Morphology Based Natural Language Processing tools for Indian Languages

Manish Shrivastava
Department of Computer Science and Engineering
Indian Institute of Technology
Powai, Mumbai
manshri@cse.iitb.ac.in

Nitin Agrawal
Department of Computer Science and Engineering
Indian Institute of Technology
Powai, Mumbai
nitina[at]cse.iitb.ac.in

Bibhuti Mohapatra
Department of Computer Science and Engineering
Indian Institute of Technology
Powai, Mumbai
bibhuti[at]cse.iitb.ac.in

Smriti Singh
Department of Computer Science and Engineering
Indian Institute of Technology
Powai, Mumbai
smriti[at]cse.iitb.ac.in

Pushpak Bhattacharya
Department of Computer Science and Engineering
Indian Institute of Technology
Powai, Mumbai
pb[at]cse.iitb.ac.in

Abstract— **The morphological strength of Indian Languages warrant the use of thorough morphological analysis. Morphological analysis should be the first step towards any Indian language processing task. In the process of working towards a Rule-based Part-of-Speech tagger for Hindi we have developed tools such as Stemmer and Morphological Analyser. We discuss here how these tools and in-depth morphological analysis of the language assist in Part-of-Speech tagging task. Also, we discuss how these tools help other NLP tasks by describing their integration with other systems. The results of the Stemmer and Morphological Analyser available so far are very encouraging. The Stemmer and Morphological Analyser are currently integrated with Hindi Wordnet Project (due to be released soon), Hindi Generation Project and aAqua Question Answering project. The results of the Stemmer and Morphological Analyser available so far are very encouraging.**

I. Introduction

"The ultimate goal of research on Natural Language Processing is to parse and understand language.For this reason much of the research in NLP has focussed on intermidiate tasks. One such task is Part-of-Speech tagging" **[**1**]**.**Part-of-Speech tagger** is the building block for all Natural Language Processing (NLP) tools and applications. Part of Speech tagging problem is defined as the task of providing the correct grammatical information for words in a sentence. For example:

Input: राम खा रहा है
Output: राम_NP खा_VB रहा_VDM है_VC

This is the first step towards understanding a language.It can be viewed as a case of limited syntactic disambiguation. Many words in a language represent different Part-of-Speech depending on their usage. In the text and in tagging we try to determine which of these categories is most appropriate for a particular use of a word. Though, the information gained by tagging is limited, it is very useful for many Natural Language Processing tasks like Information extraction, Question Answering, Word sense disambiguation etc.

Various techniques have been explored for Part-of-Speech tagging **[**2**]**. Some of these are completely automated while others need a lot of human input.The primary step towards development of a Rule Based Part-of-Speech tagger for any language demands an in-depth understanding and analysis of that language **[**3**]**. Analysis of the language then helps in developing the computational model to handle data. From the initial analysis and results of brute-force approaches, we felt the need of more modular approach towards the analysis of language and development of Part-of-Speech tagger. From earlier studies it was found that:

1. Indian languages are morphologically rich and the best way to approach the problem of tagging is to first analyse the data.
2. The Linguistic informtion should be encoded in a way that it is independent of the code and is easy to change for a linguist.

At this stage on linguistic front, we are ready with Noun and Verb analyses that are very detailed. In Noun analysis, the Nouns are classified into 20 paradigms based on GNPC

Paradigm	Gender	No.	Del char	Suffix	Case	form
राजा	Male	sg	-	-	direct	राजा
राजा	Male	sg	-	-	oblique	राजा
राजा	Male	pl	-	-	direct	राजा
राजा	Male	pl	-	ओं	oblique	राजाओं
लड़का	Male	sg	-	-	direct	लड़का
लड़का	Male	sg	आ	ए	oblique	लड़के
लड़का	Male	pl	आ	ए	direct	लड़के
लड़का	Male	pl	आ	ओं	oblique	लड़कों

TABLE I

NOUN ANALYSIS TABLE

Tense	Asp	Mod	Gen	Num	Per	Ex
Present	Stv	-	M	sg	2nd, fam	खाता है
Past	Inf	-	M	sg	2nd, fam	खाता था
Future	-	-	M	sg	3rd	खायेगा
Present	Dur	-	M	sg	3rd	खा रहा हैं
Present	Stv	Abl	M	sg	1st	खा पाता हूं

TABLE II

TAM-GMP ANALYSIS

(*gender, number, person and case*) values they carry. The analysis done will be utilised for Part-of-Speech tagging and other NLP tasks like language generation, stemming etc.

