Analyzing & Enhancing Accuracy of Part of Speech Tagger with the usage of mixed approaches for Gujarati

Pooja M Bhatt, Amit Ganatra

Abstract: Tagging an accurate grammar to the specific phrase in sentences could be very crucial undertaking for specific Indian languages .Part of speech tagging is a fundamental manner for one of a kind natural language processing applications like machine translation, speech Recognition etc. Part Of Speech is used for assigning tag the usage of the grammatical statistics of every word of a sentence. We have used statistical approach like Hidden Markov Model (HMM) and rule based method to investigate the accuracy of a part of speech tagger for Guajarati language. In the paper we discussed available tagging strategies for numerous Indian languages. Further we discussed proposed approached with the use of BIS tag set that includes 11 fundamental tags and more than 25 sub tags. Further we practice HMM model for Sports and amusement information set, we are getting accuracy 70% and 56% respectively. After applying rule based approach we achieved 76% accuracy for sports activities and 80% for Entertainment dataset. After that we have used leisure information set with 95614 phrases and we were given 52% accuracy with hmm and 83 % accuracy with the aid of after making use of rules with hmm.

Index Terms: Hidden Markov model, Natural Language Processing, Part of Speech tagging, Statistical models, Rule based approach.

I. INTRODUCTION

Natural language processing (NLP) is related to the area of human-machine interplay. It is part of Artificial Intelligence. ‘Part of Speech tagging is the process of attaching the best grammar tag to every phrase of a sentence. A phrase in a sentence can act as a adjective, verb, adverb, conjunction, preposition, noun, pronoun, and so forth.

POS is used for assigning tag using the grammatical facts of every word of a sentence. While we are assigning a tag it's far vital to decide the context of the word i.e. Whether or not it's adjective verb, noun and so forth. Sometime it may takes place a phrase can act as a verb in one sentence and noun in every other sentence. So we need to take care that in what context the word is used earlier than selecting a POS tag for a phrase**.** For Indian languages like Guajarati, Hindi, Marathi, it is a complicated venture to assign the ideal tag to every word in a sentence because of its a few unknown phrases and morphology. Gujarati has 3 genders (masculine, feminine and neuter), three instances (nominative, oblique/vocative and locative) for nouns and two numbers (singular and plural). POS tagging is helpful in numerous NLP applications like Information Retrieval, Machine Translation, Information Extraction, Speech Recognition and so on. There are three categories for POS tagging techniques referred to as statistical or probabilistic, Rule based and Hybrid. In statistical, we use some statistical models or probability theory to decide the tag for word..In Rule based tagging done via guideline that used are hand – written. In hybrid we integrate the above approaches for tagging phrase.

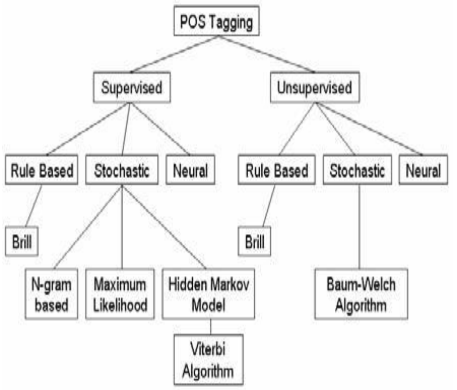

Fig 1. Pos tagger Approach classification [22]

A) Supervised POS Tagging

The model requires tagged dataset that is used for mastering details about rule sets, word-tag frequencies, tag units, and so on. The performance of supervised pos tagger fashions increase with enhancement of corpus’s length.

B) Unsupervised POS Tagging

The Model does not require tagged dataset. They apply computational methods like the transformation rules, algorithm to automatically generate tag sets, etc. Based on the information, they either develop the contextual rules which will be used by rule-based approach or calculates the probability by the stochastic approach.

C) Stochastic Based Approach

This technique consists of opportunity, frequency or records. The method diagnosed commonly used tag for a phrase inside the annotated education statistics that applies to become aware of phrase’s tag within the unannotated textual content.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author

Pooja M Bhatt*,Coputer Engineering Department, CHARUSAT uniersity, Changa,Gujarat,India.
Dr. Amit Ganatra, Coputer Engineering Department, CHARUSAT uniersity, Changa,Gujarat,India.

N-gram approach is used to calculate the chance of a given series of tags. It calculate the opportunity which occurs with the n preceding tags, wherein n is set to 1 is called Unigram ,2 is referred to as Bigram or three is called Trigram for realistic purposes. Viterbi Algorithm is the general set of rules for put in force an n-gram approach for tagging enter textual content statistics and which keep away from the polynomial enlargement of a BFS(breadth first search) with the aid of looking at every degree of tree with the assist of pleasant m MLE (Maximum Likelihood Estimates) in which m is the wide variety of tags of the subsequent phrase .

There are different approaches available for POS tagging, some of which are described below.

i) Hidden Markov Models

Hidden Markov Model [11]s is called Hidden because for a phrase sequence, we cannot discover the exact collection of tags. It is called Markov as it is decide the tag using Markovian assumption that the current tag can be decided by previous m tags. The HMM model [10] trained on labeled facts to discover the emission and transition possibilities. For a chain of phrases W, HMM reveals the series of tags T using the system:

$$T = \text{argmax } P(W, T) \qquad (1)$$

For example: red car
Adverb noun

At time T if we have a word car the tag for car is depend on previous tag that is of red. If red is adverb then after adverb there is always a noun. So tag for car is noun.

ii) Conditional Random Fields

Its an undirected graphical model[24]. It segments and labeled the sequence of data. Segmentation may create many problems. The probability of the label sequence can be decided by focusing on non-independent features of sequence instead of focusing on distribution of the dependencies.

The probability of a transition between labels depends on the current, past as well as future observations.

