International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
August 2015, PP 63-68
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
www.arcjournals.org

Hybrid Approach Based POS Tagger for Hindi Language

Pravesh Kumar Dwivedi[1]
Ph.D. Research Scholar
Center for Technology Studies,
School of Language
MGIHU, Wardha (Maharashtra)
dpraveshkumar@gmail.com

Pritendra Kumar Malakar[2]
Ph.D. Research Scholar
Center for Technology Studies,
School of Language
MGIHU, Wardha (Maharashtra)
pritmalakar@gmail.com

Abstract: *In this research paper, we present an implementation of newer Hindi POS tagger based on Hybrid approach (combination of Rule-based and Probability based) uses some linguistic resources, several rules and database with some predefine list of possible prefixes, suffixes and other required and related data for the Hindi language. This process is regarded as a simplified form of morphological analysis. Morphologically analysis gives the grammatical information of a word on the bases of word structure. This program only deal with assigning a POS and tag to given surface from word.*

Keywords: *POS, Tagger, Morphological analyzer, NER, Chunker, Token, Hindi Language, Devnagari Script.*

1. INTRODUCTION

A POS (Part-Of-Speech) Tagger is a piece of software program that takes natural language sentences as input and assigns Parts-of-Speech to each word such as- noun, verb, adjective etc. as output. POS tagging also known as Grammatical Tagging or Word Category Disambiguation. Hindi Language is considered morphologically rich and free order language so it is the biggest problem in Machine Translation, Language Learning & Teaching, Natural Language Generation, etc. to assigning correct part-of-speech to each word of a given input text depending on the context, POS tagger plays leading role to solve such problems. Many Hindi POS tagger currently available doesn’t work properly and correct POS tagging in Hindi sentences because Morphophonemic changes are major problems in Hindi text. In this POS tagger consists of subprogram like- Morphological Analyzer, NER, Token, Chunker, etc. so that this program is using less data and produce many words in deferent grammatical categories.

2. PROBLEM OF STATEMENT

Many words are ambiguous in their part of speech. For example, "सोना" can be a noun or a verb. However, when a word appears in the context of other words, the ambiguity is often reduced: in "राम को रात भर सोना है" the word "सोना" can only be a Verb. POS tagger is a system that uses context to assign parts of speech to words. Automatic text tagging is an important first step in discovering the linguistic structure of large text corpora. Part-of-speech information facilitates higher-level analysis, such as recognizing noun phrases and other patterns in text.

Application of POS tagging:

Partial parsing: Syntactic analysis

Machine Translation: POS tagger is playing important role in machine translation. This system is analysis of source language text to base on target language text.

Information Extraction: tagging a partial parsing help identify useful terms and relationships between them.

Information Retrieval: noun phrase recognition and query-document matching based on meaningful units rather than individual terms.

Question Answering: analyzing a query to understand what type of entity the user is understand what type of entity the user is looking for and how it is related to other noun phrases mentioned in the question.

POS tagging Approaches:

1. **Rule Based Approaches:** The earliest POS tagging systems are rule-based system, in which a set of rules is manually constructed and then applied to a given text. Probably the first rule-based tagging system is given by Klein and Simpson (1963), which is based on a large set of handcrafted rules and a small lexicon to handle the exception. The rules were then used to tag the words for which the left and right context words were unambiguous. The main drawbacks of these early systems are the laborious work of manually coding the rules and the requirement of linguistic background resources.
2. **Statistical Based approach:** The rule-based methods used for the POS tagging problem began to be replaced by statistical models in the early 1990s. The major drawback of the oldest rule-based systems was the need to manually compile the rules, a process that requires linguistic background. Moreover, these systems are not robust in the sense that they must be partially or completely redesigned when a change in the domain or in the language occurs. Latter on a new paradigm, statistical natural language processing, has emerged and offered solutions to these problems. As the field became more mature, researchers began to abandon the classical strategies and developed new statistical models.
3. **Hybrid Based Approach:** Consists in combining Rule-Based method with a Statistical method used to assign the best tag for each of the words of input text. The construction of this tagger contains a trained machine learning which includes approximated rules.
4. **Maximum Entropy Approach:** Strong independence assumption and less use of contextual information are limitation for classification tasks such as POS tagging. For statistical POS tagging, we usually assume that the tag of a word does not depend on previous and next words, or a word in the context does not supply any information about the tag of the target word. Maximum entropy models provide us more flexibility in dealing with the context and are used as an alternative to probability base model in the domain of POS tagging.

