CHAPTER 4

PART OF SPEECH TAGGING IN BENGALI

This chapter presents an overview of part of speech tagging, its different paradigms and standard approaches. It also presents some applications of part of speech tagging in the field of computational linguistics especially in Multiword Expressions Detection in Bengali.

4.1 Introduction

The process of marking up morph syntactic categories of each lexical item including punctuation mark in text documents according to the context is called part of speech tagging. Due to rich morphological nature of Bengali (Bangla), it is a language with a high inflectional system. Inflections include postpositions, number, gender and case markers on nouns, and inflections on verbs include person, tense, aspect, honorific, non-honorific, pejorative, finiteness and non-finiteness. Seeing as syntactical bracketing is a job of shallow processing and size of the tagset is one of the important factors, only postpositions, accusative and possessive case markers on nouns have been built-in in this tagset. To reflect only these characteristics of morphology, a separate category of 'suffixes' has been included to denote the inflections. When a noun or a pronoun is inflected by a suffix, the base form and inflections are separated by a plus sign (+) (Altaf *et al.*,2009). Verbs are categorized based on their form such as finite, non-finite etc.

Tagset development forms a foundation of any computational processing effort. It is generally acknowledged that, as a introduction to syntactic analysis of natural language by computers, a text must be annotated with tags indicating the POS. The major troubles in POS tagging task arise from the fact that a word can take different lexical categories based on its context of use. The tagger has to resolve this ambiguity and determine the best sequence for a sentence. Tags are also applied to punctuation markers, thus tagging for natural language is the same process as tokenization for computer languages, although tags for natural languages are much more ambiguous that vary from language to language.

In 1963, Klein and Simmons introduced a computational method for grammatical coding of English words. Their primary goal was to reduce time for constructing a very large dictionary. Their algorithm uses a set of 30 POS categories. It first looks for each word in dictionaries, then checks for suffixes and special characters as clues. Finally, the context frame tests are applied. This algorithm properly and unambiguously tags about 90% of the words in numerous pages of the Golden Book Encyclopedia[1].

Generally, POS is a set of specially designed codes term as 'tags' carrying grammatical information are assigned to words to indicate their parts-of-speech with regard to their use in the text (Leech and Garside, 1982). In POS, a well-defined set of linguistic rules are used to identify POS and POS tags are assigned in words to determine their lexical, syntactic and semantic grammatical functions in the text.

4.2 Related Works

Part of speech tagging is the pre-processing task for developing a multiword expressions detection tool. The research area of part of speech tagging using computational techniques has been carried out over the last few decades by several researchers. It is old like linguistics. Since its initiation in the middle sixties and seventies (Haris,1962),a lots of new concepts have been introduced to enhance the efficiency of the tagger and developing new part of speech taggers for different languages. At first, people manually engineered rules for tagging. Linguistic taggers added the knowledge as a set of rules or constraints written by linguists. Presently several statistical or probabilistic models have been developed and used for the POS tagging work for providing portable adaptive taggers. Several well advanced machine learning algorithms have been developed that are capable of achieving vast information.

Usually, the statistical models tend to achieve manually tagged corpora to learn the underling language model, which is difficult to acquire for a new language. Thus, some of the recent works focus on semi-supervised and unsupervised machine learning models to resolve the problem of unavailability of the annotated corpora.

[1] **http://www.ldcil.org/standardsTextPOS.aspx**

We incorporate several sources of for tagging purpose for our research work. This chapter presents brief review of the prior work in part of speech tagging along the review of MWEs .Although, our focus has been made on part of speech taggers of Indian languages in connection with Automated Multiword Expression Detection in Bengali language.

In the year 1992 Eric Brill developed a rule based POS tagger with the accuracy rate of 95-99% (Brill, 1992). Use of combine hand-crafted rules and statistical learning, POS tagging of languages like Turkish, Czech was attempted by (Oflazer and Kuruiz, 1994), (Hajic *et al.,* 2001).

For the development of Bengali language some of the POS tagger was developed using different approaches. Dandapat et al.,(2007) proposed two stochastic based taggers using HMM and Maximum Entropy (ME) approaches. Ekbal Asif also developed a POS tagger for Bengali language using Conditional Random Fields (CRF). Asif and Bandyopadhyay (2008) developed another POS tagger using SVM algorithm based on machine learning approach. In the year 2010, Hammad Ali proposed an unsupervised Parts-of-Speech Tagger for the Bangla language. In 2011, Debasri Chakrabarti of CDAC Pune proposed a Layered Parts of Speech Tagging for Bangla (Antony and Soman, 2011).

For Bengali Language three different types of stochastic POS taggers were developed. Based on 40 tags in a tagset, a supervised and semi supervised bigram HMM & a ME based was examined (Kumar and Josan, 2010; Bharati and Mannem, 2007). It is seen that supervised model perform better performance compared to other model and enhancement can be done using morphological analyzer.

The second POS tagger for Bengali is based on Conditional Random Fields (CRF) in which where features selection plays a vital role in the development of POS tagger (Chakrabarti, 2011; Kumar and Josan, 2010). To develop such POS tagger a tagset of 26 tags was used. System accuracy was achieved 90.3%.