On computational front, we have developed a **Stemmer** and a **Morphological analyser**, which will be used as modules for Part-of-Speech tagger. Stemmer provides the root, suffix and grammatical category of the input word and Morphological analyser is the tool which gives detailed analysis of a word based on its inflection and the context of its usage. The morphological analyser will provide information like person, number, gender, aspect etc. for the word and this information is used for deciding the Part-of-Speech tag. For example:

Input : लड़की Output : Noun

Input : लड़कियाँ Output : Noun plural, direct

II. HINDI MORPHOLOGICAL ANALYSIS

Linguistic analysis forms the platform on which all NLP tools are developed. The analysis provides computer scientists with an insight into the development of the system. The efficiency and accuracy of the system depends on the richness of the analysis.

The analyses done so far can be categorised as follows :

A. *Noun analysis*

Nouns are categorised into 20 different paradigms based on the following criterion:

1. Vowel ending.
2. Valid suffix of a word.
3. Gender, Number, Person and Case information.

A snapshot of the analysis in shown in table 2.1.

There are 20,000 Nouns classified in 20 such paradigms. Initial paradigm lists were obtained from IIIT, Hyderabad. The analysis helped in the formation of Suffix-Replacement(S-R) rules for Hindi language. This helps in reducing a lot of ambiguities in the process of stemming. For example,

Root:लड़की
Plural Direct form:लड़कियाँ
Suffix:ईयाँ.
Paradigm:लड़की
Suffix-Replacement: ईयाँ/ई

and,

Root:चिड़िया
Plural Direct form:चिड़ियाँ
Suffix:ईयाँ.
Paradigm:चिड़िया
Suffix-Replacement: ईयाँ/इया

Thus, the analysis using the paradigms helps in increasing the accuracy by returning only the correct root. The paradigm analysis is also used by Morphological analyser to correctly analyse suffixes. It is noted that the morpheme analysis of a suffix varies depending on the paradigm.

For example,

Paradigm : खर्च Suffix : ए Analysis : plural, direct and,
Paradigm : लोहा Suffix : ए Analysis : singular oblique and plural oblique

Presently we have following resources for Nouns:

1. Suffix-Replacement rules for all paradigms.
2. Noun List divided into paradigms.
3. Exhaustive Noun list from WordNet.

All these resources are very useful for building NLP tools.

B. *Verb Analysis*

The Verb Group represents the following grammatical properties:

1. Tense : Present, Past and Future.
2. Aspect: Durative, Stative, Infinitive, Habitual and Perfective etc.
3. Modal: Abilitive, Deontic, Probabilitative etc.
4. Gender: Male, Female, Dual.
5. Person: 1st , 2nd and 3rd.

These values formed the basis to list Verb Groups according to their TAM-GNP values. A TAM-GNP matrix having all possible VGs is developed. Presently there are 622 unique paradigms in the TAM-GNP matrix.

Linguistic resources developed with the help of TAM-GNP analysis are:

1. Suffix list of verb
2. Morpheme analysis for VG
3. Formation of disambiguation rules for Morphological analyser and ultimately for Part-of-Speech tagger.

The analysis and the linguistic resources help in the following activities:

1. Identifying Verb Group in a sentence: Verb Group follows a structured word order. Using these paradigms developed from this analysis, any Verb Group can be

identified in a given sentence. For example:
Input sentence :“राम खाता रहता है ”
Verb Group :“खाता रहता है ”
and,
Input sentence :“राम घर मे रहता है”
Verb Group :“रहता है”.

