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.uccs.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 is 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

The Department of Computer science & Engineering, Tezpur University is funded by university grant commission (UGC)'s departmental research support (DRS) Phase-I under special assistance programme (SAP).

978-0-7695-4288-1/10 $26.00 © 2010 IEEE
DOI 10.1109/IALP.2010.64

TABLE I
REPORTED POS TAGGING RESULTS WITH INDIAN LANGUAGES

Author	Approach	Language	Accuracy
[6]	HMM	Bengali	84.37%
[7]	CRF	Hindi Bengali Telegu	70.67% 65.47% 65.85%
[8]	HMM	Hindi Bengali Telegu	69.98% 67.52% 68.32%
[9]	CRF	Hindi Bengali Telegu	71.65% 80.63% 53.15%
[10]	HMM	Assamese	85.64%
[11]	CRF HMM	Manipuri	72.04% 74.38%
[12]	Rule Based	Manipuri	69%

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 -জন (*jan*) and -জনী (*janI*) 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].

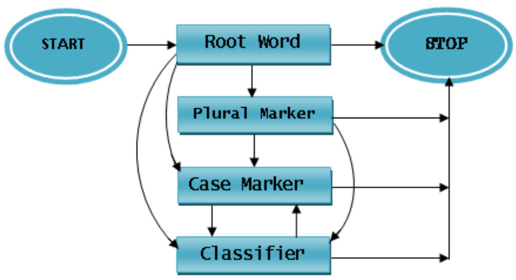

Fig. 1. Assamese noun inflection model

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 বহ (*bH : to sit*). Table IV shows some inflectional forms of verb কৰ (*kr : to do*). An Assamese verb conjugator

TABLE II
FORMATION OF COMPOUND WORD IN ASSAMESE

Stem(Category)	Stem(Category)	New word(Category)
কথা (NN) (*talk : kathA*)	ছবি (NN) (*picture : Cbi*)	কথাছবি (NN) (*cinema : kathACbi*)
কৃষ্ণ (NN) (*dark : krishna*)	পক্ষ (NN) (*fortnight : pakhya*)	কৃষ্ণপক্ষ (NN) (*dark fortnight : krishnapakhya*)

TABLE III
FORMATION OF DERIVATIONAL NOUN AND VERB IN ASSAMESE

Prefix	Root	Category	Suffix	New word	New category
-	আঙুলি	NN	আ	আঙুলিয়া	VB
-	কৰ	VB	আ	কৰা	VB
-	চল	VB	অন	চলন	NN
-	ডাঙৰ	ADJ	জনী	ডাঙৰজনী	NN
ন	হয়	VB	-	নহয়	VB
ন	কৰ	NN	এ	নকৰে	VB

is available in *http://www.tezu.ernet.in/∼nlp/res.htm*. Table IV summarizes of some inflectional form of verb কৰ (*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 corpus[1] (5300 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 নতুনপাৰা and নতুন পাৰা. পাৰা (*pArA*) and বাৰী (bArI) are among most popular Assamese suffixes placed after village or neighbourhood names. In নতুন পাৰা, নতুন is considered separate adjective word, which qualifies পাৰা whereas actually নতুনপাৰা 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.'ৰ or U.G.C.-ৰ (*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/

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

কৰ (*kr : to do*)	1^{st} Person	2^{nd} Person (Familiar)	2^{nd} Person (Respect)	3^{rd} Person
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 Perfect	কৰিছোঁ (*karicho*)	কৰিছ (*karicha*)	কৰিছে (*kariche*)	কৰিছা (*karichA*)
Past Perfect	কৰিছিলোঁ (*karichilo*)	কৰিছিলি (*karichili*)	কৰিছিল (*karichil*)	কৰিছিলা (*karichilA*)
Causative	–	কৰাবা (*karAbA*)	কৰোঁৱাওক (*karowAok*)	কৰোঁৱা (*karowA*)
Future Conditional	কৰিমচোন (*karimson*)	কৰিবিচোন (*karibison*)	কৰিবচোন (*kkaribason*)	কৰিবাচোন (*karibAson*)

TABLE V
EXAMPLE OF SUFFIXES WITH CATEGORIES IN ASSAMESE

Case Marker	-ক, -ৰ, -ত, -ই, এ, -ৰে, -লৈ, -ৰপৰা etc.	মানুহৰ
Plural suffix	-বোৰ, -হত, -মখা, -সোপা -সমুহ, -গন etc.	মানুহবোৰ
Classifiers	-জন, -জনী, -কণ, -খন -ডাল, -পাত -টো, -টা etc.	মানুহজন
Verbal suffix	-ইছিলোঁ, -ইছিলি, -ইবি, -ইম, -আ etc.	কৰিছিলোঁ

correct. For example
নাতিনীয়েককেইজনীমানেহে (*nAtinIyekkeijanImAnehe*) $\rightarrow$
নাতিনী + য়েক + কেই + জনী + মান + এ + হে
Noun+ inflected form of kinship noun[2] + infix + 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 -আখি (*aAkhi*) is always placed after কল (*kal : banana*) that is *kalaAkhi*. 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 most of 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.
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 মানুহবোৰৰ, it will first identify the suffix sequence বোৰৰ (বোৰ+ৰ: plural marker + genitive marker), which is a noun suffix and hence মানুহবোৰৰ is tagged as *genitive plural noun* and মানুহ as noun, in suffix free form. We find that the suffix 'এ' (Nominative case marker for noun and the endings of 2^{nd} person (familiar) past and present perfect tense marker) is the most ambiguous as it applies to both nouns and verbs.

