RESEARCHERID
THOMSON REUTERS

ISSN: 2277-9655
[Maheshwari, Arya, Rokadiya * *et al.*, 6(11): November, 2017] Impact Factor: 4.116
IC™ Value: 3.00 CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# PROPOSED ARCHITECTURE TO RECOGNIZE THE WORDS USING NATURAL LANGUAGE PROCESSING

**Kartik Maheshwari[*1], Kanishka Arya[2] & Prateek Rokadiya[3]**
[*1]Computer Science, Institute of Engineering and Technology, DAVV, Indore, India
[2&3]Computer Science & Engineering, Acropolis Institute of Technology & Research, Indore, India

**DOI**: 10.5281/zenodo.1054623

## ABSTRACT

In this project work we have worked on Natural Language processing. As we know that there are certain words whose meaning is different in different sentences. To identify such words from the sentences and correct them is the tedious task. To remove the language barriers among the citizens specifically in agricultural domain. The Google translator that is majoritarily used by people around the world still doesn't have the perfect accuracy in machine translation. For example: rose goes to school. - input word. The translated work looks like गुलाबस्कूलजाताहै. The wrong translation could lead to chaos. Several other translators like Babylon translator software also lacks in accuracy.There arises a lexical ambiguity and the problem of over stemming in Google translator, the extraction of root word is an ambiguous process.For example: 1) मेरादिलसोनेकाहै-here सोने means "gold. 2)मेरादिलसोनेकाहै-here सोने means "to sleep". In this research paper, we have proposed architecture to find out such words from the sentences and their appropriate meaning

**KEYWORDS**: Natural Language Processing, Stemmer, Lemmatizer

## I. INTRODUCTION

The NLP project for Indian Languages is mainly focused on two main tools, i.e. stemmer and lemmatizer. We have created these two tools for Hindi and Malayalam. Stemming, as the name suggests, stem means root so we extract root word from the given input word, as a result we get some root word. The root word extracted can be meaningful and sometimes non-meaningful too. For this we create lemmatizer rules, which help in extracting meaningful root words. The first process was to collect the corpus which was related to agricultural domain, then tokenizing the corpus so that we can study individual words ,after then we remove suffixes and prefixes and check whether they are meaningful or not. There arises two conditions, firstly if we get meaningful words, they are displayed as it is, and secondly, if they are not meaningful then we store them in database as exception words. As the project Stemmer and lemmatizer is able to Stem or lemmatize only the words of agricultural domain and the future scope to it is making use of this to create a translator so the project is deliverable to the users associated with the agricultural domain.

## II. SOFTWARE DEVELOPMENT LIFE CYCLE MODEL

The spiral model is a risk-driven process model generator for software projects. Based on the unique risk patterns of a given project, the spiral model guides a team to adopt elements of one or more process models, such as incremental, waterfall, or evolutionary prototyping. The spiral model combines the idea of iterative development with the systematic, controlled aspects of the waterfall model. This Spiral model is a combination of iterative Development process model and sequential linear development model i.e. the waterfall model with a very high emphasis on risk analysis. It allows incremental releases of the product or incremental refinement through each iteration around the spiral.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

