A Statistical Method for Evaluating Performance of Part of Speech Tagger for Gujarati

Pooja M. Bhatt, Amit Ganatra

***Abstract**: Part of Speech Tagging has continually been a difficult mission in the era of Natural Language Processing. This article offers POS tagging for Gujarati textual content the use of Hidden Markov Model. Using Gujarati text annotated corpus for training checking out statistics set are randomly separated. 80% accuracy is given by model. Error analysis in which the mismatches happened is likewise mentioned in element.*

Index Terms: **BIS tag set,** ***Hidden Markov Model, Natural Language Processing, POS tagging, Statistical approach.***

I. INTRODUCTION

Guajarati language is one of the maximum spoken languages in Gujarat. POS tagging performs a pivotal position inside the development of Natural Language processing programs like Parser and Morphological analyzer. Numbers of articles are available within the literature on POS tagging undertaking for Indian languages like Hindi, Marathi, Odia, Punjabi and so on. But Gujarati textual content is problem. Various literature survey on POS tagging for Indian languages is noted in [1,2].Stochastic approaches are used for exclusive morphological wealthy languages[3,4,5].The arcitle contains five parts. After Introduction, segment 1 intricate the hidden markov model(HMM) and viterbi set of rules. Section three protected literature survey approximately statistical technique used for unique langauges. Section 4 protected the element of tag set used for experiment. Experiment and outcomes are discussed in Section 5 and Section 6 encompass end. Paper cease with references detail.

II. BACKGROUND

A) Pos Tagging Approaches

POS tagging method [1] divides in 3 categories known as Statistical or probabilistic, Rule based totally and Neural. In statistical approaches we use some statistical version or probability principle to decide the tag for word.In Rule based totally tagging the rule that used are hand – written.

i) Supervised POS Tagging

The version requires tagged dataset this is used for studying details about rule sets, word-tag frequencies, tag gadgets, and so forth. The overall performance of supervised pos tagger fashions boom with enhancement of corpus's length.

ii) Unsupervised POS Tagging

The Model does now not require tagged dataset. By applying computational techniques together with the transformation rules, Algorithm to robotically generate tag units and so forth. Based at the facts, they either increase the contextual policies with a purpose to be utilized by rule-based technique or calculate the chance by the stochastic approach.

iii) Stochastic Based Approach

This method includes opportunity, frequency or facts. The method identified normally used tag for a word within the annotated schooling facts that applies to come to be aware of word's tag within the unannotated text. N-gram approach is used to calculate the hazard of a given series of tags. It calculate the possibility which happens with the n preceding tags, in which n is set to at least one is called Unigram ,2 is referred to as Bigram or 3 is known as Trigram for sensible purposes. Viterbi Algorithm is the overall set of guidelines for put in force an n-gram technique for tagging enter text statistics and which avoid the polynomial enlargement of a BFS(breadth first search) with the useful resource of looking at each degree of tree with the help of nice m MLE (Maximum Likelihood Estimates) wherein m is the extensive type of tags of the following phrase .

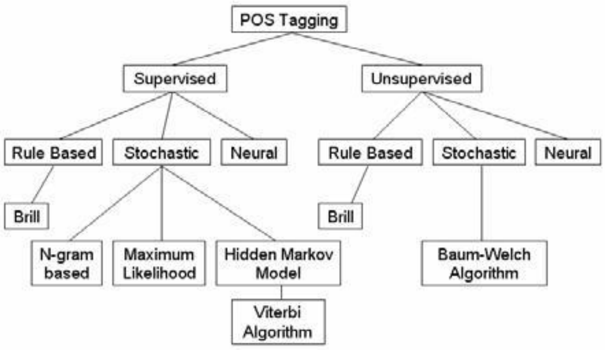

Fig1. Pos tagger Approach classification

B) Hidden Markov Model

In the mid 1980s, researchers in Europe started to use Hidden Markov Model to disambiguate components of speech, while working to tag the Lancaster-Oslo-Bergen_Corpus. HMMs involve counting cases (which includes from the Brown Corpus), and creating a desk of the probabilities of certain sequences. For example, once you've got seen an editorial along with 'the',

Revised Manuscript Received on 30 July 2019.
* Correspondence Author

Pooja M Bhatt*, Computer Engineering Department, CHARUSAT University, Changa, Gujarat, India.
Dr. Amit Ganatra, Computer Engineering Department, CHARUSAT University, Changa, Gujarat, India.

possibly the subsequent word is a noun 40% of the time, an adjective 40%, and a number of 20%. Knowing this, a application can determine that "can" in "the can" is far more likely to be a noun than a verb or a modal.There are 3 basic troubles that the HMM have to remedy for use for any sensible reason. They are as follows:

1. Possibility of the commentary sequences, given the model?
2. How do we pick a corresponding nation collection which is top-quality in some significant experience?
3. How will we modify the version parameters to maximize?

