er, -Er, -rprA |
| Verbal suffix | -m, -aA |

TABLE V
SOME INFLECTIONAL FORM OF kr VERB WITH RESPECT TO TENSE AND PERSON.

<table>
<thead>
<tr>
<th>Present</th>
<th>1st Person</th>
<th>2nd Person (Familiar)</th>
<th>2nd Person (Respect)</th>
<th>3rd Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>किर (kari)</td>
<td>किर (kar)</td>
<td>किरिल (karile)</td>
<td>किरिल (karile)</td>
</tr>
<tr>
<td>Past</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
</tr>
<tr>
<td>Future</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
</tr>
<tr>
<td>Present Perfect</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
</tr>
<tr>
<td>Past Perfect</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
</tr>
<tr>
<td>Causative</td>
<td>-</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
</tr>
<tr>
<td>Future Conditional</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
<td>किरिल (karilo)</td>
</tr>
</tbody>
</table>

IV. RESULTS AND DISCUSSION
We are not able to categories some most frequent verbs with the method describe above. For example the inflectional form of या (ja : to go) verb is not the same as the inflectional form of गा (ga : to sing) verb. Therefore to increase the accuracy we add 300 most frequently used verb root words in the form of a tiny dictionary. Table VIII shows obtained results. We mentioned above that the principle applied to noun can also be applied to pronoun and adjective. The inflected pronouns and adjectives in the corpus are tagged as nouns. As the result the tagging accuracy of nouns comes down. If we can embed the contextual information in our method, we hope, it will help increase POS tagging accuracy.

Ours is one of the earliest work on Assamese. The performance values of the few earlier works are not available.

V. CONCLUSION
We have implemented a suffix based noun and verb tagging approach for Assamese. We find that the performance of this method is better than stochastic approaches, such as HMM technique should be useful where required linguistic knowledge is available, but resources to prepare a large tagged corpus for training are not available. It will be interesting to compare results of this approach with those of stochastic approaches for other inflectional languages.

As the word order of Assamese is relatively free, we can not use positional information like in fixed word order languages. So a morpho-syntactic approach gives better results in comparison to [6], [10]. Another important observation from this experiment is that though Assamese is relatively free word order, some parts of speech do not occur in the initial or final positions of the sentence. As a future work we will try to
TABLE VI
Some Inflected Form of Assamese Kinship Noun with Person.

<table>
<thead>
<tr>
<th>Kinship Noun</th>
<th>1st Person</th>
<th>2nd Person (Familiar)</th>
<th>2nd Person (Respect)</th>
<th>3rd Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>edtA (deutA: father)</td>
<td>edtA (deutA)</td>
<td>edtAr (deutAr)</td>
<td>edtArA (deutArA)</td>
<td>edtAk (deutAk)</td>
</tr>
<tr>
<td>bhAi (deutA: younger brother)</td>
<td>bhAi</td>
<td>bhAyer (deutAyer)</td>
<td>bhAyerA (deutAyerA)</td>
<td>bhAyk (deutAyk)</td>
</tr>
<tr>
<td>sAhu (deutA: mother in law)</td>
<td>sAhu</td>
<td>sAhuWer (deutAhuWer)</td>
<td>sAhuwerA (deutAhuwerA)</td>
<td>sAhuwek (deutAhuwek)</td>
</tr>
<tr>
<td>kakA (deutA: grand father)</td>
<td>kakA</td>
<td>kakAr (deutAr)</td>
<td>kakArA (deutArA)</td>
<td>kakAk (deutAk)</td>
</tr>
</tbody>
</table>

TABLE VII
Statistics of Used Corpus.

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Total tokens</th>
<th>Total nouns</th>
<th>Total verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>5300</td>
<td>48027</td>
<td>28923</td>
<td>2429</td>
</tr>
</tbody>
</table>

TABLE VIII
Precision, Recall and F-Measure of Our Approach.

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>0.87</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>Verb</td>
<td>0.91</td>
<td>0.87</td>
<td>0.88</td>
</tr>
</tbody>
</table>

embed the linguistics word agreement rules in tagging nouns and verbs in Assamese.

REFERENCES


