International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072

A Review on Part-Of-Speech Tagging on Gujarati Language

Mr.Vipul Gamit[1], Rutva Joshi[2], Ekta Patel[3]

[1,2,3]*Babu Madhav Institute of Information Technology, Bardoli, India*

Abstract - *Parts of speech tagging (pos) is an important tool for processing the natural language. This review paper includes the various methods used for pos tagging and aim of this paper is to review the implementation of Part of Speech (POS) Tagger for Gujarati language which will help in building accurate corpus for Gujarati Language. This paper describes different methods used for Guajarati parts of speech tagging. It is simplest and most stable model for Natural Language Processing application. It is the process of marking up the words in the corpus according to particular parts of speech like noun, pronoun, verb, Adverb, adjective, conjunction, preposition. It is very essential task and pre-processing step for all the natural language processing activities. A POS tagger takes a sentence from input data and assigns a unique parts of speech tag to each lexical item of the sentence. There are many challenges in POS tagging like Ambiguities, foreign words, un-annotation etc.*

***Key Words*: POS, Gujarati NLP, Tag-set, Ambiguity, Stop words, Rule-base, Stochastical, Machine Learning.**

1. INTRODUCTION

Natural language processing (NLP) is currently an active research area today. Different aspects of NLP have been subdivided into separate topics, one of them is POS tagging **[13]**. POS tagging plays a vital role in the development of Natural Language Processing applications. It simply means labelling words with their appropriate Part-Of-Speech. It is one of the simplest as well as most stable and statistical model for much NLP application, POS Tagging is an initial stage of information extraction, summarization, retrieval, machine translation, speech conversion. It is the process of tagging the words of a running text with their categories that best suits the definition of the word as well as the context of the sentence in which it is used.

Parts of speech have two different categories:

Open class: - This is the class where you can add new words anytime and things are correspond to new words

E.g.:-Nouns, Verbs, Adverb and Adjective

Closed class: - This is the class where you cannot add new words.

E.g.:- Conjunction, Determiners, Pronouns

Work in this field is usually either Stochastical, machine learning based, or rule based **[3]**.Some of the model that uses the first approach are Hidden Markove Model (HMMs), Conditional Random Fields (CRFs), Maximum Entropy Markove Model (MEMMs) etc.
POS tags are also known as word classes, morphological classes, or lexical tags to choose correct grammatical tag for word on the basis of linguistic feature.

1.1 Current needs of POS tagging:-

1. Part-of-speech tagging is only a first necessary step in understanding what a text is about.
2. POS tags have been used for a variety of NLP tasks and are extremely useful since they provide linguistic signal on how a word is being used within the scope of a phrase, sentence, or document.
3. POS is very useful in cases where it distinguishes the word sense (the meaning of the word).
4. It is used for information retrieval, classification.
5. To check off the words and punctuation in a textual matter having suitable POS labels of Gujarati text **[4]**.

1.2 Architecture of pos tagger:-

The figure demonstrated below presents the user interface architecture of the POS taggers developed for the Gujarati languages.

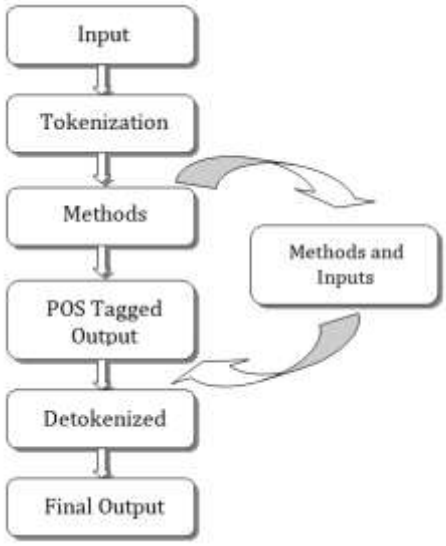

Figure 1: Architecture of POS Tagger [11]

1. The users provide the input in Guajarati language and then select Algorithm.
2. The taggers internally tokenize the input data and process it.
3. They send the input text to the respective algorithms and process the tagged output.
4. Finally, the tagged data is detokenized and the final output is shown on the display

2. Classification of pos tagging

A Part-Of-Speech Tagger (POS Tagger) is defined as a part of software which assigns parts of speech to every word of a language that it reads. The approaches of POS tagging can be divided into three categories; rule-based tagging, hybrid tagging and Stochastical tagging **[4]**.