2. Identifying Main verb and Auxiliary verb in Verb Group: In the example above रहता was present in both the sentences. In first sentence खाता is the Main verb and रहता is an Auxiliary verb whereas in second sentence रहता is the Main verb.
3. Identifying suffixes: For example if the word is खेलता the stemmer can identify that the suffix is ता and the morphological analyser will give the output that त gives Aspect being Stative and आ gives the analysis that the Gender is Male.
4. Generation of VG: The analysis might be very helpful in the process of Hindi generation. Suppose the information provided is as follows:
   Verb : खा
   Aspect : Abilitive , Stative
   Gender : Male
   The Verb Group generated will be: “खा पाता है ”

All these resources are being used for building NLP tools like Stemmer, Morphological Analyser and Part-of-Speech tagger.

III. Stemmer

Stemming is guided by **morphology** which is the study of rules for forming admissible words. The words are made of morphemes and **morphemes** are minimal units which have a meaning or a grammatical function. For example,
the word खेलेगा comprises following morphemes खेल, ए, ग and आ.
In stemming we remove only Inflectional morphemes. Traditionally, Stemmers remove only the longest suffix **[4]** and return only the stem. The stem in the part used for further higher level tasks. We found out that with the help of the analysis it is possible to return the complete root of the word. The replacement rules recreate the root word from the stem, which is better suited for task described below.

A. Application

Stemmer is presently being used for following applications:

1. **Part of Speech Tagging** : In Part-of-Speech tagger, stemmer will be functioning as initial tagger by providing the grammatical category of the input word. The stemmer helps in finding the category of the unknown word (words not present in the dictionary) by looking at the attached suffix. For Example,
   Input : नाचता
   Root : नाच
   Suffix : ता
   then,
   Category : Verb We can make a guess about the category of a word given its suffix.
2. **Hindi WordNet query** : WordNet is a machine readable lexical database. Presently Hindi and Marathi WordNet are being developed at CFILT, IIT Bombay. WordNet is used in various applications like Word Sense Disambiguation, Machine Translation, Information Extraction etc. For all these applications, queries are send to WordNet for retrieval of synsets, hypernymy, hyponymy etc. *The information in WordNet is stored in the root form*. So, if the query word is inflected then the null output will be returned. Therefore, it is very important to find the root of the word before the search can be made. For example:
   Suppose the person searches for word मकानों then the search system is first required to find the root of the word. The word stored in WordNet is in the root form, i.e the word stored will be मकान and if the search is made using मकानों then there will be no output.
3. **Word Sense Disambiguation** : A word can be used in multiple senses. Word sense disambiguator finds the correct sense of the word in the sentence based on its usage. The Word sense disambiguator being developed uses WordNet for finding different possible senses of the word. For this purpose the stemmer will provide the root of the word for firing query to database.

B. System Design

Stemmer uses following information to get the root for a given inflected word:

1. List of all the possible suffixes along with their category information.
2. The replacements to be made after removal of suffix so that valid root can be formed and
3. Word-lists for different grammatical categories.

The list of possible suffixes were divided in the categories of Verb, Noun, Adverb and Adjective. For Nouns the list is divided in the form of paradigms. All the suffixes and replacements are stored in rule file. The format of the rule file is as follows :

```
# Category name
suffix1/replacement1 $ example1
suffix2/replacement2 $ example2
```

Rulefile example :

```
# Verb
ता/                    $खेलता
ती/                    $खेलती
ते/                   $खेलते
ना/                    $खेलना
नी/                    $खेलनी
ने/                   $खेलने

# लड़का
ए /आ                      $लड़के
ओं/आ                      $लड़कों
```

Adjective	6988
Adverb	872
Verb	2680
Noun	19948

TABLE III

WORD LIST STATISTICS

Presently we have a rule file with 51 Verb, 32 Noun (divided in 20 paradigms) and 3 Adjective suffixes/replacement rules.

The word-list is used to check whether the input word is inflected and if inflected then the root formed after the removal of suffix and addition of replacement is valid or not. Statistics of word-list presently we have is shown in table III. For the stemmer to be used as initial tagger for Part-of-Speech tagging, we are including the list of those categories which do not undergo inflection. Such list includes Case, post-positions, pronouns etc.

Compound Nouns and Proper Nouns have not been handled yet.

1) Working of the system: The stemmer returns the root, suffix and the category for all possible matched suffixes.