For supervised POS tagging, we solve problem 2, in which we attempt to uncover the hidden part of the model, i.e. to find the “correct” state sequence. For POS tagging, HMM is used to model the joint probability distribution P(word, tag). The generation process uses a probabilistic Finite State Machine (FSM). The states of HMM correspond to the tags, it has an alphabet which consists of the set of words, the transition probabilities $P(Tag_i|Tag_{i-1})$ Eq. (1) and the emission probabilities $P(Word_i|Tag_i)$ (2)In HMM, for a given (word, tag) pair we have the probability:

$P(w, t) = \Pi_i$ P(Tagi|Tagi-1) * P(Wordi|Tagi) (3)

It is known as Markov as it is primarily based at the Markovian assumption that the current tag depends simplest on the preceding n tags. The HMM version trains on annotated corpora to discover the transition and emission probabilities. For a sequence of words w, HMM determines the collection of tags t the use of the system:

t = argmax P(w, t) (4)

The computation of this formulation is very high priced as all feasible tag sequences are required to be checked so that you can find the collection that maximizes the possibility. So a dynamic programming approach called the Viterbi Algorithm is used to find the gold standard tag sequence.

III. LITERATURE SURVEY

There are diverse strategies available for element-of speech tagging and distinct researchers have advanced POS taggers for various language like Arabic, English, different European languages than Indian languages. Indian Languages for which Part Of Speech taggers had been advanced are Marathi, Urdu, Hindi, Bengali, Panjabi and Tamil. “Parts Of Speech Tagging for Indian Languages: A Literature Survey” [1] by Antony P J Dr. Soman K P present Survey on concepts of POS tagging for Indian languages like Bengali, Panjabi , Hindi and Dravidian languages. All proposed mthods advanced by various organization and individuals and POS taggers have been primarily based on in different Tagset.They present a range of developments in POS-tagset and Part of speech taggers for different Indian language, that is extremely important computational linguistic apparatus useful for NLP applications. “Part of Speech Taggers for Morphologically Rich Indian Languages “[2] by Dinesh kumar Gurupreet sing Josan present in their paper that present Survey on hindi, punjabi,bengali,telugu with different POS tagging approaches. For hindi93% for hmm,89% for CRF and for Punjabi 80% and for telugu rule based achived 98% accuracy they achieved. “Mix Hidden Markov Model Based Part-of-Speech Tagging for Urdu in Limited Resource Scenario”[3] via M. Humera khanam , K.V. Madhumurthy and Md.A. Khudhus. They proposed HMM base totally stochastic algorithm intended for part of speech tagging and with HMM they used morphological Analyzer and stemmer to improve overall performance of tagger. They conclude that using morphological attribute is specifically beneficial to increase a reasonable POS tagger whilst tagged sources are constrained. Even though HMM performs fairly well for component-of-speech disambiguation project, it makes use of handiest nearby features (cutting-edge phrase, preceding 1 or 2 tags) for POS tagging. Uses of most effective nearby functions may not work properly for a morphologically rich & relatively unfastened order word language Urdu. “HMM based POS Tagger and Rule-based Chunker for Bengali “[4] by Sivaji Bandyopadhyay, Asif Ekbal, Debasish Halder , this paper work describes a Part Of Speech tagger based on the HMM(Hidden Markov Model) with a rule-based chunker for Bengali language. The Part Of Speech tagger changed into educated on the training sets ANNOT-A and ANNOT-B collectively along with 40956 tokens. The taggerwas examined at the improvement check set ANNOT-D along with 5967 tokens and confirmed 85.42% accuracy. Finally, the tagger became examined on the unannotated check set which includes 5129 token sand tested 79.12% accuracy. “Sanskrit Tag-sets and Part-Of-Speech Tagging Methods”[7] by Sulabh Bhatt, Krunal Parmar and Miral Patel. They provide brief introduction to various approaches and the working of two most famous statistical methods used for POS tagging: Conditional Random Fields (CRF) and Hidden Markov Model (HMM). “Morphological Richness Offsets Resource Demand- Experiences in Constructing a POS Tagger for Hindi”[8] by Smriti Singh, Kuhoo Gupta, Manish Shrivastava, Pushpak Bhattacharyya. They work on building a POS tagger for a morphologically rich language like Hindi. The theme of the research is to vindicate the stand that- if morphology is robust and arnessable, then lack of education corpora isn't tiring. A main power of the work is the learning of is ambiguation rules, which in any other case could have been hand-coded, as a consequence disturbing exhaustive evaluation of language phenomena. Attaining an accuracy of near 94%, from corpora of virtually 15,562 phrases lends credence to the notion that morphological richness can offset resource scarcity. “A Study on Different Part of Speech (POS) Tagging Approaches in Assamese Language”[9] by Bipul Roy, Bipul Syam Purkayastha. Syntactic parsing is a critical assignment that's essential for Language Processing application which include POS tagger. For the enrichment and improvement of languages, a POS tagging performs a very crucial function. POS, mainly for the close by Indian languages can deliver universal technique. For a domestic language like Assamese, POS tagging has emerge as a good deal vital for the general flourishment of the language. The linguistic professionals have evolved various type of tagging methods such as Stochastic based, Neural Network , Rule based technique, and plenty of others. In this paper authors intention is to in brief evaluation the computational tasks which has been carried out until date by the usage of the linguists within the area of tagging of Assamese language. Discovered on this observe that each one of the three NLP processes are inexperienced and incredible, but best for the easy Assamese sentences. Thus, on this regard plenty of effort has to be accomplished to deal with complicated Assamese sentences with distinct systems.