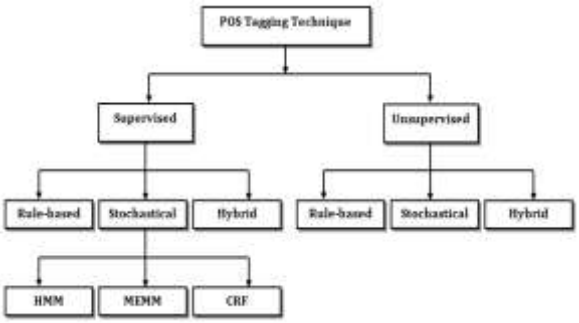

Figure 2: Classification of pos tagging

2.1 Supervised Technique: The supervised POS tagging models require pre-tagged corpora which are used for training to learn information about the Tag-set, word-tag frequencies, rule sets etc. [12]. It is the technique in which we deal with the data which is labelled. If you have labelled class column, then the analysis is supervised. Pre-tagged models are required by the supervised POS tagging models as they are used to learn information about the tag-set, word-tag frequencies, rule sets etc. for training. Increase in the size of corpora generally increases the performance of the models come supervised technique, the predicted output is compared with desired output based upon accuracy and we will have various performance measurement.

Supervised technique has other two parts

1 Classification: - When we have to predict the class that particular word falls in which class.
2 Regression:- Deciding the next based on the previous action

Working of supervised technique:-

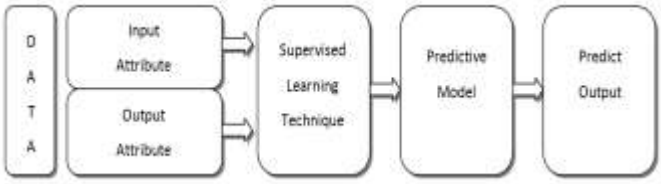

Figure 3: Process of Supervised Technique

Unsupervised Techniques: The unsupervised POS tagging models do not require pre-tagged corpora. Instead, they use advanced computational methods like the Baum-Welch algorithm to automatically induce tagsets, transformation rules etc. Based on the information, they either calculate the probabilistic information needed by the stochastic taggers or induce the contextual rules needed by rule-based systems or transformation based systems **[12]**. In this conclusion, opinion is formed based on the information analysis.

Un-supervised techniques have other two parts:-

1 Dimension reduction technique(trying to reduce no of variables from data)
2 Clustering(trying to reduce no of records, cases).

Working of the unsupervised technique:-

Figure 4: Process of Unsupervised Technique

2.2 Rule based Technique:-

Rule-based tagger use linguistic rules to assign the correct tags to the words in the sentence or file. A set of hand written rules were applied and also contextual information was used in order to assign POS tags to words in the rule based POS tagging. These rules are generally known as context frame rules **[4]**.

The tagger is divided into two stages:

1. It searches word in dictionary.
2. It assign a tag by removing disambiguate of words using linguistic features of word.

The module reads the Gujarati corpus and split the sentence into words according to the delimiter. The system finds the words in the database and assigns the appropriate tag to the words. Rule based approach required less amount of data and vast knowledge about the language. Rule based system is usually difficult to develop.

The main disadvantages of the rule based systems are the necessity of a linguistic background and manually constructing the rules **[4]**.

The system mainly works in two steps-firstly the input words are found in the database, if it is present then it is tagged. Secondly if it is not present then various rules are applied.

Algorithm:

1. Input the text using file upload button or manually enter by user.
2. Tokenize the input text word by word.
3. Normalized the tokenized words. I.e. separate out the punctuation marks and the symbols from the text.
4. Search the number tag by using Regular Expression. For Example: - ૨૦૧૨, ૧-૨, ૧૨મી etc.
5. Search the date tag by using regular expression. For Example: - ૧૭/૧૦/૧૯૧૭ etc.
6. Search the time tag by using regular expression. For Example: - ૧૭: ૧૦, ૧૦: ૧૦: ૧૦ etc.
7. Search for the abbreviation using regular expression. For Example: - એ. આર. કે etc.
8. Search in database for different input words and tag the word according to corresponding tag.
9. Then different rules are applied to tag the unknown words.
10. Display the tagged data to the user.

The disadvantage of this system is that it doesn't work when the text is not known. The problem being that it cannot predict the appropriate text. Thus in order to achieve higher efficiency and accuracy in this system, exhaustive set of hand coded rules should be used **[4]**.

