Review of Stochastic POS tagging techniques used in Bengali

Abul Kalam Md. Rajib Hasan
School of Engineering and Computer Science, Chittagong Independent University, Bangladesh

ABSTRACT

In this paper, we describe different stochastic methods or techniques used for POS tagging of Bengali language. We have shown a generalized stochastic model for POS tagging in Bengali. We reviewed kinds of corpus and number of tags used for tagging methods. In the study it is found that as many as 45 useful tags existed in the literature. There are four useful corpus found in the study. As Bengali is a morphologically rich language we outlined a feature list that could be used with different training algorithms. We found that a hybrid HMM model used with a morphological analyzer work best in Bengali with an accuracy of 96.3%.

General Terms

Natural Language Processing (NLP), Machine Learning.

Keywords

Natural Language Processing (NLP), Machine Learning.

1. INTRODUCTION

Many Natural Language Processing (NLP) task such as sentiment or opinion mining requires appropriate assignment of POS (Part of Speech) tags (i.e. whether a word belongs to Noun, Pronoun, Verb etc) for previously unseen text. A number of large corpora with pre annotated tags already exist for languages like English. One of such corpora is Brown Corpus which is seen embedded with some NLP tools or applications (For Example NLTK [1]). These tools use statistical learning algorithms to learn from the corpus in order to assign tags automatically for a document. This paper is a review of stochastic statistical techniques applied for POS tagging of Bengali Language. Bengali is a language of more than 2 billion people around the world. A good number of Bengali websites and blogs have already been published and counting more. Phonetic typing made it easy for people writing Facebook status updates, comments, tweets in Bengali. As a result mining Bengali text would be interesting for researchers and organizations. Text mining requires automatic POS tagging of words as the beginning step of the process.

In this paper we have outlined a number of corpuses available in Bengali, tag set and different statistical model used for training the system to automatically identify POS tag of a word. Any approximation and parameter classification of the model is avoided and we compared their accuracy in tagging Bengali Language text that was described in different research.

2. RELATEED WORK

A comparative study of was found in [2] where comparison of Uni gram, Bi gram, Hidden Markov Model and Brill's POS tagger is done with respect to South Asian Language The study in [2] has detail literature survey of POS tagging of South Asian Language. Dash [3] has detailed morphological analysis of Bengali Language and explained tag set in different layers. In [4][5][6] we have found different stochastic processes used for tagging Bengali texts.

3. GENERALIZE POS TAGGGING STOCHASTIC MODEL

A POS tagger takes a sentence as input and assigns a unique part of speech tag (i.e. Noun, Pronoun, Verb etc) to each lexical item of the sentence. Generally the rule for POS tagging is learned from a pre tagged text corpus or rules from lexicon and then train the system to tag untagged text corpus. A generalized POS tagging model is shown in figure 1.

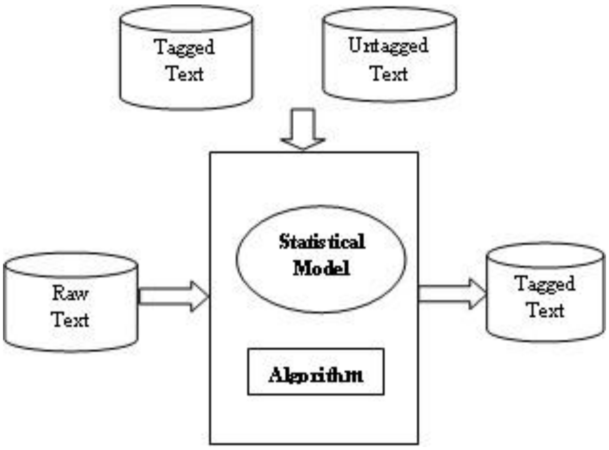

Fig 1: Generalized POS tagging stochastic model

4. CORPUS AVAILABLE IN BENGALI

Bengali is a morphologically rich language. It is very difficult to find out exact rules for construction of sentences and as such rule based approaches did not result well in many cases which we shall see in our analysis in later section as well. That is why we found that most of the processes deployed for POS tagging are stochastic processes. But stochastic processes require long corpora for learning and the accuracy of the system mostly depend on the corpora used for training the system. In our study we have found four types of corpora used by the researchers when it came to Bengali languages. They are:

