Approaches of POS Tagging Algorithm for Bangla Corpus

Maksuda Sultana[1] and Francis G. Balazon[2]
[1]College of Computer Study Department, AMA University, Philippines
[2]College of SGS Department, AMA University, Philippines
[1]mak.sultana@gmail.com, [2]fbalazon@yahoo.com

***Abstract* —** Parts of speech is the process of classifying words into their parts of speech and labeling them accordingly in lexical categories and by using this POS tagging it is very easy to identify the words as nouns, verbs, adjectives etc. in each word in a natural Language sentence. For building lemmatizers which we are used to reduce a word to its root form in natural processing language, the POS tagging is essential part. The text analysis, machine translator, information retrieval and text to speech synthesis etc. POS tagging is initial stage in NLP application. Now a days to implement POS tagger various approaches have been proposed. In this paper Trigram and HMM methods are using to develop the tagger in general statistical approach and present a clear idea about this algorithm and also represent tag set with Indian corpus for tagging Bangla text for trying to find the accuracy of taggers output. This paper also presents the various development in POS taggers and POS-tag-set for Bangla language, which is very important computational verbal tool needed for natural language processing (NLP) presentation [1].

***Keywords*—** Tag-set, Ambiguity, Trigram, HMM, NLP, Token, Corpus, Bangla Language.

I. INTRODUCTION

Bengali is the fourth spoken language in internationally. Almost 210 million people in spoken by this language. With some a hundred million Bengali speakers in Bangladesh, about 85 million in India, usually inside the states of West Bengali, Assan and Tripura and good sized immigrant communities within the United Kingdom, the Unites States, and the Middle East. Therefore, for this large number of people, in addition to expanding the field of Bangle language research it is necessary to enhance modern artificial intelligence technology. Natural language processing is subpart of Artificial Intelligence and very much important because any type of language is processed by natural language processing. It is almost ended to discover all the modern methods for all the important languages of the world but in comparison, the Bengali language is a little behind.

Part of Speech tagging is the process of assigning a part of speech tag to each word which we called token in the sentence. Identification of the parts of speech such as nouns, verbs, adjectives, adverbs for each word of the sentence helps in analyzing the role of each word or token in a sentence. There are eight parts of speech is stated in in the Bengali grammar book: Noun, Pronoun, Adjective, Adverb, Finite Verb, Non-finite Verb, Postposition and indeclinable. People also easily distinction among plural and singular nouns and pronouns; phrases marked with case makers and inflections in texts; grammatical gender observed with phrases; nouns and adjectives marked with gender and variety markers; verbs connected with man or woman, number, gender, worrying, factor, modality, and different markers; adjectives marked with suffixes of degree; and so forth. Information of tagging of these kind of types is required in an algorithm evolved for automated POS tagging of phrases in a language text through a computer system. Else a POS tagging algorithm will fail to hint numerous precise linguistic information of words, while a POS tagged text will fail to show off the finer linguistic features of phrases the statistics of which required in next linguistic studies and development activities associated with each theoretical and applied linguistics. It is quite tough for developing Part of Speech tagger for Bangla language because of morphological richness and large annotated corpora and also lack of weird linguist.

II. CLASSIFICATION OF POS TAGGER

The collection of tags used for a particular task in Parts is Speech tagging is known as a tagset. We can say a Part-Of-Speech Tagger or POS Tagger is a piece of software and this software can reads text in some language and assigns parts of speech to each word and other token, for example as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural' [3]. There are various types of POS tagging and we divided into three categories: a) rule based, b) statistical tagging and c) hybrid tagging. Rule-Based techniques are used along with Lexical Based approaches to allow POS Tagging of words that are not present in the training corpus but there are in the testing data. Frequency and probability are includes in statistical approach. This method assigns the POS tags based on the probability of a particular.

sequence occurring. To assign a POS Tag Conditional Random Fields. (CRFs) and Hidden Markov Models (HMMs) are probabilistic approaches.