The third POS tagger for Bengali is statistical approach using a supervised machine learning algorithm called SVM (Kumar & Josan, 2010; Bharati and Mannem, 2007). Here the whole corpus is divided into two half training corpus and development set. Accuracy was achieved 86.84% (Antony & Soman, 2011)

For Hindi Language ,Central Institute of Indian Language (CIIL) proposed a tagset based on Penn tagset for Hindi language based on Penn tagset (Sathish and Kishore, 2007). This tagset was designed to increase the lexical categories in comparision with IIIT-Hyderabad. It contains 36 tags.

Another tagset developed by IIT- Kanpur for Bengali language which consists of 40 tags. There is one more tagset called CRBLP tagset. It consists of a total of 51 tags, where 42 tags are general POS tags, and 9 other tags are intended for special symbols (Parak *et al.,* 2011; Antony & Soman, 2011).

Stochastic models of part of speech tagging Dermatar & George (1992), Manju *et al*. (2009) have been used exclusively for simplicity and language independence of the models. Out of stochastic models, bi-gram and tri-gram of Hidden Markov Models (HMM) are quite popular. Brants (2000) gives the idea of TnT as a widely used efficient statistical trigram HMM tagger which uses a suffix analysis technique to estimate lexical probabilities from known tokens based on properties of the words in the training corpus which share the same suffix. Stochastic tagger requires large amount of annotated text to make corpus more powerful and at the same time tagger to be more efficient. More than 95% word-level accuracy of stochastic taggers have been developed for English, German and other European languages, for which large tagged data are available. Simple HMM models are not efficient when small amounts of tagged data are used to estimate the model parameters. To achieve high accuracy of POS tagging in HMM model, additional information is coded Cutting *et al.,* (1992).

To find out the accuracy between linguistic and statistical taggers some author perform comparison with favorable conclusion (Chelliah, 1997)[2].

In this way, Indian languages like Hindi, Bengali, Punjabi, Marathi, Tamil, Telegu and Malayalam languages have many POS taggers. Several earlier work on Indian language of POS tagging have been seen in Bharati *et al.* (2008). They represent a framework for Indian languages where POS tagging is implicit and emphasized with the parsing problem in their work on computational Paninian parser.

Different POS tagging works have been done on Hindi Language by different researcher. A POS tagging methodology has been proposed by Smriti Singh on Hindi

[2]http://research.microsoft.com/en-us/groups/mls/

language which can be used by a language having low resources (Kumar & Josan, 2010).

POS tagging can be achieved using rule-based systems, probabilistic data-driven systems, neural network systems or hybrid systems. Various work have been done in Bengali POS tagging among them Statistical based Hidden Markov Model (HMM) and Conditional Random Field (CRF) (Ekbal *et al*., 2007), Maximum Entropy Model (Dandapat *et al.*, 2011).

The most important advantages of POS tagging can be found in the following three categories:

a. Lexical level

It is the formal property of a word-type determining its acceptable uses in syntax. It provides representation of lexical information and acquisition of lexical information.

b. Orthographic level

The orthographic level indicates the degree to which a written language deviates from simple one-to-one letter phoneme association. It depends on how easy it is to predict the pronunciation of a word based on its spelling, shallow orthographies are easy to pronounce based on the written word, and deep orthographies are complex to utter based on how they are written.

c. Syntactic level

It allows identifying syntactic grammatical functions of words to assign their POS. A syntactic category is a type of syntactic unit that considers theories of sysntax. Word classes, basically related to traditional parts of speech (e.g., noun, verb, preposition, etc.) are syntactic categories. In phrase structure grammars, the phrasal categories (e.g. noun phrase, verb phrase, prepositional phrase, etc.) are also syntactic categories.

Corpus based Natural Language Processing (NLP) tasks for popular languages like English,French etc. have been much worked on with accomplishment in POS. Many works on POS have been done on languages like Bengali which are in growing level in the NLP realm. One of the main reasons is the absence of annotated corpus

for such languages. Corpus annotation is the observation of adding interpretative particularly linguistic information to a text corpus by coding, added to the electronic representation of the text itself.

Once performed manually, methodology of POS tagging in the context of computational linguistics now involves the use of algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags.

The development of an automatic POS tagger requires a set of linguistically provoked rules or a large annotated corpus. But such rules and corpora have been developed for a few languages like English and some other languages. POS taggers for Indian languages are not readily available due to lack of such rules and large annotated corpora.

According to N.S.K. Das (2013), we can differentiate tag between plural and singular nouns and pronouns, tag words identified with case makers and inflections in texts, tag grammatical gender is seen with words, tag nouns and adjectives marked with gender and number markers, person, number, gender, tense, aspect, modality, honourification, and other markers are seen in tag verb, tag adjectives marked with suffixes of degree, tag adjectives behaving like nouns in texts etc. To carry out all these types of tagging information, it is required to develop an algorithm for automatic parts of speech tagging of words in a language text by a computer system. Sometimes, POS tagging algorithms may fail to trace out unique linguistic information of words and in some cases a POS tagged text will fail to find out the finer linguistic functions of words which are essential to linguistic research and development activities concerned with theoretical and applied linguistics field of Natural Language Processing. Based on these difficulties various POS taggers have been developed to enhance POS in different languages.

The first precondition for automated POS tagging in a tagset is that it is a set of comprehensive categories into which any token of the language can be placed. While the nature of the language is that there will always be words that are hard to classify, or are ambiguous between two categories, the tagset categories should be designed in such a way so as to minimize such difficulties.