1. The EMILLE Corpus has been constructed as part of a collaborative venture between the EMILLE project (Enabling Minority Language Engineering), Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India. EMILLE is distributed by the European Language Resources Association. This corpus has as many as 14 Indian Languages including Bengali and can be used for research and teaching purpose [7].
2. A corpus developed by Microsoft Research (MSR) as a part of the Indian Language Part-of-Speech Tag set(IL-POST) project, a collaborative effort among linguists and computer scientists from MSR India, Anna University, Chennai, Delhi University, IIT Bombay, Jawaharlal Nehru University (Delhi) and Tamil University (Tamilnadu).The corpus has 7168 sentences (102933 words) of manually annotated text from modern standard Bengali sources including blogs, Wikipedia and a portion of the EMILLE/CILL corpus[8].
3. NLTK (Natural Language Toolkit) Indian languages corpus has less amount of sentences and words in their corpus. Following is the statistics captured from NLTK[1]:

   Total number of words in Bangla is 10281
   Hindi 9408
   Marathi 19066
   Telugu 9999

   Total number Sentences in Bangla 899
   Hindi 541
   Marathi 1197
   Telugu 994

Many Researchers using NLTK with this small corpus does not get the best result for unknown words and multi words.

4. SPSAL (Shallow Parsing for South Asian Languages) workshop on 2007 had released a corpus of 5000 sentences and 20000 words. The corpus was tagged in IIT-H tag set in Shakti Standard Form(SSF).Many of the papers described in this paper had used this corpus though any online version of the corpus was not found through GOOGLE.

5. POS TAG SET

Most of Indian Languages uses 26 Tag set for POS tagging. In our study we have found many variations of number of POS tags used in different training corpora. Though 26 Tag sets are used for above mentioned Corpora we have found out few more Tags for Bengali Languages.

BIS tag set proposed for Indian Languages has 45 tags [3]. For example, while the tag ‘NN’ stands for singular common nouns, ‘NS’ stands for plural common nouns, ‘NP’ stands for singular proper nouns, etc. In table 1 we outline top level 15 tags used in Bengali with example found in [3].

Table 1: Top level POS tags with example in Bengali

POS Categories	Label	Example
Noun	[NN]	বালক (bālak), শহর (śahar), কথা (kathā)
Pronoun	[PR].	আমি (āmi), তুমি (tumi), সে (se)
Demonstrative	[DM]	যে (ýe), এই (ei), ওই (oi), তাই (tāi), etc
Finite Verb	[FV].	করছি (karchi), করতাম (kartām)
Non-Finite Verb	[NF]	করলে(karle), করতে (karte),গেলে (gele),
Adjective	[AD]	ভাল (bhāla), মন্দ (manda) etc
Adverb	[AV]	হঠাৎ (haṭhāt), বাবদ (bābad) etc
Postposition	[PP]	পারে (pare), কাছে (kāche), আগে(āge), নিচে(nice), etc.
Conjunction	[CN]	তবে (tabe), যদি (ýadi), নইলে(naile), যাতে (ýāte), etc.
Indeclinable	[IN]	কিন্তু (kintu), অথবা (athabā), বরং (baraṃ), আর (ār), etc.
Particle	[PT]	ই (i), ও (o), তো (to), না (nā), নে (ne), নি (ni), etc.
Quantifier	[QT]	এক (ek), দুই (dui), প্রথম (pratham), পয়লা (paylā),
Reduplication	[RD]	বনে বনে (bane bane), কত কত (kata kata),যে যে (ýe ýe), etc.
Punctuation	[PN]	., : ; - / …, !, ? ( ), [ ], { }, etc.
Others	[OR]	Mathematical symbols, +, -, x, >, <, $, #, @, ^, &, * etc.