In this paper the author discusses the different types of statistical tagging approaches which are Trigram and HMM and also shows the evaluation done and the comparative study of their result. Different approaches are works on POS tagging.

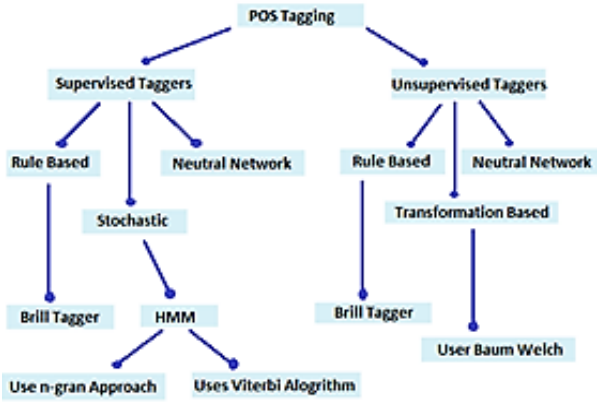


Fig-1 shows the different models of POS tagging approaches

2.1 IMPORTANCE OFA POS TAGGED CORPUS

Corpus is one of the first things required for natural language processing (NLP) responsibilities. Corpus (literally Latin for body) refers to a set of texts in linguistics and NLP. This sets of texts may be formed of a single language of texts, or can span multiple languages. Actually corpora are completely used for statistical linguistic evaluation and hypothesis testing.

In natural processing applications for developing system for grammar checkers, sentence parsing, text understanding, query addressing, information retrieval this is the first step. A tagged text corpus is also useful for machine learning, extraction of linguistic properties. In the area of applied linguistics and mainstream linguistics a POS tagged corpus ls very useful for the frequency calculation of words, type token analysis of word, lemmatization, lexical sorting etc.

2.2: PROBLEMS IN BANGLA CORPUS

The unknown and ambiguous words specially those words whose more than one tag can exit is the main problem in of Bangla tagging process [5]. By emphasizing on context rather than single words we can solve this problem. But this process is easy task for humans but not easy for the automatic word taggers. And also Bangla corpus are limited. So in the POS tagging process sometimes we can find such words those have different tag categories. These are called lexical ambiguity. Ambiguous words are the main problem in Bangla corpus when we are doing part of speech tagging. Some words have different meaning in different context but they have same POS taggings [6].

In English the word sound can be tagged as a noun, an adjective (e.g., a sound decision), as well as a verb (e.g., he sounds rational) based on the context of its use in sentences. Similarly, in Bengali, the word হাত (hāt) 'hand', can be tagged as a noun, as a finite verb, as well as a non-finite verb based on its use in different contexts.

Some example are given below:

(1) তার হাতে অনেক টাকা এসেছে

(tār hāte[NN] onek ṭākā eseche)

" He gained lots of money"

(2) সে অনেক টাকা হাতিয়েছে

(se onek ṭākā hātiyeche[FV])

"He grabbed lots of money"

(3) সে সব টাকা হাতিয়ে নিয়েছে

(se sob ṭākā hātiye[NFV] niyeche)

"He stole all the money"

In sentence (1) the word হাতে (hate) is a noun(NN), in sentence (2) the word হাতিয়েছে(hatiyeche) is finite verb(FV) and in sentence(3) the word হাতিয়ে (hatiye) is nonfinite verb(NFV).

But these words are actually derived from the noun হাত (hat) 'hand' by adding different suffixes. But any Bengali speaker can easily indicate these word which are noun, finite verb or non-finite verb and can perform easily.

So in grammatical and semantic analysis of these words innately based on his internalized linguistic rules and grammar, a computer system which is being trained to tag words automatically in natural text corpora, needs elaborate linguistic rules and conditions to perform the task of identifying parts-of-speech of words in texts.

