ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol4, Issue 6, June 2017

Part of Speech Tagging for Konkani Corpus

[1] Meghana Mahesh Pai Kane
Assistant Professor, Dept CSE, AITD College, Goa, India

***Abstract*— The wide spectrum of languages are been used for communication around the world , utilization of world wide web for searching information requires computational linguistics because majority of the search engines uses bag of words that causes problem in extracting of the information due to use of Multi words . This has made to think beyond the boundaries about what kinds of query a human can submit and also its interpretation in forms of its annotation could be used to obtain good result. The essential st ep in the Natural Language Processing resides in obtaining the grammatical information of the words used in the input as per it appearance in the text .POS taggers for several other Indian languages have been developed but assumption of unavailability of the POS tagger for the Konkani language aims at developing the same. Further POS tagging to do manually is much tougher job due to huge content of data. This paper aims at part of speech tagging for Konkani corpus.**

***Index Terms*— Konkani corpus,Multi words, Natural Language Processing, POS tagging.**

I. INTRODUCTION

In linguistics, Part-Of-Speech Tagging (POST or Pos Tagging) is called grammatical tagging or word-category, Disambiguation is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context such as its relationship with adjacent and related words in a phrase, sentence, or paragraph. This paper aims at developing a Konkani POS tagger that can not only tag a sentence, but a corpus or a file in text editor format containing the data as per contextual use of words in it. POS-tagging algorithms fall into distinctive groups rule-based and stochastic/probabilistic approach.

II. RELATED WORK

Review Stag Many words have got ambiguities associated with respect to its part of speech [3]. If a word "bank" if taken into consideration it could be a verb or a noun. Part of speech for a word helps for analyzing text at a higher-level, such as for above example the word “Bank ” is recognized a noun or a verb. Konkani is a language spoken in State of Goa, due to unavailability of a search system in Konkani language the process of Konkani Part-of-speech tagger could be utilized to invent a Konkani language search engine.

Konkani Part-of-speech tagger is process to identify and analyze lexical categories for existing Konkani words which are been manually tagged based on context [3]. The categories for Tagging of each word taken through Konkani BIS Tag sets [16] such as Noun, Pronoun, Demonstrative , Verb ,Adjective, Adverb, Postposition , Conjunction, Particles, Quantifier and Residuals.

The primary goal resides in obtaining the grammatical information of the words used in the input as per it appearance in the text. This forms an essential step in the Natural Language Processing .POS taggers for several other Indian languages have been developed but assumption of unavailability of the POS tagger for the Konkani language aims at developing the same. An architecture is proposed for Konkani Part-of-speech tagger at next step[11][12].
Then we discuss the results obtained. Finally, A conclusion is notes with its future work.

III. SYSTEM DESCRIPTION

The Part-of-speech tagger system which is been designed is useful to linguists, decrease use of human work for manual tagging each words for documents and it can be used for introducing Konkani search engine.

The Main objective includes:

• Accepting a sentence in Konkani language and tagging each word therein with the most appropriate and most likely POS tag depending upon the context in which the word occurs in the sentence.
• This will be done by making a probabilistic comparison in the output and also with dictionary of words which already have POS tags assigned to it.

The Sub goals of the system include:

- Automatic tagging of Konkani text, with suitable POS tag having acceptable accuracy.

Performance of the tagger depends on:

- The amount of training data
- The tag set
- The difference between training data and test data
- The occurrence of unknown words in the test data

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol4, Issue 6, June 2017

A. User Requirements-

The Part Of Speech Tagger system scan the documents of a Konkani corpus, then extract the sentences from documents and words present into following sentence Konkani corpus. And then finally display each word with its unique Tag such as declared for Konkani BIS Tag sets such as Noun, Pronoun, Demonstrative, Verb ,Adjective, Adverb, Postposition , Conjunction, Particles, Quantifier and Residuals.

B. POST System Development Steps –

The following details brief overview of the system
Activity:

- Scanning of the Konkani corpus.
- Extract the sentences from the Konkani corpus and words are been identified as per delimiter.
- Part of speech tagger is been build .
- Forming Highest frequency rules for the identification of Part of Speech Tag for each unique words, numbers, Punctuations and when No tag is provide for a particular word.
- Implement the system for marking and providing tags to particular words with its various types POST categories.
- Analysis of output POST data, such as each word can be provided with only one highest frequency tag.
- Develop a Graphical User Interface.
- Test the system and then it evaluation

C. Description of Modules in detail-

C.1 Konkani File Read

Browsing of Konkani documents, reading the content in Unicode format is done through this module. It counts total number of documents selected to be processed, extracts overall total lines found for every document , total number f Konkani sentences, unique words for following Konkani corpus .

Input: Konkani (Unicode) text documents

Processing: This module reads number of files selected for *processing* , number of lines, sentences, unique words and shows the path were file is been browsed .

