ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 7, January 2018

VEA Model in Word Formation Process of Maithili MT

Saroj Kumar Jha[1], Piyush Pratap Singh[2], Vijay Kumar Kaul[3]
Dept. of Computational Linguistics, School of Languages,
Mahatma Gandhi Antarrashtriya Hindi VishwaVidyalaya,Wardha (MS)

Abstract: ***Morphological analysis is the most remarkable stage for the development of Maithili-English-Hindi MT system under NLP. This paper is motivated to design a morph analyzer for Maithili language and as add-on of Maithili MT system for appropriate analysis at the morphological level. The research is contributing through derivational process of analyzing word attached with affixes. This paper has reviewed most of the methods of MA at different level. Among all the development of MA, suffix striping and FSA are commonly practiced. Some of them are using lemma based approach for analyzing morph (Nikhil et.al. 2012). There are several linguists and computer scientists have discussed the statistical approach and Hybrid approach with FSA, probability based model (Rinju et. al. 2013) and several other approaches of machine learning to develop Maithili Morphological Analyzer (MMA). But rule based approach is friendly enough to use with all the machine learning models. VEA is one of the emerging modal in word formation process which is somehow adequate with Maithili language. Overall linguistics approach is core of MA in Maithili for developing MT system. This research is introducing linguistically a friendly and bit new with machine learning and proficient model for analyzing words and generating multiple words on the basis. The discussion also covered the concatenation with root word to suffix and prefix. Maithili MA is demonstrating a small concept with rule based model and we are designing it with hybrid modal including corpus based approach. This design is incorporating the lexicon tables, suffix list, prefix list and the Vowel Ending Approach (VEA) to justify that how does concatenation take place. In the above table POS category of Noun, Adjective and Verb are shifting to another category of POS after concatenation of suffixes. And it's also focused that how the words end with their vowels and how does suffix connects on the basis of its vowel ending mechanism.***

Keywords: **NLP, Morphology, POS, MA, MT.**

I. STATE OF ARTS

[1] The research paper on *"Unsupervised Improvement of Morphological Analyzer for Inflectionally Rich Languages"* written by (Bharti, Akshar. et. al. 2001). This paper presented an algorithm for unsupervised learning of morphological analysis and generation of inflectionally rich language like Hindi, given a low coverage morph and a corpus of raw text. The result of the algorithm are encouraging with the coverage of primitive morph going up from 32% to about 63% and that of an advanced morph going up from 96% to about 97%.

[2] The research paper on *"Hindi Derivational Morphological Analyzer"* is written by (Kanuparthi, Nikhil et.al. 2012). In this paper the authors present their Hindi derivational morphological analyzer. Their algorithm upgrades an existing inflectional analyzer to a derivational analyzer & primarily achieves two goals. First it successfully incorporates derivational analysis in the inflectional analyzer. Second, it also increases the coverage of the inflectional analysis of the existing inflectional analyzer. The authors pursued the five steps approach for building their derivational analyzer – studying Hindi derivations, derivational rules, finding majority of properties, using Wikipedia data for confirming genuineness, Develop an algorithm for derivational analysis. The algorithm uses the principle of Porter's stemmer & Krovetz stemmer.

[3] The paper on *"Morphological Analyzer for Malayalam: Probabilistic Vs Rule Based Method"* is written by (Rinju, O.R. et. al. 2013). This paper presents a Morphological Analyzer for Malayalam, by considering the noun and verb categories of a word. The proposed morph analyzer returns the morpheme along with the grammatical information such as Gender, Number and case information of noun and tense aspect for verb. A probabilistic and rule based method is used for analysis using inflection and suffix list, which is created using look up tables. The result shows that the rule based method is more accurate than the others.

[4] The research paper on *"Hindi Morphology Analyzer & Generator"* written by (Vishal et. al. 2008). This paper presents the morphological analysis and generator tools for Hindi language using paradigm approach for windows platform having GUI. Hindi is very rich in inflectional morphology can be witnessed from the fact that is English usually there are maximum 7-8 inflected word forms of noun but in Hindi it can be up to 40 and even more than that. This morphological analyzer gives preference to the time taken to search for a word in the database to know its grammatical information & also accuracy of returned results. In the database used by this tool all the possible word forms of all root words are stored. Though it takes a bit more space but the search time is very less.