Retrieval Number: B1492078219/19©BEIESP
DOI: 10.35940/ijrte.B1492.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

As of fairly free phrase order traits and several confusing phrases, tagging is hard venture. The brought trouble in tagging is of unavailibity of annotated dataset and predefined tag set it truly is past public get entry to. Authors future planning is to create an unlabeled dataset and an Syntactic Analyzer through thinking about the rich morphology of Assamese language to provide their minute

contribution to the useful resource for poor Assamese language.“POS tagging for Gujarati using CRF[11] “ by Chirag Patel and Kartik Gali, Using CRF model and find error then generate rules and again apply CRF and try to improve efficiency and also consider suffix, prefix, but cant identified unknown words. There are 26 tags in tag set. Labeled 600 sentences’ and 5000 not labeled statements are used for learning. The authors achieved 92% accuracy using Gujarati dataset.“POS tagging and Chunking for Indian Languages” [12] by Himanshu Agrawal ,they used CRF with Knowledge database and respective gold standard POS tagset in training data present approach for a chunker and POS tagger for South Asian Languages. He has worked on enhancing the machine learning performance. They didn’t apply language specific tools like morphological analyzers , dictionaries, etc. They use a large raw unannotated text and achieved average accuracy 92% for chunking and 79.13% for POS tagging. “Segmental HMM based Pos tagger” [13] by Mohammad Hadi, Hossein Sameti, Mohammad Bahrani, Bagher Babaali presents modify viterbi algorithm with HMM for parsian languages which consider semi space to resolve the problem where a word can be composed of several tokens. The system has a built-in tokenizer that find out words boundaries and also its matching tag sequence by letting the states of model to output more than one token. “POS for Hindi corpus” [14] by Nidhi mishra, Amit mishra ,POS System read Hindi corpus , tokenize the sentences and words and display tag for each word. Easy to use and user friendly interface but more training data require for future work.They Achieved accuracy of 92%.They remove the disambiguation of word-tag by relative information available within text. “HMM based POS tagger for hindi”[15] by Nisheeth joshi Hemant Darbari, Iti mathur, they used trigram approach for Marathi language .Trigram approach is use to explore the most probable tag for a token based on specified details of preceding two tags by cunning probabilities to find out which is the best sequence of tag.Using this approach they get 91.63% accuracy and used test corpus of 2000 sentences. “POS tagging and chunking with HMM and CRF”[16] by Pranjal Awasthi Dilip Rao Balaraman Ravindran they In this paper we suggest an method Intial tagging withTnT tag set and follow rule for error correction and for each generation new education data generated.They attain accuracy with errors 0.74% and without error 79.66%.