Output: Displays browsing path and total unique words.

C.2 Tokenization

C.2.1 Extract Sentences

This module extracts Konkani (Unicode) corpus into the sentences.

Input: Konkani Corpus

Processing: Splits Konkani corpus into the sentences according to its delimiter.

Output: Displays the Konkani sentences.

C.2.2 Word Tokenization

This module extracts Konkani (Unicode) and splits the sentence into the unique word according to the space delimiter.

Input: Extracted Konkani sentence

Processing: Split each sentences into unique words according to the space delimiter.

Output: Display the unique Konkani words.

C.3 Highest frequency rules-

This module helps to extract each unique word with its tag based on highest frequency because ambiguity may be observed were one word may have two or more tags. Hence highest frequency rules provide an unique tag for unique word.

Input: Extract Konkani corpus which is been already manually tagged.

Processing: Processes the input data to find two or more tags for each unique word and sort the words based on its highest frequency and appropriate tag.

Output: Displays the sorted highest frequency file.

C.4 Tagging

This module tags each Konkani word with their related tags like Noun, Pronoun, Demonstrative, Verb ,Adjective, Adverb, Postposition , Conjunction, Particles, Quantifier and Residuals. If Konkani word does not come into any of categories of POST then it is been by default tagged as "No_Tag".

Input: Extracted Konkani sentence with its unique words.

Processing: Tag each word of input sentence.

Output: Display the tag output Konkani corpus.

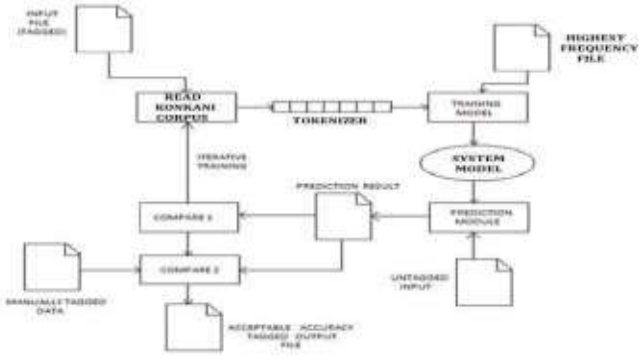

Fig. 1. Architecture of POST for Konkani Corpus

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol4, Issue 6, June 2017

IV. EVALUATION AND EXPERIMENTS

The efficiency and validity of POST system is judged by parameters of user requirements and throughput. For evaluation both software testing methods like black box Testing and white box testing are used.

Test Cases:

MODULE 1 (KONKANI CORPUS READ)

Sr. No	Test Case ID	Input	Description	Actual Result	Exp. Result	Pass/ Fail
1.	1.1	Konkani(Unicode) Corpus is Browsed txt file	Click on "Read" button	Read corpus	corpus	Pass
	1.2	Konkani (not in Unicode) Corpus is Browsed txt file	Click on "Read" button	Should not Read corpus	Char not recogniz ed	Fail
	1.3	Konkani (Unicode) corpus not entered	Click on "Read" button	Msg "no data found"	Msg "no data found"	Fail

MODULE 2 (TOKENIZER: SENTENCE TOKENIZER)

Sr. No	Test Case ID	Input	Description	Actual Result	Exp. Result	Pass/ Fail
2.	2.1	Konkani(Unicode) Corpus is Browsed And delimiter selected	Click on "Sentence Tokenize" button	Read corpus	Extract corpus	Pass
	2.2	Konkani (not in Unicode) Corpus is Browsed	Click on "Sentence Tokenize" button	Should not Extract corpus	Char not recogniz ed	Fail
	2.3	Konkani (Unicode) corpus not entered	Click on "Sentence Tokenize" button	Msg "no data found"	Msg "no data found"	Fail

MODULE 3 (TOKENIZER: WORD TOKENIZER)

Sr. No	Test Case ID	Input	Description	Actual Result	Exp. Result	Pass/ Fail
3.	3.1	Konkani(Unicode) Corpus	Click on "Word Split" button	Read corpus	Tokenize corpus	Pass
	3.2	Konkani (not in Unicode) Corpus is Browsed	Click on "Word Split" button	Should not Extract corpus	Char not recogniz ed	Fail
	3.3	Konkani (Unicode) corpus not entered	Click on "Word Split" button	Msg "no data found"	Msg "no data found"	Fail

V. RESULTS AND DISCUSSIONS

The overall result of the Konkani Part of Speech Tagger is Discussed below:

A. Konkani File Read
This module Read Konkani (Unicode) corpus and count total number of lines, words and sentences in selected Konkani corpus and displays the whole corpus with path and file name if user browse the file.

Input:
Konkani Corpus:
उदक चड पियेवचें. तोंड सुकतकच बॅक्टेररया नेटान हल्लो करतात.