Let F as a factor graph over B. So the distribution P(b/a) factorizes according to F and P(b/a) is a CRF for any fixed a. For example

Tags : QF NN NN VM
Literal: All people house in went.
Sentence: Badha loko gam ma gya.

Here, noun i.e NN is for "loko" and Quantifier i.e QF is fr "bdha" .NN is also assign to" house" and VM for "gya".

iii) Maximum Entropy Model[27]

The basic principle of Maximum Entropy Model [18] is Maximum Entropy, which states that when choosing between a numbers of different models for a set of data.

D) Rule Based Approach

Rule based totally method [26] for POS uses linguistic regulations for figuring out which POS tag to be assigned to the enter word Rule based totally technique requires extensive knowledge of language. To apply context policies, in POS tagging, it necessitates sizable linguistic knowledge.

II. LITERATURE SURVEY

There are numerous strategies to be had for part-of speech tagging and various researchers have developed POS taggers for numerous languages like Arabic, English and other European languages have more POS taggers than Indian languages. Indian Languages for which POS taggers have been developed are Marathi, Urdu, Hindi, Bengali, Panjabi and Tamil.

"POS tagging for Gujarati using CRF"[1] by Chirag Patel and Kartik Gali.Using CRF model and finding errors then generate rules and again follow CRF and attempt to enhance efficiency and also remember suffix, prefix, but cant diagnosed unknown words. The general Indian Language (IL) tag set consisting 26 tags. Both 600 tagged sentences' and 5000 untagged sentences are used for studying. The authors achieved an accuracy of 92% for Gujarati language.

"POS tagging and Chunking for Indian Languages"[2] by Himanshu Agrawal. They used CRF with Knowledge database and respective gold standard POS tag set in training statistics present method for a chunker and a part of speech tagger for South Asian Languages. Author used a large raw unannotated text. He has worked on enhancing the machine learning by excluding other language evaluation tools like morphology analyzer, dictionaries, and many others. They gain average accuracy 79.13% for POS tagging and for 92% chunking.

"Segmental HMM based Pos tagger" [3] by Mohammad Hadi, Hossein Sameti, Mohammad Bahrani, Bagher Babaali.Paper provides modify viterbi algorithm with HMM for Parsian languages which consider semi space to clear up the trouble wherein a word may be made from numerous tokens. The system has a integrated tokanizer with it that indicates phrases limitations and additionally its matching tag series by using allowing the states of model to output more than one token.

"POS for Hindi corpus" [4] by Nidhi mishra, Amit mishra.POS System study Hindi corpus, tokenize the sentences and words and display tag for every phrase. Easy to apply and consumer friendly interface however greater learning data require for future work. They achieved accuracy of 92%.They remove the disambiguation of word-tag via contextual data available within the text.

"HMM based POS tagger for hindi"[5] by Nisheeth joshi ,Hemant Darbari, Iti mathur**.** They used trigram technique for Marathi language. The major use of Trigram is to find out the maximum in all likelihood tag for a token based on given information of previous tags by calculating probabilities to find out that's the quality collection of tag. Using this technique they get 91.63% accuracy and used test corpus of 2000 sentences.

"POS tagging and chunking with HMM and CRF"[6] by Pranjal Awasthi Dilip Rao Balaraman Ravindran. In this paper authors recommend an approach Initial tagging with TnT tag set and follow rule for error correction and for each iteration new training statistics generated. They achieve accuracy with error 80.74% and without mistakes 79.66. "Part of Speech Taggers for Morphologically Rich Indian Languages" [7] by Dinesh The trouble of tagging in herbal language processing is to discover away to tag every word in a textual content as a selected a part of speech, e.g.

proper pronoun. POS tagging is a totally critical preprocessing task for language processing activities. This paper reviews about the Part of Speech (POS) taggers proposed for diverse Indian Languages like Malayalam, Punjabi, Telugu, Hindi and Bengali. Various part of speech tagging procedures like Hidden Markov Model (HMM), Support Vector Model (SVM),Rule based approaches, Maximum Entropy (ME) and Conditional Random Field (CRF) had been used for POS tagging. Accuracy is the prime element in comparing any POS tagger so the accuracy of every proposed tagger is also mentioned on this paper.

For Hindi, 4 taggers were proposed based totally on HMM, ME, CRF and a morphology driven technique. The average accuracy as reported by various authors is 93.05%, 89.34%, 82.67% and 93.45% respectively. A rule-based POS tagger turned into proposed for Punjabi. This is the most effective tagger available for Punjabi. The accuracy of 80.29% together with unknown word and 88.86% apart from unknown phrases become accomplished with the aid of the proposed tagger. In case of Telugu, rule based, Brills tagger based and Maximum Entropy based tactics had been used for the development of tagger. The accuracy performed with the aid of a lot of these taggers is 98%, 90%, 81.78% respectively. From this have a look at, it is observed that the Indian Languages are morphologically rich languages. Thus, morphological analyzer performs a vital role in developing a POS tagger. Further, machine learning based methods offers truly better outcomes as compared to other techniques. Very confined work has been finished on Indian Languages for Part of speech tagging. Hence, specific processes may be used for the improvement of efficient tagger.

"Parts Of Speech Tagging for Indian Languages: A Literature Survey" [8] 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.

"Vishit: A Visualizer for Hindi Text "[9] by Priyanka Jain, Hemant Darbari and Virendrakumar C. Bhavsar. It is an application of pos tagger with hindi tagset. It takes the sentence(Hindi language) as input then process it and capture knowledge like role identification, background detail and object creation. At last it creates scene synthesis and generation.

"HMM based POS Tagger and Rule-based Chunker for Bengali "[13] 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.