IV. BIS (BUREAU OF INDIAN STANDARD) TAG SET

BIS tag set[10] is taken into consideration for the standardization in the place of morpho-syntactic annotation for all of the Indian Languages. It is having eleven primary tags and similarly it divides in sub tags. Main classes of dataset are Demonstrative (DM),Post Position (PSP),noun (NN), Adverb (RB), Verb (V), Pronoun (PR),Conjunction (CC), Particles (RP), Quantifier (QT), Adjective (JJ),Residual(RD).

V. EVOLUTION AND RESULTS

Evolution is done for enhancing the performance of system on domain of arts and culture data. The system was evaluated on 84808 words for arts and culture data. These test set is collected from multilingual Guajarati text available on TDIL. Following table shows the different test cases for testing. First, we apply Hidden Markov Model and got the accuracy 80%.

Table1.Test Case

TEST NO.	DOMAIN	NO. OF WORDS	ACCURACY
1	ARTS & CULTURE DATA	84808	80%

a) *Precision, Recall, F-measure for all tags*

The evaluation metrics for the data set is F-Measure, precision and recall. These are defined as following:-

F-Measure =Recall *Precision / Recall + Precision

Precision =Number of Correct answer / Total number of words.

Recall = No. of correct answer specified by system / Total number of words.

Now, in figure 1 we can see accuracy of all tags available for tagging in hmm based approach. We have measured accuracy with the parameters via recall, f-measure and precision.

b) *Error Analysis*

For analyzing the error in tag assignment we have taken dataset which shows what tag system has assigned and what is actual tag should be. We have used BIS tag set for evolution. Here we are having 30 tags for assignment.11 tags are main tags like verb, noun, adjective etc and others are subtypes of it. Table 2 showing error analysis using HMM model.

Table 2 shows the error analysis of the words whether mismatch takes place. In some cases like Noun, Verb, HMM model assigned wrong sub tag of same category.

Table 2. Error Analysis of data set using HMM

ACTUAL TAG	ASSIGNED TAG	ERROR
DM_DMD	DM_DMR	296
V_VAUX	V_VM	262
V_VAUX_VNP	V_VM	260
CC_CCD	RP_RPD	254
V_VM	V_VAUX_VNP	166
N_NN	N_NNP	162
PR_PRP	DM_DMD	128
PSP	RD_SYM	122
N_NN	JJ	116
N_NST	RD_SYM	100
QT_QTC	RD_SYM	92

RB	RD_SYM	82
DM_DMI	PR_PRI	80
N_NN	V_VM	76
V_VM	JJ	70
QT_QTF	RD_SYM	64
DM_DMD	PR_PRL	54
JJ	QT_QTF	52
N_NN	PSP	48
N_NN	N_NST	48
V_VAUX	RD_SYM	46
JJ	V_VM	46
V_VM	N_NN	38
JJ	N_NST	38
JJ	PSP	36
DM_DMD	PR_PRP	34
JJ	RB	34
DM_DMD	RD_SYM	32
JJ	N_NN	28
PR_PRP	RD_SYM	28
V_VAUX_VNP	V_VAUX	28
QT_QTO	RD_SYM	26
JJ	QT_QTO	26
V_VM	PSP	26
PSP	PR_PRL	26
PSP	N_NST	24
PR_PRP	N_NNP	24
RP_RPD	RB	22
JJ	QT_QTC	22
RB	JJ	20
PSP	DM_DMD	18
N_NN	QT_QTC	18
V_VM	PR_PRP	18
N_NNP	N_NN	16
JJ	V_VAUX_VNP	16