3. RESEARCH METHODOLOGY

In this POS Tagger we used linguistics resources so applied linguistic analytical methods, statistical methods for probability of each word and assign correct tag, Qualitative method to usefor quality of collecting data and empirical method to use for applying all derived rules and probability of each word to each tag.

1.	NCM	Common Noun	मेज,विद्यालय,जहाज,घर
2.	PDM	Demonstrative	यह, वह, ये, इसने, यही, इतना,उतना
3.	VMV	Main Verb Finite	पढ़, लिख, खा,पी,रो,सो
4.	VAX	AuxiliaryVerb	रहा_है, चुका_था,होगा
5.	ADJ	Adjective	अच्छा,बुरा,बड़ा,सुंदर,चतुर,अलग
6.	PRT	Particle	भी, तो, भर, ही, तक
7.	ADV	Adverb	दूर, पास, कल, ऐसा, अब,ऊपर, सामने, बाहर

Tag-Set Development: We have created a tag-set which are represent short form of grammatical category for Hindi Language. Many tag-set are available but they are borrowed from English language. In this tag-set we complete analysis of syntax and create of tag-set which are based on BIS tag-set.

POS tagger lexicon generation: Hindi is very rich Language in morphological level and it's have more complexity faced on Morphophonemic changes. When join root and its possible suffix then Root's last character and suffix's first character are join together. Here we analysis of Hindi text with full morphology and derived various type of morphophonemic rules which are exist particular grammatical category.

Paradigm of Noun word: Following table is representing Root/steam, Prefix and suffix based on morphological analysis. Here we can saw how to change grammatical category when root/steam derive different prefixes or suffixes.

Prefix	Rood/Word	Suffix	TAG
प्रति	राष्ट्र	वाद	NNPP
परि	रूप	मान	NNPP

Paradigm of Verb Word: Following table is representing verb analysis, Verb morphology is also complex. Here, which suffix are add with root verb is declared by his information like Gender, Number, Person, Tense, Aspect, Mood.

Root Verb	Suffix	Gender	Number	Person	Tense	Aspect	Mood
जा	ता है	M	S	T	**Present**	-	-
खा	ती है	F	S	T	**Present**	-	-

Paradigm of other words: Here we are defining words paradigm in various form like direct, oblige with singular and plural lexicon information. Every word contain particular grammatical category with provided by lexicon feature. Here lexicon feature is important part of paradigm.

Words	Tag
लड़का	NCM
धीरे	ADV
अच्छा	ADJ

Rule Development: In this POS Tagger we have derived different type of following rules-

1- **Noun Morphophonemic rules:**

- Prefix + Root word + Suffix
- Prefix + Rootword + Suffix1+ Suffix1
- Prefix + Rootword + Suffix1 + Suffix2 + Suffix3
- Prefix + Root word
- Rootword + Suffix1
- Rootword + Suffix1 + Suffix2
- Rootword + Suffix1 + Suffix2 + Suffix3

2- **Verb Morphophonemic rules:**

- root+suffix+intensifier+Aux1+Aux2+TAux
- root+suffix+intensifier+Aux1+TAux
- root+intensifier+Aux1+TAux
- root+suffix+intensifier+TAux
- root+intensifier+TAux
- root+TAux

3- **Disambiguation rules:**

- If Tag = "VBG"[1] and Tag-1 = “POS[2]/PDM”[3] then Tag= “NN”[4]
- If Tag == "POS"[2] and Tag+1 = “VMAIN”[5] Then Tag = “NN”[4]
- If Tag == "VBG"[1] or Tag == "VMAIN"[5] or Tag == "VMAIN2"[6]Then Tag= “NN”[4]

Statistical Method: Statistical method is the most studied formalisms for probability in the POS tagging program. Let W=w_1, w_2, w_3….w_n be a sequence of words and T= t_1, t_2, t_3…t_n be the corresponding POS tags. The problem is finding the optimal tag sequence corresponding to the given word sequence and can be expressed as maximizing the following conditional probability:

$$P\ (T/W)$$

Applying Bayes' rules, we can write

$$P\ (T/W) = \frac{P(W/T)P(T)}{P\ (W)}$$

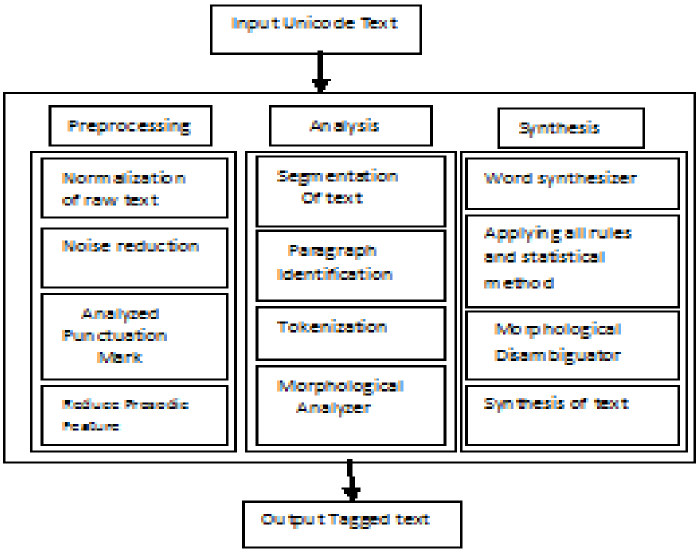
The problem of finding the optimal tag sequence can then be stated as follows:

$$\text{Argmax}_T\ P\ (T/W) = \text{argmax}_T\ \frac{P\ (W/T)\ P\ (T)}{P\ (W)}$$

$$= \text{argmax}_T\ P\ (W/T)\ P\ (T)$$

Where P (W) term was eliminated since it is the same for all T.

Working procedure:



Algorithm: We are design to following algorithm for Hindi POS tagger-

Step first-Read the input text and assign the same on a string type variable.

Step second-Removing unwanted junked characters form the string that is normalizing the text.

Step third-Breaking the text into sentences and further every sentence breaks into words and taking one by one word from the text and goes to next process.

Step fourth-Checking the word properties just like root/steam.

Step fifth-At this step we analysis the words root or steam existence. If it these exist then the go for tagging and applying derived rules otherwise go to the morphological analysis.

Step sixth-At this step we apply prefix, suffix on root/steam.

Step seventh-At this step we apply morphological synthesizer on word and tag the word.

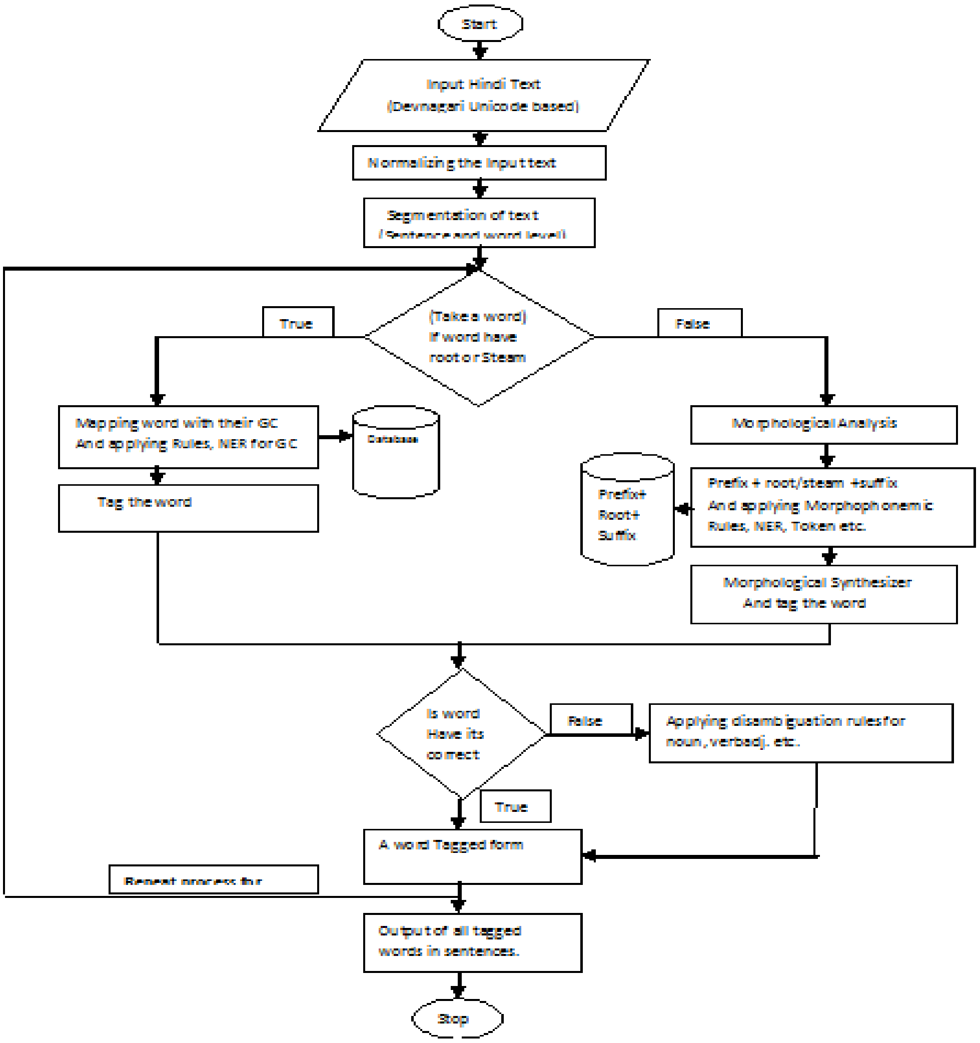
Step eighth-If word has its correct GC then tag the word and display the result. Otherwise applying disambiguation rules for noun, verb, adjective etc. and tag the word.