The stemmer works in the following manner :
(We will be taking the word खेलता as an example for better understanding of the system)

- The input word is searched in all the word-lists to check if the input word is in the root form. खेलता is not present in any of the word-lists indicating that its not a root.
- The input word is matched against all the possible suffixes one by one. The valid suffix present in खेलता are ता and आ both the suffixes are from Verb group and replacement for both the suffixes is null strings.
- The matched suffix is removed and the replacement, as specified in rule, is attached. The root formed is looked into the word-list of category of the matched suffix. If the root is found in the list then the suffix is added to the list of possible suffixes along with the category information. The search continues for other possible suffixes. In our example :
  Possible root : खेल
  and,
  Possible Root : खेलत (not found in Verb wordlist, hence dropped)
- At the end either the root word or the longest matched suffix is returned (this is how unknown word is handled). In case of example word if खेल and खेलत both are not present in the word-list then the suffix returned will be ता because it is the longest matched suffix.

C. Results

The categories returned are Verb, Adjective and paradigm of Noun. Presently the system is not handling Compound nouns and Conjunct verbs. The system was tested on a corpus and the statistics are as shown in table IV . Here in the table ambiguous output means more then one root returned by the system, this does not mean that the output is incorrect.

Number of words	12000
Total outputs	6578
Correct output	6578
Ambiguous output	428
Single output	6150
Precision	100 %
Ambiguity	6.5 %
Recall	0.5481
F-score	0.3540

TABLE IV

STEMMER OUTPUT STATISTICS

Stemmer was unable to produce results for a large number of words because of the following reasons :

1. Word not present in word-list
2. Unclean data : mostly spelling errors

IV. MORPHLOGICAL ANALYSER

The **morphological analysis** is the process of providing grammatical information about the word on the basis of properties of the morpheme it contains **[**5**]**. For Example:

In English :
Input word = Jumped
Category = Verb
Root = Jump
Suffix = -ed
Tense = Past, presence of -ed
In Hindi :
Input word = रहेगा
Category = Verb
Root = रह
Suffix = एगा
Person = 3rd person, presence of ए
Tense = Future, presence of ग
Gender = Male, presence of आ

Our approaches of morphological analysis for Hindi language are:

1. Phrase level Analysis
2. Word level Analysis and then its extension to phrase level

V. MORPHOLOGICAL ANALYSER AT PHRASE LEVEL

The structure of sentence has been seen as the combination of phrases. As seen in fig 1 a sentence comprises a Verb phrase and a Noun phrase. We need to have paradigms/rules/patterns, for representing a phrase, to identify such occurance in any given input.

For verbs, 622 paradigms have been identified. Each paradigm represents a unique Verb grouping. The format of the paradigm is shown below.

Paradigm 1 : (Verb-root)(fct)(gnr)(SPACE)(cop)(pnr)

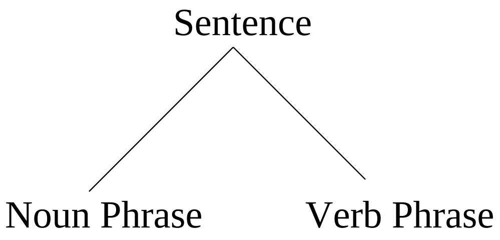


Fig. 1. Sentence Structure

Sentence represented : (Verb-root)ता हैं

Paradigm 2 : (Verb)(SPACE)(vpt)(gnr)(SPACE)(cop)(sbjn)(pnr)(fut)(gnr)
Sentence represented : (Verb) रहे होंगे
Verb-root : खेल , नाच, हँस

Given a sentence " वह खाना खाता है " paradigm 1 will catch " खाता है". Now we have the Verb group and paradigm under which it falls. The morphological analysis for this group is stored in TAM-GNP analysis and will be retrieved from there. The analysis of the sentence above will be " Verb, aspect stative, gender male".
Thus the paradigms help in identifying:

1. Verb group.
2. Main verb.
3. Auxiliary verbs.

The identification of Main verb is an important step in Verb Group morphological analysis. There are certain verbs like "रह , पा, लग " which can occur as Main verbs as well as Auxiliary verbs. Paradigms help in resolving such ambiguities. For example:

In sentence "मुझे लग रहा है ", "लग" is the Main verb and "रह " is the Auxiliary verb.