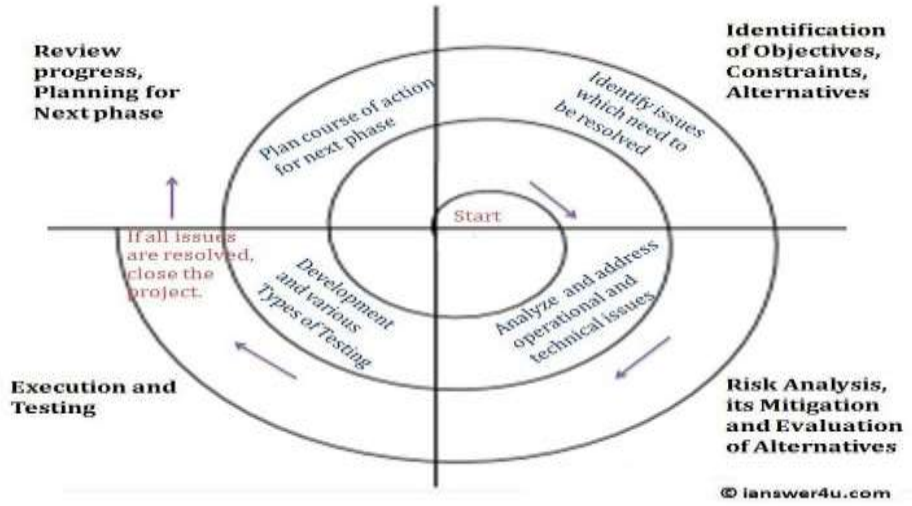


***Figure 1: Spiral Model used for NLP***

**Reason for use**

The development team in Spiral-SDLC model starts with a small set of requirement and goes through each development phase for those set of requirements. The development team adds functionality for the additional requirement in every increasing spirals until the application is ready for the productionsphase.The reasons for using spiral model are:

1. Additional functionality or changes can be done at a later stage
2. Cost estimation becomes easy as the prototype building is done in small fragments
3. Continuous or repeated development helps in risk management
4. Development is fast and features are added in a systematic way
5. There is always a space for customer feedback

## III. PROBLEM DOMAIN

1. To remove the language barriers among the citizens specifically in agricultural domain.
2. The Google translator that is majoritarily used by people around the world still doesn't have the perfect accuracy in machine translation.
   For example:rose goes to school.- input word
   i. □□□□□□□□□□□□□□□□□□ – translated word
3. The wrong translation could lead to chaos.
4. Several other translators like Babylon translator software also lacks in accuracy.
5. There arises a lexical ambiguity and the problem of over stemming in Google translator, the extraction of root word is an ambiguous process.

For example:

1) मेरादिलसोनेकाहै-here सोने means "gold
2) मेरादिलसोनेकाहै-here सोने means "to sleep"

## IV. SOLUTION DOMAIN

| Word | Root | Suffix | Prefix |
|---|---|---|---|
| दूधवाला | दूध | वाला | - |
| निम्नानुसार | निम्न | ानुसार | - |
| दादागिरी | दादा | गिरी | - |

Stemmer for Hindi

| Word | Root | Suffix | Prefix |
|---|---|---|---|

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

| | | | |
|---|---|---|---|
| लडको | लडक+ ा | ो | - |
| लडकियों | लडक+ ी | ियों | - |
| पद्धतियों | पद्धत+ ी | ियों | - |

