RESEARCH ARTICLE OPEN ACCESS

Paradigm based Part of Speech Tagging with priorities: Implantation for Gujarati Script

Kapadia Utkarsh N [1], Deasi Apurva A [2]

[1] Department of Computer Science, Veer Narmad South Gujarat University, Surat - India
[2] Department of Computer Science, Veer Narmad South Gujarat University, Surat - India

ABSTRACT

Part of Speech Tagging is an important aspect of any natural language processing application like grammatical inference, information retrieval, and machine translation. The problem of tagging in is to assign most appropriate tag for each word present in a sentence based on its lexical and contextual aspects. There are mainly two types of approaches available, supervised and unsupervised. Unsupervised approach is not based on certain pre-defined rules or labelled text. We have presented here a hybrid Part-of-Speech (POS) tagger for Gujarati. Rules are devised with language experts and native language speakers. We have evaluated performance with dictionary of 30,050 words on 12,637 sentences for 30 different standard part of speech tags for Gujarati. Evaluation of the system is done on text from various domains of Gujarati. These domains includes news, essays and short stories. Our system has achieved accuracy of 82.52%

Keywords:- Part Of Speech(POS), Tagging, Gujarati Script.

I. INTRODUCTION

POS tagging is the problem of assigning natural language sentences with the most appropriate POS tag like noun, verb, preposition, pronoun, adverb and adjective or other lexical class for each word present in a sentence. Two types of approaches are mainly found in literatures, for POS tagging namely supervised and unsupervised. Unlike unsupervised, supervised tagging uses tagged corpus. Size of corpus increases the accuracy of tagging in supervised approaches. Both approaches can be further classified in to stochastic, rule based and hybrid approach. Recently Neural network based approach are being explored.

For any POS tagger, set of tags called standard tag set is inevitable. Standard tag set contains tags for major grammatical categories and sub tags according to morph-syntactic feature of language. Tag set may also vary according to scope and objective of project. For Indian languages several tag-sets are existing and most prevailing tag-set is developed with help of ILMT (Indian Language Machine Translation) guidelines [20]. There are two set of standard POS tag sets developed by LDC-IL (Language Data Consortium for Indian Language) and BIS (Bureau of Indian Standard) for major Indian languages including Gujarati. Tag set used is shown in Appendix A.

Indian languages are morphologically and inflectionally rich, and there is scare of large annotated corpus in most Indic languages, so it is difficult to carry out work. Mostly work on Indian languages in literatures are found in Hindi, Bengali, Tamil, Telugu, Punjabi and Marathi etc. but very less work is found on Gujarati. POS tagger design should be done considering all morph-syntactic categories that can occur in particular language [8]. There are two aspects important for arriving to syntactic category of word in sentence. First being lexical aspect without referring to context of the word. Second is contextual aspect to assign syntactic category to each word in a sentence. Second aspects helps in disambiguation when a particular word appears in different contexts.

Gujarati being Indo Aryan language being inflectionally and morphologically rich. Gujarati has fifty letters, out of them thirty four consonants and sixteen vowels as per Devanagri characters but out of them only eleven vowels and twenty nine consonants are used commonly. Gujarati words can be classified mainly in five classes, also known as Part of Speech. They are Noun, Pronoun, Adjective, Verb, and others. Noun admits inflection to express number, gender and case. Two numbers are singular and plural, and genders are masculine, feminine and neuter. There are seven cases in Gujarati speech omitting vocative, they are nominative, agentive, dative, genitive, instrumental and locative. Gujarati nouns are mostly ending in vowels e.g. અ, આ, ઇ, ઉ, એ, ઓ, ઐ etc but less nouns ending in consonants e.g. ખ, ઠ, શ. Gujarati nouns are formed by: Noun stem + Gender Market + Number Marker + Case Marker. E.g. છોકરાઓને (for boys) can be expressed by: છોકર + ◌ા + ઓ + ને.In Hindi, case marker comes separately in sentence e.g.लडकोने though both languages belongs to same Indo-Aryan family.