Only six years ago, arguably the first generic POS tagset for Bengali was designed by an individual to tag manually a text database of nearly hundred thousand words of modern Bengali prose for academic and training purposes (Dash 2005a). [8]

III. METHODOLOGY

3.1: Corpus Creation:

Generally, a corpus is a large collection of data. It gives grammarians, word specialists, and other invested individuals with better descriptions of a language. Indian corpus contains a collection of Bangla, Hindi, Marathi, and Telugu language data. To work with any language, first we have to import NLTK (Natural Language Tool Kit). Then the Indian corpus has to be imported from the NLTK.

Corpus package automatically creates a set of corpus reader instances that can be used to access the corpora in the NLTK data package.

3.2 Tag-set finder

Each word is assigned a set of tags in tag-set finder. By providing information required to determine word feature the tag-set finder supports fetching word information.

The module of Tag set finder restrain information about words observed in the corpus.. So we can say by providing information required to determine word feature, any tag-set finder can bear fetching word information.

3.3 Tag analyzer

The function of tag analyzer is split the corpus into sentence and then split the sentence into words and after complete this store those words into lexicon table which we call Disk. After then the tagger tags the words in a sentence with their related tags. And finally completion of tagging words the tester module provides us the final test result.

3.4: Trigram

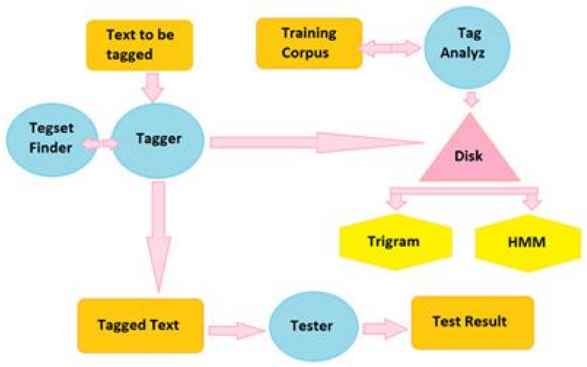
In computational linguistics, a trigram tagger is a statistical method for automatically identifying words as being nouns, verbs, adjectives, adverbs, etc. based on second order Markov models that consider triples of consecutive words. ... The description of the trigram tagger is provided by Brants (2000). For Trigram Model of POS tagger, we focus to perform POS Tagging to determine the most likely tag for a word, given the previous two tags. So if tag sequence aret1, t2 …tn and corresponding word sequence are w1, w2…wn then the following equation explains this fact- P (ti/wi) = P (wi/ti). P (ti/ti-2, ti-1)..... (1)

Where ti indicate tag sequence and wi indicate word sequence. P (wi/ti) is the probability of current word given current tag.

Here, P (ti | ti−2ti−1) is the probability of a current tag given the previous two tags.

This provides the transition between the tags and helps capture the context of the sentence. These probabilities are computed by following equation.

P (ti/ti-2, ti-1) =f (ti-2, ti-1, ti)/f (ti-2, ti-1)…… (2)



Each tag transition probability is computed by calculating the frequency count of two tags which come together in the corpus divided by the frequency count of the previous two tags coming in the corpus.

3.5: HMM Tagger

The hidden Markov model (HMM) based tagger assigns POS tags by searching for the most likely tag for each word in a sentence. However, a HMM based tagger finds a tag sequence for a sentence as a whole, rather than finding a tag for each word separately. Given a sentence w1,…,wn, a HMM based tagger chooses a tag sequence t1,.....,tn that maximizes the following joint probability:

$(t1…tn,1…..wn)=(t1…tn)P(w1…wn|t1…tn)$

In practice, it is often impractical to compute P (t1: :: tn). Therefore many different taggers have been proposed to simplify this probability computation. TnT, one of the most commonly used HMM based tagger, uses second order Markov models to simplify the computation; it assumes that the tag of a word is determined by the POS tags of the previous two words. Tree tagger is another popular HMM based tagger, which leverages decision trees to get more reliable estimates of parameters in Markov models [11].

The following equation created based on this model-

P (ti/wi) = P (ti/ti-1). P (ti+1/ti). P (wi/ti)……. (3)

P (ti/ti-1) is the probability of current tag given previous tag.

