An Efficient & Learning Approach of POS Tagging using Rule-Based for Devanagari Script

Madhuri Sharma, Medhavi Malik, Neha Gupta

***Abstract:**In this paper, we present a novel approach for Devanagari Script which can help in various applications like:- Named Entity Resolution, Sentiment Analysis, and Question Answer. This framework works on the following approaches: Learning Task, Performance Efficiency Improvement Task. This system will automatically learn its rules and the main focus is on the tagger which automatically identifies, & resolves its shortcomings which will incrementally increased the performance.*

The basic problems with POS tagging for Indian Languages are very difficult due to some constraints in there. But with some rules and overcoming of limitations restrictions this can be solved. User provides the sentence which goes through a series of stages. In each stage, system also learns the new rules and provides the better results.

***Keywords:**Part of Speech Tagging, Devanagari Script*

I. INTRODUCTION

POS tagging is a significant tool for various NLP applications. There are no efficient POS tagger frameworks still available. Uses of wide interworking lead to the emergence of this field. The main concern is to design a system which adapts it automatically and have an improvement over a result. The main objective of this paper is to automatically tag the words and if there is any error then it automatically corrects it.

II. IMPLEMENTATION

This framework inputs the text from the user. Frame the inputted into number of sentences separated by पूर्णविरामचिन्ह(|). Tokenize the inputted text and remove all the inconsistencies like: - spelling checking, punctuation marking, grammar checking etc. If there is any mistake in grammar and spelling it automatically correct and give the suggestions to the user. Identify the special words like: - numbers, date, time, abbreviations, special symbols (like: - ॐ). Then check all the words by the pre-defined dictionary. If there are no such words found, then various approaches are applied on it. If any word remaining there and further tagged words and have to be suggested and corrected by the user.

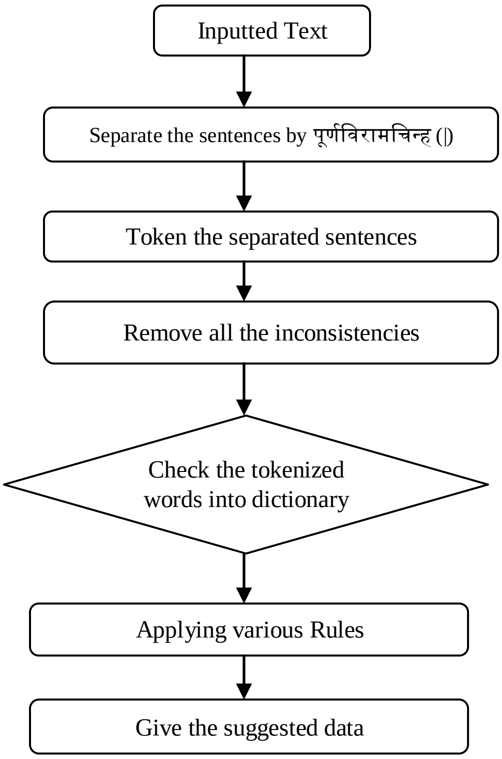
Inputted Text
Separate the sentences by पूर्णविरामचिन्ह (|)
Token the separated sentences
Remove all the inconsistencies
Check the tokenized words into dictionary
Applying various Rules
Give the suggested data

Figure: Schematic Diagram of PoS Tagging

III. RULES IDENTIFICATION

The problem of identifying the Proper Noun (व्यक्तिवाचकसंज्ञा) and Common Noun (जातिवाचकसंज्ञा) is very difficult.

Example: देशमें श्रवणकुमारों कीकमीनहींहै।

पंडितजी भारतकेप्रधानमंत्रीहुए।

Here, in first sentence श्रवणकुमारों is considered as Common Noun while in second sentence पंडितजी is considered as Proper Noun.

But this feature can be rectified by having the checking of the sentence formations.

On the basis of Subject (कतf), Verb (क्रिया)& Object they can be identified.

Example: रामहिंदीबोलताहै।

Subject (कत*f*):- राम

Verb (क्रिया):- बोलता

Object: - हिंदी

Revised Manuscript Received on May 29, 2019.

Madhuri Sharma,Department of CSE, SRM-IST, Chennai, T.N, India. (Email: madhurisharma44@gmail.com)

Medhavi Malik,Department of CSE, SRM-IST, Chennai, T.N, India. (Email: medhavimalik28@gmail.com)

Neha Gupta,Department of CSE, Dr. KNMIET, Uttar Pradesh, India. (Email: n88gupta@gmail.com)

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

In Devanagari script, differentiation can be done on the basis of Gender and its form.