VI. Morphological Analysis at word level

To analyze an input word, the Morphological analyser at word level needs the following information:

1. Grammatical category of the word.
2. Suffix present.
3. Morphemes present in suffix.

The Stemmer satisfies the first two requirements.

A. Resources required

The resources used by the Morphological analyser are as follows:

1. Suffix list
2. Word list/ Dictionary
3. Morpheme analysis

The first two resources are used by stemmer which is a part of the morphological analyser. The morpheme analysis provides following feature information based on the grammatical category of the suffix:

1. Verb : Tense, Aspect, Marker, Gender, Number and Person.
2. Noun : Gender, Case, Number.
3. Adjective : Number.

The table representing the analysis is in the following format:

#category
feature name - morpheme(along with position information) - % analysis.

We will see why category information and position information is vital while discussing the working of Suffix analyser.

1) System Design: We will workout with an example to understand the design and working of the system.
Suppose the input to the system is word "खेलेगा"

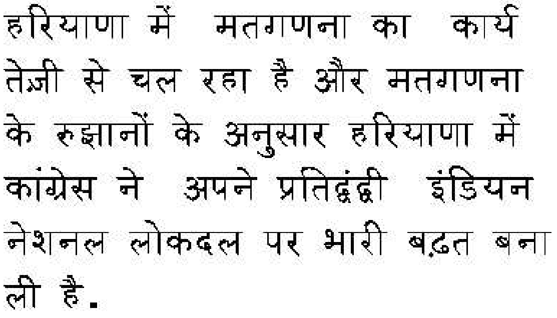
- **Stemmer**: Input :"खेलेगा " Output :
  Root : खेल
  Suffix : एगा
  Category : Verb

- **Suffix Analyser**: The analyser will check for all the morphemes in the Verb category. We saw in the morpheme analysis list that we store the position information of the morpheme, as the same morpheme may perform different function at different word positions.
  For example,
  Suffix : एगा
  Category : Verb Analysis (with position information not stored) :
  ए : 2nd Person , dual plural
  ग : Future
  आ : male
  Analysis (with the position information included) :
  ए : 2nd Person
  ग : Future
  आ : male
  The duality information is given by suffix ए only when it comes at the end of word like in "जाते".

  The analyser, with the help of available morpheme analysis, searches for all possible morphemes and stores the analysis. After analyzing all the morphemes we make sure that only the analyses for the longest suffixes are returned.

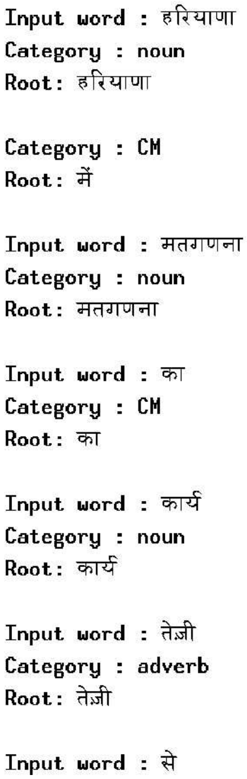
In this approach, Stemmer works as Initial tagger, the Morphological analyser will help us in determining the properties of the word which will be useful in deciding the tag. Some results of the morphological analyser on live data are here. The corpus is taken from http://www.bbc.co.uk/hindi/. This is a news item, dated 27/02/05, on Haryana Elections.

The result shows that in order to get a more detailed morphological analysis, to resolve the ambiguities and to move towards Part-of-Speech tagging, we need to include the context information. The information provided by the suffix is not sufficient to determine the property of the word. The reasons for extending word level morphological analyser to phrase level morphological analyser are:



हरियाणा में मतगणना का कार्य तेज़ी से चल रहा है और मतगणना के रुझानों के अनुसार हरियाणा में कांग्रेस ने अपने प्रतिद्वंद्वी इंडियन नेशनल लोकदल पर भारी बढ़त बना ली है.