Lemmatizer for Hindi

## V. PROPOSED METHODOLOGY & UML DIAGRAM

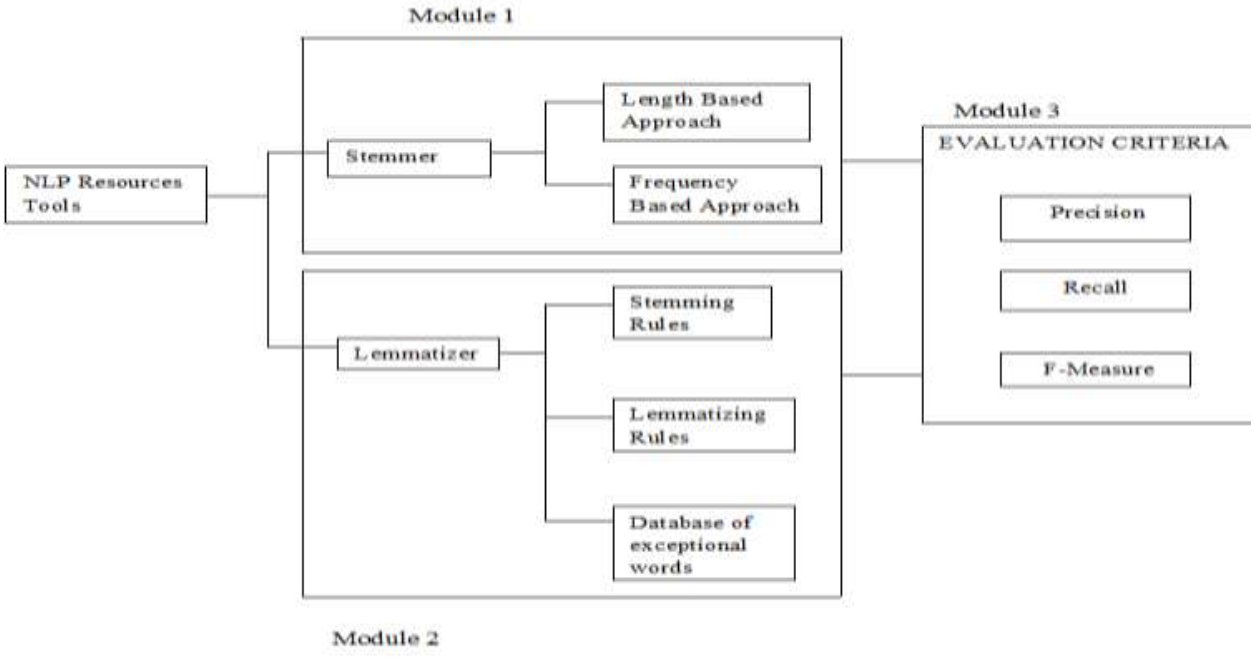


***Figure 2: Proposed Architecture for NLP***

## VI. RESULTS ANALYSIS

AnyHindi word with suffix
दार,शाली,कारी,वाला,शील,विधि,अनुसार,वान,वट,बाज़,नीति,दार,वी,वे,इक,इत,हीन,वार,कोण,
त्मक,ता,इति, यों,ओं,ए,ऐ,

1. Any word in Hindi without the suffix list mentioned above.
2. No other Language other than Hindi is supported.

We had taken total 50 words for stemming of which all the words were correct and the accuracygained was:

$$Accuracy = \frac{\text{Matched Output}}{\text{Total Number of words}} \times 100$$

i.e. $Accuracy = \frac{50}{50} \times 100 = 100\%$

For Lemmatization , we had taken 50 words of which about 40 words were correct and accuracy gained was:

$$Accuracy = \frac{\text{Matched Output}}{\text{Total Number of words}} \times 100$$

i.e. $Accuracy = \frac{40}{50} \times 100 = 80\%$

## VII. CONCLUSION & FUTURE SCOPE

The main limitation of "modern NLP technologies" is their dependency on huge computing power and the alignment and language modeling have been a challenging issue for researchers to improve MT.

1. Requires clarification on dialogue
2. May require more keystrokes
3. May not show context
4. Is unpredictable

Natural-language processing (NLP) is an area of artificial intelligence research that attempts to reproduce the human interpretation of language. NLP methodologies and techniques assume that the patterns in grammar and the conceptual relationships between words in language can be articulated scientifically. The ultimate goal of NLP is to determine a system of symbols, relations, and conceptual information that can be used by computer logic to implement artificial language interpretation.

## VIII. REFERENCES

[1] Jiandani, Kartik Suba Dipti, and Pushpak Bhattacharyya. "Hybrid inflectional stemmer and rule-based derivational stemmer for gujarati." Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2011). 2011.

[2] Khan, Sajjad Ahmad, et al. "A light weight stemmer for Urdu language: a scarce resourced language." 24th international conference on computational linguistics. 2012.

[3] Thangarasu, M., and R. Manavalan. "A literature review: stemming algorithms for Indian languages." arXiv preprint arXiv:1308.5423 (2013).

[4] Riddhi Dave, PremBalani "Survey paper of Different Lemmatization Approaches ."International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue 1st International Conference on Advent Trends in Engineering, Science and Technology "ICATEST 2015", 08 March 2015.

[5] Plisson, Joël, Nada Lavrac, and DunjaMladenic. "A rule based approach to word lemmatization." Proceedings C of the 7th International Multi-Conference Information Society IS 2004. Vol. 1. No. 1. 2004.

## CITE AN ARTICLE

**Maheshwari, K., Arya, K., & Rokadiya, P. (2017). PROPOSED ARCHITECTURE TO RECOGNIZE THE WORDS USING NATURAL LANGUAGE PROCESSING. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *6*(11), 409-412.**