P (ti+1/ti) is the probability of future tag given current tag.

P (wi/ti) Probability of word given current tag

It is calculated as-

P (wi/ti) = freq (ti, wi)/ freq (ti)……. (4)

Elements of a Hidden Markov Model (HMM)

A hidden Markov model, ΦΦ, typically includes the following elements [10]:

- Time: t={1,2,...,T}t={1,2,...,T};
- NN States: Q={1,2,...,N}Q={1,2,...,N};
- MM Observations: O={1,2,...,M}O={1,2,...,M};
- Initial Probabilities: πi=p(q1=i), 1≤i≤Nπi=p(q1=i), 1≤i≤N;
- Transition Probabilities: aij=p(qt+1=j|qt=i), 1≤i,j≤Naij=p(qt+1=j|qt=i), 1≤i,j≤N;
- Observation Probabilities: bj(k)=p(ot=k|qt=j) 1≤j≤N,1≤k≤Mbj(k)=p(ot=k|qt=j) 1≤j≤N,1≤k≤M.
- The entire model can be characterized by Φ=(A,B,π)Φ=(A,B,π),where A={aij}A={aij}, B={bj(k)}B={bj(k)} and π={πi}π={πi}.

The states are "hidden", since they are not directly observable, but reflected in observations with uncertainty.

3.6. Tester

Based on three different domain of test corpus tester perform the testing and produce the result from tagged data.

IV. BANGLAE TAG SET FOR PART OF SPEECH TAGGING

For developing tagger, the first requirement is annotating a corpus based on a tag-set. We are going to used IL POS tag-set[12] proposed by Bharti et. Al. Table:2 describe the tags which tags are used. In their paper they have almost 20 semantic tags and 15 syntactic tags. For POS tagging and chunking of the Indian languages a common tag-set has been designed. They are 26 lexical tags are used in this tag set.

Sl NO	Top Level	Subtype/ Grammatical word	TAG	Example
1	Noun (N)	Common	NN	কলম(kolom), চশমা(Chasma),
2		Proper	NNP	মহন(Mohon), রবি (Robi)
3		Verbal	NNV	Not required for Bangla
4		Nioc	NST	উপরে(Upore), নিচে(Niche), ভিতরে(Vitore)
5	Pronoun(PR)	Personal	PRP	আমি(Ami), সে (she), তুমি (tumi), আমরা(amra)
6		Reflexive	PRF	নিজেকে(Nijeke)
7		Relative	PRL	যে(ze), যারা(zara), যাদের (Zader)
8		Reciprocal	PRC	পরস্পর(porospor)
9		Wh-word	PRQ	কে(ke), কাকে (kake), কারা (kara), কাদের(kader)
10	Demonstrative (DM)	Deictic	DMD	ে(a), এই(ai),সে (se), সেই(sei), ও(o), ঐ(oi)
11		Relative	DMR	যেje), যেই(jei), যাহা(jaha)
12		Wh-word	DMQ	কোন(kon), ্কোন(kono)
13	Verb(V)	Finite	VF	যাব(zabo), গেল(gelo), করেছিলাম (korechilam)
14		Non-finite	VNF	করে(kore), করলে(korle), খেয়ে(kheye), খেতে(khete)
15		Infinite	VINF	করতে(korte),যে তে (jete), খেতে(khete)
16		Gerund	VNG	যাওয়া(zaoya), আছি(asa), পরে(pora), দেখা (dekha)
17		Auxiliary	VAUX	ছিল(chilo), হবে(hobe), আছে(ache)
18	Conjunction (CC)	Coordinative	CCD	আর(ar), েবং(ebong)
19		Subordinate	CCS	কিন্ত(kinto), নইলে(noile)
20				
21		Default	RPD	তো(to),যে (ze),