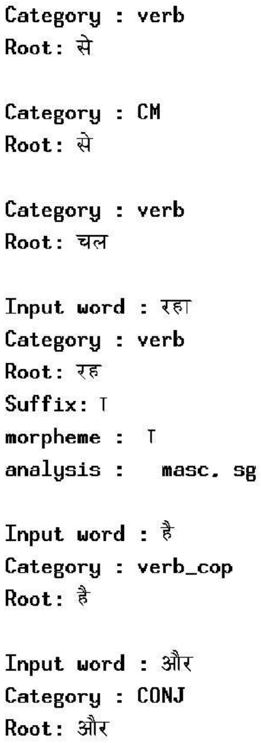
Fig. 2. A sentence from news dated 27/02/05



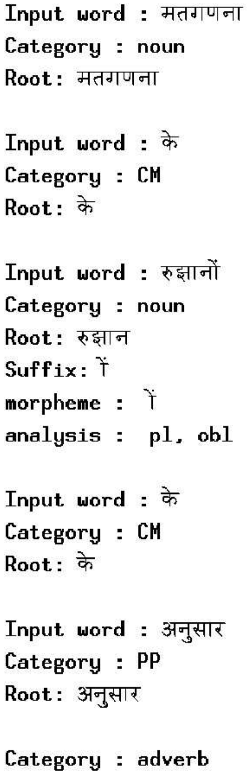
```
Input word : हरियाणा
Category : noun
Root: हरियाणा

Category : CM
Root: में

Input word : मतगणना
Category : noun
Root: मतगणना

Input word : का
Category : CM
Root: का

Input word : कार्य
Category : noun
Root: कार्य

Input word : तेज़ी
Category : adverb
Root: तेज़ी

Input word : से
```

1. Multiple analyses of the same morpheme: Many of such ambiguities are handled by the paradigm information of the morpheme. But, there are ambiguities that cannot be handled at word level and require context information and they will be handled by morphological analyser at phrase level. Example:
   Analysis of ए matra in Nouns depends on the case information.
   In sentence **लड़के ने गाय को मारा** suffix ए of **लड़का** indicate singular direct
   In sentence **लड़के खेल रहे हैं** the suffix ए of **लड़का** gives plural information.

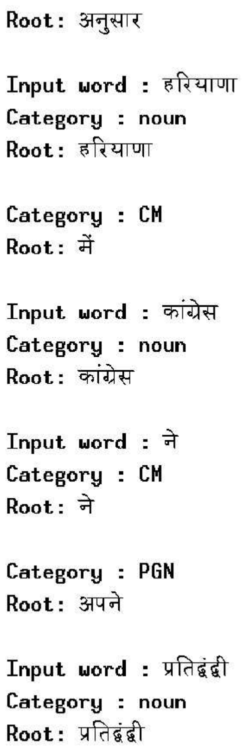
2. Analysis depends on context: There are few Aspect and Modality markers in Verb Groups which may also occur as root words. For example,



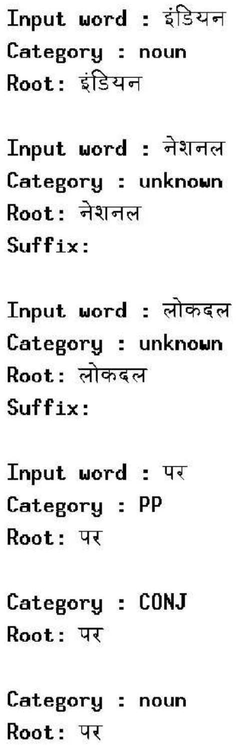
```
Category : verb
Root: से

Category : CM
Root: से

Category : verb
Root: चल

Input word : रहा
Category : verb
Root: रह
Suffix: ा
morpheme :  ा
analysis :   masc. sg

Input word : है
Category : verb_cop
Root: है

Input word : और
Category : CONJ
Root: और
```



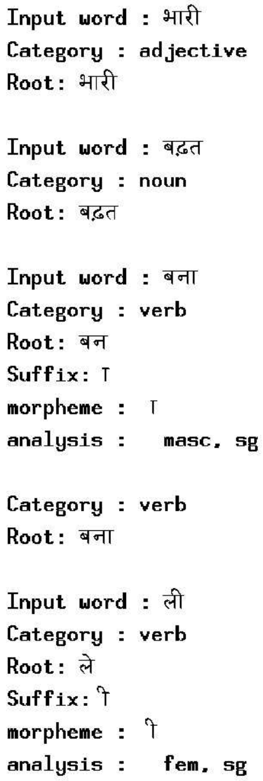
```
Input word : मतगणना
Category : noun
Root: मतगणना

Input word : के
Category : CM
Root: के

Input word : रुझानों
Category : noun
Root: रुझान
Suffix: ों
morpheme :  ों
analysis :  pl, obl

Input word : के
Category : CM
Root: के

Input word : अनुसार
Category : PP
Root: अनुसार

Category : adverb
```