6. FEATURE SELECTION IN BENGALI

In many studies it is found that only statistically learning techniques not enough for morphologically rich Bengali language. For Example if we consider two sentences below:

a. সময় এখন অনেক **বদলে/VVF** গেছে।
“Time has indeed changed so much by now”

b. তুমি আজকের **বদলে/PSP** কালকে আস।
“Come tomorrow instead of today”

We can see that the word **বদলে (bodole)** has two tags VVF and PSP in two sentences. The word without the suffix is “বদল” (bodol) which means change and is a VB (Verb Base). Since suffix “ে”(e) is added the tag is changed to VF(Verb Finite) from VB in sentence “a”. On the other hand in sentence “b” because of the position the word became a Preposition (PSP).

For a more complex word analysis in Bengali we can divide a word into diffferent morphens.For example word অনাধুনিকতার(“anAdUnIktAr”) can be divided into the morphemes“an” (PREFIX), “AdUnIk” (ROOT), “tA”(DERIVATIONAL SUFFIX) and “r” (INFLECTIONAL SUFFIX).

We have found the following major features used by the researchers for morphological analysis. The features are also used for training the system in different statistical learning algorithm.

Word Suffix: A number of characters used after a word.

Word Prefix: A fixed length of characters surrounding the word.

POS tag of previous word: POS tag of the previous word is helpful to identify the tag of the current word.

Infelction List: Inflection list does not the tag of the word rather ascertain the tag. Asif Ekbal, Rejwanul Haque, and Sivaji Bandyopadhyay [4][5] used the following inflection list for training their maximum entropy and conditional random field based model:

a. Noun Inflection List(27 entries)

b. Adjective Infelction List(194 entries)

c. Verb Inflection List(214 entries)

7. POS TAGGING TECHNIQUES

Most of the POS tagger falls in two categories:

1. Supervised POS Tagging
2. Un Supervised POS Tagging

Supervised techniques require a pre tagged corpus written in the language to be processed where as such corpora is not required for the unsupervised techniques.

Both supervised and unsupervised techniques again can be of Rule based , Stochastic and Neural Network.

In our study we have compared among different stochastic based techniques used for processing Bengali text. Figure 2 depicts different techniques used for POS tagging Bengali Text.

7.1 Hybrid HMM POS Tagger

Given a sequence of words, Part of Speech Tagger is interested in finding the most likely sequence of tags that generates that sequence of words. In order to accomplish this, HMM Part of Speech Tagger makes two simplifying assumptions:

1. The probability of a word depends only on its tag. It is independent of other words and other tags.

2. The probability of a tag depends only on its previous tag. It is independent of next tags and tags before the previous tag.

Thus in a given sentence of words $w_1 w_2 w_3 \ldots\ldots\ldots w_n$ most likely sequence of tags $T_1 T_2 T_3 \ldots\ldots\ldots T_n$ is

$T_1 T_2 T_3 \ldots\ldots\ldots T_n =$

$$\arg\max_{T_1 T_2 T_3 \ldots\ldots T_n} P(T_1 T_2 T_3 \ldots\ldots T_n | w_1 w_2 w_3 \ldots \ldots . . w_n)$$

$$= \arg\max_{T_1 T_2 T_3 \ldots\ldots T_n} \prod_{i=1}^{n} P(W_i | T_i) \prod_{i=2}^{n} P(T_i | T_{i-1}) P(T_1) \quad (6.1)$$

Of course if we run the equation (6.1) without using an algorithm that will be inefficient. So for efficient running of the model Viterbi Algorithm [9] is used.

Sandipan Dandapat, Sudeshna Sarkar and Anupam Basu [10] presented a hybrid hidden markov model(HMM) for POS tagging Bengali sentences in 2007. It used hidden markov mode l(HMM) tagger initially. The training method is based on partially supervised learning. Initially it used 500 tagged sentences for supervised training (method 1) and then 50000 raw words for re estimating the parameters (method 2).In the third method they used a morphological analyzer to estimate the tags of a word in a sentence. Morphological analyzer outputs the best possible tag set for a word rather all the tags from the tag set. The brief description of the methods has given below [10]:

Method 1: POS tagging using only supervised learning

Method 2: POS tagging using a partially supervised learning and decoding the best tag sequence without using Morphological Analyzer restriction

Method 3: POS tagging using a partially supervised learning and decoding the best tag sequence without using Morphological Analyzer restriction.