"Hindi POS Tagger Using Naive Stemming : Harnessing Morphological Information Without Extensive Linguistic Knowledge" [14] by Manish Shrivastava ,Pushpak Bhattacharyya ICON 2008, authors represent HMM primarily based POS tagger, that employs a longest suffix matching stemmers and a pre-processor to obtain 93.12% accuracy. This method does not require any linguistic resource apart from a list of possible suffixes for the language. This list can be effortlessly created the usage of current machine studying strategies. The aim of this method is to demonstrate that even without employing tools like morphological analyzer or resources like a pre-compiled structured lexicon, it is possible to harness the morphological richness of Indian Languages.

"Morphological Richness Offsets Resource Demand-Experiences in Constructing a POS Tagger for Hindi"[15] 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. "Part of speech tagging and shallow parsing "[16] by delip rao and david yarowsky. they proposed How performance can be stepped forward by numerous functions enhancement and improve modeling techniques for Indian languages and for chunking tasks they used CRF model with improved features. Used CRF for shallow parsing and getting enhance features with chunk tag inventory on different Indian languages and separating punctuation from linguistic phrases. Accuracy achieved 73% for hindi 64% for Bengali and 68% for telugu. "Mix Hidden Markov Model Based Part-of-Speech Tagging for Urdu in Limited Resource Scenario"[19] 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.

"Sanskrit Tag-sets and Part-Of-Speech Tagging Methods"[20] 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).

"Improving Punjabi Part of Speech Tagger by Using Reduced Tag Set "[21] by Manjit Kaur, Mehak Aggerwal and Sanjiv Kumar Sharma. They conduct experiment by reduced Part Of Speech tag set (contain 36 tags)which was proposed via Technical Development of Indian Languages (TDIL). That has been used to enhance the tagging accuracy of HMM based Part Of Speech tagger.

lastly the end result has been evaluated by hand from a linguistic knowledge and they achieved accuracy among 90 to 95%. "Survey of various POS tagging techniques for Indian regional languages"[28] by Sharvari Govilka and Shubhangi Rathod. POS is an important tool for processing any natural languages. It is best as well as maximum stable and statistical variations for many NLP applications. This is an approach of marking up a word in a corpus as just like a specific POS such as verb, adverb, adjective and noun. There are various demanding situations in POS tagging like ungrammatical enter statistics, foreign phrases and ambiguities.In this paper, assessment of several POS tagging techniques for Indian languages has been cited elaborately.

"Parts of speech tagging for hindi languages using hmm" [29] by Rajesh Kumar, Sayar Singh Shekhawat, The paper describes the Part of Speech tagging for Indian Languages "HINDI". Part of Speech tagging is the one of the most basic troubles of Natural language processing NLP. Part of speech tagging is the way of assigning a tag or different lexical class marker to every and each phrase in a sentence. A lot of POS tagging work has been completed through the researchers for various languages the usage of specific processes SVM(Support Vector Machine), HMM (Hidden Marcov Model) and ME (Maximum Entropy). HMM techniques concerned for POS tagging of sentences written in Hindi languages are mentioned on this paper. This paper also discussed a hybrid based totally approach, for tagging Hindi language.

"Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" [30]by Christopher D. Manning. He observes what might be crucial to transport element-of-speech tagging normal overall performance from its modern level of about 97.3% accuracy to close to 100%. He recommends that it need to despite the fact that be viable to noticeably increase tagging overall performance and study some beneficial upgrades which have currently been made to the Stanford Part-of-Speech Tagger. However, evaluation of some of the remaining errors indicates that there may be restricted similarly mileage to be had each from better device getting to know or higher features in a discriminative sequence classifier.

The potentialities for further profits from semi supervised gaining knowledge of additionally appear quite confined. Rather, he advocate and start to illustrate that the most critical opportunity for further improvement comes from enhancing the taxonomic basis of the linguistic belongings from which taggers are educated. That is, from superior descriptive linguistics. However, he concludes through suggesting that there are also limits to his system. The reputation of a few phrases may not be able to be efficaciously captured by way of assigning them to surely one in all a small variety of categories. While conventions can be utilized in such instances to enhance tagging consistency, they lack a robust linguistic basis.

"Part of speech tagging for Gujarati text"[31]by Pandya Abhinay, Dave Mainak. Part-of-speech (POS) tagging is a process of assigning the lexicon class to each lexicons in a given natural language sentence, that first-class suits the definition of the lexicon in addition to the context of the sentence in which it is used. Part-of-speech tagging is an essential a part of Natural Language Processing (NLP) and is useful for most NLP programs. Part-of-speech tagging is often a primary step in maximum of the NLP duties along with chunking, parsing, etc. Gujarati is the national language of Gujarat, a western state of India, and is spoken by means of 70 percent of the country's population. More than 46 million humans international don't forget Gujarati as their first language.

Apart from Gujarat, it's miles broadly spoken in the states of Maharashtra, Rajasthan, Karnataka and Madhya Pradesh and also around the world. Natural language processing of Gujarati is in its early level of existence. Gujarati POS tagger is a middle component for maximum NLP applications. Information retrieval, machine translation, shallow parsing and word experience disambiguation tasks can be work more correctly and effectively with the existence of a POS tagger. their attention is to increase an effective Gujarati text POS tagger. their foremost project of thesis is to built a system which can annotate element-of-speech for Gujarati texts mechanically, with the help of various gadget learning algorithms. they have used tag sets defined by means of IIIT Hyderabad. they have used machine mastering techniques one is Hidden Markov Model and 2nd is Conditional random Field. Since Gujarati is a morphologically reach language, they are able to use Morphological Analyzer (MA) to limit the set of feasible tags for a given phrases. Gujarati language is based totally on Paninian framework, guidelines of morphology are nicely-described. Hence we've got defined morphological policies for Gujarati. While MA facilitates us to limit the possible desire of tags for a given word, one can also use prefix/suffix information (i.e, the collection of first/last few characters of a phrase) to further improve the fashions. HMM version uses suffix statistics for the duration of smoothing procedure even as CRF uses suffixes as a feature.