Part of speech tagging, also known as morph syntactic categorization or syntactic word class tagging (Karthik *et al*., 2006) is the process of passing a part of speech or other lexical class marker to each word in a corpus. Tags are also applied to punctuation markers, thus tagging for natural language is the same process as tokenization for computer languages, while tags for natural languages are much more ambiguous.

We know list of parts-of-speech presented in grammars, school text books and dictionaries are sufficient for a language to tag words in texts. A part from these there are many fine text categories and text sub-categories in a natural language text. For processing such text using machine translation, these fine divisions require to be spelt out separately, if we desire to design a POS tagging scheme for words or Multiwords to be used in a text of a language. For example, in Bengali, it is generally explained in all grammar books that it has only eight parts-of-speech namely, Noun, Pronoun, Adjective, Adverb, Finite Verb, Non-finite Verb, Postposition, and Indeclinable. It is considered that these parts-of-speech are sufficient to identify each Part of Speech of each and every word used in language. A part from these for other text categories, like demonstratives, infinitives, gerunds, conjunctions, enclitics, quantifiers, punctuations, particles, etc., the formation and function of these we need to analyze, identify and understood properly, so that we have full command over parts of speech of words and at the same time over with grammar of the language.

The earliest work on tagset started in 1960s and early 1970s in the US and focused on English language. The most important tagset of this earliest period are those of Klein and Simmons 1963 and Greene and Rubin 1971. Later, sequences of tagsets for English have been devised such as the Penn tagset and CLAWS (the Constituent Likelihood Automatic Word-tagging System) tagset including the series C_1, C_2, C_5, C_7. The publication of Eagles recommendations for morphosyntactic annotation of corpora (Leech and Wilson 1996) was the original attempt to develop common tagset guidelines for several European languages.

The Eagles project is deals with Natural Language Processing (NLP) and thus, it has a very broad idea in NLP, and needs to furnish the large number of circumstances in which text is used**.** The aims of Eagles guidelines were to standardize the tagset used in different languages to attain cross linguistic

compatibility, reusability and interchangeability and a wide range of understandability in the field of NLP.

Part of speech tagging has been studied extensively in the last two decades and lot of work has been done in various European languages including many Indian languages like Hindi, Urdu, Bengali, Sanskrit, Tamil and Kannada.

Part-of-speech gives most important information about the word and its neighbours which can be useful in a language model for different speech and natural language processing applications. Development of a Bengali POS tagger will manipulate several pipelined modules of natural language understanding system including: information extraction and retrieval, machine translation, partial parsing and word sense disambiguation. The existing POS tagging technique seems that the development of a good accuracy POS tagger having requires either increasing an exhaustive set of linguistic rules or a large amount of annotated text.

4.3 Annotation

The POS tagging is another form of text annotation, which is considered to be the first step of a more comprehensive process where multiword expressions, such as, compound words, reduplicated forms, idiomatic expressions, proverbial expressions, set phrases, and others used in a text that are assigned with chunking markers to assignment of phrase markers to each of the sentences used in a text or corpus.

The POS tagging helps in following types of text annotations:

i. **Syntactic annotation** : Addition of information about how a given sentence is parsed in terms of analysis into such units such as phrases and clauses
ii. **Semantic annotation**: Addition of information about the semantic category of word. The noun cricket as a term for spot.
iii. **Phonetic annotation:** Addition of information about how a word in a spoken corpus is pronounced.
iv. **Prosodic annotation:** Again, in a Spoken corpus, adding information about prosodic features such as stress, information (modulation of the voice) and pauses.

v. **Pragmatic annotation**: Addition of information about the kinds of speech act that occur in a spoken dialogue.

vi. **Discourse annotation**: Addition of information about anaphoric links in a text.

vii. **Stylistic annotation:** Addition of information about speech and thought presentation (direct, indirect speech, indirect thought etc).

The use of POS tags in a text makes the text very difficult to read for a normal reader who has no linguistics knowledge compared to human beings, the text becomes maximally suitable for providing linguistic information needed by a computer for differentiating between words used in different parts-of- speech (Leech and Eyes, 1993).

Some of the POS taggers as explained by Siddiqui and Tiwary (2008) are:

1. Standford Log-Linear Part of Speech (POS) tagger[2]
2. Part of speech Tagger for English[3]
3. TnT Tagger[4]
4. Brill Tragger[5]
5. Tree Tagger[6]

4.3.1 Standford Log-Linear Part of Speech (POS) Tagger

This is basically based on Maximum entropy Markov Models. The key features of the tagger are as follows:

i. It allows to use both the previous and subsequent tag contexts via a dependency network representation.

ii. It uses a wide range of lexical features.

iii. It uses priors in conditional log-linear models.

The accuracy report of this Tagger on the Penn Treebank WSJ is 97.24%, which is amounts to an error reduction of 4.4% on the best previous single automatically learned Tagging result (Tuatanova *et al.*, 2003).

[2] http://nlp.standford.edu/software/tagger.shtml
[3] http://www.tsujii.is.s.u-tokyo.ac.jp
[4] http:////www.coli.uni-saarland
[5] http://www.cs.jhu.edu/~brill/RBT1-14.tar.Z
[6]http://www.ims.uni.stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

4.3.2 Part of speech Tagger for English

Part of speech tagger for English uses a bidirectional inference algorithm for tagging. It is based on Maximum Entropy Markov Model (MEMM). The tagging algorithm can find details account of all possible decomposition structures and highest probability sequence together with the corresponding decomposition structure in polynomial time. Experimental results of this part of-speech tagger reveal that the proposed bi-directional inference methods constantly do better than unidirectional inference methods and bidirectional MEMMs give as good as performance to that achieved by state-of-the art learning algorithms, including kernel support vector machines (Tsuruoka and Tsujii, 2005)[7].