[5] The paper on "Developing Morphological Analyzer for South Asian Languages: Experimenting with Hindi and Gujarati Languages" is written by (Ashwani, Niraj. et. al. 2010). This paper is described on morphological analyzer for the Hindi and Gujarati language. In order to demonstrate our approach's portability to other similar languages, we present our experiments for Gujarati language. The paper presents a rule-based morphological analyzer where the rules are acquired semi-automatically from corpora. The experiment proposes an approach that takes both prefixes as well suffixes into account. Given

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 7, January 2018

an inflected Hindi word, our system returns its root form. It uses a dictionary, and a monolingual corpus to obtain suffix-replacement rules. The improvement is partially depending on the GRFL list which causes variation in result.

II. INTRODUCTION

Morphological analysis is a significant process covered under Machine Translation development for Text-to-Text (T2T) and Speech-to-Speech (S2S) model. The phase of Morph Analysis is a complex task due to autographic variation of text and speech areal variation where its challenge to decide the word within semantic context. Therefore MA is called a noteworthy process of MT under Computational Linguistics. This experiment is categorized in the two main category, the one is Fixed (रूढ़) and the other is Unfixed (यौगिक). We can simply take it as the one can be breakable with its individual meaning and the other without meaning. This paper has preferably focused on the POS category of Noun, Adjective and Verb. And how all these POS category contain some suffix with them. So what are the possible conditions, where the suffixes could be attached with words and the important description of *Sandhi rules*, is the point of discussion. In first step, we have taken three category of POS like Noun, Adjective and Adverb and picked up some most used suffix with root having ("अ", "आ", "इ", "ई", "उ", "ऊ", "ए", "ऐ", "ओ" and "औ") vowel endings. Second step, we confirm the ending status and assured the most used suffix with it having same vowel initials or final. Third step we analyze the function of concatenation are according to *Sandhi* rules are followed or violated. If, *Sandhi* rules are followed then with or without modification and the last step, if violated then find the cause of violation. Here we have picked up some prominent example to justify the conditions.

A. Corpus Design

For meeting research purpose we have made the corpus in following order which could be explained in detail. First of all the root list of the noun, adjective and verb have been prepared. Then next we have made the suffix and prefix list with all their features and their switching POS and Sex category from one to another like NN-NN, NN-ADJ, NN-VB etc.

B. Analysis

This analysis is an effort of the discussion on assembling a Morph Analyzer for Maithili. While analyzing the morphological aspects in Maithili Language, several unrevealed concepts have been disclosed to us. The discussion is also leading us towards the concerned aspects encountered, which incorporate significant role in analysis and pattern which would lead us to appropriate result. The paper have classified in various segments to understand the most relevant for analysis to justify illustrations. At the first, we have filtered the root list of noun, adjective and verb. Second we have prepared the list of affixes (prefix and suffix). Further we have checked the possible pattern available in Maithili corpus, fetched from the web. Moving to further we have filtered the all the possible words having prefix or suffix. Moving to next level we have assign the possible features of suffix and prefix that, does the word class shift its POS categories with majority or exceptionally? Initially we have categorised mainly three POS category for detail description like Noun, Adjective and Verb. The few illustrations are discussed over here to look for.

S. No.	Root	VE	Suffix/Prefix	POS Shift	Maithili Word	Hindi	Remark
Noun							
1.	गुण	--	अब	NN-NN	अबगुण	अवगुण	--
2.	मान	--	अभि	NN-NN	अभिमान/गौरब	अभिमान	--
3.	मुट्ठि	"इ"	आ	NN-NN	मुठिया	मुठिया	हाथ से पकड़ने वाला
4.	नून	"अ"	गर	NN-JJ	नुनगर	अत्यधिक नमकीन	अत्यधिक नमक
5.	छौंकी	"ई"	आएब	NN-VB	छौंकियाएब	लठियाना	बांस के पतले सीक से

Example: 1

Maithili – ओकरा अपना पर अभिमान छै।
Hindi – उसे खुद पर अभिमान है।
English – He has proud on himself.

In the above example the word "अभिमान" contains the noun POS category attached with "अभि" prefix to "मान" root incorporate the meaning of "respect" but in *Sandhi* process the word have split in the "अभि+मान" format. Here the attachment takes place on the basis of "इ+आ" vowel concatenation between prefix and root. The concatenations are based on the *Sandhi* rules.

