***FINAL REPORT***

***OBJECT TITLE:*** *Automatic Extraction of Telugu Verbs through Suffixes and Post position stripping In Given Corpus Data*

***INTRODUCTION OF PROJECT*:**

The purpose of this project is to find the automatic extraction of telugu verbs through their suffix stripping by using Perl program.

**Input data:** sakshi.txt(moral stories taken from sakshinews paper)

**Total no. Of words**: 4487

**Total no. of verbs**: 1321

**Accuracy**:99%

***PROCEDURE*:**

In Telugu,verbs can be identified in two types:

1. Based on the termination of their suffixes from root

2. Based on termination of Tense+PNG(person-number-gender)

.

It can be explained in below

1. **Based on the termination of their suffixes from root:**

The simple verbs in telugu are in 3 forms

1. tam ending words ex:cheaya+tam (do)

2. dam ending words ex:raya+dam (write)

3. ta endings ex: kottu+ta (strike or beat)

4. yu ending words ex: che+yu (do)

**2. Based on termination of Tense+PNG(person-number-gender):**

In this we obsrved in 3 types:

1. Progressive present tense:

In this the word ending with the roots : nnanu,nnaavu,nnaaru,nnaadi,nnaadu,nnaamu.

Ex: kottutu+nnaanu

Kottu=verb nnanu=Tense+PNG

2. Habitual present and future tense:

In this the word ending with the roots : taanu,taavu,taaru,taadi,taadu,taamu.

ex:kottu+taanu.

3.Past tense:

In this the word ending with the roots :

tini,tivi,tiri,timi,tiri,…..etc.

ex:kottu+tini.

***Programming design:***

In this section I can do only two steps:

They are

1. Splitting, tokenization and sorting.

2. Automatic extraction of verbs based on roots of words.

***Explanation***:

Splitting can be done on following logic:

@element = split/\s+/,$line;

Tokenization can be done on following logic:

next if ($line =~ /^\s\*$/);

for($line){

s/<\/?.+?>//g;

s/\'(.+?)\'/$1/g;

s/[!.,::?()"\/]//g;

s/[a-z\/]//g;

s/[A-Z\/]//g;

s/[0-9\/]//g;

s/\s~\s//g;

s/^\s\*//;

s/\s+$//;

s/\'//;

s/\-//;

s/\=//;

}

Extracting verbs can done on regular expressions

In this I can use only while and if loop, logics in regular expressions

While used for assigning the input line to each operation in loop.

If loop is assigned in while loop for checking the given condition is true or false.

***Use utf8:*** this function can be encode and decode the given data.

Regular expressions logics are:

***i$line =~ /(\w+)చాడు$|(\w+)చారు$|(\w+)చావు$|(\w+)చాను$/***

#matching and extract the root words chaadu,chaaru etc.

$line =~ /(\w+)టాడు$|(\w+)టాను$|(\w+)టాడు$|(\w+)టాము$|(\w+)టారు$|(\w+)టవి$/

#matching and extract the root words taamu,taadu etc.

$line =~ /(\w+)చ్చారు$|(\w+)చ్చాడు$|(\w+)చ్చును$/) #pattern matching

my $root2 = substr$line,0,-7; #extracting

$base3 = $root2."వ్వటం"; #adding root

Remaining part of the program can explained on same logic(shown on above).

***Issues*:** some accuracy missing on root word (exact verb not

Comes, comes only rootword)

***Thanks☺***