22	Particles(RP)	Classifier	CL	(খানাkhana),
23		Interjection	INJ	আরে(are), হে(hay)
24		Intensifier	INTF	খুব(khub), অতি(oti)
25		Negation	NEG	না(na), নি(ni)
26	Quantifiers (QT)	General	QTF	কিছু(kichu), অল্প(olpo)
27		Cardinals	QTC	এক(ak), দুই(doi)
28		Ordinals	QTO	প্রথম(prothom),
29	Residuals	Foreign word	RDF	Word written in script other than Bengali
30		Symbol	SYM	$, &, *, (,), etc
31		Unknown	UNK	
32		Ech-word	ECH	তোল(tol), বই-টই(boi-toi)
33	Adjective	Default	JJ	তারা(tara), সর্বোত্তম(Sorbottoma)
34	Adverb	Default	RB	খায় (khay)
35	Preposition	Default	PSP	ভিতরে(Vitore), নিচে (Niche)

V. PRACTICAL WORK

5.1: In order to perform our system we apply Trigram and HMM method for bangle text by using hindi corpus. For this purpose we developed a test corpus of 500 sentence
(13765 words) and finally we get the results of all POS taggers in term of accuracy. The accuracy was calculated by using the following formula:
Accuracy (%) = (No. of correctly tagged token/ Total no. of
POS tags in the text)*100

5.1.1: For trigram in games sentence:
System assigned the total number of correct POS tag =131187 and total number of POS tag in the text = 14150
Thus the accuracy of the system is 91.18%.

[('বাংলাদেশ', 'Unk'), ('দলের', 'Unk'), ('সাবেক', 'Unk'), ('এই', 'DEM'), ('অধিনায়ক', 'Unk'), ('আরও', 'QF'), ('বলেন', 'VM'), (',', 'SYM'), ('আইপিএল', 'Unk'), ('হলে', 'VM'), ('হবে', 'VAUX'), (',', 'SYM'), ('না', 'NEG'), ('হলে', 'VM'), ('নাই।', 'Unk'), ('এটা', 'PRP'), ('আমার', 'PRP'), ('কাছে', 'NST'), ('খুব', 'INTF'), ('একটা', 'QF'), ('ব্যাপার', 'NN'), ('না।', 'Unk'), ('বাংলাদেশের', 'Unk'), ('হয়ে', 'VM'), ('খেলার', 'Unk'), ('চাইতে', 'VM'), ('গর্ব', 'Unk'), ('করার', 'VM'), ('মতো', 'PSP'), ('কিছুই', 'PRP'), ('হতে', 'VM'), ('পারে', 'VAUX'), ('না।', 'Unk')]

5.1.2: For trigram in ICT sentence:

System assigned the total number of correct POS tag = 141587 and total number of POS tag in the text = 15150
Thus the accuracy of the system is 91.48%.

'তথ্য', 'NN'), ('ও', 'CC'), ('যোগাযোগ', 'Unk'), ('প্রযুক্তি', 'Unk'), ('(', 'SYM'), ('আইসিটি', 'Unk'), (')', 'SYM'), ('প্রতিমন্ত্রী', 'Unk'), ('জুনাইদ', 'Unk'), ('আহমেদ', 'Unk'), ('পলক', 'Unk'), ('বলেছেন', 'VM'), (',', 'SYM'), ('প্রধানমন্ত্রী', 'Unk'), ('ও', 'CC'), ('তথ্য', 'NN'), ('উপদেষ্টার', 'Unk'), ('সঙ্গে', 'PSP'), ('কথা', 'NN'), ('বলে', 'UT'), ('২০২২', 'Unk'), ('সালে', 'NNP'), ('আন্তর্জাতিক', 'Unk'), ('রোবট', 'Unk'), ('অলিম্পিয়াড', 'Unk'), ('বাংলাদেশে', 'NNP'), ('আয়োজনের', 'Unk'), ('ব্যবস্থা', 'NN'), ('নেওয়া', 'VM'), ('হবে।', 'Unk'),

5.2.1: For HMM in games sentence:

System assigned the total number of correct POS tag = 131187 and total number of POS tag in the text = 14150
Thus the accuracy of the system is 94.18%.