4.3.3 TnT Tagger

Accoring to Brants (2000)[8], Trigrams's Tags or TnT is an efficient statistical Part of speech tagger. This tagger is based on hidden Markov models (HMM) and uses some optimization techniques for detecting and handling unknown words. It performs other modern approaches, including the maximum entropy framework. TnT Tagger uses second order Markov models for part-of speech tagging. The states of the model represent tags, outputs represent the number of words. Transition state probabilities depend on the transition states, thus on pairs of tags. Output probabilities only depend on the most recent transition states.

4.3.4 Brill tagger

According to Brill (1992)[9], Brill Tagger is a trainable rule based tagger that performance task similar to that of stochastic taggers. It uses transformation based learning to automatically induce rules. The Brill tagger shares features of both tagging architectures. Like the rule based tagger, it is based on rules which determine when an ambiguous word should have a given tag. It works like the stochastic taggers, it has a machine-learning component i.e., the rules are automatically induced from a previously tagged training corpus.

[7]http://www.tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/postagger/

[8]http://www.coli.uni-saarland.de/~thorssten/tnt/

[9]http://www.cs.jhu.edu/~brill/RBT1_14.tar.Z.

A number of extensions to this rule-based tagger have been proposed by Brill *et al*. (1994). He describes method for expressing lexical relations in tagging that stochastic taggers are currently unable to express. It implements a rule based approach to tagging unknown words. It denotes how the tagger can be extended into a k-best tagger, where multiple tags can be assigned to words in some cases of uncertainty.

4.3.5 Tree-Tagger

According to Schmidt (1994) Tree-tagger is a probabilistic tagger. It eliminates the problems faced by the Markov model methods when estimating transition probabilities from sparse data by using a decision tree to estimate transition probabilities. The decision tree automatically determines the actual size of the context to be used in evaluation. In comparison with the accuracy of the Penn-Treebank WSJ corpus, the Tree tagger is above 96% of the Penn-Treebank WSJ corpus. . The tagger is available for downloading at the link[10].

Automatic part-of-speech taggers help in building automatic word-sense disambiguating algorithms and POS taggers are also used in advanced ASR Language models such as Class-based n-grams. Parts of speech are very frequently used for 'partial parsing' texts, for example, quickly finding names or other phrases for the information extractions applications.

4.4 Part of Speech Tagging in Bengali for Multiword Expressions

Part of speech tagging plays a vital role for identification of MWEs in the language. A variety of taggers are there, among them one is Rule based POS tagger and another is machine learning based POS tagger.

a. Rule based POS taggers uses manually written rules to assign tags to unknown or ambiguous words.
b. Machine learning based POS taggers use a large amount of annotated data for the development of a POS tagger in shorter time.

[10]http://www.ims.uni-stuttgart.de/projekte/projekte/corplex/Tree Tagger/Decision Tree Tagger.html

Due to no tagged corpus was available to us for the development MWEs and its detection. Thus, there is a necessity for tagged corpus using an automatic Part-of-Speech tagger for Bengali. With this we cover major goal of the thesis.

4.5 General Framework for Bengali Part of Speech Tagging

Bengali POS Tagging follows the Eagles Guidelines and the Penn tree bank tagset. Many other Indian tagging guidelines like IL-POST, ILMT and Sanskrit tagset were taken into consideration. The tagging schema for Bengali is designed taking into consideration both language features in general and the idiosyncratic features of Bengali. After careful consideration a hierarchical tagset was favored in case of MWEs.

Part of Speech tagging for the final system was performed as follows. First we split the training data randomly into two halves. The first half is used to train the TnT tagger and the second half is used for testing.

Any error in this process results in learning of appropriate transformation rules. These transformation rules are then used to correct the results produced by the TnT tagger on the test set. We account the performance measures by averaging over four random 50:50 splits of the training data.

The design of the Bengali tagset for MWEs is based on three distinct features into which the grammatical schema is distributed. The features are:

I. Category
II. Type
III. Attribute

Categories involve major grammatical categories like Nouns, Verbs etc. The type includes the type of those grammatical categories like Common Noun and Proper Noun for Noun category, Main Verb, Auxiliary Verb etc. for Verb category, and so on. The attribute level takes features within each type like Gender (masculine, feminine), Number (singular, plural), Case (dative, ergative, ablative, etc.), Tense and Aspect etc. into consideration. The category list includes all Bengali categories that

can occur. The type list within a category includes all types of the category that can occur. The attribute list includes all possible attributes of the type that can occur.