"A comprehensive survey on parts of speech tagging approaches in Dravidian languages"[32] by Merin Francis. Parts of speech tagging is the process of assigning tag to every phrase in a document a tag that corresponds to that means of the phrase within the unique context. It is significant and act as a essential step in lots of language processing software from phrase experience disambiguation to speech identification. As of variations in grammatical construct and morphological differences, the techniques for tagging in different languages are broadly distinct.

The theoretical methods encompass supervised studying techniques as CRF based taggers, SVM based taggers and HMM based Model or unsupervised techniques as rule based taggers. The languages considered are having rich morphology Dravidian languages as Tamil, Kannada and Malayalam. Various techniques are compared on their accuracy and analysis is performed. In paper, components of speech tagging strategies discussed for Dravidian languages. Different tagging methodologies are discussed and also include comparative accuracy study of various approaches.

"A Survey of Ontology Learning Approaches" [35] by Maryam Hazman, Samhaa R. El-Beltagy, Ahmed Rafea.The trouble that ontology attending to cope with is the information acquisition bottleneck, that is to say the difficulty to certainly version the data associated with the domain. Ontology is the medium with the aid of way of which we are able to model and proportion the knowledge amongst several packages in a selected domain.

Thus many research advanced several ontology getting to know methods and structures. In the paper, authors represent a survey for the exceptional procedures in ontology reading from semi-structured and unstructured data. "Survey: Part-Of-Speech Tagging in NLP" [37]. by Nidhi Adhvaryu, Prem Balani. For assigning special tags to each word of the sentence part of speech tagging (POST) is used. POST having primary two methods: Supervised tagging and unsupervised tagging. These techniques are similarly divided into categories: Rule based, Statistical method and Transformation based method. In rule based technique, policies are generated manually and regular with all rules sentence may be tagged. In statistical method, three types of techniques are used: MEMM, HMM and CRF which are corpus based techniques. Transformation based method is used to observe rules and fed features and tag the unannotated corpus. They discovered that Rule based technique rule primarily based approach is complicated as the rules are generated manually. MEMM And HMM having the label bias problem this is solved with the useful resource of CRF method. Transformation based learning is use the function choice method to enhance the rules of tagging.

"Part-of-speech tagging based on dictionary and statistical machine learning"[43] by Zhonglin Ye, Zhen Jia, Junfu Huang, Hong feng. Part-of-speech tagging is the basis of NLP, and is broadly utilized in statistics retrieval, textual content processing and machine translation fields. The conventional statistical machine learning techniques of POS tagging rely upon the excessive trained factual information, but obtaining the training statistics is very time-eating. The methods of POS tagging primarily based on dictionaries ignore the context information, which lead to decrease overall performance. This paper proposed a POS tagging method which mixes methods based totally on dictionaries and traditional statistical device learning. The experimental outcomes show that the technique no longer most effective can clear up the hassle that the training facts are inadequate in statistical methods, however can also improve the performance of the strategies based totally on dictionaries. The People's Daily corpus in January 1998 is used as test data and the accurate price of POS tagging achieves 95.80%. For the anomaly phrase POS tagging, the accuracy achieves 88%.

"An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging"[46] by Canasai Kruengkrai and Kiyotaka Uchimoto and Jun'ichi Kazam .Authors present a discriminative phrase-character hybrid version for combined Chinese POS tagging and word segmentation. The model gives high overall performance because it can manage both unidentified and known words. The techniques that yield excellent balance for mastering the characteristics of known and unknown lexis and advocate an mistakes-driven coverage that supplies such balance via obtaining examples of unknown phrases from unique errors in a training corpus. Authors use a proficient framework for training the proposed model based totally at the Margin Infused Relaxed Algorithm; examine proposed technique at the Penn Chinese Treebank, and display that it achieve advanced performance as compared to the present day tactics reported inside the literature.

"A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic" [47] by Hitham M Abo Bakr, Khaled Shaalan. Recently the cost of written colloquial text has increased dramatically. It is getting used as a medium of expressing ideas specially across the WWW, typically in the form of blogs and partly colloquial articles. Most of these written colloquial has been within the Egyptian colloquial dialect, that's considered the maximum widely dialect understood and used all through the Arab world. Modern Standard Arabic is the reputable Arabic language taught and understood all over the Arab global. Diacritics play a key position in disambiguating Arabic textual content. The reader is predicted to deduce or expect vowels from the context of the sentence. Inferring the full form of the Arabic word is also useful whilst growing Arabic NLP tools and applications. Authors introduce a general technique for converting a written Egyptian colloquial sentence into its corresponding diacritized Modern Standard Arabic sentence which can without problems be prolonged to be implemented to different dialects of Arabic. In spite of the non-availability of linguistic Arabic assets for this undertaking, authors have advanced techniques for lexical acquisition of colloquial words which are used for transforming written Egyptian Arabic into Modern Standard Arabic. They efficiently used Support Vector Machine approach for Arabic text.

"An end-to-end discriminative approach to machine translation"[48] by Percy Liang Alexandre Bouchard-Cote Dan Klein Ben Taskar. Paper present a perception-fashion discriminative method to device translation in which big characteristic sets may be used. Unlike discriminative re ranking techniques, proposed technique can take gain of learning abilities in all stages of interpreting. Paper introduction talk is about several challenges to errors-driven discriminative techniques. In specific, they explore various methods of updating parameters for given a training data. Authors discover that building frequent however minor update is optimal to developing fewer however big updates. Furthermore, paper shows an array capabilities and display both how they increase BLEU score quantitatively and interact on specific examples qualitatively. One particular feature checks out is a unique manner is to introduce getting to learn knowledge of the preliminary phrase extraction approach, which has formerly been absolutely heuristic.

