International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 11 | Nov 2020 www.irjet.net p-ISSN: 2395-0072

Survey on Part of Speech Tagger for Hindi Language using Rule based Approach

[1]Priyanka Lohe, Vikas Pandey[2]

[1]M.Tech Scholar, Dept. of Information Technology, Bhilai Institute of Technology, Durg
[2]Associate Professor, Dept. of Information Technology, Bhilai Institute of Technology, Durg

---

Abstract - *In this paper, we are discussing a Rule Based Part of Speech Tagger for Hindi Language. POS Tagger is an important tool used to create language translator and data extraction as an input it assigns Part of speech to the words in a sentence. As indicated by its quality in the given content POS tagger tag every word in a text. Tagging is the first step in the development of any Natural Language Processing application (NLP). It is the process of increasing a token in a sentence as a specific POS tag or lexical having a place with a specific class (noun, adverb, pronoun etc) based on its context in the sentence, its morphological information and its definitions. (POS) tagger is the basic building block for different NLP tools.*

Key Words*: *Part of Speech, Natural Language Processing, NLP.

1. INTRODUCTION

Part of Speech is a significant use of Natural language processing .Natural Language Processing is a rapidly growing technology at present and with the help of imposing some queries and keywords it fetches information from collection of huge amounts of data. Part of Speech tagging is fundamentally a training set of doling out language explicit punctuation labels and a grammatical feature like thing, action word, relational word, pronoun, verb modifier, descriptor or other lexical class producer to each word in a sentence of language-explicit information text, as per word's appearance in the content [1].

Natural Language Processing is a field of computer science, artificial intelligence and linguistics it has interactions among PCs and human language. It is a process of extracting information from natural language. The cycle of POS labeling comprises of these process Tokenization, Assign a tag to tokenized word and search for Ambiguous word. Text Tagging is a complex task as we get words which have different tag categories many times as they are used in different context. This phenomenon is termed as lexical ambiguity [3]. Ambiguity occurs in most words in text associated with it in terms of their part of speech. For example, "bank" can be treated as a noun or a verb. Rule based approach use linguistic rules to relegate the right labels to the words in the sentence or document. The rule based POS tagging model requires a ton of physically composed sets and uses pertinent information to allot POS labels to words and also cannot predict appropriate tags. C-DAC and TDIL IIT Bombay played important role on research venture "POS tagger for Hindi". In Tagger some normal labels: [V] Verbs, [N] Nouns, [PR] Pronouns, [JJ] Adjectives, [RB] Adverbs, [PP] Postpositions, [PL] Participles, [QT] Quantifiers, [RP] Particles, [PU] Punctuations.

2. LITERATURE SURVEY

There are a number for approaches for Part of Speech Tagger in various languages for Hindi Language additionally there exists various execution of POS taggers. AnnCorra, abridged for "Annotated Corpora," is a task of Lexical Resources for Indian Languages (LERIL), is a synergistic exertion of a couple of gatherings. They developed a framework utilizing factual system, which gives syntactic and semantic data. [4][5].

Singh and Tripathi had proposed a "Hindi language text search: a literature review ".The survey centers around the serious issues of Hindi content looking over the web. Furthermore, they presume that various issues despite everything exist in the zone of interpretation including the Hindi language [6].

Sharma and Motlani had proposed a "POS Tagging For Code-Mixed Indian Social Media Text : Systems from IIIT-H for ICON NLP Tools Contest". Their framework uses to train the CRF sequence labeling algorithm and had three language sets, specifically Hindi-English (Hi-En), Bengali-English (Bn-En) and Tamil-English (Ta-En).Their framework for Hi-En played out the best with 80.68% precision [7].

Modi and Jain proposed "Part-of-Speech Tagging of Hindi Corpus Using Rule-Based Method". Their aim is to increase automaticity and maintain high precision. The system achieves 91.84 % of average precision and 85.45 % of average accuracy. [4].