II. RELATED WORK

Considerable work in literatures found particularly on part-of-speech tagging, are mostly based on statistical methods for morphologically rich languages [8]. POS taggers are available in various Indic languages like Assamese, Bengali, Hindi, Kanadda, Marathi, and Telugu etc.

Aniket et al. in 2006[2] have presented Part-of-Speech tagger based on Maximum Entropy (ME) Markov Model. They have trained system with annotated corpus for Hindi and assigns tags to previously unseen tags. They have used various features together to predict tag for a particular word. Feature set was based on word feature, dictionary feature, context-based features and corpus-based features. Except context-based feature all other features are language dependent. They have used corpus of NLPAI-ML with 35000 words annotated with 29 tags. They could achieve accuracy of 89.34% for POS tagging.

Smriti Singh et al.[3] in 2006 have demonstrated method of POS tagging which can be used for low resource languages. They have used locally annotated corpora of 15,562 words and lexicon database with high overage with 42000 entries in 26 categories based on decision tree based learning algorithm (CN2). They have used corpus of BBC news site. System was evaluated with 4- fold cross validation. Simply lexicon lookup approach gave 61.19% accuracy while applying morphological rule it gave accuracy of 73.62% which was further improved to 82.63% by applying disambiguation rules which they call it BL (Baseline) approach. Accuracy of POS tagger was reported about 93.45%.

Himanshu Agarwal[4], 2006 presented Conditional Random Field based approach for Hindi POS tagging. Morphological analyser was used to provide root and possible POS tag information for training. They have trained system on 21000 words. CRF based tagger was 82.67% accurate and with chunking accuracy was improved to 90.89%. Reason for not very good performance of CRF based tagger was small size of training data as CRF being discriminative in nature it requires larger amount of training data.

Chirag Patel[4], 2008 developed machine learning algorithm for Gujarati Part of Speech Tagging. Machine learning part is performed using CRF model. Features were given to CRF are chosen keeping the linguistic aspect of Gujarati in mind. As Gujarati being resource poor language, manually tagged data of 600 sentences were used. Their tagset contains 26 different tags which are standard Indian Language (IL) tagset. Both tagged 600 and untagged 5000 sentences are used for learning. Their algorithm achieved 92% accuracy on limited training corpus of 10,000and test corpus of 5,000 words.

Manish Shrivastava[6], 2008 simple tagger for Hindi based on Hidden Markov Model (HMM) was presented. It uses naïve (longest suffix matching) stemmer as pre-processor to achieve reasonably good accuracy. This method do not require any other linguistic resource apart form a list of possible suffixes for the language. It was also generated from machine learning technique. It increases probability of correct choice while decreasing the ambiguities of selection. As a pre-processor they have employed longest suffix matching stemmer and achieved 93.12% of accuracy.

Most of work done in POS is based on statistical approach or rule-based approach, and by increasing size of lexicon accuracy can be improved. Summary of some of major work done in POS tagging is mentioned in below table.

Table I. Indian language POS tagger Summary

Source / Year	Lexicon Size# / Method	Test Data#	Script	Accuracy (%)
IIT Bom 2006	35000 Max Entropy		Hindi	89.34
IIT Bom 2008	42000 Decision Tree	15562	Hindi	86.77
IIIT Hyd 2006	21000 CRF		Hindi	82.67
IIIT Hyd 2008	10000 CRF	5000	Gujarati	91.94
IIT Bom 2008	HMM		Hindi	93.12
Punjabi Uni 2012	18249 Rule Based	26149	Hindi	87.55
CDAC Pune 2013	358288 HMM		Hindi	91.63

So from above table it can be inferred that limited amount work is done in Gujarati, and that mostly based on statistical methods. Our approach is considering grammar rules specific to the language which can help to lead to higher accuracy for a particular language. Since rules formation requires native language speaker and grammar experts, work related to rule based implementation is less reported in literature

III. SYSTEM DESCRIPTION

Proposed system is based on rule based approach with 30 different part of speech (POS) tags which are given in Appendix. Tags are prescribed by Department of Information Technology Ministry of Communication & Information Technology with few other tags which are time, date and number tags etc. Collection of 30,052 words were tagged with various tags used as lexicon. The system works by first finding input word in the database; if the input word is present then it is tagged. If word is present but more than one tag is found, more than one tag is produced by tagger. But if the given word is not matching with any of the word in database then rules are also applied in similar way would lead to more than one result. Algorithm for the same is given in below Table 2:

