Layered Parts of Speech Tagging for Bangla

Debasri Chakrabarti

CDAC, Pune

debasri.chakrabarti@gmail.com

Abstract-In Natural Language Processing, Parts-of-Speech tagging plays a vital role in text processing for any sort of language processing and understanding by machine. This paper proposes a rule based Parts-of-Speech tagger for Bangla with layered tagging. There are 4 levels of Tagging which also handles the tagging of Multi verb expressions.

I. Introduction

The significance of large annotated corpora is a widely known fact. It is an important tool for researchers in Machine Translation (MT), Information Retrieval (IR), Speech Processing and other related areas of Natural Language Processing (NLP). Parts-of-Speech (POS) tagging is the task of assigning each word in a sentence with its appropriate syntactic category called Parts-of-Speech. Annotated corpora are available for languages across the world, but the scenario for Indian languages is not the same.

In this paper I have discussed a rule based POS tagger for Bangla with different layer of tagging. The paper also shows how the layered tagging could help in achieving higher accuracy.

The rest of the paper is organized in the following way- Section 2 gives a brief overview of Bangla and the process of tagging with examples, Section 3 discusses layered POS Tagging and section 4 concludes the paper.

II. POS Tagging in Bangla

Bangla belongs to Eastern Indo-Aryan group, mainly spoken in West Bengal, parts of Tripura and Assam and Bangladesh. Bangla is the official language of West Bengal and Tripura and the national language of Bangladesh. It is a morphologically rich language, having a well-defined classifier system and at times show partial agglutination. In this section I propose a rule-based POS tagging for Bangla using context and morphological cue. The tag set are both from the common tag set for Indian Languages (Bhaskaran et al.) and IIIT Tag set guidelines (Akshar Bharti). For the top level following tags are taken as given in Table 1. This includes the 12 categories that are identified as the universal categories for the Indian languages from the common tag set framework.

Table 1

Top Level Tagging

	TAGSET	DESCRIPTION
1.	NN	Noun
2.	NNP	Proper Noun
3.	NUM	Number
4.	PRP	Pronoun
5.	VF	Verb finite
6.	VB	Verb Base
7.	VNF	Verb Nonfinite
8.	JJ	Adjective
9.	QF	Quantifier
10.	RB	Adverb
11.	PSP	Postposition
12.	PT	Particle
13.	NEG	Negative
14.	CC	Coordinating
15.	UH	Interjection
16.	UNK	unknown
17.	SYM	Symbol

After the top level annotation there is a second level of tagging. The tag sets are shown in Table 2.

Table 2

Second Level Tagging

	TAGSET	DESCRIPTION
1.	CM	Casemarker
2.	CL	Classifier
3.	CD	Cardinal
4.	CP	Complementizer
5.	DET	Determiner
6.	INTF	Intensifiers
7.	QW	Question Word
8.	SC	Subordinating Conjunction

A. Approaches to POS Tagging

POS tagging is typically achieved by rule-based systems, probabilistic data-driven systems, neural network systems or hybrid systems. For languages like English or French, hybrid taggers have been able to achieve success percentages above 98%. [Schulze et al, 1994]. The works available on Bangla POS Tagging are basically statistical based- Hidden Markov Model (HMM) [Ekbal et al.], Conditional Random Field (CRF) [Ekbal et al.], Maximum Entropy Model [Dandapat]. In this paper we talk about a Rule Based POS Tagger for Bangla. The aim is to proceed towards a hybrid POS Tagger for the language in future.

B. Steps to POS Tagging

The first step towards POS tagging is morphological analysis of the words. For this a Noun Analysis and a Verb Analysis had been done. Nouns are divided into three paradigms according to their endings, these three paradigms are further classified into two groups depending on the feature ± animate. The suffixes are then classified based on number, postposition and classifier information. Verbs are classified into 6 paradigms based on morphosyntactic alternation of the root. The suffixes are further analysed for person and honourofic information. Noun Analysis is shown in Table 1 and Verb Analysis is shown in Table 3.

