Parts of Speech Tagging: *a hybrid approach*

Knowledge Sharing Event - 4
POS Tagging

Dr. Aparupa Dasgupta

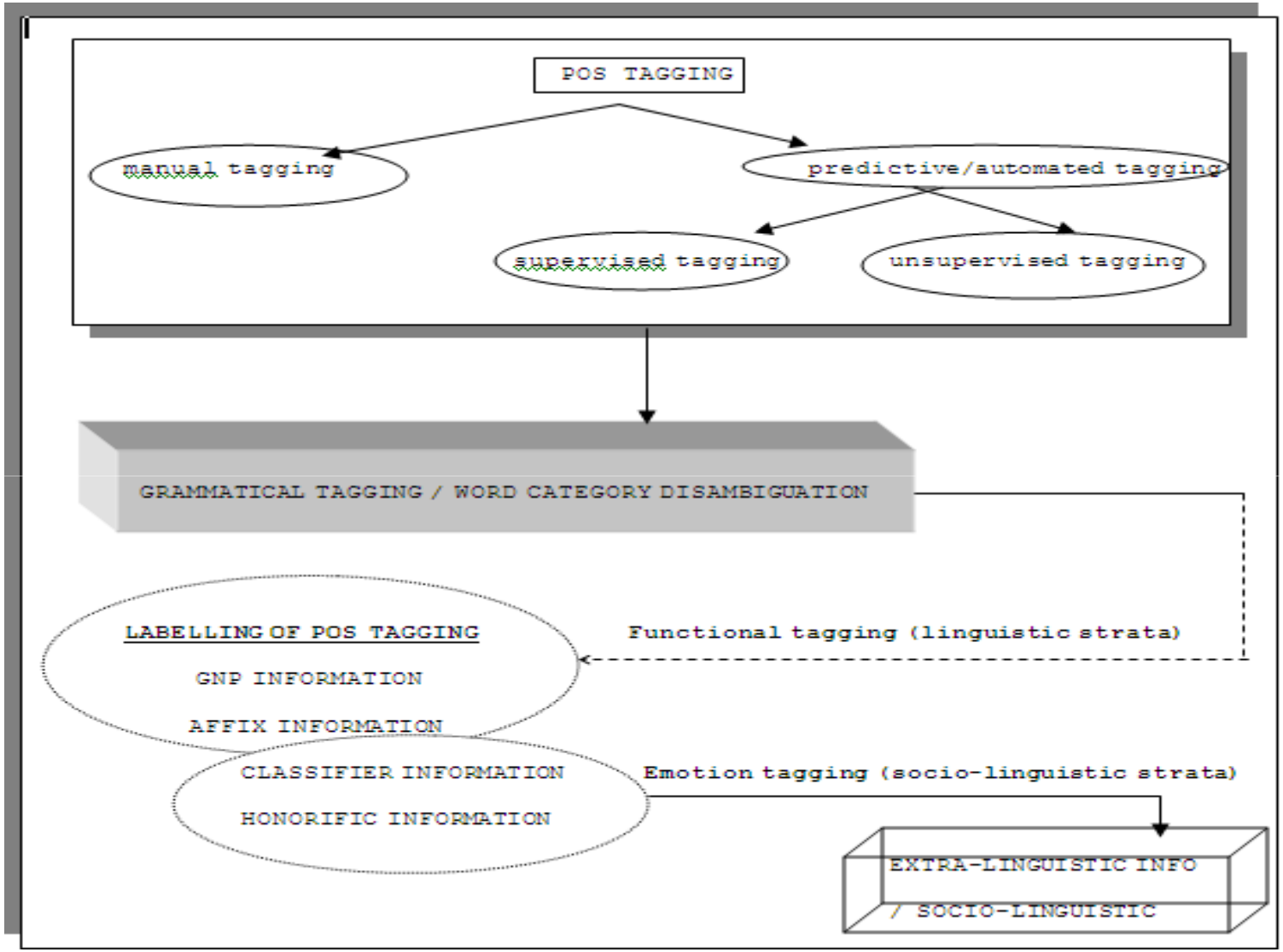
Dr. Debasri Chakrabarti

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. In each of the quarter of machine translation, information retrieval or speech processing, it becomes obligatory for the analysis thereafter. This paper proposes a rule based Parts-of-Speech tagger for Bangla with different stages of tagging. The paper also suggests one level higher tagging after identifying the parts of speech. We call this level as emotion tagging.

The following diagram represents the proposed labeling of POS tagging in this paper.

POS Tagging: *the proposed labeling*



Introduction

1. The significance of large annotated corpora is a widely known fact.

2. 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).

3. 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.

4. Annotation of corpora can be done at various levels viz, Part of Speech, Phrase/clause level, Emotion Tagging etc. Part of speech tagging can form a basic step towards building an annotated corpus.

5. This level of tagging can be further extended to higher levels of annotation.

Introduction

6. Annotation (POS, Semantic, lexical etc.), sub-categorization, domain dependency, morpho-syntactic or semantic tagging and discourse resolution etc. Along with the 22 official languages existing in India that associates contractiveness and complementariness among Indian languages, which will definitely provide a platform to design core NLP resolution Tool or a collator and discourse resolution tool.