Adjective						
पाकल	"अ"	"आहा"	JJ-JJ	पकलाहा	पका हुआ	कोई फल या सब्जी
सरल	"अ"	"आहा"	JJ-JJ	सरलाहा	सड़ा हुआ	कोई फल या सब्जी
मरन	--	"आ"	NN-JJ	आमरन	आमरण	मरने तक
नेन	"अ"	"बा"	JJ-JJ	नेनबा	--	छोटा सा
पातर	"अ"	"ई"	JJ-JJ	पतरकी	पतली	पतला सा

Example: 2

Maithili – ई सरलाहा भट्टा किया अनलहूँ?

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 7, January 2018

Hindi – ये सड़ा बैगन क्यों लाये है।

English – Why did you bring this rotten brinjal?

In the above example the word are shifting from adjective to adjective and adjective to other. The word "सरलाहा" has been broke up in "सरल+ आहा" which follows the *Sandhi* rules of "अ+आ" ending where root ending contains "अ" and suffix initial containing "आ" vowel ending.

Verb					
"अ"	"अल"	VB-NN	बानहल	बंधा हुआ	रस्सी या किसी तार से
"अ"	"ब"	VB-NN	दौगब	दौड़ना	दौड़ना
"अ"	"ई"	VB-NN	झंपनी	ढकनी	कोई वस्तु ढकने हेतु
"अ"	"हार"	VB-NN	लेनिहार	लेने वाला	लेने वाला व्यक्ति
"अ"	"औनी"	VB-NN	परहौनी	पढौनी	पढ़ाया गया

Example: 3

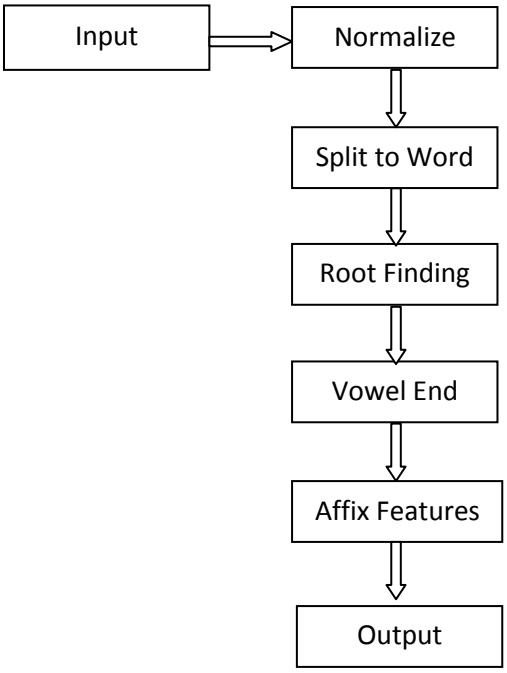
Maithili – ओकरा दौगब निक लगे छैक।

Hindi – उसे दौड़ना अच्छा लगता है।

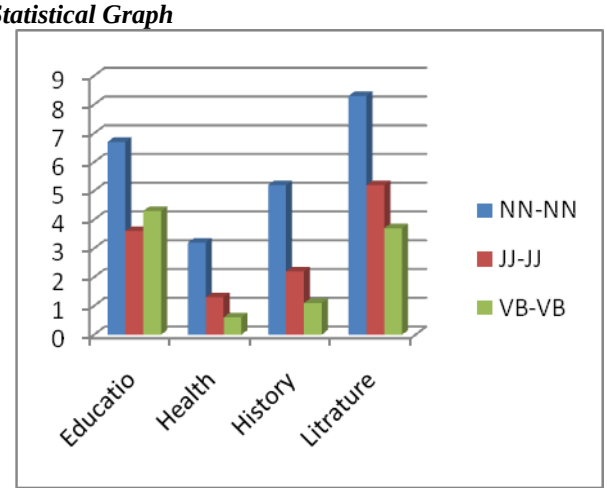
English – He loves to run.

