International Advance Journal of Engineering, Science and Management (IAJESM)
ISSN -2393-8048, January-June 2019, Submitted in April 2019, iajesm2014@gmail.com

Approaches Review for Part Of Speech Tagging

Alok Kumar, Department of Computer Science and Engineering, University Institute of Engineering and Technology, Chhatrapati Shahu Ji Maharaj University, Kanpur-208024, India. akumar.uiet@gmail.com
Deepak Kumar Verma , Department of Computer Science and Engineering, University Institute of Engineering and Technology, Chhatrapati Shahu Ji Maharaj University, Kanpur-208024, India. deepak300572@gmail.com
Ruchika Rastogi, Department of Management, Pranveer Singh Institute of Technology, Kanpur 209305, India. rastogi.ruchika@gmail.com

Abstract:

In a natural language text, a word can occur with different lexical tags in different contexts. First step in all natural language applications is to get the accurate syntactic category of each word in a sentence based on its role in the text. This is also called part of speech tagging (POST). There are different approaches for assigning a part of speech (POS) tag to each word of a natural language sentence. Broadly the classification of approaches can be done as Rule based, Statistical based and Transformation based. Subsequent researches have added various modifications to these basic approaches based on Maximum Entropy, Decision tree, Conditional Random Field, Support Vector Machine, Neural Network etc.. Each approach has its advantages and disadvantages. In this paper, we elaborate the functionality of different approaches and try to present a comparison of latest approaches. The comparison shows that due to availability of large amount of data through internet, statistical techniques with some contextual and morphological knowledge of language are more powerful than the regular grammar based technique.

Introduction:

Part of speech (POS) Tagging is a process that attaches a suitable tag with each word of a sentence from a given set of tags. A word can have multiple categories and meaning. The main challenge in POS tagging involves resolving this ambiguity of possible POS tags for a word.

For example:

In Hindi: (1) वह आम आदमी है। (2) उस आदमी को आम खाना पसंद है।

In sentences (1), syntactic category of word "आम" is "**adjective**" and in sentences (3) syntactic category of word "आम" is "**noun**".

In English: (3) Mohan is reading a book. (4) Agents book railway tickets.

Similarly in sentences (3) and (4), the word "**book**" has role of "**noun**" and "**verb**" respectively.

In sentences (1) to (4) we can resolve the ambiguity on the basis of sentence structure. But in some cases, even the pattern of the sentence is not helpful in resolving the ambiguity and it can be resolved only on the basis of context. For example in the sentence "eq>s l¨uk pkfg;sA", syntactic category of the word "l¨uk" can be resolved only on the basis of context.

Tag of a word gives lexical and syntactic information of word i.e. syntactic Category, Number, Person, Gender etc.. The set of all tags is called Tagset. There are some standard tagsets for each Language and it's size varies from language to language. Part of Speech Tagging plays significant role in NLP, accuracy of tagging influences the performance of different applications of NLP. A number of approaches have been proposed by researchers. In this paper, we have made an effort to understand and analyze the existing approaches for POS tagging in Natural language.

Approaches Used for POS Tagging [1][2][3]**:**

The area of automated Part-of-speech tagging has been enriched over the last few decades by contributions from several researchers. We studied following approaches.

Rule Based[1][2][3]**:**

Initially, people manually engineered rules for tagging where knowledge was incorporated as a set of rules or constraints written by the linguists. The linguistic rules ranged from a few hundreds to several thousands, and it usually required years of labour. Rule based Tagger works in two phases. In first phase all possible tags are assigned using lexicon or dictionary and in second phase linguistic rules are imposed to get single correct tag of each word.