1 Lot of manual work: The Rule Based system demands deep knowledge of the domain as well as a lot of manual work **[2]**.
2 Less learning capacity: Here, the system will generate the result as per the rules so the learning capacity of the system by itself is much less **[2]**.
3 Time consuming: Generating rules for a complex system is quite challenging and time consuming **[2].**
4 Complex domains: If an application that you want to build is too complex, building the rule based system can take lot of time and analysis. Complex pattern identification is a challenging task in the rule based approach **[2].**

2.3 Stochastical Technique:-

A stochastic approach includes frequency, probability or statistics. The simplest stochastic approach finds out the most frequently used tag for a specific word in the annotated training data and uses this information to tag that word in the unannotated text **[12]**.

An alternative to the word frequency approach is known as the n-gram approach that calculates the probability of a given sequence of tags. It determines the best tag for a word by calculating the probability that it occurs with the n previous tags, where the value of n is set to 1, 2 or 3 for practical purposes. These are known as the Unigram, Bigram and Trigram models.

There are different models that can be used for stochastic POS tagging, some of which are described below:-

1. Hidden Markove Model (HMM)
2. Maximum Entropy Marcov Model (MEMM)
3. Conditional Random Field (CRF)

Disadvantages: - disadvantage of this system is that some sequences of tags can come up for sentences that are not correct according to the grammar rules of a certain language.

Hidden Markov Model

The HMM is a sequence model. A sequence model is a model whose job is to assign a label or class to each unit in a sequence. It is a probabilistic sequence model: given a sequence of units (words, letters, morphemes, sentences, whatever), it computes a probability distribution over possible sequences of labels and chooses the best label sequence **[15]**.

The model has a number of interconnected states connected by their transition probability. A transition probability is the probability that system moves from one state to another. A process begins in one of the states, and moves to another state, which is governed by the transition probability. An output symbol is emitted as the process moves from one state to the next. These are also known as the Observations **[16]**.

A first-order hidden Markov model instantiates two simplifying assumptions **[15]**.

1. First, as with a first-order Markov chain, the probability of a particular state depends only on the previous state:

   Markov Assumption:
   P (qi |q1...qi−1) = P(qi |qi−1)

2. Second, the probability of an output observation oi depends only on the state that produced the observation qi and not on any other states or any other observations:

   Output Independence:
   P(oi |q1 ...qi ,...,qT ,o1,...,oi ,...,oT)=P(oi |qi)

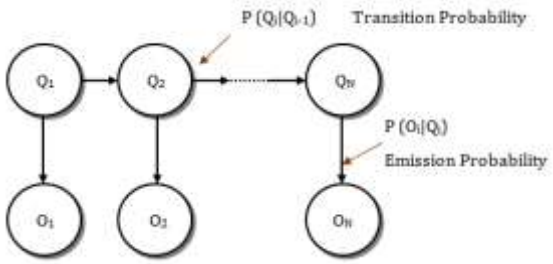

Figure 5 : Hidden Markove Model [20]

Maximum Entropy Marcov Model

MEMM are conditional probabilistic sequence models **[19]**. This Model determines the probabilities based upon constraints. Upon the application of constraints the most probable sequence of tags is produced. These constraints are determined from the preparation information, keeping up connection between the history and probable Outcomes **[16]**.

Conditional Random Fields

CRF stands for Conditional Random Field. It is a type of discriminative probabilistic model. It has all the advantages of MEMMs without the label bias problem. CRFs are undirected graphical models (also known as random field) which are used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes **[17]**.

Hybrid Technique:-

This approach combines the advantages of both of the above rule based approach and stochastic approach. Words in this technique are first tagged probabilistically and then as post processing, linguistic rules are applied to tag tokens. Accuracy of taggers based on this approach generally gives good results than other techniques **[9].**

Table 1 Advantage-Disadvantage of Approaches

Sr. No	Approach	Advantage	Disadvantage
1.	Rule-based Method	• High Precision. • It can effectively remove ambiguous tags **[10]**. • It can tag the words which have never been encountered **[10]**. • It has the potential to tag almost any sentence.	• Lot of manual work: The Rule Based system demands deep knowledge of the domain as well as a lot of manual work **[2].** • Time consuming: Rules for a complex system is quite challenging and time consuming **[2].**