Table II: Algorithm for replacing affixes from word

```
Input: Gujarat Unicode text file
Output: Gujarati text with each word followed by tag
While Not (End of file(EOF)) is reached
Split the file into sentences by End of Sentence '.' marker
Repeat for each sentence
 Tokenize each sentence in to words by
          word separator ' '
   Repeat
   search if word in root table for corresponding tag(s)
          if word is present then
          result_tag = result_tag + search_result
          else
          for each case marker suffix do
          if word ends with suffix then
          remove suffix from word
          (word = word – matching case suffix)
          sub_tag = sub_tag + case_info
          end if
          end for
          for each plural marker suffix do
          if word ends with suffix then
          remove suffix from word
          (word = word – matching plural suffix)
          sub_tag = sub_tag + plural_info
          end if
          end for
          for each gender marker suffix do
             if word ends with suffix then
                remove suffix from word
               (word = word – matching gender suffix)
               sub_tag = sub_tag + gender
            end if
          end for
          search if the word in root table for corresponding tag(s)
          if word is present then
          result_tag = result_tag + matched_tag +sub_tag
          end if
          for each verb tense suffix do
           if word ends with suffix then
               remove suffix from word
              (word = word – matching verb suffix)
              sub_tag = sub_tag + verb-class
            end if
          end for
          end if
   Until each word in the sentence is processed
  Until all sentences are processed
End While
```

Working of above algorithm is depicted in following table of examples. Total suffix would be forty eight for nouns if we take all combinations of number, case and gender which are reduced to thirteen by dividing them on the basis of appearance in word. Similarly total verb affixes based on each of fifty form of word 300 to 55 by applying in priorities.

Table III: Examples

Input word = અધિકારીઓએ Case Marker = એ Number Marker = ઓ Fem. Gender marker = ી Entire Suffix =ીઓએ Stem = અધિકાર **Result:**Category = NNM.PL.	Input word = રમતો Case Marker = NULL Number Marker = ો Gender marker = NULL **Result1**: Category: NNF.PL. Verb Rule: તો Stem = રમ **Result2**: Category: VM.SPST.SG