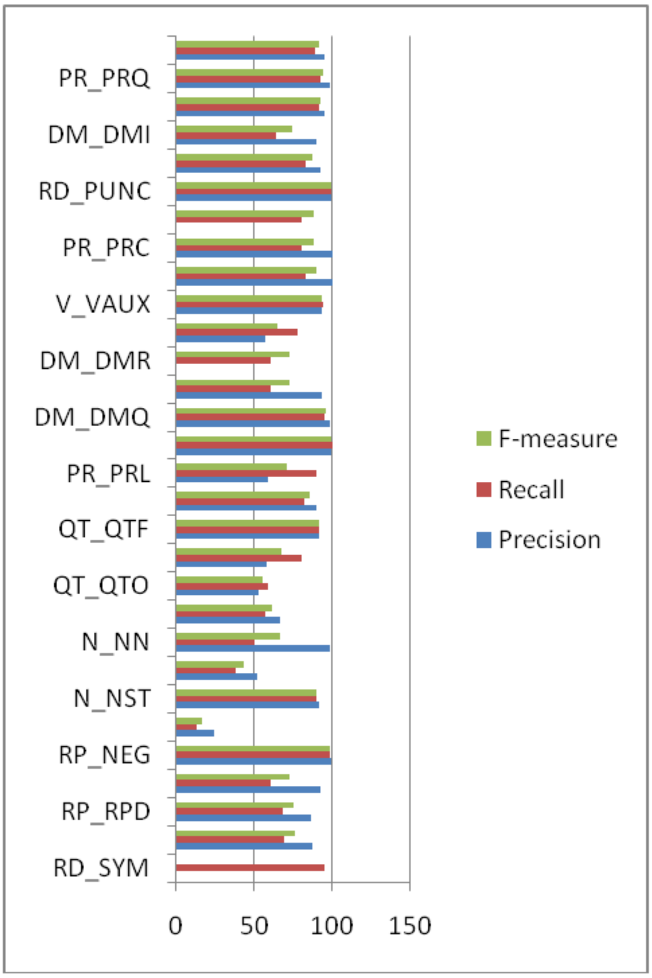

Fig 2. Per tag Accuracy using HMM

VI. CONCLUSION

Pos tagging is a primary step for any NLP applications. In this paper we used HMM version for evaluating the accuracy of pos tagger for Guajarati language. We have completed 80% accuracy for art and tradition dataset which is having approximately 80000 phrases. The accuracy may be ameliorated by using taking extensively big database in addition to via applying some advance strategies like recurrent neural network.

REFERENCES

1. Survey: POS tagger for Indian languages by Antony P J Dr.Soman K P IJCA, 2011.
2. POS for morphologically rich Indian languages: A survey by Dinesh kumar,Gurupreet sing Josan IJCA 2010.
3. Mix Hidden Markov Model Based Part-of-Speech Tagging for Urdu in Limited Resource Scenario by M. humera khanam. K.V. Madhumurthy and Md.A. Khudhus in IJARCSSE august 2013.
4. HMM based POS Tagger and Rule-based Chunker for Bengali by Sivaji Bandyopadhyay,Asif Ekbal,Debasish Halder 2008.
5. Learning Hidden Markov Model structure for Information extraction by Kristie seymore, Andrew McCallum,Ronald Rosenfeld in AAAI Technical Reports 2011.
6. Comparison of Different POS Tagging Techniques (n-grams, HMM and Brill‟s Tagger) for Bangla", by Hasan F.M., UzZaman N, Khan M. 2006 International Conference on Systems, Computing Sciences and Software Engineering (SCS2 06) of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CIS2E 06)

7. Sanskrit Tag-sets and Part-Of-Speech Tagging Methods by Sulabh Bhatt, Krunal Parmar and Miral Patel in IJIERE 2015.
8. Morphological Richness Offsets Resource Demand- Experiences in Constructing a POS Tagger for Hindi by Smriti Singh, Kuhoo Gupta, Manish Shrivastava, Pushpak Bhattacharyya ACL-2006.
9. A Study on Different Part of Speech (POS) Tagging Approaches" in Assamese Language by Bipul Roy, Bipul Syam Purkayastha International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.
10. http://www.ldcil.org
11. POS tagging for Gujarati using CRF by Chirag Patel and Kartik Gali IJCNLP 2008.
12. Segmental HMM based Pos tagger by Mohammad Hadi, Hossein Sameti IEEE 2010.
13. POS tagging and chunking for indian languages by Himanshu agrawal IJCNLP 2008.
14. POS for Hindi corpus by Nidhi mishra, Amit mishra IEEE 2011.
15. HMM based POS tagger for Hindi by Nisheeth joshi ,Hemant Darbari, Iti mathur CSIT 2013.
16. POS tagging and chunking with HMM and CRF by Pranjal Awasth, Dilip Rao ,BalaramanRavindran, IJCAI – 2007.