4.6 Bengali Part of Speech Tagset (LDC-IL based)

A detailed account schema for Bengali POS is given below based on Linguistic Data Consortium for Indian languages (LDC-IL)

Table 4.1 List of POS tagging in Bengali

Sl No.	Tag Description	Level	Examples
1	Common Noun	N_NNC	বালক (bālak), শহর(śahar), কথা (kathā), মানুষ (Man), লোক (lok) 'man', (mānuṣ),দুধ (milk) etc.
2	Proper Noun	N_NNP	করিম(Karim),দিল্লি(Delhi)
3	Material Noun	N_NNM	কলম (kalam), pen
4	Nloc noun	*N_NST*	উপরে (upare)
5	Temporal Noun	*N_NNT*	গতকাল (yesterday), আজ (today), এখন (now)
6	Verb root	*N_NNV*	গোসল করা (taking bath), পান করা (drink)
7	Locative noun	*N_NNL*	উপর (up), নিচে (down), আগে (front)
8	Question locative noun	*N_QNL*	কোথায় (where), যেখানে (relative 'where')
9	Question temporal noun	N_QNT	কখন (when), যখন (relative 'when')
10	Collective noun	N_NNL	দল (dal) 'party'
11	Abstract noun	N_NNA	ভয়(bhay) 'fear'
12	Verbal noun	N_NNV	গ্রহণ (grahaṇ) taking, নাইস্ (nice) ভিতেরে(bhitare)
13	Pronoun	PR	আমি (ami), তুমি (tumi),সে(se), তারা (tar ā), তু ই (tui), etc.
15	Personal Pronoun	PRP	আমি (ami),সে (se) তুমি (tumi), আমরা (amra)
16	Reflexive Pronoun	PRF	নিজেকে(nijeke)
17	Relative pronoun	PRL	যে(ýe), যারা (ýārā), যাদের (ýāder), যাকে (ýāke)

18	Reciprocal pronoun	PRC	পরস্পর (paraspar)
19	Wh-word Pronoun	PRQ	কে (ke), কাকে (kāke), কারা (kārā), কাদের (kāder)
20	Question Pronoun	QPR	কে (who), কারা (plural 'who'), যে (relative 'who')
21	Demonstrative	DM	যে (ýe),এই (ei),ওই (oi), তাই (tai), etc.
22	Deictic Demonstrative	DMD	এ(e),এই (ei),সে (se),সিই (sei),ও (o),ওই (o)
23	Relative Demonstrative	DMR	যে(ýe),যেই (ýei) যাহা (ýāhā), যা (ýā)
24	Wh-word Demonstrative	DMQ	কানো (kano), কোন্না(kona)
25	Finite Verb	FV	করিছ (karchi), করতাম (kartām),গেলা (gela), যাবে (ýā be), etc.
26	Non-Finite Verb	NFV	করলে (karle), করতে (karte),গেলে (gele), গিয়া(giye), etc.
27	Non finite perfective verb	VBT	করা (doing), করানো(causative 'doing'), পড়া (reading)
28	Subjunctive verb	VBC	করেল (if done) ,
29	Auxiliary Verb	VBX	করে ফেললাম/VBX (have done), হেসে উঠলো/VBX (burst into laughter)
30	Finite Existential	VBE	হয় (be), হবে (will be)
31	Nonfinite Existential	VBEF	হেত (to be)
32	Adjective simple	AD	ভাল (bhāla), মন্দ (manda), সুন্দর (sundar) (beautiful), সাদা (sādā), লাল (red), শ্রেষ্ট(best), শ্রেষ্টতম (the best)etc
33	Verb root adjective	JJV	লাল/JJV হওয়া/VBM (to redden), দুর্বল /JJV হওয়া/VBM (to weaken)
34	Question Adjective	QJJ	কেমন (how), যেমন (relative 'how')

35	Adverb	AV	হঠাত্ (haṭhāt), বাবদ (bābad), কারণে (kāraṇe), etc
36	Question Adverb	QRB	কেন (why), কিভাবে (how), যেভাবে (relative 'how')
37	Postposition	PP	পের(pare), কাছে (kāche), আগে (āge), নাইচ (nice), দারা (by), থেকে (from), জন্য (for), চাইতে (than) etc.
38	Conjunction	CN	তেব (tabe), যদি (ýadi) নইলে (naile), যাতে (ýāte), etc.
39	Coordinating Conjunction	CC	এবং (and), অথবা (or), নতুবা (nor)
40	Compound coordinating	CCC	না/CCC হয়/CC(neither)
41	Suspecion Conjunction	CN	যদি (if), পাছে (if)
42	Eternal joining Conjunction	CET	যেমন/CET ... তেমন/CET (like … like), যখন/CET ... তখন/CET (when … then)
43	Subordinating Comjunction	CS	যে (Complementizer 'that'), এইজন্য (for this)
44	Compound Coordinating Conjunction	CSC	তাই/CSC বেল/CS (that's why), এই/CSC কারনে/CS (for this reason)
45	Interjection	UH	ওহ! (oh!), হায়! (alas!)
46	Indeclinable	IN	কিন্তু (kintu),অথবা (athabā), বরং (baraṃ), আর(ār), etc
47	Particle	PT	ই (i),ও(o),তা (to),না (nā), নে (ne), নি (ni), etc.
48	Question Particle	QPT	কি (question particle)
49	Quantifier	QT	এক (ek),দুই (dui),প্রথম (pratham),পয়লা (paylā), etc.
50	Reduplication	RD	চা টা(cha ta),বেন বেন (bane bane), কত কত (kata kata) যেযে ,(ýe ýe), etc.
51	Foreign Word	FW	যেকোন বিদেশী শব্দ (any foreign word)
52	Postpositional Suffices	SFON	এ, য়, তে

Table 4.1: List of POS tagging in Bengali (contd…)

53	Accusative Postposition	SFAC	কে, রে, এরে, দিগেরে
54	Possessive Postposition	SF$	এর, দের
55	Punctuation Marks	PN	., : , |(full stop, !, ? ( ), [ ], { }, ‘,’ etc and Others Mathematical symbols like , +, -,- , x, >, <, $, #, @, ^, &, * etc.