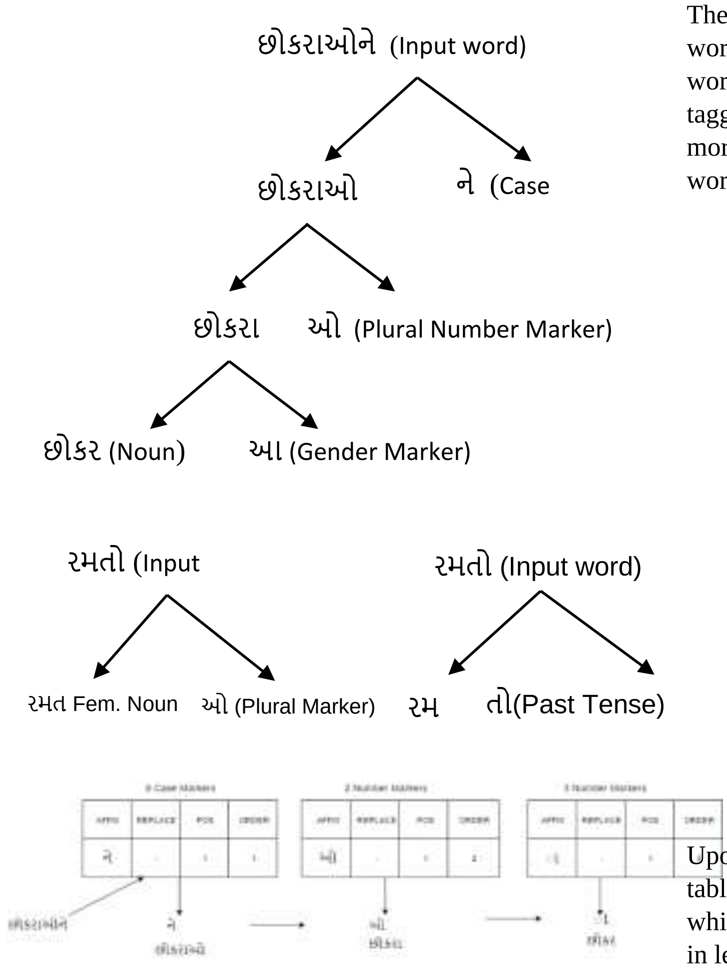

Fig 1: Working of Affix removal

IV. POS TAGGING RULES

Total 150 replacement rules were derived with help of native speakers. Some of the rules to replace affixes with their priorities are listed in below table4.

Table IV: Sample Suffix Replacement Rules

	Order	Loc	Affix	Replacement	Example
1	1	Right	નું	-	છોકરાનું ->છોકરા
2	1	Right	નો	-	છોકરાનો ->છોકરા
3	2	Right	ઓ	-	છોકરાઓ ->છોકરા
4	2	Right	◌ા	◌ુ◌ં	માટલા ->માટલું
5	3	Right	◌ા	◌ુ◌ં	છોકરા->છોકરો
6	4	Right	◌્યું	વું	રમ્યું ->રમવું

There were 19,968 unique words in corpus, and 10,437 words from those were assigned single tag. While 2,953 words were assigned more than one tag, in other words tagger could not come to single tag only based on morphological and lexical analysis. Frequency of top 10 words in corpus are given in following table.

Table V: Most frequent words

Word	Frequency
છે	4386
નયન	1065
જ	889
પણ	888
આ	857
એક	798
અને	773
એ	687
હતો	659
તે	630
નથી	567
હતી	522
હોય	515

Upon applying algorithm, we have got results as given in table 1. We have got 74,010 words tagged out of 82,334 which comes to 89.88%. Untagged words were either not in lexicon or rules were not formed to derive correct root.

Table VI: Results of Tagger

Total Sentences	12637
Total Words	82334
Tagged Words	74010
Tagged by tagger	89.88%
Single Tagged	10437
Untagged	6334
Ambiguously tagged	2953
Unique Words	19968

Rules were established to further resolve ambiguities based on contextual aspect of word, which we could not resolve by morphological analysis of word.

Tag disambiguation rules

Rules are formed to derive to single tag for each word to which more than one tags were assigned by tagger. Rules are categorized into Noun identification, Verb identification, and Adjective identification rules etc.

Rule 1 (verb): If current word is identified as either Noun or Verb and next word is Auxiliary verb then current word shall be main verb.

બાળકો રમતો રમે છે

NNN - NNF/VM- NNM/VM-VAUX (before applying rule)

NNN - NNF/VM –VM –VAUX (after applying rule)

નયન વિચારે છે.

NNP/NNM- NNM/VM-VAUX (before applying rule)

NNP/NNM-VM-VAUX (after applying rule)

Above rule found to be holding in 423 sentences out of 12637 sentence.

Rule 2 (noun): if current word is identified as either noun or verb and next word is main verb then current word shall be noun.

બાળકો રમતો રમે છે

NNN - NNF/VM- VM-VAUX (before applying rule)

NNN- NNF -VM-VAUX (after applying rule)

Rule 3 (noun): If current word is either Pronoun or Verb and next word is verb then current word shall be pronoun.

પ્રશ્ન મારે ઉકેલવાનો હતો.

NNM-PP/VM-VM-VAUX/VM (before applying rule)

NNM-PP-VM-VAUX/VM (after applying rule)

If current word is either Pronoun or Verb and next word is Auxiliary verb then current word shall be main verb otherwise pronoun

વર્ષાઋતુ આપણને પાણી આપે છે.

NNF - PP- NNN- PP/VM-VAUX (before applying rule)

NNF - PP- NNN- VM-VAUX (after applying rule)

Above rule found to be holding on 88 sentences out of 12638 sentences.