First very popular rule based tagger was TAGGIT [1], which was used for initial tagging of the Brown Corpus. There are 3300 manually created context based rules and 71 tags to determine the correct tag of a word. Rules in TAGGIT consider maximum two left and right tags. Rules are written as:

Tag_i Tag_{i+1} Tag_{i+2}? Tag_{i+3} → Result Tag------------------------(1)

Tag_i Tag_{i+1} Tag_{i+2}? Tag_{i+3} → Not Tag --------------------------(2)

Tag_{i+2} in Rule1 and Rule2 is determined on the basis of left and right context tags (Tag_i ,Tag_{i+1} , Tag_{i+3})

Tag1 ? Tag2→ Not_Plura _Noun -------------------------------(3)

Where Tag2 is Third_Person_Singlar_Verb, Rule 3 says that Plural_Noun can not followed by Third_Person_Singlar_Verb.

Creating a complete set of such rules for any natural language is almost impossible and as the number of rules the rule base tends to become ambiguous.

A lot of effort has been devoted to improve the quality of the tagging process in terms of accuracy and efficiency. The constraint grammar formalism has also been applied for other languages like Turkish. The development of ENGTWOL (an English tagger based on constraint grammar architecture) can be considered one of the most successful efforts. Rule based approach is a natural and classical method for tagging but it is a very lengthy and time taking process. Identification and representation of all linguistic rules are the main bottlenecks in rule based approach.

Statistical Based Part of Speech Tagging[1][2][3]**:**

Statistical POS Tagging approaches can be divided into two parts Supervised and Unsupervised. Supervised POST approaches uses pre tagged corpus. Corpus is a collection of thousands of natural language sentences in which each word is assigned a pre decided label / tag. Using the statistical information of words, tags and tag sequences in given corpus, tag sequence of new sentence is determined. HMM, Maximum Entropy, Conditional Random Field are example supervised approaches. In unsupervised approaches, there is no need of pre decided tagset and tagged corpus. Instead these systems make use of advance algorithms like Baum- Welch algorithm and machine learning based clustering techniques to induce tagset and rules from untagged corpus, example of this approach is Brill Tagger.

Unigram:

Unigram is simplest statistical method of POS tagging. In this method most probable tag is assigned to words. Most probable tag of a word is computed by using conditional probability of word with different possible tags.

P(T1 | Word) = # Word appear labeled with T1 in corpus / # word appear in corpus, where: T1 is any possible tag of tagset.

HMM [4][5]**:**

Hidden Markov Model is a statistical tool for modeling time series data like problem. HMM is an extension of Markov Process. A Hidden Markov model can be describe with the help of following four parameters {S, A, B, Π}.

Where:

S is the set of all possible N states in system. These states are hidden not visible to user. Users can see only possible M observations on each state. For POS tagging problem states are all possible tags in tagset and all possible distinct words in corpus are M observations.

A[ ] is a transition matrix of size N*N, it store all possible probabilities from one state to all other states. For POS Tagging problem it is calculated with the help of pre tagged corpus, it represents the probabilities of a tag followed by other tags.

B[ ] is emission matrix of size N*M, it store the probability of each observation from each state. It represents how much time a word is used as noun, verb..etc. it is computed with corpus.

Π[ ] is initial state matrix of size N for each state. It represents how much time sentences start with different tags.

HMM consider POS tagging problem as a sequential data problem. POS tagging formally can be express as P(S|O), where $S(S_1,S_2,S_3…S_N)$ is the sequence of states / tags and O $(O_1,O_2,O_3…O_M$)is sequences of observation / words.

S*= Argmax(P(S|O) : Find such S which maximize P(S|O)

Computation of P(S|O) is very difficult by using bayes theorem and chain rules it can be simplified as

S*=Argmax(P(O|S) x P(S) / P(O)), denominator P(O) can be ignored, it is same for all sequences of states.