Rajesh Kumar Sayar and Singh Shekhawat "PARTS OF SPEECH TAGGING FOR HINDI LANGUAGES USING HMM" 2018 paper describes the Part of Speech (POS) tagging for Indian Languages "HINDI" .They prefer Hindi POS tagging using Hindi WordNet dictionary and HMM. HMM approaches concerned for POS tagging of sentences written in Hindi languages are discussed in their paper The performance analysis has been carried out for Precision, Recall and F1-Measure. We obtained 93.17 % precision,96.46 % Recall and 90.13 % F-measure[8].

Mohnot, Bansal.Singh and Kumar proposed "Hybrid approach for Part of Speech Tagger for Hindi language". Their system evaluated corpus of 80,000 words. System achieved an accuracy of 89.9% [1].

Shachi Mall et al., 2011 designed a system using a Rule based approach. The module reads the Hindi corpus and split the sentence into words as indicated by the delimiter. The system finds the words in the database and assigns the appropriate tag to the words[9].

3. POS TAGGER

POS tagging is utilized as an early phase of text analysis in numerous applications, for example, subcategory procurement, text to speech sysnthesis and alignment of parallel corpora. POS tagging is a necessary pre-module and building block for various NLP tasks like Machine translation, Natural language text processing and summarization, User interfaces, Multilingual and cross language information retrieval, Speech recognition, Artificial intelligence, Parsing, Expert system and so on [11].POS tagging is a basic step for language processing and can work as the first phase in other language processing tasks. The work on Part-of-Speech (POS) tagging for natural language tagging has begun in the early 1960s. For Indian languages researcher, it's difficult to write linguistic rules using rule based approaches because of morphological richness [10].

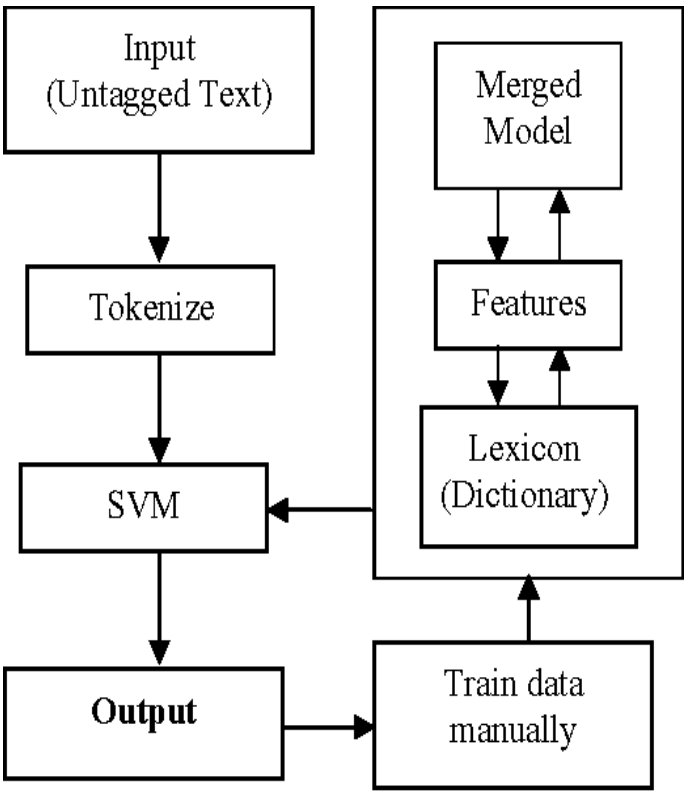

Figure 1: Architecture for POS tagging

4. DETAILS OF IDENTIFIED MODULES

A) Tokenizer

Tokenizer breaks the sentence into words, punctuation marks and other symbols also called tokens. Tokens are separated by white-space characters, line breaks or punctuation markers. Special tokens are handled separately to avoid wrong tokenizations. In computer science, lexical analysis, lexing or tokenization is the process of converting a sequence of characters (such as in a computer program or web page) into a sequence of tokens (strings with an assigned and thus identified meaning).

B) Tagging

The tagging module assigns tags to tokens and also search for ambiguous words and according to their type assign some special symbols to them. Tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph.