All modules mentioned here are developed in Java. Table VII shows the statistics of test corpus.

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 যা (*jA : to go*) verb is not the same as the inflectional form of খা (*kha : to eat*) or গা (*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 a 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 batter 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

[2]All relational nouns in Assamese are added with inflected form য়েক in 3^{rd} person. For example in 3^{rd} person relational noun ভাই (*bhAi : younger brother*) is inflected to ভায়েক, ককাই (*kakAi : elder brother*) is inflected to ককায়েক etc. [15] reported that Assamese has the highest number of kinship noun among Indo-Aryan languages. Table VI shows some inflected form of Assamese kinship noun.

TABLE VI
SOME INFLECTED FORM OF ASSAMESE KINSHIP NOUN WITH PERSON.

Kinship Noun	1^{st} Person	2^{nd} Person (Familiar)	2^{nd} Person (Respect)	3^{rd} Person
দেউতা (*deutA : father*)	দেউতা (*deutA*)	দেউতাৰ (*deutAr*)	দেউতাৰা (*deutArA*)	দেউতাক (*deutAk*)
ভাই (*bhAi : younger brother*)	ভাই (*bhAi*)	ভায়েৰ (*bhAyer*)	ভায়েৰা (*bhAyerA*)	ভায়েক (*bhAyk*)
ককাই (*kakAi : elder brother*)	ককাই (*kakAi*)	ককায়েৰ (*kakAyer*)	ককায়েৰা (*kakAyerA*)	ককায়েক (*kakyek*)
শাহু (*sAhu : mother in law*)	শাহু (*sAhu*)	শাহুৱেৰ (*sAhuWer*)	শাহুৱেৰা (*sAhuwerA*)	শাহুৱেক (*sAhuwek*)
ককা (*kakA : grand father*)	ককা (*kakA*)	ককাৰ (*kakAr*)	ককাৰা (*kakArA*)	ককাক (*kakAk*)

TABLE VII
STATISTICS OF USED CORPUS.

Sentences	Total tokens	Total nouns	Total verbs
5300	48027	28923	2629

TABLE VIII
PRECISION, RECALL AND F-MEASURE OF OUR APPROACH.

Category	Precision	Recall	F-measure
Noun	0.87	0.79	0.82
Verb	0.91	0.87	0.88

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

REFERENCES

[1] Daniel Jurafsky and James H. Martin, *SPEECH and LANGUAGE PROCESSING, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Pearson Education, 2000.

[2] Eric Brill, "Transformation based error-driven learning and natural language processing: A case study in part of speech tagging," *Computational Linguistics*, 1995.

[3] Thorsten Brants, "Tnt - a statistical part-of-speech tagger," in *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, 2000.

[4] Christer Samuelsson and Atro Voutilainen, "Comparing a linguistic and a stochastic tagger," in *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. 1997, pp. 246–253, ACL.

[5] Taner DINCER, Bahar KARAOGLAN, and Tarik Kisla, "A suffix based part-of-speech for Turkish," in *Proceedings of 5th International conference on Information Technology: New Generations*, 2008.

[6] Sandipan Dandapat and Sudeshna Sarkar, "Part-of-speech tagging for Bengali with Hidden Markov Model," in *Proceedings of NLPAI ML workshop on Part of speech tagging and Chunking for Indian language*, 2006.

[7] Dilip Rao and David Yarowsky, "Part of speech tagging and shallow parsing of Indian languages," in *Proceedings of IJCAI-07 workshop on Shallow Parsing for South Asian Languages (SPSAL)*, 2007.

[8] G. M. Ravi Sastry, Sourish Chaudhuri, and P Nagender Reddy, "A HMM based part-of-speech and statistical chunker for 3 Indian languages," in *Proceedings of IJCAI-07 workshop on Shallow Parsing for South Asian Languages (SPSAL)*, 2007.

[9] Asif Ekbal, Samiran Mandal, and Sivaji Bandyopadhyay, "POS tagging using HMM and rule based chunking," in *Proceedings of Workshop on Shallow Parsing for South Asian Languages (SPSAL)*, 2007.

[10] Navanath Saharia, Dhrubajyoti Das, Utpal Sharma, and Jugal Kalita, "Part-of-Speech Tagger for Assamese Text," in *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, 2009.

[11] T. D. Singh and S. Bandyopadhyay, "Manipuri POS tagging using CRF and SVM: A language independent approach," in *Proceedings of International Conference on Natural Language Processing(ICON)*, 2008.

[12] T. D. Singh and S. Bandyopadhyay, "Morphology driven Manipuri POS tagger," in *Proceedings of IJCNLP-08 workshop on NLP for Less Privilege Languages*, 2008.

[13] Utpal Sharma, *Unsupervised Learning of Morphology of A Highly Inflectional Language*, Ph.D. thesis, Tezpur University, 2007.

[14] Utpal Sharma, Jugal K. Kalita, and Rajib K. Das, "Acquisition of morphology of an Indic language from text corpus," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 7, 2008.

[15] Lilabati Saikia Bora, *Asamiya Bhasar Ruptattva*, M/s Banalata, Guwahati, 2006.