"A fuzzy ontology and its application to news summarization"[49] by Chang-ShingLee, Zhi-Wei Jian, Lin-Kai Huang.In this paper, authors focus on fuzzy ontology and its utility to information summarization. The fuzzy ontology with fuzzy necessities is an extension of an ontology area with standards. It is the suitable to explain the area expertise than place ontology for solving the uncertainty reasoning issues First, the domain ontology with numerous occasions of information is predefined via domain experts. The record preprocessing mechanism will generate the significant terms primarily based completely at the statistics corpus and the Chinese information dictionary described through the vicinity expert. Then, the significant phrases can be categorized consistent with the activities of the records via the term classifier. The fuzzy inference mechanism will generate the club ranges for each fuzzy idea of the fuzzy ontology. Every fuzzy idea has a membership levels associated with various events of the domain ontology.

Moreover, a news agent primarily based on the fuzzy ontology is also evolved for news summarization. The information agent carries five modules a file preprocessing mechanism, a retrieval agent, a sentence creator, a sentence route extractor and a sentence filter to carry out information summarization. In addition, They construct an trial internet site to check the proposed method The experimental outcomes show that the data agent based on the fuzzy ontology can correctly operate for information summarization.

III. PROPOSED APPROACH

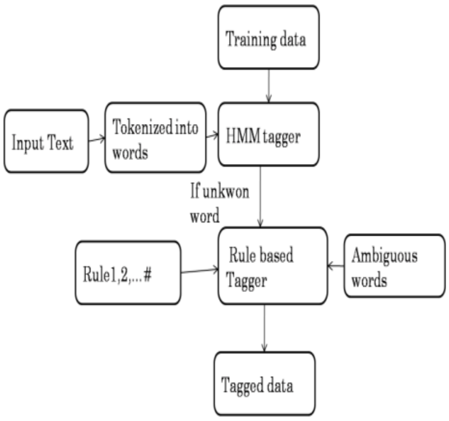

Fig 2.Proposed approach

Proposed approach is work as follow:

1. Input text data.
2. Tokenized input text in word by word.
3. Apply stochastic (Hidden Markov) Approach to assign appropriate tag the word.
4. If unknown words found which is not in dataset then apply different rules to tag unknown words.
5. We will get Output as a tagged data text.

Phase 1

In this phase we are taking input as text untagged data.

Phase 2

In this phase resources which are essential for tagging the text that is generated. So in this phase data cleaning and data tokenization are done.

Text MY NAME IS POOJA.”

```
>>> nltk.word_tokenize(“TEXT”)
[‘TEXT’]
>>> import nltk
>>> nltk.word_tokenize(text)
 [‘
[‘MY’ ,NAME’,’POOJA’,’.’]
```

Phase 3

In this phase we are applying HMM model to compute the probability of tags from the corpus and predict best sequence.HMM predict the probability of next word’s tag based on previous tag. Here we have used BIS tag set and 8000 sentences of arts and social database. After applying hmm will get tagged data as output text.

There are different types of HMM model available. It is a class of probabilistic models that presume that we can predict the probability of some future model without using the past information. Gram approach, used in the proposed system, which looks into previous *n-1* words.

There are simple 3-tuple in models Π is for Initial Probabilities, B is for Emission Probabilities, A is for Transition Probabilities .For input sequence of words W, we can assign a tag sequence T such that P(W,T) is maximized.

$$P(W,T)=\Pi^{N}_{i} \qquad (2)$$
$$[P(w_i|t_{1,\text{naïve}},w_{1,i-1})P(t_i|t_{1,i-1},w_{1,i-1})] \qquad (3)$$

Where P= prbabiity,W=w1…wn (sequence of words) and T=t1…..tn(sequence of tags).

Phase 4

In this phase we are applying rule based method to enhance the tagger accuracy by using figuring out ambiguous words. We also can resolve trouble of name entity reorganization by way of this technique. Some of the rules for noun, verb, adjective, pronoun are as below:

1.<u>Noun rule</u>

<u>Rule 1</u>: If current word in sentence is relative pronoun, then there is highest probability that subsequent word will be noun.

For Example:
This is that palace where ram was staying.
In above example that and where is relative pronoun and palace and ram is noun.

<u>Rule 2</u>: If given word in given sentence is adjective, then there is highest probability that subsequent word will be noun.
For Example
He is a true soldier.
In above example true is adjective and soldier is noun.

<u>Rule 3</u>: If word in given sentence is reflexive pronoun, then there is highest probability that subsequent word will be noun.
For Example
He gone at his home.
In above example his reflexive pronoun and home is noun.

<u>Rule 4</u>:If word in given sentence is personal pronoun then there is highest probability that subsequent word will be noun.
For Example
This is our village.
In above example our is personal pronoun and village is noun.

<u>Rule 5</u>: If a word in given sentence is post position, then there is highest probability that prior word will be noun.
For Example
He throw ball in the water.
In above example water is noun and in is post position.

Rule 6:If current word in given sentence is verb, then there is highest probability that prior word will be noun.
For Example
He was eating food.
In above example food is noun and eating is verb.

2.Proper noun rule

Rule 1: If present word is name and subsequent word is surname in given sentence, and then we tagged them as proper single noun.
For Example
Mukeshkumar
In above example Mukesh is name and Kumar is surname. They both are tagged as single proper noun

Rule 2: If present word is not tagged and subsequent word is tagged as proper noun in given sentence, then there is highest probability that present word will be proper noun.
For Example
rani jha
In above example rani, jha are tagged as proper noun.

3.Verb rule

Rule 1: If current word is not tagged and next word tagged as a auxiliary verb in given sentence, then there is highest probability that current word will be main verb.
For Example
He is eating food.
In above example food is main verb and Eating is auxiliary verb.

Phase 5

We will get tagged data as output using linguistic rules.