Rule 4 (noun): If current word is noun or proper noun (NNF/NNP or NNM/NNP) and next word is auxiliary verb, then current word will be noun. Consider following examples:

મારા પણ એક ગુરુ છે

PP-CC/NNN-JJ/NNM-NNM/NNP-VAUX (before apply rule)

PP- CC/NNN-JJ/NNM – NNM –VAUX (after applying rule)

Above rule was found to be holding at 80 places in 12637 sentence corpus.

Rule 5 (pronoun): If current word is personal pronoun and next word is pronoun then next word will be relative pronoun or possessive. Consider following examples:

હું તેનું સ્થાન ગ્રહણ કરવા માંગું છું.

PP-PP/JJ-NNN-NNN-NNM/VM-VM-VAUX (before applying rule)

PP-PPR- NNN-NNN-INF-VM-VAUX (after applying rule)

હું તારો પક્ષ નહી ખેંચુ.

PP-NNM/PP-NNM/NNP-NEG-VM (before applying rule)

PP-PP-NNM/NNP-NEG-VM (after applying rule)

આ અમારું ઘર છે.

PP-PP/JJ-NNN-VAUX (before applying rule)

PP- PPR-NNN-VAUX (after applying rule)

Above rule was found to be holding at 271 sentences in 12637 sentences.

Rule 6 (Adjective): If current word is either verb or adjective and next word is noun then current word shall be adjective. Consider following examples:

ખરો પ્રશ્ન મકાનનો હતો.

JJ/VM-NNM-NNN-VAUX (before applying rule)

JJ -NNM-NNN-VAUX (after applying rule)

Rule 7 (Infinitive): If current word is verb and next word is also verb then current word shall be infinitive. Consider following examples:

તેઓ ખરીદી કરવા ગયા હતા.

PP-NNF/VM-VM/INF-VM-VAUX (before applying rule)

PP-NNF/VM- INF-VM-VAUX (after applying rule)

Above rule found to be holing at 106 sentences out of 12837 sentences.

Rule 8 (Noun): If current word is either noun or verb and next word is infinitive or verb, then current word will be noun. This rule should also check if noun is not present in any other tag in simple sentence Consider following examples:

તેઓ ખરીદી કરવા ગયા હતા.

PP-NNF/VM- INF- VM-VAUX (before applying rule)

PP-NNF- INF- VM-VAUX (after applying rule)

Rule 9 (Postposition / Adverb / Adjective): If current word is either adverb, adjective or postposition and if previous word is noun, then current word will be adjective. If previous word is personal pronoun, then current word will be postposition.

તે પોતાના ભાઇ પાસે ઊભો હતો.

CC/PP/X-REFP-NNM-RB/PSP/JJ-VM-VAUX (before)

CC/PP/X-REFP-NNM-JJ-VM-VAUX (after applying rule)

તેની પાસે મોટર છે.

PP-RB/PSP/JJ-NNF-VAUX (before applying rule)

PP-PSP-NNF-VAUX (after applying rule)

Rules for unknown tags

Rule 1: (Postposition): If current word is unknown and next word is postposition then current word will be common noun, pronoun or proper noun

ઘરના છાપરા પરથી ફેંકવા લાગી.

NNN-?-PSP-VM-VM (before applying rule)

NNN-NNN-PSP-VM-VM (after applying rule)

Rule 2: (Adjective/Noun): If current word is unknown and next word is common noun and previous word is pronoun then current word will be noun or adjective.

આ કાળી પેન્સિલ છે.

PP-?-NNF-VAUX (before applying rule)

PP-JJ/NN-NNF-VM (after applying rule)

હું પુરી કોશિશ કરીશ.

PP-?-NNF-VM (before applying rule)

PP-JJ/NN-NNF-VM (after applying rule)

V. EVALUTION & RESULT

Evaluation of Part-Of-Speech (POS) tagger was done to improve the performance on different domains of news and short stories and small essays. The system was evaluated using lexicon size of 30050 words. The overall accuracy of achieved by our system is 82.52%. Various data sets were prepared from different sources. We have collected data set from various sources like Gujarati news websites, Gujarati short stories and essay books etc.