C) Ambiguity Resolution

This is also called disambiguation. Disambiguation is based on information about word such as the probability of the word. Disambiguation is also based on related information or word/tag sequences. For example, the model might prefer noun analyses over verb analyses if the preceding word is a preposition or article. Disambiguation is the most difficult problem in tagging. The ambiguity which is identified in the tagging module is resolved using the Marathi grammar rules.

D) WordNet

The main relation among words in WordNet is synonymy. WordNet is an electronic database which contains parts of speech of all the words which are stored in it. It is trained from the corpus for higher performance and efficiency. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The majority of the WordNet's relations connect words from the same part of speech (POS). Thus, WordNet really consists of four sub-nets, one each for nouns, verbs, adjectives and adverbs, with few cross-POS pointers. Cross-POS relations include the "morph semantic" links that hold among semantically similar words sharing a stem with the same meaning [10].

E) Corpus

For correct POS tagging, training the tagger well is very important, which requires the use of well annotated corpora. Annotation of corpora can be done at various levels which

include POS, phrase or clause level, dependency level etc. Corpus linguistics is the study of language as expressed in samples (corpora) of "real world" text. Corpus is a large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based. The plural form of corpus is corpora. Some popular corpora are British National Corpus (BNC), COBUILD/Birmingham Corpus, IBM/Lancaster Spoken English Corpus.

F) Tagset

Apart from corpora, a well-chosen tagset is also important. The language tagset represents parts of speech and consist on syntactic classes. According to contextual and morphological structure, natural languages are different from each other [6].In the top level the following categories are identified as universal categories for all ILs and hence these are obligatory for any tagset. Some common tags: [N] Nouns [V] Verbs, [PR] Pronouns, [JJ] Adjectives, [RB] Adverbs, [PP] Postpositions, [PL] Participles, [QT] Quantifiers, [RP] Particles, [PU] Punctuations.

5. ALGORITHM FOR TAGGING AND SPLITTING

When a user gives the input path of the corpus, the Hindi sentences are tokenised based on the technique of finding delimiter. In case of Hindi sentences, Purnviram (‗|') is the delimiter. The final untagged words of split sentences are stored in a separate file. The pseudo code for splitting and tagging id shown below:

Input Hindi text (Unicode)

Read Hindi Text T source

T Temp=T source text

Reg X Parse sentence parse [ ] =Split sentence ( )

For i = 0 to Regular sentence Parse length-l

a (i) = Reg Sentence Parse (i)

```
Sentence = at length
Words [ ] =split words ( )
Open file F .txt
 Ib I [ ]=a I [ ]
for i = 0 to Ibl-l
Sentence id = i+ 1
for j = 0 to word count -1
 word id = (j+l)
word [ ]= words [j]
next
write word [ ] to F.txt
display F.txt
```

S is the user input the source text and i is the length of the sentence.

6. RULE BASED TECHNIQUE

The rule based POS tagging approach that uses a set of hand written rules. Rule base taggers depend on word list or lexicon or dictionary to assign appropriate tag to each word. The tagger divided into two stages. First, it search words in dictionary and second, it assigns a tag by removing disambiguity of words using linguistic features of word [13]. On the basis of level rule divided as lexical rules act in a word level, each sentence splits into small words called lexeme or token And, the context sensitive rules act in a sentence level, to check the grammar for the sentence [12]. The transformation based approach is similar to the rule based approach in the sense that it depends on a set of rules for tagging. The transformation based approaches use a pre-defined set of handcrafted rules as well as automatically induced rules that are generated during training [14]. The main drawback of rule based system is that it fails when the text is not present in lexicon. Therefore the rule based system cannot predict the appropriate tags.