Table 3

Noun Paradigm

Paradigm	No	Animate	Honourofic	Del Char	Classifier	Case	Form
chele 'boy'	Sg	+	+	0	-	Direct	chele 'boy'
chele 'boy'	Sg	+	+	0	Ti	Oblique	cheleTi 'boy'
chele 'boy'	PL	+	+	0	rA	Direct	cheleraa 'boys'
chele 'boy'	PL	+	+	0	der	Oblique	cheleder 'boys'
chele 'boy'	PL	+	-	0	gulo	Oblique	chelegulo 'boys'
phuul 'flower'	Sg	-	-	0	-	Direct	phuul 'flower'
phuul 'flower'	Sg	-	-	0	TA	Oblique	phuulTA 'flower'
phuul 'flower'	Sg	-	-	0	Ti	Oblique	phuulTi 'flower'
phuul 'flower'	PL	-	-	0	gulo	Direct	phuulgulo 'flowers'
phuul 'flower'	PL	-	-	0	gulo	Oblique	phuulgulo 'flowers'

Verb analysis based on Tense, Aspect, Modality, Person and Honouroficity (TAMPH) matrix is shown in Table 4.

Table 4

Verb Paradigm

Tense	Asp	Mod	Per	Hon	Eg.
Present	fct	-	1st	-	kor-i 'I do'
Present	fct	-	2nd	-	kar-o 'You do'
Present	fct	-	2nd	+	kar-un 'You (Hon) do'
Present	fct	-	3rd	-	kar-e 'He does'
Present	fct	-	3rd	+	kar-en 'He (Hon) does'
Past	Inf	-	2nd	-	kar-ar chilo 'was to be done'
Future	-	-	3rd	+	kor-be-n 'He (Hon) will do'
Present	Dur	-	3rd	-	kor-che 'He is doing'
Present	fct	Abl	3rd	-	kor-te pare 'He can do'

Based on this analysis a MA will return the following for the sentence *'ekjon chele boigulo diyeche'*

1. *ekjon (NN,CD) chele (NN) boigulo (NN) diyeche (VF)* 'A boy gave the books'

These are the simple tags that a MA can give. To reduce the ambiguity we need linguistic rules. The ambiguity here is between a Cardinal and a Noun. *ekjon* 'one' can be both- a Noun and a Cardinal. To resolve this sort of ambiguity following rule is given

> Noun vs. Cardinal: if the following word is a noun without a suffix and the token to be processed can qualify the succeeding noun, then the processing token is a cardinal, otherwise it is a noun. [eg. in *ekjon chele*, *ekjon* can be a cardinal or noun, but as it can qualify *chele*, and *chele* is without a suffix it will be an cardinal, not a noun]

The POS tagger will go through 3 stages. At the first stage preliminary tags will be assigned with the help of MA and disambiguating rules. Stage 2 will do a deeper level analysis and provide information like Classifier, TAMPH, Postposition etc. Stage 3 or final stage will run a local word grouper and give the noun group and verb group information. Fig.1. shows stage by stage output of the POS Tagger of the sentence *ekTi shundori meye nodir dhare daNRiye ache* 'One beautiful girl is standing on the bank of the river'