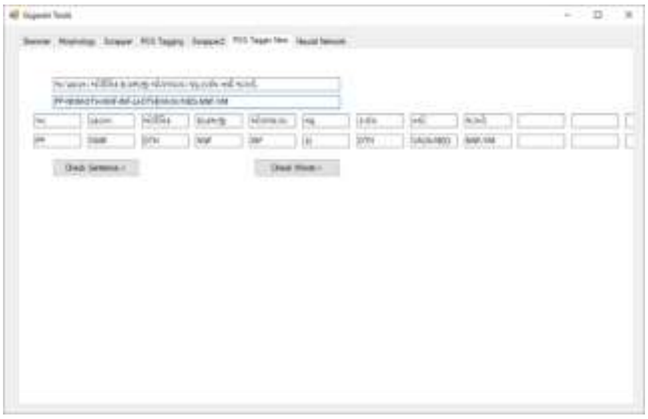

Fig 2: POS Tagger Screenshot

Following table shows statistics for each type of domain in corpus:

Table 9: Testing on datasets

Test Set	Source	Domain	No of words
Dataset1	Newspaper WebSites	News	48985
DataSet2	History Books	Short Stories	2
DataSet3	Books	Essays	7103

Total sentences of each length in corpus:

Table 10: Frequency distribution of sentence length

Length	No of words
1-3	2642
4-6	6794
7-8	3202

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

Recall = Number of correct answer given by system / Total no of words tagged by system.

Precision = Number of Correct answer by system / Total number of tags present

F-Measure = (β2 + 1) PR / β2 R + P

β is the weighting between precision and recall and typically β = 1.

Table 11: Precision, F-Score and Recall

Test Set	Precision	Recall	F-Measure
Dataset1	0.7815	0.8214	0.8009
DataSet2	0.7625	0.8155	0.7881
DataSet3	0.7921	0.8021	0.7970

VI. IMPORTANT ISSUES IN TAGGING

Major tagging issues in Gujarati is observed in Gerund (Verbal Noun), Participle (Verbal Adjective), and repetitive words.

Gerund (Verbal Nouns)

Verbal nouns are derived from verbs which are generally called gerunds. They generally has suffix – (nuM) which is assigned to verb that makes verbal noun which are sometimes also called infinitive verb. We can easily distinguish between gerund and infinitive by observing context if it is followed by post-positions.

e.g. મને તરવું છે (I want to swim) where તરવું is Verb infinitive but in the sentence તરવું એ સારી કસરત છે. (Swimming is good exercise) in this sentence તરવું is Verbal Noun or Gerund. At present we are tagging all Gerund and Verb Infinitive as main verb (VM).

In another example, જમવાની ઉતાવળ/NNF કરશો નહી. Here જમવાની is verbal noun also in ખાવા માટેનું ફળ, ખાવા is also verbal noun which can be distinguished using postposition.

Participle (Verbal Adjectives)

Sometimes verbs inflected and take place of adjectives called verbal adjectives or participle.

For example દોડતો છોકરો(running boy), દોડતો(running) is participle as it is actually used here as adjective to noun (boy) but it was previously tagged as verb. They are called verbal adjective or participle. They can also be inflected for gender, number or person also can take tense marker.

Repeated Words

For example in this sentence, તે રમતાં રમતાં સુઈ ગયો in this sentence second and third words are identified as main verb as there is no tag in tag set for repeated verbs.

Complex Word

Consider following example: અહીં અનાજ ઉત્પન્ન થાય છેwhere ઉત્પન્ન is adjective or noun? Also in sentence મને આ વસ્તુ પ્રાપ્ત થઈ, where it is required to be decided for પ્રાપ્ત as adjective or noun. So they should be treated as adjective.

VII. CONCLUSION AND FUTURE WORK

We have discussed Part of Speech tagger with hand crafted suffix replacement rules. First tokenization is performed then words are searched in the database and if not found then appropriate rules are applied. Sometimes when we apply rules then system may tag the words with wrong POS tags.