Example: लड़का(Masculine Singular)
लड़के(Masculine Plural)
लड़की(Feminine)

Also, Morphology can be done on a particular single word which provides the entity to it. Devanagari Script have only regular noun. If no words have been found; then it is to be considered as Proper Noun. Also, after they have been suggested by the end-user.

Individual words are checked whether they are Inflectional Morphology or Derivational Morphology. Inflectional Morphology belongs to the same class members while Derivational belongs to different class. Inflection leads to add the specifications to the word only while Derivational provides the different meaning. Inflectional includes how many numbers, what the tense is person and gender.

Example: पढ़ता;पढ़ती

Above these two words are different in their gender which is inflectional morphology.

There are some words in Hindi also; as there are two meanings.

Example: खाता can be used as a Ledger and Eating.

Prepositions also have an important role in joining words and make the complete structure of sentence.

Example: मेरेनीचे(First Person Singular)
आपकेनीचे(Second Person Singular)
उसकेनीचे(Third Person Singular)
हमारेनीचे(First Person Plural)
उनकेनीचे(Third Person Plural)

Example: मैकळसेकामकरूँगा।
मैगुजऱातसेआरहाहूँ।
मैऩेरामसेसुना।

The word prepositions से worked as three different ways : First reveals point of time ; Second tells the about some place while third from any source.

Some words in Hindi have exceptions that singular and plural forms are same. They cannot be determined easily.

Example: सेबफ़ल

Some suffixes that are used in conjugation of verbs make it noun and adjective.

Example:पढ़ाई = पढ़ + आई (Abstract Noun)
खिलाडी = खेल + आडी (Adjective)

IV. WORD GROUPING

The typical order group of sentences in Hindi is somewhat as: SOV (Subject, Object, Verb)

Example: मैंचावलखारहाहूँ।
Subject: मैं
Object: चावल
Verb: खारहाहूँ

Grammar: S -> NP VP
NP -> PN NN
VP -> V NP | V AUX SS
PN ->मैं
V->खा
AUX ->रहा
SS ->हूँ
NN -> चावल

S: Sentence; NP: Noun Phrase; VP: Verb Phrase;
PN: PRONOUN; NN: NOUN; V: Verb;
SS: SuchakShabd (सूचकशCn)

V. CONCLUSION AND FUTURE WORK

In this paper, an efficient and learning approach for Part of Speech (PoS) tagger using rule based technique has been discussed. First, the system confirms the inputted text into the pre-defined dictionary; if not found then various techniques are applied on it. Also, the user gives a confirmation to the specified output. The system lacks where there is a problem of identifying the common noun and proper noun. But, it always crams the sentence formation of it; so gives a better performance for the new similar sentences. Another situation is that when large number of words is not correctly identified by the system; due to the selection of which appropriate rule to be applied?

REFERENCES

1. Bharati, Akshar, VineetChaitanya and Rajeev Sangal. (1995). Natural Language Processing: A Paninian Perspective, Prentice-Hall of India, New Delhi.
2. DeepaModi and Neeta Nain ,“ Part-of-Speech Tagging of Hindi Corpus Using Rule-Based Method”, DOI 10.1007/978-81-322-2638-3_28
3. Royal Sequiera, MonojitChoudhury, Kalika Bali, “*POS Tagging of Hindi-English Code Mixed Text from Social Media: Some Machine Learning Experiments*”, 2015 Proceedings of International Conference on NLP, *December 2015,*
4. Pawan Deep Singh, ArchanaKore, RekhaSugandhi, GauravArya, SnehaJadhav, “*Hindi Morphological Analysis and Inflection Generator for English to Hindi Translation*”, International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 9, March 2013, ISSN: 2277-3754
5. ShwetaVikram,, " *Morphology: Indian Languages and European Languages*", International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013, ISSN 2250-3153
6. Chaudhary, Sharma, Gupta, Goel, “*QA Typology”,* IJSEC(International Journal of Engineering Science and Computing), ISSN 2250-1371, Volume 7 Issue No. 4, April 2017
7. Kristina Toutanova, Dan Klein, Christopher D. Manning, Yoram Singer ,“*Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*”, North American chapter of the association for computational linguistics, 2003