4.7 Bengali Part of Speech Category

Central Institute of Indian Languages (CIIL), Department of higher education ministry of human resource development, government of India, Manasagangotri, Mysore - 570 006, categorized Bengali Part of speech in the following categories (which we follow in our research work):

Category

1. Noun (N)
2. Pronoun (P)
3. Demonstrative (D)
4. Nominal Modifier (J)
5. Verb (V)
6. Adverb (A)
7. Participle (L)
8. Postposition (PP)
9. Particle (C)
10. Numeral (NUM)
11. Reduplication (RDP)
12. Residual (RD)
13. Unknown (UNK)
14. Punctuation (PU)

1. **Noun**

Table 4.2 Types of Noun

Category	Types	Attributes
Noun	Common(NC)	Number, Case, Case marker, Definiteness, Emphatic
	Proper(NP)	Number, Case, Case marker, Definiteness, Emphatic
	Verbal(NV)	Case, Case marker, Definiteness, Emphatic
	Spatio-temporal (NST)	Case, Case marker, Definiteness, Emphatic

2. Pronoun

Table 4.3 Types of Pronoun

Category	Types	Attributes
Pronoun (P)	Pronominal (PPR)	Number, Person, Case marker, Definiteness, Emphatic, Honorificity.
	Reflexive (PRF)	Number, Person, Case marker, Definiteness, Emphatic, Honorificity.
	Reciprocal (PRC)	Number, Case marker, Definiteness, Emphatic, Honorificity.
	Relative (PRL)	Number, Case marker, Definiteness, Emphatic, Honorificity.

3. Demonstrative

Table 4.4 Types of Demonstrative

Category	Types	Attributes
Demonstrative (D)	Absolute (DAB)	Emphatic , Dimension
	Relative Demonstrative (DRL)	Emphatic
	Wh-demostrative (DWH)	Emphatic

4. Nominal Modifier

Table 4.5 Types of Nominal Modifier

Category	Types	Attributes
Nominal Modifier (J)	Adjective(JJ)	Emphatic
	Quantifier(JQ)	Definiteness, Emphatic, Numeral
	Intensifier	Emphatic

5. Verb

Table 4.6 Types of Verbs

Category	Types	Attributes
Verb(V)	Main verb	Person, Tense, Aspect, Mood, Finiteness, Emphatic, Negative, Honorificity
	Auxiliary Verb (VA)	Person, Tense, Aspect, Mood, Finiteness, Emphatic, Negative, Honorificity

6. Adverb

Table 4.7 Types of Adverb

Category	Types	Attributes
Adverb(A)	Manner	Definiteness, Emphatic
	Location(ALC)	Case marker, Definiteness, Emphatic

7. Participle

Table 4.8 Types of Particle

Category	Type	Attributes
Participle(L)	Verbal(LP)	Emphatic
	Conditional(L)	Emphatic , Realis

8. Postposition

Table 4.9

Category	Type	Attributes
Postposition (PP)	----------	Case marker , Emphatic

9. Particle

Table 4.10 Types of Particle

Category	Types	Attributes
Particle(C)	Coordinating(CCD)	----------
	Sub-Coordinating (CSB)	Emphatic
	Interjection(CIN)	----------
	(Dis) Agreement(CAGR)	Emphatic
	Classifier (CCL)	Definiteness, Emphatic
	Similative (CSIM)	Emphatic
	Interrogative (CIQ)	
	Others(CX)	

10. Numeral

Table 4.11 Type of Particle

Category	Types	Attributes
Particle(C)	REAL(NURM)	Case maker, Emphatic
	SERIAL(NUMS)	Case maker
	Calendric (NUMC	Case maker
	Ordinal (NUMO)	Case maker , Emphatic

11. Reduplication

Table 4.12 Reduplication

Category	Type	Attributes
Reduplication	**-------**	**----------**

12. Residual

Table 4.13 Residual

Category	Type
Residual	**------------**

13. Unknown

Table 4.14 Unknown

Category
Unknown(UNK)

14. Punctuation

Table 4.15 Punctuation

Category
Punctuation (PU)

4.8 Some examples of POS with corresponding Tag

Part of Speech: Compound Common Noun

Tag: NNC

Category: Noun

Examples

ভারতের/NNP+SF$ প্রতোকটি/DM জেলায়/NN+SFON রহেছেনVB একজন/QFNUM জেলা/NNC প্রশাসক/NN

“There is one district commissioner at each of the district of India”

বিষয়টি/NN স্বরাষ্ট/**NNC** মন্ত্রালয়ে/**NN+SFON** পেশ/NNV করা/VBM হয়েছে/VBE

“The matter has been submitted to home ministry”

Part of Speech: Proper Noun

Tag: NNP

Category: Noun

Example: কারিম **/NNP** একজন/QFNUMযুদ্ধা /NN

“Karim is a warrior”

Part of Speech: Compound Proper Noun

Tag: NNPC

Category: Noun

Example: কাজী **/NNPC** নজরুল/**NNPC** ইসলাম/NNP

“Kazi Nazrul Islam”