7. Proposed Schema for POS Tagging. In this paper we discuss on the Proposed Schema Representation of POS tagging for Bangla.

8. Tagging and Indian Language e-corpus. For a corpora, may it be, digitized or non-digitized have a requirement for POS tag so that the resources can be utilized for either dictionary/ thesaurus or MT, IE-IR and speech technology.

9. Morpho-syntactic and semantic tagging. Such morpho-syntactically tagged corpora can be of direct input to any MT system.

Introduction

10. Corpus tagging need not necessarily be just POS tagging, it also can be discourse references inclined towards pragmatic resolutions.

11. Thus, discourse resolution can be actually tagged for word-sense recognizer, dependency marker and annotating the connectors (contextual) and clause-level boundaries in Indian

language corpora more explicitly

12. Benefitting the above discussion on POS tagging in Indian scenario, we discuss the POS tagging of Bangla from linguistic point of view one level higher annotation, which can also be called emotion tagging.

POS TAGGING IN BANGLA

1. In connection to above discussion on POS tagging in Indian scenario, we suggest one level higher annotation, the layers can be termed as **rule based tagging** (suffix or classifier) and **emotion tagging**.

2. It is a morphologically rich language, having a well-defined classifier system and at times show partial agglutination. In this section we propose a rules-based POS tagging of Bangla using context and morphological cue. The tagset are both from the common tagset for Indian Languages.

3. The top 12 POS tags that are identified as universal tags:
[N] Nouns, [V] Verbs, [PR] Pronouns, [JJ] Adjectives, [RB] Adverbs, [PL] Participles, [PP] Postpositions, [DM] Demonstratives, [QT] Quantifiers, [RP] Particles, [PU] Punctuations, [RD] Residual

4. After a top level annotation we go into one level deep tagging:

[WQ] Question Words, [QC] Cardinals, [CL] Classifiers, [INTF] Intensifiers, [INJ] Interjections , [NEG] Negative, [C] Compounds, [RDP] Reduplication, [ECH] Echo words

POS TAGGING IN BANGLA – different approaches

1. 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].

2. 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].

3. 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.

STEPS TO RULE BASED POS TAGGING IN BANGLA

1. The first step towards POS tagging is morphological analysis of the words. For this we have done a Noun Analysis and a Verb Analysis.

2. 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.

3. Verbs are classified into 6 paradigms based on morphosyntactic alternation of the root. The suffixes are further analysed for person and honorific information. Noun Analysis is shown in Table 1 and Verb Analysis is shown in Table 2.

RU LE BASED POS TAGGING IN BANGLA FOR NOUN (AFFIX / CLASSIFIER) – table 1

Paradigm	No	Animate	Honorific	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	chelederr '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'

RULE BASED POS TAGGING IN BANGLA FOR VERB – table 2

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'

POS TAGGING IN BANGLA – the morphological analysis

Based on this analysis a Morphological Analyzer (MA) will return the following for the sentence -

ekjon chele boigulo diye gelo ‘A boy gave the books’

ekjon (N,QT) chele (N) boigulo (N) diye gelo (V)

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

Given a noun / quantifier ambiguity, if the following word is a noun without a suffix and the token to be processed can qualify the succeeding noun, then the ambiguity will be resolved in favour of the quantifier, otherwise it will be resolved in favour of the noun. [eg. in ekjon chele, ekjon can be a quantifier or noun, but as it can qualify chele, and chele is without a suffix it will be an quantifier, not a noun]