S*=Argmax(P(O|S) x P(S)

$=P(S_1,S_2,S_3…S_N)$ x $P(O_1,O_2,O_3…O_N \mid S_1,S_2,S_3…S_N)$)

$=P(S_1)$x $P(S_2|S_1)$ x $P(S_3|S_2,S_1)$ x $P(S_4|S_3,S_2,S_1)$……$P(S_N|_{SN-1}S_{N-2}…..S_1)$ x $P(O_1| S_1,S_2,S_3…S_N)$ x $P(O2| S_1,S_2,S_3…S_N,O_1)$ x $P(O_3| S_1,S_2,S_3…S_N,O_1O_2)$……… $P(O_N) |S_1,S_2,S_3…S_N, O_1,O_2,O_3…O_N)$.

Using Markov assumptions of order1 (present state depends on only previous state and independent on other states) and Independent assumption (observations depends on only current state and it is independent on other states). Expression can be simplified as

$=P(S_1|S_0))$x $P(S_2|S_1)$ x $P(S_3|S_2)$ x $P(S_4|S_3)$……$P(S_N|_{SN-})$ x $P(O_1| S_1)$ x $P(O2| S_2)$ x $P(O_3| S_3)$……… $P(O_N| S_N,)$.

$=\prod_1^n \mathrm{P(Si|Si - 1)\ x\ P(Oi|Si)}$

This expression is computed for all possible state combinations, which state sequences has highest value is the f and it is time taking process, its computation time is $|S|^{|O|}$. By using Viterbi algorithm Computation time can be reduced from $|S|^{|O|}$ to |S|x|O|.

HMM is an important mile stone in statistical based natural language processing, it is used in all statistical based techniques as a base. Many tagger developed using this approach with accuracy more than 90% [4][5]**:**. Accuracy of HMM is depends on the corpus size and size of tagset.

Maximum Entropy Based Model For POS Tagging [6][7][8]**:**

POS tagging can be considered as a classification problem. In classification problems data are separated into a number of classes on the basis of features (attributes) of data. Maximum entropy model uses multiple features at the same time to predict the tag of a word. For POS tagging language specific features like word features, dictionary features and corpus features in addition to context features added in HMM.. In which objective is to find an estimation function f (A,B) which assigns x □ A (where x denote word and its contextual and morphological information) it's appropriate class y □ B, where y is the one of the class / tag of predefined classes or possible tags.

Contextual information of input x can be expressed with the help context predicate Boolean function. It returns true if the input has required context and returns false if not, context predicate functions are as

CP1(input) → { true, false }

Word-Ends-ing(word) → { true, false }, Last_chars_Ly(friendly) → { true, false }

context predicate function Word-Ends-ing () returns true if input is 'walking' and returns false if input is 'book', A set of such context predicate functions (CP1,CP2,….CPN) are used to capture complete lexical , morphological and local contextual information of a input word.

Feature functions f:(A x B) is Boolean function that map class and context predicate. Feature functions takes two arguments one is possible class and other is set of context predicate functions (CP1,CP2,….CPN). Feature functions are as.

f:(A x B) → {0,1}

$$f(a,b) = \begin{cases} 1 & if\ class =='\ a'\ and\ \mathrm{Word - Ends - ing(b)} = \mathrm{true} \\ 0 & otherwise \end{cases}$$

Where Word-Ends-ing(b) is a context predicate function as given above. A set of language feature functions are identified. Weights are assigned to each feature according the correctness of feature in given corpus during the learning phase of model.

Maximum entropy model is also called unbiased constrained based model, above said feature functions are considered as constrained on the system. Word Entropy is used to measure uncertainty of the system. Entropy is directly proportional to uncertainty in system, entropy with uncertainty. Maximum entropy means most uncertain or uniform distribution. So it is called unbiased system. This model takes advantages of statistics of corpus as HMM and knowledge of linguistics in form of feature functions. Taggers developed using this approach has achieved accuracy up to 96% .

Conditional Random Field [9][10]**:**

Lafferty proposed a constrained based statistical discriminative model called conditional Random Field (CRF), it is an extension of both HMM and Maximum entropy model. In generative models (Figure 1) data is generated from hidden states/tags while in discriminative model tags / hidden states are generated from data.