The evaluation results are given in table 2 [10]:

Table 2: Result from different methods used with HMM

Method 1	Method 2	Method 3
64.31	67.6	96.28

7.2 POS Tagger using Maximum Entropy

Ratnaparkhi [11] has derived a maximum entropy model for POS tagging which is adopted by Asif Ekbal1, Rejwanul Haque, and Sivaji Bandyopadhyay [4]. The basic model is as follows:

$$\rho(h,t) = \pi\mu \sum_{j=1}^{k} \alpha_j^{f_j(h,t)} \qquad (6.2)$$

Where π is a normalized constant ,$\{\pi, \alpha_1, \alpha_2, \alpha_3, \ldots \ldots . \alpha_n\}$ are positive model parameters and $\{f_1, f_2, f_{3,} \ldots \ldots . , f_k\}$ are the known features and $f_j(h,t) \in \{0,1\}$.Each parameter α_j corresponds to a feature f_j.

The features are binary/multiple valued functions, which associate a POS tag with various elements of the history. For example:

$$f(h,t) = \begin{cases} 1, \ \text{word(h) = রাজীব and t = NNP} \\ 0, otherwise \end{cases}$$

Asif Ekbal, Rejwanul Haque and Sivaji Bandyopadhyay [4] used above model for finding out the best tag for a word in a given sentence.

The number of POS tags used for the experiment is 26 and the training corpus had 72341 words .This POS tagger had 88.2% accuracy for a set of 20K words.

7.3 POS Tagger using Conditional Random Field (CRF)

Conditional Random Fields (CRFs) [12], the undirected graphical models, are used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes. The conditional probability of a state sequence given an observation sequence is calculated as:

$$P(S|O) = \frac{1}{Z_o} \exp\left(\sum_{k=1}^{T} \sum_k \lambda_k f_k(S_{t-1}, S_t, O, t)\right) \qquad (6.3)$$

where, $f_k(S_{t-1}, S_t, O, t)$ is a feature function
Whose weight λ_k is to be learned via training. The
values of the feature functions may range between
-∞ to +∞, typically they are binary [5].

They used same corpus and with the same features used for maximum entropy model and the accuracy of the system was 90.3%.

7.4 Memory based POS Tagger

Kamal Sarkar and Arup Ratan Ghosh [6] proposed a memory based learning for POS tagger. Classification of tags is done using K nearest neighbor algorithm. A number of features similar to the features we discussed in section 5 developed. During training the feature vector of a tagged word is constructed and added to the memory. At the time of testing the unlabeled test pattern is compared with the labeled patterns stored in the memory and the distances are computed using Euclidean measure. Finally, the input is labeled as the class that is the mode of the classes of *K nearest* pattern vectors selected from the memory. The system achieved accuracy of 80.3%.

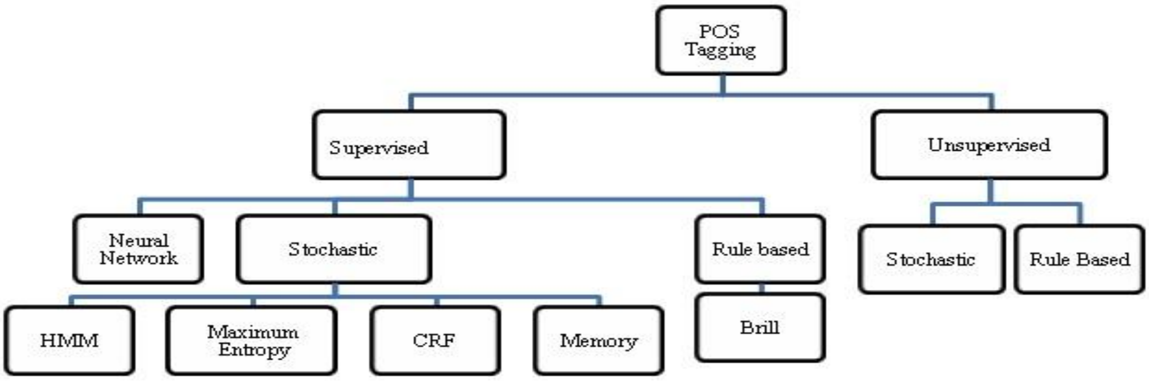

Fig 2: POS Tagging Techniques illustrated

8. COMPARISON OF MODELS

In this section we compare the models studied in this paper in terms of percentage of accuracy. We also have outlined number of words in the corpus and number of words used for testing. Table 3 has the details:

Table 3: Stochastic models comparison

Model	No of Tags	No of Words in Training Corpus	Number of Words Tested	Reported Accuracy
Hybrid Hidden Markov Model	26	72441	20K	96%
Maximum Entropy Model	26	72441	20K	88.2%
Conditional Random Field Model	26	72441	20K	90.3%
Memory Based	26	895	--	80.3%

9. FUTURE WORK

In our study we have not included unsupervised approaches to train the POS tagger system. This is due to the fact that not much research has been done is this field because of the unavailability of large corpus and less computational power. In different studies of other languages it was found that Brill's tagger [13] produce good result. We would like to find out result these approaches in our next study. Neural Network based approaches also not found in the literature. We are looking forward to results from algorithms incorporating Neural Network based approaches.

10. CONCLUSION

We have compared the accuracy of four stochastic based approaches for predicting POS tag of Bengali word from different researches [10][4][5][6].We have found that hybrid HMM model[10] works best in Bengali.

We also found that accuracy of the models depend on the size of the training corpora. In Bengali large training corpora is rare .So in near future we must build large Bengali corpus for Natural Language Processing (NLP) task.

11. EFERENCES

[1] NLTK 3.0: Natural Language Toolkit. http://www.nltk.org/

[2] Antony P J, Dr. Soman K P: Parts Of Speech Tagging for Indian Languages: A Literature Survey in International Journal of Computer Applications (0975 – 8887) Volume 34– No.8, November 2011.

[3] Dash, Niladri Sekhar ,"Part-of-speech (POS) Tagging of Bengali Written Text Corpus". Bhasa Bijnan o Prayukti: An International Journal on Linguistics and Language Technology.Vol. 1, No. 1, Jan-Jun 2013, Pp. 53-96.

[4] Asif Ekbal, Samiran Mandal and Sivaji Bandyopadhyay (2007), "Maximum Entropy Based Bengali Part of Speech Tagging", Workshop on shallow parsing in South Asian languages, shiva.iiit.ac.in/SPSAL2007/proceedings.php.

[5] Asif Ekbal, Samiran Mandal and Sivaji Bandyopadhyay (2007), "Bengali Part of Speech Tagging using Conditional Random Field", Workshop on shallow parsing in South Asian languages, shiva.iiit.ac.in/SPSAL2007/proceedings.php.

[6] Kamal Sarkar ,Arup Ratan Ghosh. A Memory Based POS Tagger for Bengali. http://www2.cse.iitk.ac.in/~iwml/2013/papers/116.pdf Accessed on 10/08/2014.

[7] The EMILLE/CIL Corpus http://catalog.elra.info/product_info.php?products_id=696

[8] Kalika Bali, Monojit Choudhury, Priyanka Biswas. Indian Language Part-of-Speech Tagset: Hindi. https://catalog.ldc.upenn.edu/LDC2010T24 . Accessed on 08/08/2014.

[9] G. D. Forney, Jr., "The Viterbi algorithm," Proc. IEEE, vol. 61, pp. 268–278, March 1973.

[10] S Dandapat, S Sarkar, A Basu : A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali. International conference on computational intelligence, 169-172

[11] Ratnaparkhi, A.: A maximum entropy part-of - speech tagger. In: Proc. Of EMNLP'96. (1996)

[12] Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. of the 18th ICML’01, 282-289.

[13] Eric Brill, “A Simple Rule-Based Part-of-Speech Tagger”, In Proceeding Of The Third Conference on Applied Natural Language Processing, Trento, Italy, 1992, pp. 152-155