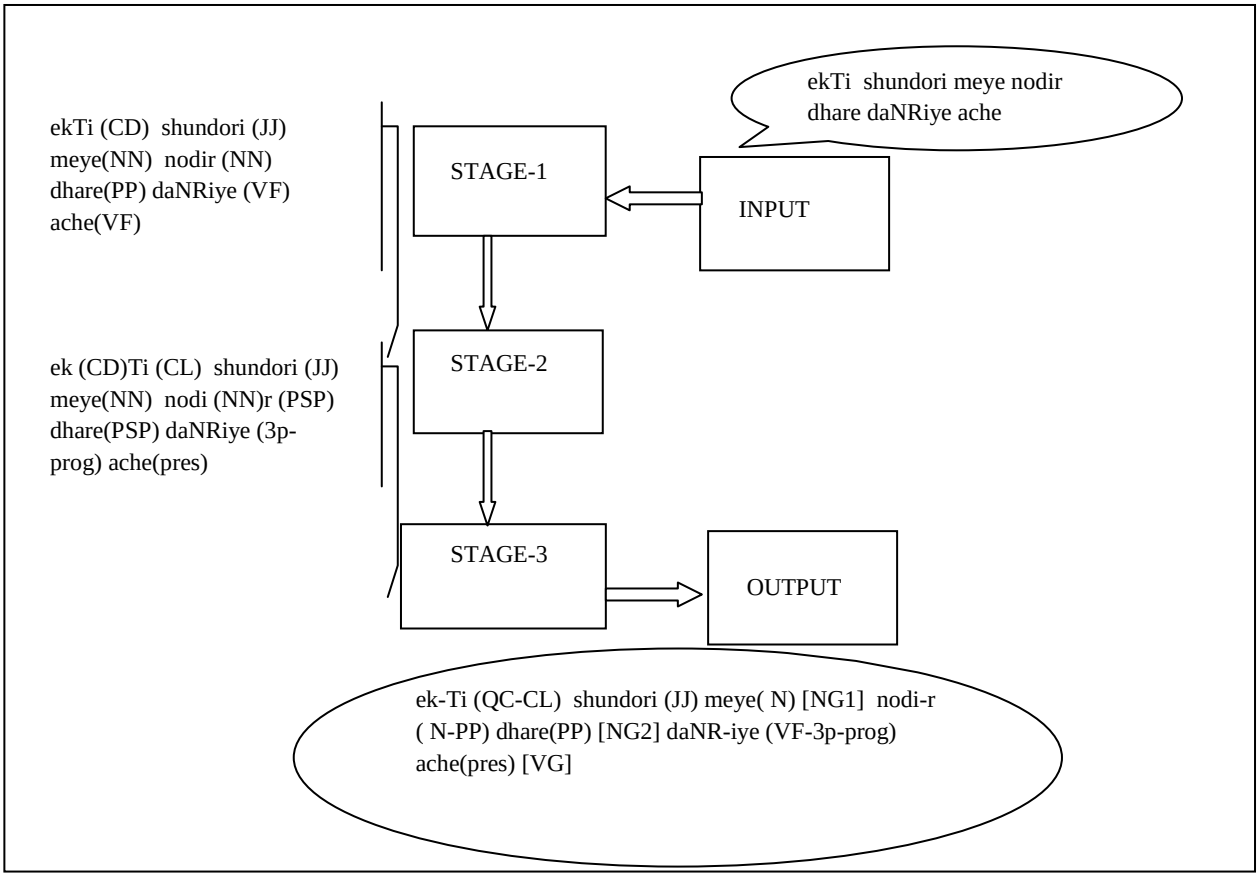

Fig. 1. Stages of POS Tagger

III. Handling Multi Verb Expressions

The POS Tagging process described in this paper till now will be able to tag and group simple verbs. Multi verb expressions (MVE) are not taken care here. MVEs are very frequent in South Asian Languages. These MVEs can be of two types-

a. Noun+Verb Combination, e.g., aarambha karaa 'to start'

b. Verb+Verb Combination e..g., kore phæla 'to do'

The former type of constructions is commonly known as Conjunct Verbs while the latter is called Compound Verb. The Tag set explained here does not include tags for this sort of combination. Therefore, examples like 2 and 3 will have the following tagging-

2. chelegulo kaajTaa aarambha koreche 'The boys started the work'

NN	NN	NN	VF
NN-CL	NN-CL	NN	VF-3p-pt.
[NG1]	[NG2]	[NG3]	[VG]

3. kaajTaa bandho hoyeche 'The work stopped'