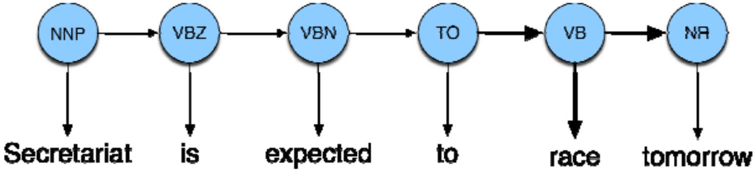

Figure 1: Generative Model

In generative models, like HMM, observation elements are represented as isolated units, independent from the other elements in an observation sequence. Dependencies are shown in figure 3.

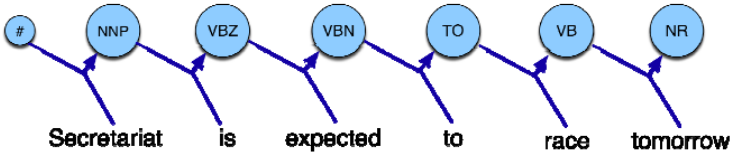

Figure 2: Discriminative model

More precisely, In generative models like HMM, Maximum entropy model, the observation element at any given instant of time may directly depend on the state / tag. This assumption is appropriate for simple data sets, however best representation of observation sequence are done in term of multiple interacting feature and dependencies which are of long range between the observation elements. The most important advantage of discriminative model (Figure 2) like CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Like Maximum entropy, CRF also avoid the label bias problem observed in HMM.

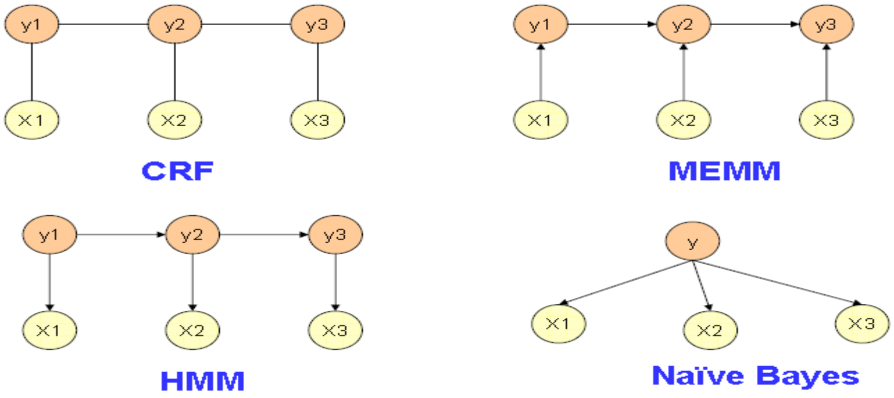

Figure 3: Graphical representation of HMM, Maximum Entropy

Conditional distribution is captured with the help of feature functions of data / word with class / tag and a linear combination of feature functions with their weights is used to identify the best sequence.

Mathematically problem can be expressed as.

$Y^* = \text{argmax}_y\, p(y|x:w) = \text{argmax}_y \sum_j wj\ Fj(x, y)$

$$P(y^*|x^*:w) = \frac{\boldsymbol{exp}(\sum_i \sum_j \boldsymbol{wjfj}(\boldsymbol{yi-1,yi,x*,i}))}{\sum_y \boldsymbol{exp}(\sum_i \sum_j \boldsymbol{wjfj}(\boldsymbol{yi-1,yi,x*,i}))}$$

By using this formula P(y*|x*:w) is estimated that is the linear sum of all feature function with their weights considering all other words in word sequence for a tag sequence over linear sum of all feature function with their weights considering all other words in word sequence for all tag sequences. This is very good approach for POS Tagging and its accuracy are better than Maximum Entropy and HMM based approach. But its training is very slow and feature function finding is a lengthy process. POS Tagger implemented using this approach claimed accuracy around 97%.