[('বাংলাদেশ', 'NN'), ('দলের', 'Unk'), ('সাবেক', 'Unk'), ('এই', 'DEM'), ('অধিনায়ক', 'NN'), ('আরও', 'QF'), ('বলেন', 'VM'), (',', 'SYM'), ('আইপিএল', 'Unk'), ('হলে', 'VM'), ('হবে', 'VAUX'), (',', 'SYM'), ('না', 'NEG'), ('হলে', 'VM'), ('নাই।', 'VN'), ('এটা', 'PRP'), ('আমার', 'PRP'), ('কাছে', 'NST'), ('খুব', 'INTF'), ('একটা', 'QF'), ('ব্যাপার', 'NN'), ('না।', 'Unk'), ('বাংলাদেশের', 'Unk'), ('হয়ে', 'VM'), ('খেলার', 'Unk'), ('চাইতে', 'VM'), ('গর্ব', 'Unk'), ('করার', 'VM'), ('মতো', 'PSP'), ('কিছুই', 'PRP'), ('হতে', 'VM'), ('পারে', 'VAUX'), ('না।', 'Unk')]

5.2.2: For trigram in ICT sentence:

System assigned the total number of correct POS tag = 141587 and total number of POS tag in the text = 15150
Thus the accuracy of the system is 94.48%.

'তথ্য', 'NN'), ('ও', 'CC'), ('যোগাযোগ', 'Unk'), ('প্রযুক্তি', 'Unk'), ('(', 'SYM'), ('আইসিটি', 'NN'), (')', 'SYM'), ('প্রতিমন্ত্রী', 'Unk'), ('জুনাইদ', 'NN'), ('আহমেদ', 'Unk'), ('পলক', 'NN'), ('বলেছেন', 'VM'), (',', 'SYM'), ('প্রধানমন্ত্রী', 'Unk'), ('ও', 'CC'), ('তথ্য', 'NN'), ('উপদেষ্টার', 'Unk'), ('সঙ্গে', 'PSP'), ('কথা', 'NN'), ('বলে', 'UT'), ('২০২২', 'PR'), ('সালে', 'NNP'), ('আন্তর্জাতিক', 'Unk'), ('রোবট', 'NN'), ('অলিম্পিয়াড', 'NNP'), ('বাংলাদেশে', 'NNP'), ('আয়োজনের', 'Unk'), ('ব্যবস্থা', 'NN'), ('নেওয়া', 'VM'), ('হবে।', 'Unk'),

*5.3:*The programming approach by using NLTK a corpus is a large collection of data. It gives grammarians, word specialists, and other invested individuals with better descriptions of a language. Indian corpus contains a collection of Bangla, Hindi, Marathi, and Telugu language data. To work with any language, first we have to import NLTK(Natural Language Tool Kit). Then the Indian corpus has to be imported from the NLTK.

Define **tnt** from **nltk.tag** for tagging each token in a sentence with supplementary information. **TnT()** is a statistical tagger which follows second-order Markov model. This model is used for probability prediction of time series and sequence.

Then we place a variable(**tagged_set**) where pre-trained Indian corpus is stored(**bangla.pos**).From Bengali corpus read the Bengali sentence and put them variable **word_set**. Using a **for** loop count all sentences which present in the corpus. **startswith()**-function is used to check the string is started with String **" ' "**. Here set the training percentage is **0.96** since the dataset is not sufficient.

train()-method is used for explicitly use of **TnT**. After the train data using evaluate() method check the performance of the trained dataset. Our evaluation score was **0.51** for using methods of Bengali data.

For the test, result user needs to provide a Bengali text here. Which is stored in a variable. Then using **word_tokenizer()** split the sentences and check parts of speech of the words [13].