IV. EVOLUTION AND RESULTS

Evolution is done for enhancing the performance of system on different domains of sports and Entertainment data. The system was evaluated on 11208 word for sports data and 12809 words for entertainment data. We have considered 3 test case with 95614 words. These test sets are collected from multilingual Guajarati text available on TDIL.Following table shows the different test cases for testing

Table1 Test Cases

Test No.	Domain	No. of Words
1	Sports related data	11208
2	Entertainment related data	12809
3	Entertainment related data	95614

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

F-Measure =Recall *Precision / Recall + Precision

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

Recall = Number of correct answer specified by system / Total number of words.

First we apply Hidden Markov Model and got the output as following.

Table 2 Accuracy of system on different test cases' using HMM

Set	Precision	Recall	F-Measure	Accuracy A
1	72	72	72	70%
2	57	57	57	56%
3	51	51	51	52%

Next, we apply Rule based approach and got the output as following.

Table 3 Accuracy of system on different test cases using rule based approach and HMM

Set	Precision	Recall	F-Measure	Accuracy A
1	75	75	75	76%
2	72	72	72	80%
3	74	74	74	83%

A) Error Analysis

For analyzing the error in tag assignment we have taken Entertainment 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 tages are main tags like verb, noun, adjective etc and others are subtypes of it.

From table 4 and table 5 we can see that for Noun tag incorrect tag assignment is 1502 using HMM while using rule with hmm we are getting incorrect assignment 560 which is less compare to hmm. Same as for Verb 286 using HMM while 154 using hybrid approach. For adjective 218 using hmm while 124 using hybrid approaches.

Table 4 showing error analysis using hidden markov model for 30 tags.

Table 4 Error Analysis of data set using HMM

Actual Tag	Assigned Tag	Error
N_NN	RD_SYM	882
N_NNP	RD_SYM	620
JJ	RD_SYM	208
V_VAUX_VNP	RD_SYM	152
V_VM	RD_SYM	118
QT_QTC RD_SYM	RD_SYM	90
RD_PUNC	RD_SYM	48
N_NN	N_NNP	46
PSP	RD_SYM	32
N_NN	JJ	20
RB	RD_SYM	16
PR_PRP	RD_SYM	14
PR_PRP	DM_DMD	12
QT_QTO	RD_SYM	12
V_VAUX	V_VM	12
N_NNP	N_NN	12
DM_DMD	PR_PRP	10

N_NNP	JJ	10
CC_CCD	RP_RD 1	10
V_VAUX	RD_SYM	10
QT_QTF	RD_SYM	6
DM_DMI	PR_PRI	6
JJ	QT_QTO	6
V_VM	V_VAUX	6
N_NNP	QT_QTO	4
JJ	N_NST	4
N_NN	N_NST	4
DM_DMI	PR_PRI	4
N_NN	N_NNP	4

Now next table 5 shows the error analysis for 30 different tags by hybrid approach that is combination of Hidden Markov Model and rule based approach.

Now from the figure 3 we can see that accuracy using HMM for sports data is 70% which is improved by 6% using Hybrid approach.

For Entertainment data set accuracy achieved using HMM is 56% while accuracy achieved by hybrid approach is 80% that is 14% higher than using HMM.

Now, in figure 4 and figure 5 we can see accuracy of all tags available for tagging in hmm and hmm with rule based approach. The accuracy of tags assign in HMM is lower compare to Hmm with rule based approach tat we can observe from graph.

Table 5 Error Analysis of data set using rule based approach and HMM(hybrid)

ACTUAL TAG	ASSIGNED TAG	ERROR
N_NN	RD_SYM	298
N_NNP	RD_SYM	262
QT_QTC	RD_SYM	162
DM_DMD	DM_DMR	114
JJ	RD_SYM	114
V_VAUX_VNP	RD_SYM	60
PR_PRL	PR_PRQ	40
V_VM	RD_SYM	32
V_VM	V_VAUX	26
QT_QTO	RD_SYM	16
V_VAUX	JJ	16
PSP	RD_SYM	14
V_VAUX_VNP	V_VM	12
PSP	JJ	8
V_VAUX	V_VM	8
PR_PRP	DM_DMD	6
N_NST	RD_SYM	6
JJ	N_NNP	6
V_VAUX	RD_SYM	6
PR_PRL	RD_SYM	6
N_NST	PSP	6
PR_PRP	RD_SYM	6
DM_DMD	PR_PRF	4
JJ	PSP	4
QT_QTC	N_NNP	4

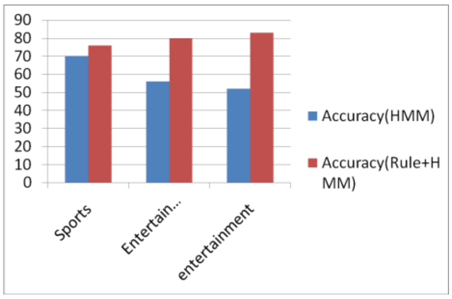

Fig 3. Comparison of accuracy using HMM and hybrid approach for both sets

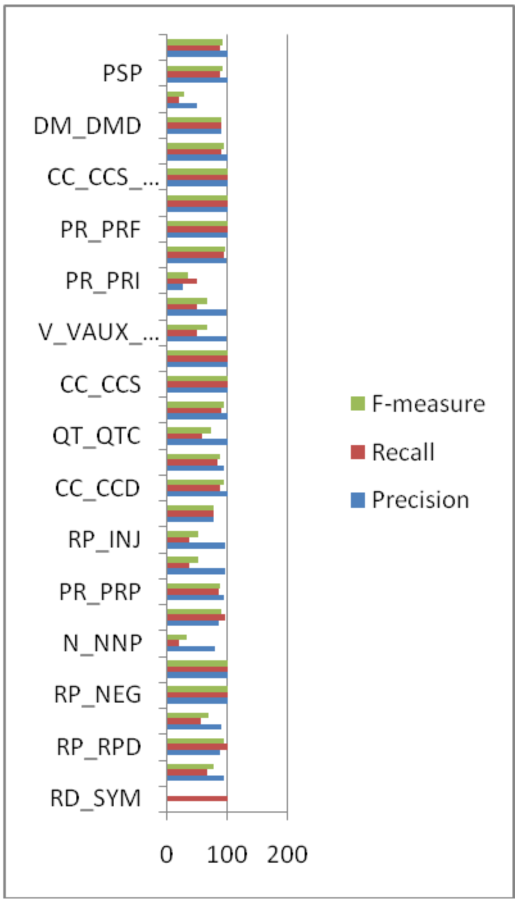

Fig 4. Per tag Accuracy using HMM

We have measured accuracy with the parameters like precision, recall and f-measure.