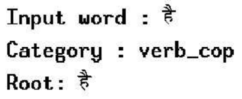
```
Root: अनुसार

Input word : हरियाणा
Category : noun
Root: हरियाणा

Category : CM
Root: में

Input word : कांग्रेस
Category : noun
Root: कांग्रेस

Input word : ने
Category : CM
Root: ने

Category : PGN
Root: अपने

Input word : प्रतिद्वंद्वी
Category : noun
Root: प्रतिद्वंद्वी
```



```
Input word : इंडियन
Category : noun
Root: इंडियन

Input word : नेशनल
Category : unknown
Root: नेशनल
Suffix:

Input word : लोकदल
Category : unknown
Root: लोकदल
Suffix:

Input word : पर
Category : PP
Root: पर

Category : CONJ
Root: पर

Category : noun
Root: पर
```



```
Input word : भारी
Category : adjective
Root: भारी

Input word : बढ़त
Category : noun
Root: बढ़त

Input word : बना
Category : verb
Root: बन
Suffix: ा
morpheme :  ा
analysis :   masc, sg

Category : verb
Root: बना

Input word : ली
Category : verb
Root: ले
Suffix: ी
morpheme :  ी
analysis :   fem, sg
```



```
Input word : है
Category : verb_cop
Root: है
```

In sentence **खाना खा पा रहा था** and **मैने सुख पा लिया** the पा in first sentence is having the property of abilitive modal where as in second sentence पा is the main Verb.

The extension to phrase level morphological analysis is a step toward the Part-of-Speech tagger in the following way:

1. Tag disambiguation for multiple categories returned by the stemmer.
2. Tag selection with the help more accurate morphological analysis.

VII. Part-of-Speech tagging using word level morphological analyser

The block diagram for Part-of-Speech tagger using morphological analyser at word level is shown in 3. The working of each module/block is as follows:

1. **Stemmer:** As discussed in Section III.
2. **Suffix analyser:** As discussed in Section IV.
3. **Disambiguator:** There will be two disambiguation modules. First module will handle the following ambiguities
   (a) Multiple categories returned by the stemmer : For the condition
   Input : हजारों
   output 1
   Category: Noun

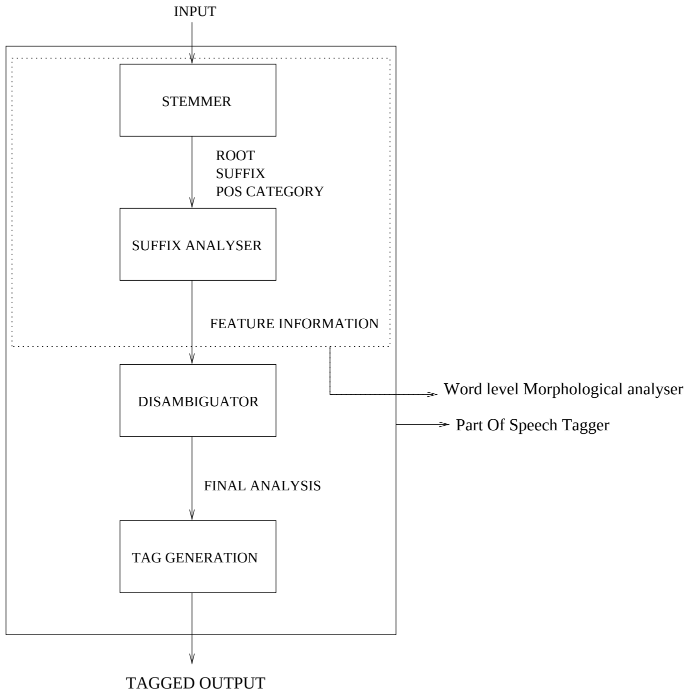


Fig. 3. Part-of-Speech Tagger using word level morphological analyser

Root: हजार

output 2
Category: Adjective
Root: हजारों

If the word *w* is followed by a Case-Marker and the Stemmer returns Adjective and Noun as its possible grammatical categories, then from linguistic analysis we can claim that word will be Noun and not Adjective as Adjective is never followed by a Case-Marker.