If a sentence containing more than half of words that are unknown, then system fails to tag them. The reason behind it is where system is to decide which rules should be applied first as word tagging resolution is based on context of words. So sometimes it becomes hard to determine tags when many tags are unknown. Accuracy of part of speech tagger can be increased by increasing the size of database. There are some chances more than one rule applicable at the same situation but outcome of that rule is different tag for same word, in this scenario system may fails, so this it is one where we can set either priority of rules to decide most effective tag for word.

REFERENCES

[1]. Brill, Eric. "A simple rule-based part of speech tagger." In Proceedings of the workshop on Speech and Natural Language, pp. 112-116. Association for Computational Linguistics, 1992.

[2]. AniketDalal, Kumar Nagraj, Uma Sawant and Sandeep Shelke. 2006. Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach, In Proceeding of NLPAI Machine Learning Competition, 2006.

[3]. Smriti Singh, Kuhoo Gupta, Manish Shrivastava, and Pushpak Bhattacharyya. "Morphological richness offsets resource demand-experiences in constructing a POS tagger for Hindi." In Proceedings of the COLING/ACL on Main conference poster sessions, pp. 779-786. Association for Computational Linguistics, 2006.

[4]. Agarwal Himanshu, Amni Anirudh. "Part of speech tagging and chunking with conditional random fields." In the Proceedings of NWAI workshop. 2006.

[5]. Chirag Patel and KarthikGali. Jan 2008. Part-of-Speech tagging for Gujarati Using Conditional Random Fields. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages 117-122, Hydrabad, India.

[6]. Manish Shrivastava and Pushpak Bhattacharyya. "Hindi pos tagger using naive stemming: Harnessing morphological information without extensive linguistic knowledge." In International Conference on NLP (ICON08), Pune, India. 2008.

[7]. Nisheeth Joshi, HemantDarbari, and ItiMathur. "HMM based POS tagger for Hindi." In Proceeding of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC-2013). 2013.

[8]. NavneetGarg, Vishal Goyal, SumanPreet. COLING 2012. Rule Based Hindi Part of Speech Tagger. Proceedings of COLING Dec 2012, p 163-174.

[9]. SharvariGovilkar, J. W. Bakal, and ShubhangiRathod. "Part of Speech Tagger for Marathi Language." International Journal of Computer Applications 119, no. 18 (2015).

[10]. AksharBharati, VineetChaitanya, Rajeev Sangal, and K. V. Ramakrishnamacharyulu. Natural language processing: a Paninian perspective. New Delhi: Prentice-Hall of India, 1995.

[11]. PVS, Avinesh, and G. Karthik. "Part-of-speech tagging and chunking using conditional random fields and transformation based learning." Shallow Parsing for South Asian Languages 21 (2007).

[12]. Manju, K., S. Soumya, and Sumam Mary Idicula. "Development of a POS tagger for Malayalam-an experience." In Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on, pp. 709-713. IEEE, 2009.

[13]. Dandapat, Sandipan, Sudeshna Sarkar, and AnupamBasu. "Automatic part-of-speech tagging for Bengali: An approach for morphologically rich languages in a poor resource scenario." In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 221-224. Association for Computational Linguistics, 2007.

[14]. M. Selvam, and A. M. Natarajan. "Improvement of rule based morphological analysis and POS Tagging in Tamil language via projection and induction techniques." International journal of computers 3, no. 4 (2009): 357-367.

[15]. V. Dhanalakshmi, M. Anand Kumar, S. Rajendran, and K. P. Soman. "POS tagger and chunker for Tamil language." In Proceedings of Tamil Internet Conference. 2009.

[16]. Jyoti Singh, Nisheeth Joshi, and ItiMathur. "Part of speech tagging of Marathi text using trigram method." arXiv preprint arXiv:1307.4299 (2013).

[17]. Dhanalakshmi, V., G. Shivapratap, and Rajendran S. SomanKp. "Tamil POS tagging using linear programming." (2009).