Now from figure 4 and figure 5 we can see that accuracy of Verb, Adjective, Noun, Quantifiers, Symbol, Punctuation are increased in terms of precision, recall and accuracy.

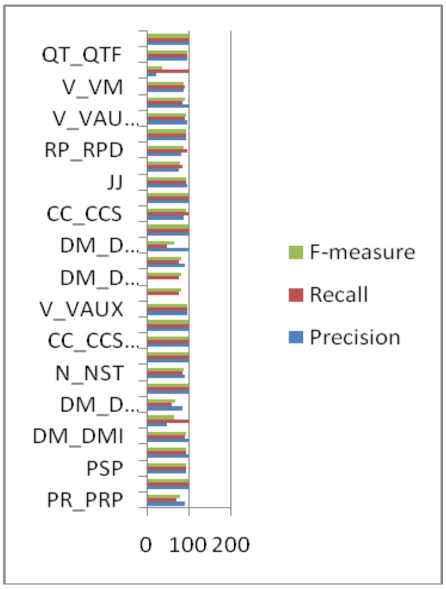

Fig 5. Per tag Accuracy using Rule with HMM

CONCLUSION

Development of highly accurate Part of speech tagger for Gujarati is an active research area of NLP. Our intention is to improving the accuracy of Gujarati POS tagger through proposed mixed technique. As in our mixed approach both rule and stochastic based models are implemented for word tagging. We designed set of rules for Noun, verb, adjective etc. We have carried out HMM and rule based methods to sports and entertainment records set. Applying HMM we have accuracy of 70% for Sports and 56% for Entertainment. After applying rule based technique we completed 76% accuracy for sports activities and 80% for Entertainment dataset.

After that we've got used enjoyment data set with 95614 words and we got 52% accuracy with hmm and 83 % accuracy through after applying rules with hmm. Rule based implementation on dataset to decide ambiguous and wrong tags assigned to a word via Guajarati Linguistic rules proposed in this paper. We have additionally seen consistent with tag accuracy for amusement records set using hybrid approach which is better than using HMM model. For future work we will enhance the accuracy of tagger via applying corpus based technique to generate stemmer for Guajarati language. We can also follow deep learning knowledge of method like Recurrent Neural Network for future work to enhance the tagger accuracy.

REFERENCES

1. POS tagging for Gujarati using CRF by Chirag Patel and Kartik Gali IJCNLP 2008.
2. Segmental HMM based Pos tagger by Mohammad Hadi, Hossein Sameti IEEE 2010.
3. POS tagging and chunking for indian languages by Himanshu agrawal IJCNLP 2008.
4. POS for Hindi corpus by Nidhi mishra, Amit mishra IEEE 2011.
5. HMM based POS tagger for Hindi by Nisheeth joshi ,Hemant Darbari, Iti mathur CSIT 2013.
6. POS tagging and chunking with HMM and CRF by Pranjal Awasth, Dilip Rao ,BalaramanRavindran, IJCAI – 2007.
7. POS for morphologically rich Indian languages: A survey by Dinesh kumar,Gurupreet sing Josan IJCA 2010.
8. Survey: POS tagger for Indian languages by Antony P J Dr.Soman K P IJCA, 2011.
9. Vishit: A visualizer for hindi text by Priyanka jain hemant darbari, virendrakumar bhavsar IEEE 2014.
10. Pos tagging of punjabi languageusing hidden markov model by Sapna Kanwar, Mr Ravishankar, Sanjeev Kumar Sharma Anu books 2011.
11. Learning Hidden Markov Model structure for Information extraction by Kristie seymore, Andrew McCallum,Ronald Rosenfeld in AAAI Technical Reports 2011.
12. Comparative study of various Machine Learning methods For Telugu Part of Speech tagging by Karthik Kumar G, Sudheer K, Avinesh Pvs International Institute of Information Technology at IJCAI – 2007.
13. HMM based POS Tagger and Rule-based Chunker for Bengali by Sivaji Bandyopadhyay,Asif Ekbal,Debasish Halder 2008.
14. Hindi POS Tagger Using Naive Stemming: Harnessing Morphological InformationWithout Extensive Linguistic Knowledge by Manish Shrivastava ,Pushpak Bhattacharyya ICON-2008.
15. Morphological Richness Offsets Resource Demand- Experiences in Constructing a POS Tagger for Hindi by Smriti Singh, Kuhoo Gupta, Manish Shrivastava, Pushpak Bhattacharyya ACL-2006.
16. Part of speech tagging and shallow parsing by delip rao and david yarowsky in SPSAL 2006.
17. Comparison of different POS tagging techniques for bangla by fahim hasahan , naushad uzzamand and Mummit Khan in BARC University.
18. Part of POS tagging and chunking with maximum entropy model by sandipan dandapat in SPSAL in 2007.
19. 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.
20. Sanskrit Tag-sets and Part-Of-Speech Tagging Methods by Sulabh Bhatt, Krunal Parmar and Miral Patel in IJIERE 2015.
21. Improving Punjabi Part of Speech Tagger by Using Reduced Tag Set by Manjit Kaur, Mehak Aggerwal and Sanjiv Kumar Sharma in IJCAIT December 2015.
22. Stemming technique and naïve approach for gujarati stemmer by Jikitsha Sheth and Bankim Patel ICRTITCS 2012
23. Comparision and mapping of two tagsets for gujarati language by purva dhakiya and Mohamed Yoonus
24. "Part of Speech Tagging and Chunking with Conditional Random Fields" by Aggarwal H, Anirudh Amni in the proceedings of NLPAI Contest 2006.
25. "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)
26. "Rule Based Hindi Part of Speech Tagger" by Garg N. ,Goyal V.and Preet S.(2012) Proceedings of COLING : Demonstration Papers, pages 163–174.
27. "A maximum entropy model for part-of-speech tagging"by Ratnaparkhi, Adwait. in Proceedings of the Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania, pp. 133–142.
28. "Survey of various POS tagging techniques for Indian regional languages "by Shubhangi Rathod, Sharvari Govilkar
29. "Parts of speech tagging for hindi languages using hmm" by Rajesh Kumar, Sayar Singh Shekhawat in International Journal Of Scientific Research 2018.
30. "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" By Christopher D. Manning in CICLing2011.
31. "Part of speech tagging for Gujarati text" by Pandya Abhinay, Dave Mainak. Part of speech tagging for Gujarati text. Dhirubhai Ambani Institute of Information and Communication Technology, viii, 44 p. (Acc.No: T00319) ,2011
32. "A comprehensive survey on parts of speech tagging approaches in Dravidian languages" by Merin Francis presented in The IIER International Conference, Beijing, China, 26th July 2015, ISBN: 978-93-85465-57-4.
33. "Opinion Mining and Sentiment Analysis –An Assessment of Peoples' Belief: A Survey" by S Padmaja and Prof. S Sameen Fatima in International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.4, No.1, February 2013.
34. Graph based Representation and Analysis of Text Document: A Survey of Techniques" by S. S. Sonawane ,Dr. P. A. Kulkarni in International Journal of Computer Applications (0975 8887) Volume 96 - No. 19, June 2014.