हाका लागून स्वासाःतल्यान घाणी येविाक लागतात.चड उदक पियेतकच जेवणाचे बारीक बारीक कण पनवळ जातात, ते भायर लाळू य तयार जाता. चचगम चाबडायल्यार लाळ तयार जाता.

Output:
Display Konkani Corpus:
उदक चड पियेवचें. तोंड सुकतकच बॅक्टेररया नेटान हल्लो करतात.

हाका लागून स्वासाःतल्यान घाणी येविाक लागतात.चड उदक पियेतकच जेवणाचे बारीक बारीक कण पनवळ जातात, ते भायर लाळू य तयार जाता. चचगम चाबडायल्यार लाळ तयार जाता.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol4, Issue 6, June 2017

Number of Lines : 4
Number of words : 40
Number of sentences : 5
FilePath:C:\DocumentsandSettings\ Desktop\KonkaniPOSTagger\konkanicorpus\health_1.txt

B.2 Word Tokenization
Input:
Konkani sentence
जेवतकच दर खेिे उदकान तोंड धुवचें.

Output:
Display the Splitted Words
1. जेवतकच
2. दर
3. खेिे
4. उदकान
5. तोंड
6. धुवचें
7. .

C. Highest frequency rules

Number of Lines : 13903
. / RD_PUNC VALUE=1002
, / RD_PUNC VALUE=441
आनी /CC_CCD VALUE=346
वा / CC_CCD VALUE=159
उणी/QT_QTF VALUE=95
घेवचो/V_VM_VNF VALUE=90
लागून/PSP VALUE=85
जाता/V_VM_VF VALUE=79
)/RD_PUNC VALUE=76 (/RD_PUNC VALUE=76
येता/V_VM_VF VALUE=72 चरबी/N_NN VALUE=63
घेवचो/V_VM_VF VALUE=62 उणी/QT_QTF

D. Tagging

This module tags each of the Konkani words using Highest frequency rules and their corresponding Tag sets such as Noun, Pronoun, Demonstrative, Verb ,Adjective, Adverb, Postposition , Conjunction, Particles, Quantifier and Residuals.
If Konkani word does not have a POS tag category then tag by default "NO_TAG" is been assigned.

Input
Konkani sentence
उणी चरबी आपिल्लो आहार घेवचो.

Output उणी/QT_QTF चरबी/N_NN आपिल्लो/V_VM_VF आहार / NO_TAG घेवचो/V_VM_VNF
./ RD_PUNC

Two domains were taken Health and Tourism ,were from Health domain there were 20 files used for training which contained each 1,000 Konkani sentences and that were manually tagged ,it was tested by training for one or more files as shown into the Tables ,similarly same was done for Tourism domain, basically considering both domains 40,000 Konkani sentences were used for training of the data to obtain highest frequency file.For testing 10 files containing 10,000 Konkani sentences were used to obtain accuracy. Various combination of number of files was used to check the change into accuracy of data. Finally a graph was obtained to check accuracy increase as number of files for training varies.

TABLE I. TESTING FOR ONE SINGLE FILE FOR HEALTH DOMAIN

Domain from which files are used	Feature Vector Selected	Number of sentences used for Training	Number of sentences used for Testing	% Accuracy
Health	For every word with its appropriate tag	1,000	1,000	28%
Health	For every word with its appropriate tag	5,000	1,000	36%
Health	For every word with its appropriate tag	10,000	1,000	44%
Health	For every word with its appropriate tag	15,000	1,000	66%
Health	For every word with its appropriate tag	20,000	1,000	89%

TABLE II. TESTING FOR ALL FIVE FILE HEALTH DOMAIN

Domain from which files are used	Feature Vector Selected	Number of sentences used for Training	Number of sentences used for Testing	% Accuracy
Health	For every word with its appropriate tag	1,000	5,000	17%
Health	For every word with its appropriate tag	5,000	5,000	23%
Health	For every word with its appropriate tag	10,000	5,000	39%
Health	For every word with its appropriate tag	15,000	5,000	52%
Health	For every word with its appropriate tag	20,000	5,000	65%