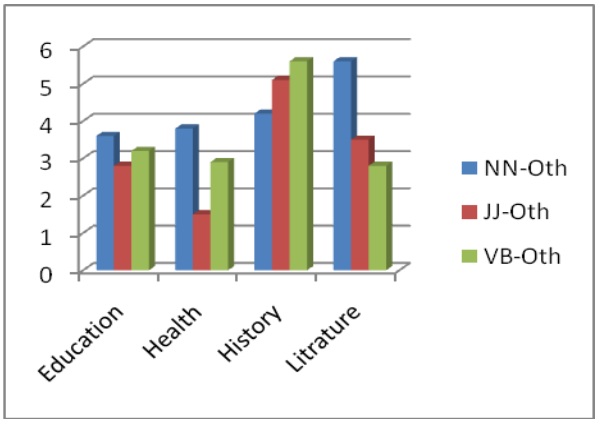
In the above examples have POS category VB as Main Verb which shift to another POS like NN. The one of the above example "दौगब" can be split like "दौग+ब". In this the vowel ending process would contain "अ" ending of "अ" initial of suffix resulted the obvious sequence. At some of the place we found such words which snag us to take appropriate decision but we are in process to overcome such issues. The above analyses have been divided in two prominent categories, the one fixed and the other is unfixed. The above analysis is based on unfixed part and fixed are counted as an exceptional till date. The exceptions are kept apart due to meaningless breakup entities and the work is under progress.

III. APPLICATION FLOW



Statistical Graph





IV. RESULT & DISCUSSION

The above analysis has stated the prominent aspects of Maithili morph analyzer currently under development for Maithili-Hindi-English Machine Translation System. The above discussed aspects are prominent enough to keep

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 7, January 2018

under consideration in support of VEA concatenation process. The graphical representation has indicated the maximum availability of suffixes concatenation among all the four domains. Therefore the research is closure to achieve satisfactory output from the initially covered of 400k words corpus and the process will continue in this direction to enhance the present current accuracy 71-74%, and the target to enhance it up to 95%.

REFERENCES

[1] Bharati Akshar, Chaitanya Vineet, Sangal Rajeev. Natural Language Processing: A Paninian Perspective. Prentice-Hall of India. (1995).

[2] Kanuparthi Nikhil, Inumella Abhilash, Misra Sharma Dipti. Hindi Derivational Morphological Analyzer (2012).

[3] O R Rinju, R R Rajeev, "Morphological Analyzer for Malayalam: Probabilistic Method Vs Rule Based Method", IJCNLP (2013).

[4] Vishal Goyal, Gurpreet Singh Lehal, "Hindi Morphological Analyzer and Generator", First International Conference on Emerging Trends in Engineering and Technology, USA, pp.1156– 1159, 2008.

[5] Niraj Aswani, Robert Gaizauskas, "Developing Morphological Analyzers for South Asian Languages: Experimenting with the Hindi and Gujrati Languages", Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valleta, Malta pp.811-815, May, 2010.

[6] Rastogi Mayuri, Khanna Pooja. Development of Morphological Analyzer for Hindi. IJCA, 2014.

[7] Malladi Deepak Kumar, Mannem Prashanth,. Statistical morphological analyzer for hindi. In Proceedings of 6th International Joint Conference on Natural Language Processing, 2013.

[8] Muley Aditi, et al., "Morphological Analysis for a given text In Marathi language", International Journal of Computer Science & Communication Network, Vol-4 (1), 13-17, 2014.

[9] Deepak Kumar Malladi and Prashanth Mannem. Context based statistical morphological analyzer and its effect on hindi dependency parsing. In Fourth Workshop on Statistical Parsing of Morphologically Rich Languages, volume 12, page 119, 2013.

[10] Agarwal Ankita, Pramila, Singh Shashi Pal, Kumar Ajai, Darbari Hemant,. Morphological Analyser for Hindi – A Rule Based Implementation. In proceeding of International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-4 Number-1 Issue-14 March-2014.

AUTHOR BIOGRAPHY

Mr. Saroj Kumar Jha, is perusing research in computational linguistics where the prime work on the development of Maithili-Hindi MT system in under the progress.

Dr. Piyush Pratap Singh is working in CILE as an Assistant Professor to promote the development of technology with Hindi and other Indian languages.

Prof. Vijay Kumar Kaul, is directing and heading the NLP activities on various languages to bridge the gap between language & technology. The research is mostly focused Hindi to make it empowered and globle.