		• This system is cost efficient and accurate of its end result **[2].**	• It do not improves with data its answer is always fixed.
2.	Stochastical Method	• Requires minimal human efforts • Can be created for any language pair with enough training data • Can prototype a new system quickly at a very low cost • Resolves linguistics uncertainty problems by a solid mathematical basis. • Extract knowledge from corpus **[22].**	• Some sequences of tags can come up for sentences that are not correct according to the grammar rules of a certain language.
3.	HMM	• Strong statistical foundation. • Efficient learning algorithms. • Can handle inputs of variable length – most flexible generation of sequence profiles **[21]**.	• Large number of unstructured parameters. • Limited by first-order markov property. • They cannot express dependencies between hidden states **[21]**.
4.	MEMM	• Increased freedom in choosing features to represent observations.	• suffer from the "label bias problem,"
5.	CRF	• It is possible to reach high quality of labelling if you choose right features • CRF is flexible enough in terms of feature selection. In addition, it is not necessary for features to be conditionally independent	• CRF is highly computationally complex at the training stage of the algorithm. It makes it very difficult to re-train the model when newer data becomes available.
6.	Hybrid Method	• It can effectively remove ambiguous tags. • It can tag the words which have never been encountered. • It has the potential to tag almost any sentence. • It has less chances of error. • It can also Tag wrong sentences • It can also tag sentences with ambiguous structure **[10]**.	—

3. Literature Review

1. Paper [1] includes the statistical approach for tagging Guajarati text, using the Guajarati dataset of 351 words and come up with 92.8% accuracy.
2. Paper [3] includes pos tagging using conditional random field method, with the dataset of 5000 words and came up with 89.90% accuracy.
3. Paper [4] includes different approaches with their comparison using the dataset of 3000 statements and in which HMM method came up with 93.38% accuracy.
4. Paper [5] statistical chunker for Indian language Guajarati used statistical method with the dataset of 5000 statements that came up with 96% accuracy.
5. Paper [6] includes pos tagging using trigram method on Marathi language, with dataset of Trigram Method that came up with 91.63% accuracy.
6. Paper [7] includes Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati using the dataset of 8,525,649 Words that came up with 90.7% accuracy.

7. Paper [9] Survey on Part-Of-Speech Tagging of Indian Languages using different methods with the different dataset Words that came up with high accuracy.
8. Paper [11] includes CRF and SVM for Hindi, Bhojpuri and Odia language using the dataset of 21000 words that came up with 82.67% accuracy, 10000 words that came up with 82% accuracy, and 90000 words that came up with 89% accuracy.
9. Paper [13] includes the statistical approach CRF for tagging Guajarati text, using the Guajarati dataset of 1000 words and come up with 92% accuracy.
10. Paper [14] includes the Rule-based approach for tagging English text, using the dataset of 40293 sentences and come up with 93% accuracy.
11. Paper [17] includes the Hybrid approach for tagging Hindi text, using the dataset of 80000 words and come up with 89.9% accuracy.

Table 2 : Literature survey on Gujarati Language

Sr. No	Approach	Corpus Size	Reference	Accuracy
1.	Stochastic approach	351 words	"part of speech tagging using statistical approach for Guajarati text"	92.87%
2.	Stochastic approach - CRF	5000 words	"Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields"	89.90%
3.	Stochastic approach	5000 Statement	"a statistical Chunker for Indian language Guajarati"	96%
4.	Rule-based and Hybrid approach	8,525,649 Words	"Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati"	90.7%
5.	Stochastic approach - CRF	10000 words	"Improve accuracy of Parts of Speech tagger for Gujarati language"	92%

Table 3 : Literature survey on other language

Sr. No	Method	Word Tagset	Reference	Accuracy	Language
1.	Statistical approach	18160 Words	"POS Tagging Approaches: A Comparison"	93.38%	Hindi
2.	Trigram Method	48635 Words	"Part Of Speech Tagging Of Marathi Text Using Trigram Method"	91.63%	Marathi
3.	Rule-Based approach	26149 Words	"A Survey on Part-Of-Speech Tagging of Indian Languages"	87.55%	Hindi
	HMM approach	1003 Words		93%	Bengali
	Trigram	2000		91.63%	Marathi

	Method	Sentence			
	Rule-Based approach	97 Category of Manipuri language		92%	Manipuri
	HMM approach	51269 Words		79.9%	Kannada
	HMM approach	70000 Words		90%	Sinhala
	CRF approach	21425 Words		77.37%	Telugu
4.	CRF approach	21000 Words	"Training & Evaluation of POS Taggers in Indo-Aryan Languages: A Case of Hindi, Odia and Bhojpuri"	82.67%	Hindi
	SVM approach	10000 Words		82%	Odia
	CRF approach	90000 Words		89%	Bhojpuri
5.	Rule-Based approach	49203 Sentences	"Part-Of-Speech Tagging From An Information-Theoretic Point Of View"	93%	English
6.	Hybrid approach	80000 Words	"Hybrid approach for Part of Speech Tagger for Hindi language"	89.9%	Hindi