Support Vector Machine[11][12]**:**

Support Vector Machine (SVM) is machine learning based approach for classification, proposed by Vapnik. This approach can handle many NLP problems like POS Tagging, Text Categorization, Speech recognition etc. with high accuracy. Initially SVMs are used for binary classification problems and it searches a hyper plane with maximum margin to separate linearly separable data. As in figure 1, there are two class linear separable data is represented in 2D space, class 1 data is represented with '-' and class 2 data is represented by '+' (Figure 4).

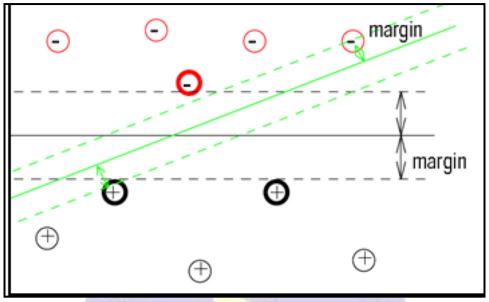

Figure 4: Linear separable Data in 2D

A number of hyper planes (lines) are existed to separate binary class data, the hyper plane which shows maximum margin from both class data.

Mathematically linear hyper plane can be expressed as

W . X +b = 0

Where W is weight vector and X is Data vector. In training, weights are changed to get hyper plane with maximum margin from both sides.

Non linearly separable data is transformed into higher dimensions and hyper plane is searched in new high dimensions data, a 2D non linearly separable data is transformed into 3D linearly separable data (figure 5).

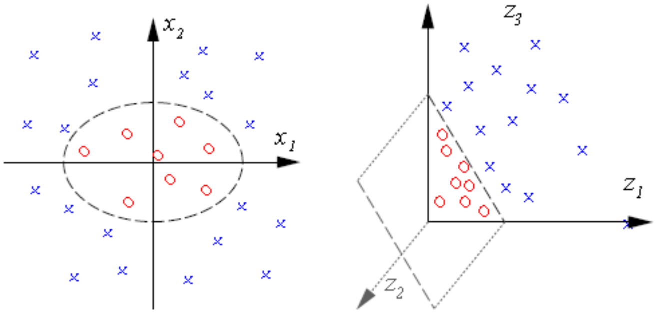

Figure 5: Transformation of Data from 2D to 3D

For K class classification problem as POS tagging, K SVMs are used, K SVM are placed in parallel and each one of them is trained to separate one class from the K - 1 others as shown in figure 7.

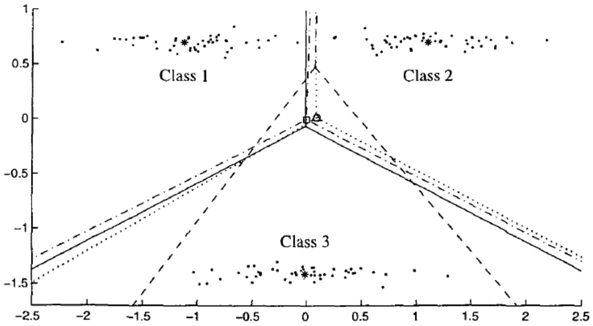

Figure 6: 3 class data

To use this model, each word is considered as a object with a number of attributes. Feature functions of word and tag, that includes word, tag, context words, lexical and morphological information of word is considered as attributes of input word. This is a powerful approach for POS tagging, 97 % accuracy are claimed by the authors. Main drawback of SVM based is its training (parameter estimation) is very slow.

Transformation Based tagging [13][14]**:**

Eric Brill proposed a rule based transformation approach for part of speech tagging (figure 8). In this approach initially most probable tag is assigned to each word of untagged corpus without considering any context and any linguistic feature. This task can be performed using simple lexicon. Unknown words are tagged using some lexical properties of words for example words end with 'ous' are tagged as adjective , capitalize unknown words are tagged as proper noun, and other unknown words are tagged as noun. After tag initialization, results are checked by linguistic expert. Wrong tagged words are corrected by the linguistic and submitted to system. From corrected corpus, system generates some replacement rules, and continues this process repeated until satisfactory results are not achieved.