VI. RESULT AND CONCLUSION

The average result of this system in Trigram is 91.18% and HMM is 94.48%. Actually if we compared the POS tagged in Bangla corpus with other POS tagged corpora such as English then we can find the rate of accuracy is very far. For example, in one million word in English text database of the American National Corpus the rate of accuracy is 97 to 98% but in ten thousand to one hundred thousand words corpora of Indian language the rate of accuracy is 85 to 90%. SO from this survey it clearly indicates that we need to take seriously initiative to develop accurate tegset to increase the rate of accuracy of Bangla POS tagged data which covering all text types for future linguistic works.

REFERENCES

[1] Antony P J, Amrita, Dr. K P Soman, "Parts Of Speech Tagging for Indian Languages: A Literature Survey", IJCA (0975-8887) Volume 34-no. 8, November 2011. IJCATM

[2] Dinesh Kumar and Gurpreet Singh Josan, (2010) "Part of Speech Tagger for Morphologically rich Indian Language: A survey". International Journal of Computer Application. Vol. 6(5). https://nlp.stanford.edu/software/tagger.shtml

[3] Brants, Thorsten (2000) "TnT- A Statistical Part-of-Speech Tagger". In the Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000), Seattle, WA, USA,

[4] Brill, Eric (1992) "A simple rule-based part of speech tagger". In the Proceedings of the Workshop on Speech and Natural Language (HLT-91), Morristown, NJ, USA:Association for Computational Linguistics. Pp. 112-116.

[5] Dhanalakshmi V, Anand Kumar1, Shivapratap G, Soman KP and Rajendran S, "Tamil POS Tagging using Linear Programming", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

[6] Gurleen Kaur Sidhu, Navjot Kaur, "Role of Machine Translation and Word Sense Disambiguation in Natural Language Processing", IOSR Journal of Computer Engineering (IOSR-JCE), May. - Jun. 2013.

[7] Akshar Bharathi and Prashanth R. Mannem (2007), "Introduction to the Shallow Parsing Contest for South Asian Languages", Language Technologies

Research Center, International Institute of Information Technology, Hyderabad, India 500032.

[8] Dinesh Kumar and Gurpreet Singh Josan,(2010), "Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey", International Journal of Computer Applications (0975 – 8887) Volume6– No.5, September, 2010, www.ijcaonline.org/ volume6/number5 /pxc3871409 .pdf..

[9] Debasri Chakrabarti (2011), "Layered Parts of Speech Tagging for Bangla", Language in India www.languageinindia.c o m, M a y 2 0 1 1, Special Volume:Problems of Parsing in Indian Languages.

[10] https://stlong0521.github.io/20160319%20-%20HMM%20and%20POS.html

[11] Nisheeth Joshi, Hemant Darbari, Iti Mathure, (2013) "HMM based Pos Tagger for Hindi". In Processing of 2013 International Conference on Artificial Intelligence and Soft Computing.

[12] Akshar Bharti, Dipti Misra Sharma, Lakshmi bai, Rajeev Sangal. AnnCorra: Annotating Corpora Guidelines for POS and Chunk with Annotation For Indian Languages , Language Technologies Research Centre IIT, Hyderabad.

[13] https://medium.com/analytics-vidhya/bengali-pos-part-of-speech-tagging-using-indian-corpus-e85f47d3ad65

[14] Antony P J, Amrita, Dr. K P Soman, "Parts Of Speech Tagging for Indian Languages: A Literature Survey", IJCA (0975-8887) Volume 34-no. 8, November 2018.

[15] Bhasa Bijnan o Prayukti: An International Journal on Linguistics and Language Technology Vol. 1, No. 1, Jan-Jun 2017, Pp. 53-96

[16] Sag, Ivan A., Timothy Baldwin, Francis Bond, Aann Copestake and Dan Flickinger (2001) "Multiword Expressions: A Pain in the Neck for NLP". In, Gelbukh, Alexander (Ed.) Proceedings of CICLING2002. Verlag: Springer. Pp. 35-41.

[17] https://www.ijariit.com/manuscripts/v2i3/ V213-1157.pdf

[18] https://www.scribd.com/document/140115676/PART-OF-SPEECH-TAGGING-OF-MARARHITEXT-USING-TRIGRA METHOD