35. ”A Survey of Ontology Learning Approaches” by Maryam Hazman, Samhaa R. El-Beltagy, Ahmed Rafea International Journal of Computer Applications (0975 – 8887) Volume 22– No.9, May 2011.
36. ”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.
37. [37]”Survey: Part-Of-Speech Tagging in NLP. International Journal of Research in Advent Technology” by Nidhi Adhvaryu, Prem Balani (E-ISSN: 2321-9637) ICATEST 2015.
38. ”A Comparative Study on the Effectiveness of Part-of-Speech Tagging Techniques on Bug Reports” by Yuan Tian and David Lo(978-1-4799-8469-5/15) IEEE 2015.
39. ”Rule-based Approach in Arabic Natural Language Processing” by Khaled Shaalan in International Journal on Information and Communication Technologies, Vol. 3, No. 3, June 2010.
40. ”Recent advances in natural language generation: a survey and classification of the empirical literature” by Rivindu Perera, Parma Nand in Computing and Informatics, Vol. 36, 2017.
41. ”Statistical analysis of part of speech (pos) tagging algorithms for english corpus” by Swati Tyagi, Gouri Shankar Mishra in International Journal of Advance Research , Ideas and Innovations in Technology.(ISSN: 2454-132X ,Volume2, Issue3).
42. ”Part of Speech Tagging Using Statistical Approach for Nepali Text” by Archit Yajnik in World Academy of Science, Engineering and Technology International Journal of Cognitive and Language Science Vol:11, No:1, 2017.
43. ”Part-of-speech tagging based on dictionary and statistical machine learning” by Zhonglin Ye, Zhen Jia, Junfu Huang, Hongfeng Yin IEEE 29 August 2016(ISSN: 1934-1768)
44. ”A Hybrid Approach to Vietnamese Word Segmentation Using Part of Speech Tags” by Dang duck Pham,Giang Binh Tran,Son Bao Pham in IEEE December 2009.
45. ”Part-of-speech tagging using decision trees” by Lluís Màrquez Horacio Rodríguez in Applications of MLSpringer June 2005.
46. ”An Error-DrivenWord-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging”by Canasai Kruengkrai and Kiyotaka Uchimoto and Jun’ichi Kazam in International Joint Conference on Natural Language Processing of the AFNLP: 2009.
47. ”A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic” by Hitham M Abo Bakr, Khaled Shaalan in Research Gate March 2008.
48. ”An end-to-end discriminative approach to machine translation” by Percy Liang Alexandre Bouchard-Cˆot´e Dan Klein Ben Taskar in ACL-44 Proceedings of the 21st International Conference on Computational Linguistics July 2006.
49. “A fuzzy ontology and its application to news summarization” by Chang-ShingLee, Zhi-Wei Jian, Lin-Kai Huang in IEEE Transactions on Systems, Man, and Cybernetics, Volume: 35, Issue: 5, Oct. 2005(ISSN: 1083-4419).

AUTHORS PROFILE

Prof. Pooja M. Bhatt is having 8 years of academic experience. She had competed her BE in computer engineering from Saurashtra University and master in computer science and engineering from Gujrat Technological University. She has published more than fifteen papers in national and international conferences as well as in journals. She is pursuing her Ph.D. from Charusat University in the area of Natural Language Processing.She is member of ISTE comeete.She has pubished two books.

Dr. Amit Ganatra is having 20 years of academic experience. Authored 100 research publications, supervised over 100 industry projects, supervised over 60 dissertations, supervising 6 Ph.D. Research Scholars, participated and developed over 5 consultancy projects. He is working as departments’ (CE/IT) coordinator for Accreditation and ISO at CHARUSAT. Member of Association of International Professor (AIP). Editorial board member of IJCSIT,IJSC as well as PCM member of the related conferences of AIRCC world wide.