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol4, Issue 6, June 2017

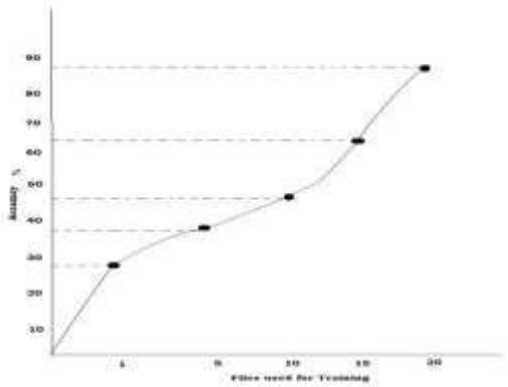

Fig. 2 Graph obtained to check accuracy of Health domain

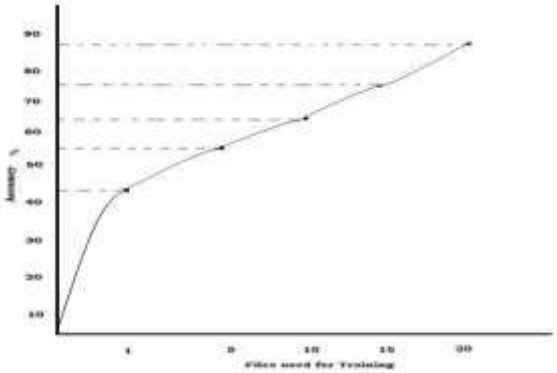

Fig. 3 Graph obtained to check accuracy of Tourism domain

TABLE III. TESTING FOR ONE SINGLE FILE FOR TOURISM DOMAIN

Domain from which files are used	Feature Vector Selected	Number of sentences used for Training	Number of sentences used for Testing	% Accuracy
Tourism	For every word with its appropriate tag	1,000	1,000	42%
Tourism	For every word with its appropriate tag	5,000	1,000	55%
Tourism	For every word with its appropriate tag	10,000	1,000	61%
Tourism	For every word with its appropriate tag	15,000	1,000	77%
Tourism	For every word with its appropriate tag	20,000	1,000	87%

TABLE IV. TESTING FOR ALL FIVE FILE TOURISM DOMAIN

Domain from which files are used	Feature Vector Selected	Number of sentences used for Training	Number of sentences used for Testing	% Accuracy
Tourism	For every word with its appropriate tag	1,000	5,000	18%
Tourism	For every word with its appropriate tag	5,000	5,000	24%
Tourism	For every word with its appropriate tag	10,000	5,000	35%
Tourism	For every word with its appropriate tag	15,000	5,000	47%
Tourism	For every word with its appropriate tag	20,000	5,000	58%

VI. CONCLUSION

The Part Of Speech Tagger System can read the Konkani corpus, Extract the Sentences and tokenize the words. Manually tagged data is been processed to obtain Highest frequency file, Later when untagged data is been provided. It gives the tagged data output of the given Konkani Corpus. The Graphical user interface is user friendly and can be understood easily Novice users shall have no problem into understanding the GUI and they will be comfortable to work on this system.

Two domains were taken Health and Tourism, were it was tested by training for one or more file and testing for one file where accuracy for Health was 89 % and Tourism was 87 % observer.

VII. FUTURE WORK

We are looking for providing a Konkani language Search Engine so that POST facilities could be used in it , also to obtain more accurate tagged data, taking condition such as prediction through previous word tag , next word tagged and also using MAXENT software enhance the Tagger in the future. By increasing the manual data for training of system, we can expect an increase in the accuracy of POST. More research work can be carried for identifying Part of Speech Tag to decrease manual Human work.

REFERENCES

[1] Ed. T. Jaynes, "Information Theory", dated 1957
http://homepages.inf.ed.ac.uk/lzhang10/maxent.html
[2] Abney, "Stochastic Attribute-Value Grammars", dated

1997 http://citeseer.ist.psu.edu/490897.html
[3] Christopher D. Manning, Hinrich Schutze, "Foundations of
statistical natural language processing".
[4] Experiences in Building the Konkani WordNet Using the Expansion Approach http://www.cfilt.iitb.ac.in/gwc2010/pdfs/54_Konkani_WordNet__Walawalikar.pdf
[5] Daniel Jurafsky and James H.Martin, "Speech and Language Processing"Adam L. Berger, Stephen A. Della Pietra and Vincent J. Della Pietra, "A maximum entropy approach to natural language processing"
[6] Stochastic Algorithm http://citeseer.ist.psu.edu/rosenfeld94adaptive.html
[7] Morphological Analyzer http://Morphadorner.northwestern.edu/morphadorner/postagger/example
[8] "A Part Of Speech Tagger For Indian Languages" http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pd f
[9] Hindi POS Tagging and Chunking Itrc.ac.in/nlpai_contest06/papers/msrindia.pdf
[10] Sanskrit Tagger, a stochastic lexical and pos tagger for Sanskrit hal.inria.fr/inria-00203467/fr/
[11] A maximum entropy model for Part of Speech tagging www.Idc.upenn.edu/acI/W/W96/W96-0213.pdf
[12] Natural Language Processing cnlp.syr.edu/publications/03NLP.LIS.Encyclopedia.pdf
[13] BIS Annotation Standards With Reference to Konkani Language – Goa university
[14] Multiword Expressions Dataset for Indian Languageshttps://www.cse.iitb.ac.in/~pb/papers/lrec16-mw-resource.pdf