Step ninth-Repeating the step from fourth to eighth for other words.

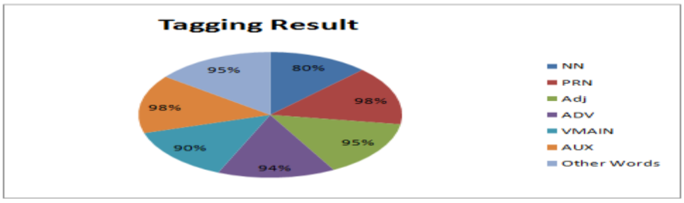
Step tenth-Output for all tagged words.

Step eleventh-Stop

Flowchart:



Result and Conclusion: we have collected 500 Hindi language sentences and analysis them. This result is finding by this POS tagger which are based on 500 Hindi sentences and evaluated following graph.



So, in summary we can say that presented POS Tagger is important task in natural language processing, and is often necessary for other processing such as Syntactic Parsing, Machine translation, Information Retrieval, Information Extraction and question answering system etc.

REFERENCES

[1]. NitinIndurkhya, Fred J. damerau. 2010. Hand book of Natural Language Processing. Second edition. CRC press.

[2]. Ahn. Y. and Y. Seo. 2007. Korean part-of-speech tagging using disambiguation rules for ambiguous word and statistical information. In ICCIT, PP. 1598-1601, Gyeongju, Republic of kore. IEEE.

[3]. David R. Dowty, LauriKarttunen and Arnold M. Zwicky. 2005. Natural Language Parsing. Psychological, computational perspectives. Cambridge university press. Combridge, New York.

[4]. Altunyurt, L., Z. Orhan, and T. Gungor. 2007. towards combining rule-based and statistical part of speech tagging in agglutinative language. Computer engineering 1(1): 66-69.

AUTHORS' BIOGRAPHY

Pravesh Kumar Dwivedi recived his Master of Science (computer science) degree from Makanlal chaturvedi National university of jounlism and communicatin, Bhopal (M.P.) in 2009. He received Master of Informatics and Language engineering from Center for Technology Studies, Mahatma Gandhi Internnaitonal Hindi University, Wardha (Maharastra) in 2012. He received M.Phil. in Computational Linguistics from MGIHU, Wardha (Maharastra) in 2013. Now he is pursuing Ph.D. in Informatics and Language engineering form Center for Technology Studies, MGIHU, Wardha (Maharastra). He is actively participated in various national, International seminar. His Research interest areas are Natural Language Processing, Speech Processing, Machine Translation and Information Retrival.

Pritendra Kumar Malakar received his B. S. degree in Computer Science from Guru Ghasidas University, Bilaspur (C.G.) of India in 2006.He received Master of Computer Application (MCA) degree from Chhattisgarh Swami Vivekananda Technical University (CSVTU), Bhilai (C.G.) in 2009.He worked for Indian Institute of Handloom Technology, Champa (C.G.) as a visiting faculty from Aug 2009 to March 2010. .He was a faculty member of the Department of Information Technology in Dr. C. V. Raman University, Bilaspur (C.G.) from March, 2010 to June 2013. Now he is pursuing Ph.D. in Informatics and Language Engineering from Center for Technology Studies, MGIHU, Wardha (Maharashtra). His research interest includes Natural Language Processing, Sentiment Analysis and Information Retrieval.