(b) Single category returned but change in type because of usage: Linguistic analysis have shown that there are some verbs that can occur as main verb as well as auxiliary verb. The stemmer will always return the category as main verb. With the help of surrounding components this can be disambiguated. This approach is similar to the one used in Brill tagger **[?]** **[**6**]**. The primary difference is the use of analysis instead of tags to disambiguate or transform the analysis of adjacent words.

The second module will be handling ambiguities arising from the analysis of suffix. A single suffix can have multiple analyses or can be incorrectly analysed. This ambiguity can be resolved looking at the context of the word. For example, in sentence "लड़के खेल रहे है " Noun लड़के is in plural form. Where as in sentence "उस लड़के ने जूता चुराया " where Noun लड़के which is followed by Case-marker ने is in singular form **[**7**]**. The results after the implementation of the complete system will provide more insight into the complexity of the language and the necessary steps to be taken for handling that.

4. **Tag Generation:** We realize that the analysis should be reflected in the tags. The tags in our approach depends directly on the analysis. We have devised a tag pattern which represent each feature as a single character in the tag. The tag patterns for various catagories is as follows:
Noun Tags : Cat_GN_C
Verb Group Tags :Cat_GNP_TAM
Adjective Tags : Cat_GN_C
Pronoun Tags : Cat_GNP_C
where,
Cat = Category such as N for Noun, VM for Main Verb, VA for Auxiliary Verb
G = Gender M,F and N
N = Number - Singular, Plural and Dual
P = Person - 1,2 and 3
T = Tense - P and F
A = Aspect - Durative (D), Stative (S)
M = Modality - Abilitive (A), Infinitive (I).
C = Case - Direct or oblique
Examples,
Input : "खेलेगा " (verb)
Output : खेलेगा _VM_M1X_PSU
and,
Input : लड़के
Output : लड़के_N_MS_O
here,
X stands for "Not applicable" and,
U stands for unidentified or ambiguous at phrase level.

VIII. Conclusion

The work for Rule Based Part-of-Speech tagger for hindi language using Morphological analysis is nearing completion. Further improvemnts would be to include phrase level analysis and disambiguation rules. It is known that Rule Based systems are not exaustive in nature. Part-of-Speech taggers for English usually take a hbrid approach for efficient tagging. Classic examples are Brill tagger **[**8**]** and CLAWS tagger **[**9**]**. These have used hybrid models using Rule-based, stochastic and morphological inputs. Though, morphology plays an important role in the Hindi tagging. Still, unknown words remain a problem. We plan to develop a hybrid system using methods to handle unknown words and to improve the overall accuracy of the system. In the meantime, more analysis will be added to the system to cover aspects which might have eluded us so far.

References

[1] C. D. Manning and H. Schutze, *Foundation of statistical Natural Language Processing*. MIT Press, 2002.

[2] L. V. Guider, "Automated part of speech tagging: A brief overview," *Handout for LING361, Georgetown University*, Fall 1995.

[3] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice-Hall, 2000.

[4] M. Porter, "An algorithm for suffix stripping," *Proceedings of SIGIR*, 1980.

[5] R. S. Akshar Bharati and V. Chaitanya, *Natural Language Processing – A Paninian Perspective*. Prentice-Hall India, 1995.

[6] E. Brill, "Unsupervised learning of disambiguation rules for part of speech taggingi," 1995.

[7] R.-A. G.Saudagar, *An Automated Generation Rule for Hindi*. MCA Dissertation, 1998.

[8] E. Brill, "A simple rule based part of speech tagger," *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992.

[9] R. G. Geoffry Leech and M. Bryant, "Automatic pos-tagging of the corpus," *BNC2 POS-tagging Manual*, 1997. [Online]. Available: http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2autotag.html