4. Conclusions

Natural Language is the medium for communication which is incorporated by every human being. One of the most important activities in processing natural languages is Part of Speech tagging **[4]**.This paper includes classification of different methods used for parts of speech tagging in Gujarat language. From our literature survey, we found that hybrid method gives higher accuracy. The methods used for pos tagging are Rule based, Stochastically, Hybrid. Rule based method use linguistic rules to assign the correct tags to the words in the sentence or file. Stochastical approach finds out the most frequently used tag for a specific word in the annotated training data and uses this information to tag that word in the annotated text. In hybrid method, words in this technique are first tagged probabilistically and then as post processing, linguistic rules are applied to tag tokens. Through these methods we will be able to tag the words given in the particular sentence.

4. Reference

[1] Yajnik, A. and Prajapati, M., 2017. PART OF SPEECH TAGGING USING STATISTICAL APPROACH FOR GUJRATI TEXT.

[3] Patel, C. and Gali, K., 2008. Part-of-speech tagging for Gujarati using conditional random fields. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.

[4] Kumawat, D. and Jain, V., 2015. Pos tagging approaches: A comparison. International Journal of Computer Applications, 118(6).

[5] Patel, C. and Ahalpara, D., 2015. A STATISTICAL CHUNKER FOR INDIAN LANGUAGE GUJARATI

[6] Singh, J., Joshi, N. and Mathur, I., 2013. Part of speech tagging of Marathi text using trigram method. Ar7Xiv preprint arXiv: 1307.4299
[7] Suba, K., Jiandani, D. and Bhattacharyya, P., 2011. Hybrid inflectional stemmer and rule-based derivational stemmer for Guajarati. In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP)(pp. 1-8)
[8] Joshi, V.C. and Vekariya, V.M., 2017. An Approach to Sentiment Analysis on Gujarati Tweets. Advances in Computational Sciences and Technology, 10(5), pp.1487-1493
[9] Mehta, D.N. and Desai, N.P., 2015. A Survey on Part-Of-Speech Tagging of Indian Languages. History, 43(198), pp.125-131.
[11] Behera, A.K.O.P., Singh, S. and Jha, G.N., Training & Evaluation of POS Taggers in Indo-Aryan Languages: A Case of Hindi, Odia and Bhojpuri.
[12] Hasan, F.M., 2006. Comparison of different POS tagging techniques for some South Asian languages (Doctoral dissertation, BRAC University).
[14] Vanroose, P., 2001. Part-of-speech tagging from an information-theoretic point of view. In 22nd Symposium on Information Theory in the Benelux.
[15] Jurafsky, D. and Martin, J.H., 2014. Speech and language processing (Vol. 3). London: Pearson.
[16] Anand, A., 2014. Parts of speech tagging using hidden Markov model, maximum entropy -dissertation).
[17] Mohnot, K., Bansal, N., Singh, S.P. and Kumar, A., 2014. Hybrid approach for Part of Speech Tagger for Hindi language. International Journal of Computer Technology and Electronics Engineering (IJCTEE), 4(1).
[**18]** Garg, N., Goyal, V. and Preet, S., 2012. Rule based Hindi part of speech tagger. Proceedings of COLING 2012: Demonstration Papers, pp.163-174.
[19] Adhvaryu, N. and Balani, P., 2015, March. Survey: Part-Of-Speech Tagging in NLP. In International Journal of Research in Advent Technology (E-ISSN: 2321-9637) SpecialIssue 1st International Conference on Advent Trends in Engineering, Science and Technology "ICATEST 2015"

Websites:

[4] https://scholar.google.co.in/
[10]http://airccse.org/journal/ijit/papers/4315ijit01.pdf
[20]http://www.davidsbatista.net/blog/2017/11/11/HHM_and_Naive_Bayes/
[21]https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781787121423/7/ch07lvl1sec71/challenges-for-the-rule-based-system
[22]https://pdfs.semanticscholar.org/d31f/e781518d98f07ae992b6c0a574ab29f127ce.pdf?_ga=2.232017819.508386308.1554821224-892306099.1554821224
[23]https://www.youtube.com/playlist?list=PLLssT5z_DsK8BdawOVCCaTCO99Ya58ryR