[18]. Bharati, Akshar, Rajeev Sangal, DiptiMisra Sharma, and Lakshmi Bai. "Anncorra: Annotating corpora guidelines for pos and chunk annotation for indian languages." LTRC-TR31 (2006)

[19]. Hardie, A. (2004). The Computational Analysis of Morphosyntactic Categories in Urdu. PhD thesis submitted to Lancaster University

[20]. Shetty, Saritha, and Savitha Shetty. "Text pre-processing and parts of speech tagging for Kannada language." Journal of Xi'an University of Architecture & Technology 12, no. II (2020): 1286-1291.

[21]. http://www.ldcil.org/standardsTextPOS.aspx

APPENDIX

Category		Label	Annotation convention	Examples
Top Level	Sub Type			
Noun		N	N	પુસ્તક,છોકરો,રાજા
	Common	NN	N_NN	પુસ્તક,ચોપડી,ચશ્માં
	Proper	NNP	N_NNP	મોહન,રવિ,રશ્મિ
	Nloc	NST	N_NST	ઉપર,નીચે,આગળ,પાછળ
Pronoun		PR	PR	
	Personal	PRP	PR_PRP	હું,તું,અમે
	Reflexive	PRF	PR_PRF	આપણે,સ્વયં,પોતે
	Relative	PRL	PR_PRL	જે,જેણે,જયારે,જ્યાં
	Reciprocal	PRC	PR_PRC	પરસ્પર,આપણા જેવું
	Wh-word	PRQ	PR_PRQ	ક્યાં,ક્યારે,કેવીરીતે
	Indefinite	PRI	PR_PRI	કોઈ,કાંઈ,કંઈ,કંઈક
Demonstrative		DM	DM	ત્યાં,જે,અહીંયા
	Deictic	DMD	DM_DMD	ત્યાં,અહીંયા
	Relative	DMR	DM_DMR	જેણે,જે
	Wh-word	DMQ	DM_DMQ	કોણ,કોણે

	Indefinite	DMI	DM_ DMI	
Verb		V	V	
	Main	VM	V_VM	રમ, જમ, હસ,..
	Auxiliary	VAUX	V_ VAUX	છું,છીએ,છે..
	Model Auxiliary	VMAUX	V_VMAUX	
Adjective		JJ	JJ	સુંદર, સારું, ખરાબ
Adverb		RB	RB	જલ્દી, ફટાફટ,..
Postposition		PSP	PSP	એને, એણે, એનાથી,એમાં
Conjunction		CC	CC	જો,તો, તથા, કારણ કે
	Co-ordinator	CCD	CC_ CCD	અને, પરંતુ, બદલે
	Subordinator	CCS	CC_ CCS	
Particles		RP	RP	
	Default	RPD	RP_ RPD	
	Interjection	INJ	RP_ INJ	
	Intensifier	INTF	RP_ INTF	બહુજ, વધારે
	Negation	NEG	RP_ NEG	ના,વ

				ગર
Quantifiers		QT	QT	
	General	QTF	QT_ QTF	થોડુંક, વધારે, કંઈક
	Cardinals	QTC	QT_ QTC	એક, બે, ત્રણ
	Ordinals	QTO	QT_ QTO	પહેલો, બીજો
Residuals		RD	RD	
	Foreign word	RDF	RD_ RDF	
	Symbol	SYM	RD_ SYM	$,&,*,(,)
	Punctuation	PUNC	RD_ PUNC	., ;, :, ?, !,
	Unknown	UNK	RD_ UNK	

AUTHOR PROFILES

Dr. Apurva A. Desai, completed his graduation and post graduation from Veer Narmad South Gujarat University. He earned his Ph.D. in the year 1997 in the field of Operation Research and Computer Science. He is a Dean of faculty of Computer Science and Information Technology and Chairman Board of Studies. He is and Editor in Chief of VNSGU Journal of Science and Technology and also serving as a member of Editorial board for some of the national and international journals. He has more than 50 research papers and four books to his credit.

Dr. Utkarsh N. Kapadia, received B.E. and M.C.A degrees from Veer Narmad South Gujarat University. He has worked as System Engineer in TCS. He has earned his PhD degree in Natural Language Grammar analysis for Gujarati. He has more than 15 years of experience in working in IT industry at various levels. He has been working as researcher in the area of Natural Language Processing.