LAYERS OF POS TAGGING IN RULE BASED

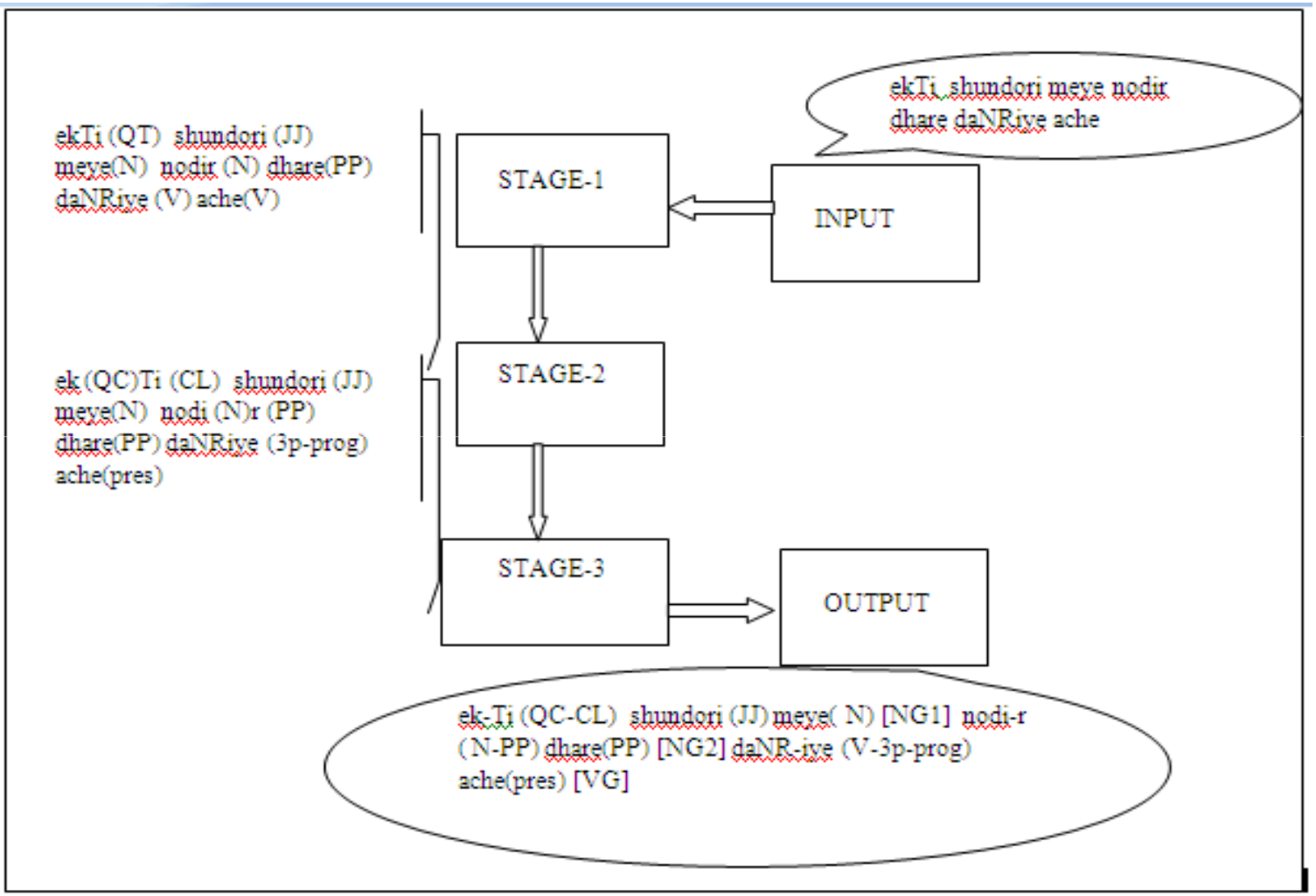
The POS tagger will go through 3 stages. At the first stage preliminary tags will be assigned with the help of MA and disambiguating rules. Stage 2 will do a deeper level analysis and provide information like Classifier, TAMPH, Postposition etc. Stage 3 or final stage will run a local word grouper and give the noun group and verb group information.

Following Figure 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'

LAYERS OF POS TAGGING IN RULE BASED



LAYERS OF POS TAGGING IN EMOTION TAGGING

In this section we will show that with the help of Ekman's (1993) six basic emotion types such as *happiness, sadness, anger, fear, surprise and disgust* how we can give one level higher tagging.

1. khela dhulo karo tumi? 'Do you play?'
   RDP V PR

2. khela phela cheRe ebar paRate mon dao 'Stop playing now concentrate on studies'
   RDP V RB V N V

3. daya kore edike ashun 'Please come here'
   V RB V

4. daya karo tumi ar cheshTa koro na 'Give a break don't try more'
   V PR JJ V

LAYERS OF POS TAGGING IN RULE BASED

In 1. and 2. /khela dhulo/ ‘play’ and /khela phela/ ‘play’ are reduplicated words and native speaker knows that /khela phela/ ‘play’ denotes a sort of anger or frustration but not /khela dhulo/ ‘play’. Similarly, in 3. and 4. /daya kore/ ‘please’ and /daya karo/ ‘please’ are the different inflected forms of the same verb /daya kara/ ‘please’ but in 4. the form is used to express disgust (or a sort of mockery) but in 3. *daya kore* ‘please’ is in a normal form denoting request. We suggest here to make a distinction between 1 & 2 and 3 & 4, the annotation should be as in 2a) and 4a).

1. **khela dhulo karo tumi?** **‘Do you play?’**
   RDP V PR

2a). khela phela cheRe ebar paRate mon dao **‘Stop playing now concentrate on studies’**
   RDP_ANG V RB V N V

3. daya kore edike ashun **‘Please come here’**
   V RB V

4a). daya karo tumi ar cheshTa koro na **‘Give a break don’t try more’**
   V_DGST PR JJ V

CONCLUSION

In this paper we have discussed a rule based POS tagger for Bangla to explain the proposed schema along with Emotion tagging.

Some linguistic rules have also been worked out for disambiguation purpose. Thus, POS Tagging for Indian languages should be guided and formalized considering the following matters according to our findings and observations:

a. Standardizing the Indian tagset

b. Formalizing the Indian tagset and coherently consider the factor of morphologically rich language families

c. To incorporate the formal tagging with sense tag and emotion tag to nurture the subtleties of natural language processing.

d. Enrich the e-corpora with above parameters for annotation in Indian languages.