NN	NN	VF
NN-CL	NN	VF-3p-pt.
[NG1]	[NG2]	[VG]

Both in 2 and 3 *aarambha koreche* 'started' and bandho *hoyeche* 'stopped' are instances of conjunct verbs. The information of conjunct verb is missing from the tagged output which is leading to a wrong verb group and Noun group identification. As of now both *aarambha* 'start' and bandho 'stop' are considered as Nouns and *koreche* 'do' and *hoyeche* 'happen' as verbs. Due to this the local word grouper

has grouped both *aarambha* 'start' and bandho 'stop' as [NG]. This will lead to wrong syntax affecting the accuracy of the system. To handle this sort of situation I suggest here to add one more layer of tagging before word grouping. The third level of tagging is shown in Table 5.

Table5.

Third Level Tagging

	TAGSET	DESCRIPTION
1.	CNJV	Conjunct Verb
2.	CPDV	Compound Verb

Fig. 2. shows the modified stage by stage output of the POS Tagger of the sentence *chelegulo kaajTaa aarambha koreche* 'The boys started the work'

IV. Conclusion and Future Work

In this paper I have discussed a rule based POS tagger for Bangla with layered tagging. There are four levels of Tagging. In the first level ambiguous basic category of a word is assigned. Disambiguation rules are applied in the second level with more detail morphological information. At the third level multi word verbs are tagged and the fourth or the final level is the level of local word grouping or chunking.