Rules are applied to identify different Tags

1. Noun Identification Rules

Rule 1: Word – Adjective -> Next word – Noun

Rule 2: Word – Relative pronoun -> Next word – Noun

Rule 3: Word – Reflexive pronoun -> Next word – Noun

Rule 4: Word – Personal pronoun -> Next word – Noun

Rule 5: Current word – Post position -> Previous word – Noun

Rule 6: Current word – Verb -> Previous word – Noun

Rule 7: Word – Noun -> Next or Previous word – Noun

1. 2. Demonstrative Identification Rules

Rule 1: Word – Pronoun, Next word – Pronoun -> First word – Demonstrative

Rule 2: Current word – Pronoun, Next word – Noun -> Current word – Demonstrative

2. 3.Proper Noun Identification Rules
3.
4. Rule 1: Current word – Not tagged, Next word – Proper Noun -> Current word – Proper Noun

7. CONCLUSION

In this paper we have presented a POS Tagging process using Rule Based Approach. Rule-based tagging assigns a word all possible tags and the uses context rules to disambiguate. So our future work is to create an automatic tagger in Hindi Language using Rule based Approach to find accuracy of tagger.

REFERENCES

1) Kanak Mohnot, Neha Bansal, Shashi Pal Singh, Ajai Kumar . "Hybrid approach for Part of Speech Tagger for Hindi language", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 4, Issue 1, February 2014,pp.25-30.

2) Navneet Garg, Vishal Goyal, Suman Preet . "Rule Based Hindi Part of Speech Tagger", Proceedings of COLING 2012: Demonstration Papers , pp .163-174.

3) Nisheeth Josh, Hemant Darbari and Iti Mathur . "HMM BASED POS TAGGER FOR HINDI", Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA - 2013 © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3639, pp. 341–349.

4) Deepa Modi and Neeta Nain. "Part-of-Speech Tagging of Hindi Corpus Using Rule-Based Method" , © Springer India 2016 N. Afzalpulkar et al. (eds.), Proceedings of the International Conference on Recent Cognizance in Wireless Communication & Image Processing, DOI 10.1007/978-81-322-2638-3_28 pp.241-247

5) Bharati, A., Sharma, D.M., and Sangal, R.: AnnCorra: An Introduction (Vol. 14), Technical Report no: TR-LTRC (2001).

6) Pratibha Singh and Aditya Tripathi, "Hindi language text search: a literature review", Annals of Library and Information Studies Vol. 64, March 2017, pp.37-43.

7) Arnav Sharma and Raveesh Motlani "POS Tagging For Code-Mixed Indian Social Media Text : Systems from IIIT-H for ICON NLP Tools Contest", Arnav Sharma and Raveesh Motlani , Systems from IIIT-H for ICON NLP Tools Contest 2016.

8) Rajesh Kumar Sayar and Singh Shekhawat "PARTS OF SPEECH TAGGING FOR HINDI LANGUAGES USING HMM", International Journal of Scientific Research Volume-7 | Issue-4 | April-2018 | ISSN No 2277 – 8179 pp.42-44.

9) Shachi Mall and Umesh Chandra Jaiswal. (2011). Hindi Part of Speech Tagging and Translation, In the proceedings of Int. J. Tech. 2011, Vol. 1: Issue 1, pp. 29-32.

10) Vijeta Khicha and Mantosh Manna "Part-of-Speech Tagging of Hindi Language Using Hybrid Approach" International Journal of Engineering Technology Science and Research Volume 4, Issue 8 August 2017 ISSN 2394 – 3386 pp. 737-741.

11) Shubhangi Rathod and Sharvari Govilkar "Survey of various POS tagging techniques for Indian regional languages", International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, pp.2525-2529.

12) Namrata Tapaswi Suresh Jain, "Treebank Based Deep Grammar Acquisition and Part-Of-Speech Tagging for Sanskrit Sentences" Software Engineering (CONSEG), 2012 CSI Sixth International Conference on.

13) Javed Ahmed MAHAR Ghulam Qadir MEMON, "Rule Based Part of Speech Tagging of Sindhi Language" 2010 proceeding of International Conference on Signal Acquisition and Processing.

14) Fahim Muhammad Hasan "Comparison Of Different Pos Tagging Techniques For Some South Asian Languages"