Part Of Speech: Nominal Verb Root

Tag: NNV

Category: Noun

Example: সে/PRP গোসল/**NNV** করেছে/VB

“He has taken a bath”

আমি/PRP এখন/NNT চা/NN পান/**NNV** করিতেছি/VB

“I am now taking tea”

Part Of Speech: Question Adjective

Tag: QJJ

Category: Adjective

আজ /NNT আবহাওয়া /NN অতো /RB সুন্দর /JJ নয় /VB যেমন /QJJ আমি /PRP ভেবেছিলাম /VB

“Today’s weather is not as that much beautiful as I thought”

4.9 Various Approaches of Part of Speech Tagging

POS tagging has different approaches. POS tagging basically classified into three categories namely Rule based, corpus based and Hybrid POS tagging. The following figure as explained by (Ghose, 2013) illustrates the different POS tagging approaches:

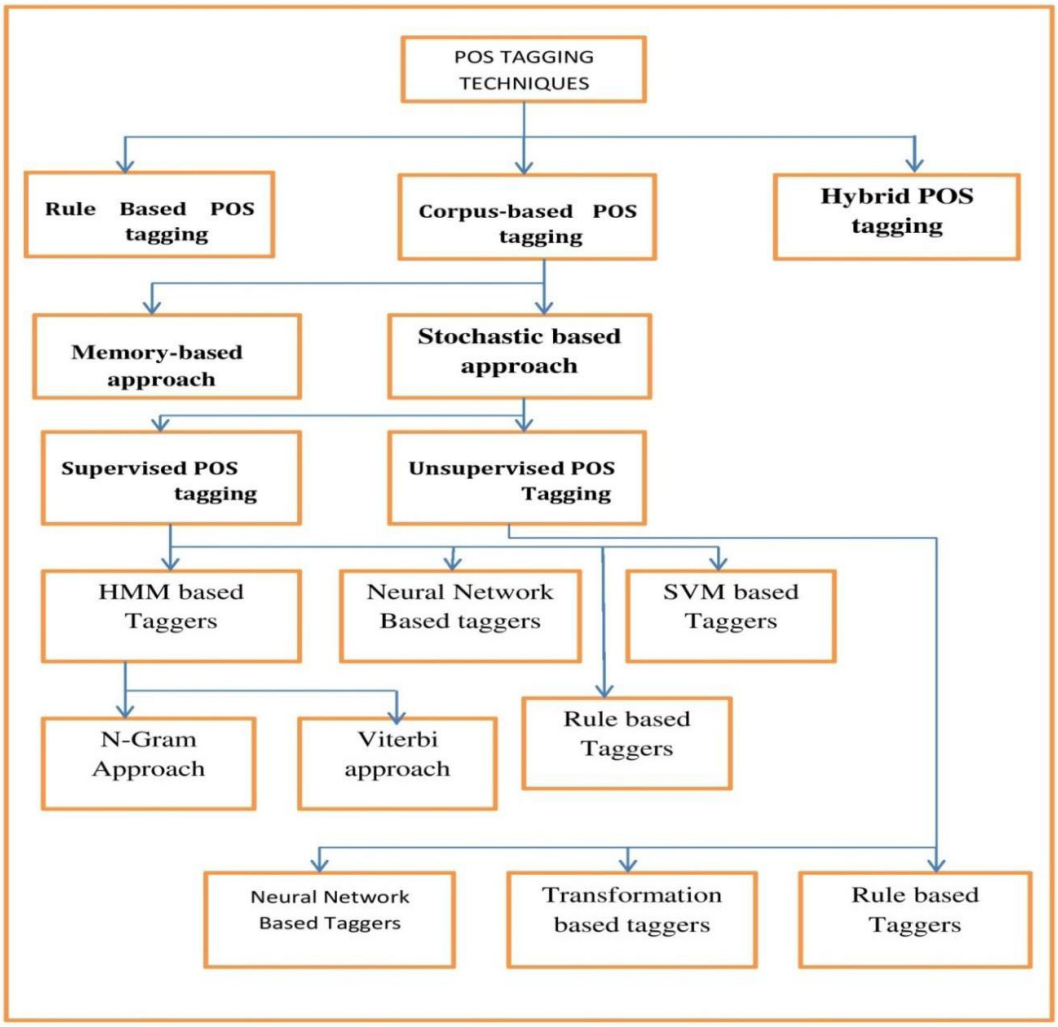

Fig 4.1: Various approaches to POS Tagging (Ghose, 2013).

To solve POS problems we briefly discuss of each approach as follows:

1. Rule Based POS tagging

Rule based taggers use hand written rules to assign tags in words. These rules use a lexicon to achieve a list of candidate tags and then use rules to discards incorrect tags. Rule based tagger based on dictionary or lexicon to get possible tags for each word to be tagged. Hand written rules are used to identify the correct tag when a word has more than one possible tag. These rules are frequently known as context frame rules. Disambiguation is made by analyzing the linguistic features of the word, its preceding word, its following word and other aspects. For example, if the preceding word is an adjective then the word in question must be adjective or noun. This information is coded in the form of rules (Barman *et al.,* 2013).