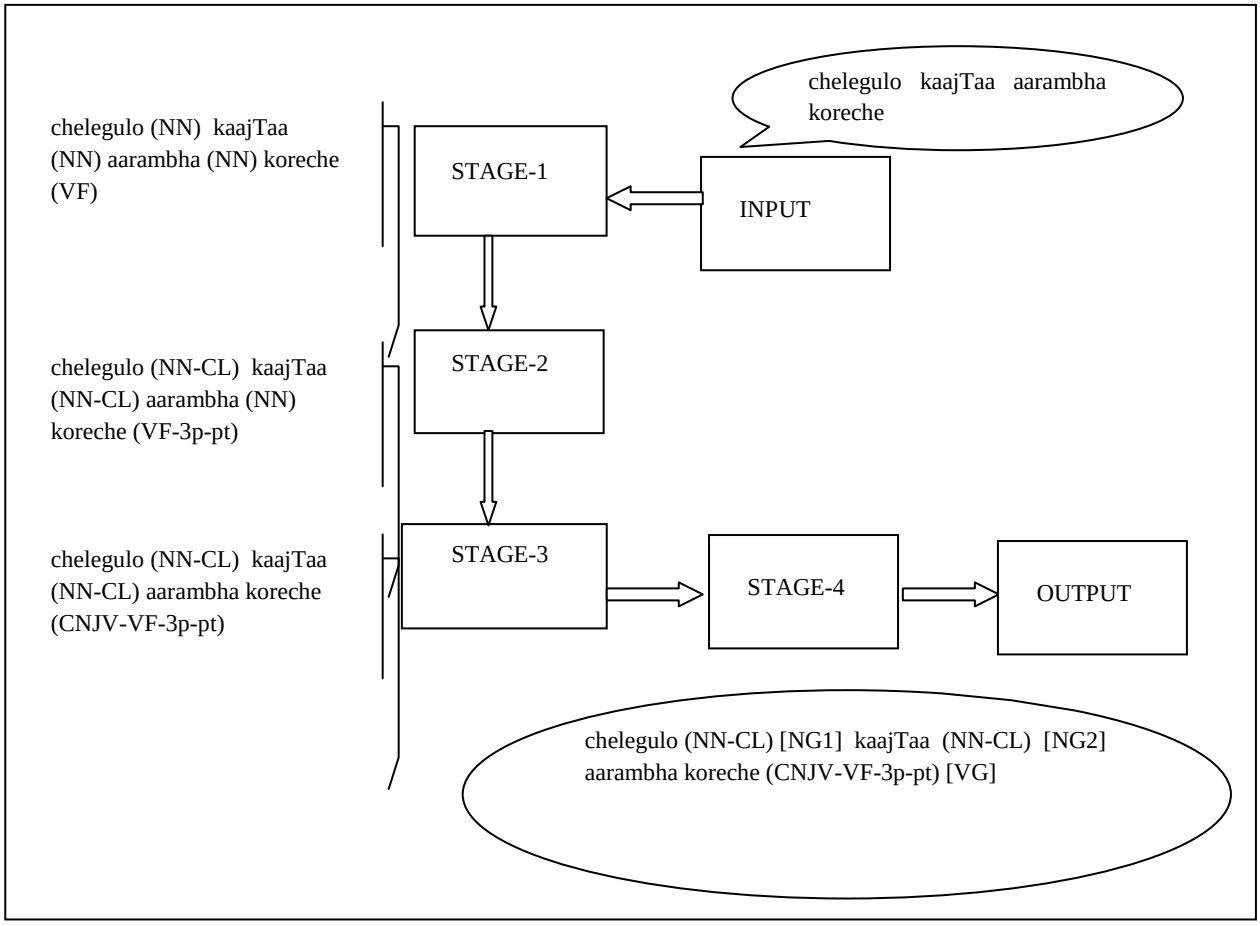

Fig. 2. Modified Stages of POS Tagger

REFERENCES

[1] Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma and Lakshmi Bai. 2006. AnnCorra:*Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages*, Technical Report, Language Technologies Research Centre IIIT, Hyderabad.

[2] ARONOFF, MARK. 1976. Word Formation in Generative Grammar. Cambridge: MA: MIT Press
SINCLAIR, J. 1991. Corpus, concordance, collocation. Tuscan Word Centre, Oxford: Oxford University Press

[3] ARONOFF, MARK. 2004. Developing Linguistic Corpora: A Guide to good practice. Oxford: Oxford University Press

[4] Banko, M., & Robert Moore, R. Part of speech tagging in context. 20th International Conference on Computational Linguistics. 2004

[5] Baskaran S. et al. Designing a Common POS-Tagset Framework for Indian Language. The 6th Workshop on Asian Language Resources. 2008

[6] Dandapat, S. Part-of-Speech Tagging and Chunking with Maximum Entropy Model. Workshop on Shallow Parsing for South Asian Languages. 2007.

[7] Dandapat, S., & Sarkar, S. Part-of-Speech Tagging for Bengali with Hidden Markov Model. NLPAI ML workshop on Part of speech tagging and Chunking for Indian language. 2006.

[8] Debasri Chakrabarti, Vaijayanthi M Sarma, Pushpak Bhattacharyya. Compound Verbs and their Automatic Extraction 22nd International Conference on Computational Linguistics, Manchester. 2008

[9] Debasri Chakrabarti, Vaijayanthi M Sarma, Pushpak Bhattacharyya. Identifying Compoud Verbs in Hindi. South Asian Language Analysis. 2006

[10] Ekbal, A., Mandal, S., & Bandyopadhyay, S. POS tagging using HMM and rule based chunking . Workshop on Shallow Parsing for South Asian Languages. 2007.

[11] IIIT-tagset. A Parts-of-Speech tagset for Indian languages. http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.

[12] Saha, G.K., Saha, A.B., & Debnath, S. Computer Assisted Bangla Words POS Tagging. Proc. International Symposium on Machine Translation NLP & TSS. 2004.

[13] Soma Paul. An HPSG Account of Bangla Compound Verbs with LKB Implementation, A Dissertation, CALT, University of Hyderabad, 2004.

[14] Smriti Singh, Kuhoo Gupta, Manish Shrivastava, and Pushpak Bhattacharyya. 2006. Morphological richness offsets resource demand – experiences in constructing a pos tagger for hindi In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 779–786, Sydney, Australia, July. Association for Computational Linguistics.