Replacement rules are like

Change tag a to tag b when:

1. The preceding (following) word is tagged z.
2. The word two before (after) is tagged z.
3. One of the two preceding (following) words is tagged z.
4. One of the three preceding (following) words is tagged z.
5. The preceding word is tagged z and the following word is tagged w.
6. The preceding (following) word is tagged z and the word two before (after) is tagged w.
7. The current word is (is not) capitalized.
8. The previous word is (is not) capitalized.

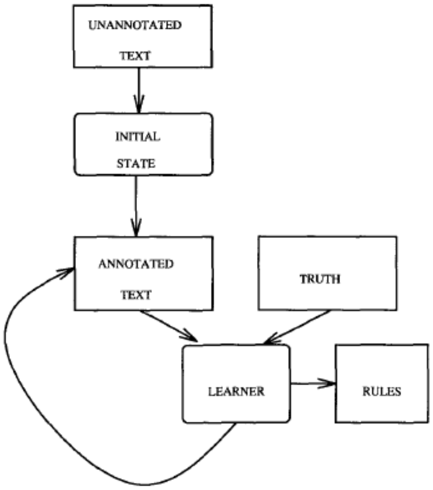

Figure 7: Transformation based model

This is put under unsupervised approach for POS Tagging. Here a large volume of pre tagged corpus is not required and there is no need to store large size contextual probability and statistics information of words also. This is a machine learning approach in which rules are automatically generated from the data. Before this approach, researchers show that NLP problems can not be modeled by the classical linguistic rule base approach and statistical approaches are the only solution for it. But Brill proved that NLP problem can be modeled with transformation based approach.

References:

[1]. *Allen, J. 2004. Natural Language Understanding. Person Education, Singapore.*

[2]. *Jurafsky, D. and J. H. Martin. 2000. Speech and Language Processing. Prentice-Hall, New Jersy.*

[3]. *Manning, C. D. and H. Schiitze. 2002. Foundations of Statistical Natural Language Processing. The MIT Press.*

[4]. *T. Brants., 2000 TnT – A statistical part-of-speech-tagger*

[5]. *Scott M. Thede and Mary P. Harper, -19993, A Second-Order Hidden Markov Model for Part-of-Speech.*

[6]. *Adwait Ratnaparkhi., 1997 A simple introduction to maximum entropy models for natural language processing. Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania.*

[7]. *Sandipan Dandapat., 2007. Part-of-Speech Tagging and Chunking with Maximu Entropy Model. In Proceedings of the SPSAL Workshop, IJCAI.*

[8]. *Aniket Dalal, Kumar Nagaraj, Uma Sawant and Sandeep Shelke., 2006 "Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach", In Proceeding of the NLPAI Machine Learning Competition.*

[9]. *Lafferty J., McCallum A. and Pereira F., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning. 282-289.*

[10]. *Sha F. and Pereira F. Shallow parsing with conditional random fields 2003. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, , Edmonton, Canada.134-141.*

[11]. *Gimenez J. and Marquez L. 2003. Fast and accurate part-of-speech tagging: The SVM approach revisited. In Proceedings of RANLP. 158-165.*

[12]. *Ekbal, A. Bandyopadhyay, S.2008, "Part of Speech Tagging in Bengali Using Support Vector Machine", ICIT- 08, IEEE International Conference on Information Technology, pp.106-111,.*

[13]. *Brill E., 1992. A simple rule-based part-of-speech tagger. In Proceedings of the 3rd Conference on Applied NLP. 152-155.*

[14]. *Brill E., 1995a. Transformation-based error-driven learning and Natural Language Processing: A case study in part-of-speech tagging. Computational Linguistics, 21(4): 543-565.*