Another approach that alternate for the word frequency is known as the *n*-gram approach. *N*-gram calculates the probability of a given sequence of tags. It determines the best tag for a word by calculating the probability that occurs with the n previous tags, where the value of n is set to 1, 2 or 3 for practical purposes. These are known as unigram, bigram and trigram models respectively. The most common algorithm for implementing n-gram approach for tagging new text is known as the Viterbi Algorithm (HSK, Corpus Linguistics (2008))., which is a search algorithm that avoids the polynomial expansion of a breath first search by edging the search tree at each level using the best Maximum Likelihood Estimates (MLE) where m represents the number of tags of the following word. The main problems with Rule Based POS tagging are that rules are realy large and rule maker should be expert with lingusitics knowledge.

2. Corpus-Based POS Tagging

After rule-based POS tagging, researchers were shifted to corpus based POS tagging methods. Literature survey reveals that the POS tagger developed is corpus-based. This approach is divided into two categories: namely memory-based approach and Stochastic based approach.

3. Memory-based approach

Memory-based learning is a supervised learning approaches of POS tagging. Supervised learning approaches are helpful when a tagged corpus is available as an

example of the desired output of the tagger. Memory-based tagging allows the relatively small tagged corpus size for training, incremental learning, details capabilities, supple integration of information in case representations, its non-parametric nature, rationally good results on unknown words without morphological analysis, and fast learning and tagging (Ghose, 2013).

4. Stochastic Part of Speech Tagging

The nature behind the stochastic part of speech tagging is to get the most often used tag for a specific word in the annotated training data and uses the same information to tag the word in the unannotated text. Stochastic taggers have data-driven approaches in which frequency based information is automatically derived from corpus and used tag words. Stochastic POS taggers uses large amount of annoted data or tagged corpus for the development of POS tagger in a shorter time. Stochastic taggers ascertain words based on the possibility that a word occurs with a particular tag. The simplest scheme is to assign the most frequent tag to each word. Hidden Markov Model (HMM) is the standard stochastic tagger. Stochastic models of part of speech tagging can be used for simplicity and language independent models. Stochastic taggers with more than 95% word level accuracy have been developed for English, German and other European languages, for which big tagged data is available. The disadvantage of this approach is that it comes with sequences of tags for sentences that are not satisfactory according to the grammatical rules of a language.

5. Supervised POS Tagging

Supervised POS tagging need pre tagged corpus for training to study information about tagged set, word tagged frequencies, rules set etc (Kumar *et al.,* 2006). The efficiency of the learning system generally increases with the increment of the training corpus set. Due to lack of annoted tagged corpus supervised POS tagging model play a vital rule in POS tagging in different languages. SVM and HMM are the examples of supervised POS tagging techniques.

6. Unsupervised POS Tagging

Unsupervised POS tagging does not require pre tagged corpus, on the other hands, unsupervised model uses the highly developed computational methods like Baum-Welch algorithm, transformation rules etc. to automatically induce tag sets. Unsupervised POS tagging is concerned with unlevel data by estimating model parameters.

7. Hybrid POS tagging

It is the hybridization of more than one process as describe above. Hybrid methods try to complement the weakness of individual approach with the strengths of others. Hybrid approaches overcome many of inherent limitations of single units. It also detects errors like stop word errors, merge errors, splits errors etc.

8. HMM based tagger

Hidden Markov Model tagger normally selects a tag sequence for a whole sentence othar than for a single word (Dias *et al.,* 2003). In a given sentence or word sequence, HMM taggers choose the tag sequence that maximizes the following formula:

P (word | tag) * P (tag | previous n tags)

To find the maximum probability HMM uses the Viterbi algorithm. The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden state called the Viterbi path that result in a sequence of experimental events, especially in the context of Markov information sources and hidden Markov models.

Generally, the natural language understanding system consists of a set of pipelined modules, each module playing a vital role in the investigation of the natural language text. As part of speech tagging is the primary step towards the understanding of natural language, it is very important to attain a high level of accuracy, else it may create problem in natural language processing.

4.10 Applications of Part of Speech Tagging

We discuss briefly some applications of POS tagging as follows:

(i) **Speech synthesis and recognition:** Part of Speech plays a significant role in information about word and its surrounding words which can be useful in language model for speech recognition. Part of speech of a word indicates how the word is pronounced depending on the grammatical category (Heeman and Allen, 1997).

(ii) **Information retrieval and extraction**: If any query is given with part of speech information from a retrieval system, that can be extracted with accurate result.

(iii) **Machine Translation (MT)**: MT allows translating words, sentences from the source language to target language automatically learned from parallel corpora, dependent on the part of speech category of the source word.

(iv) **Chunking:** Chunking is done to mark the Noun phrases and the Verb phrases since much of the key phrases are noun phrases. It help to select the subset of the words after tagging the proper grammatical categories to each word or token of the sentence.

Part of speech tagging is a useful technique in the field of NLP. It is the primary stage of natural language understanding based on which further processing like chunking, parsing etc are normally done. Part of speech tagging is used in a number of applications, including- speech synthesis and recognition, information extraction, partial parsing and machine translation etc (Jurafsky and Martin, 2009).

Apart from above cited applications point of view, POS tagging is also a useful technique that increases data retrieval process from corpora and provides basic grammatical information about words required in grammar development, semantic annotation, discourse annotation, parsing, dictionary compilation, language teaching, and language planning.

4.11 Chapter Summary

In this chapter, we discussed Part of Speech tagging in Bengali and their need in Bengali MWEs extraction and detection with different paradigms, various POS taggers, standard approaches and list of POS approved by CIIL, Mysore (India) for Bengali POS tagging ,different POS with corresponding tag with examples and finally applications of POS tagging.