I.J. Information Technology and Computer Science, **2016, 01, 33-39**
Published Online January 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.01.04

Extraction of Root Words using Morphological Analyzer for Devanagari Script

Sharvari S. Govilkar
Department of Information Technology, TSEC, Mumbai, India.
E-mail: sgovilkar@mes.ac.in

J. W. Bakal and Sagar R. Kulkarni
SJCOE, Mumbai, India
Department of Computer Engineering, PIIT, New Panvel, India
E-mail: bakaljw@gmail.com, skulkarni@mes.ac.in

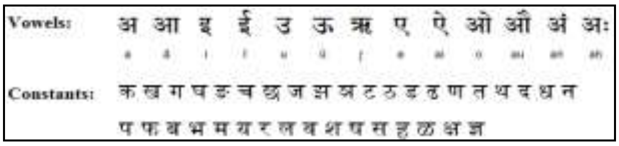
Abstract—In India, more than 300 million people use Devanagari script for documentation. In Devanagari script, Marathi and Hindi are mainly used as primary language of Maharashtra state and national language of India respectively. As compared with English script, Devanagari script is reach of morphemes. Thus the lemmatization of Devanagari script is quite complex than that of English script. There is lack of resources for Devanagari script such as WordNet, ontology representation, parsing the keywords and their part of speech. Thus the overall task of information retrieval becomes complex and time consuming. Devanagari script document always carries suffixes which may cause problem in accurate information retrieval. We propose a method of extracting root words from Devanagari script document which can be used for information retrieval, text summarization, text categorization, ontology building etc. An attempt is made to design the Morphological Analyzer for Devanagari script. We have designed CORPUS containing more than 3000 possible stop words and suffixes for Marathi language. Morphological Analyzer can acts as a preliminary stage for developing any information retrieval application in Devanagari script. We have conducted the experiments on randomly selected Marathi documents and we found the accuracy of designed morphological analyzer is up to 96%.

Index Terms—Morphological analyzer, text mining, tokenization, stop words in Devanagari, suffixes in Devanagari, stemming, removing inflections using rules.

I. Introduction

When user needs to retrieve some information from the Devanagari script document depending on the query, he/she may get some irrelevant information or may lose some important information because Devanagari script contains too many suffixes and inflections. It is also morphologically rich script.

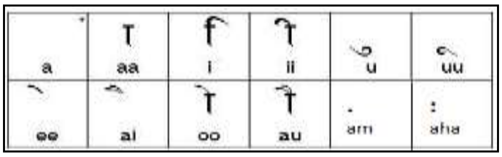
Devanagari script includes many languages such as Marathi, Hindi, Sanskrit, Prakrit etc. The Morphological analyzer operations are performed on Marathi language document. Figure1 shows that Marathi language has 13 vowels and 36 consonants. Figure2 shows the modifiers used in Devanagari script. Marathi is official language of the state of Maharashtra (India). With 300 million fluent speakers worldwide, Devanagari ranks as the second most spoken language in India and fifteenth most in the world [1].



Vowels: अ आ इ ई उ ऊ ऋ ए ऐ ओ औ अं अः
a ā i ī u ū ṛ e ai o au aṁ aḥ

Constants: क ख ग घ ङ च छ ज झ ञ ट ठ ड ढ ण त थ द ध न
प फ ब भ म य र ल व श ष स ह ळ क्ष ज्ञ

Fig.1. Vowels and Constants in Devanagari Script



a	ा aa	ि i	ी ii	ु u	ू uu
े ee	ै ai	ो oo	ौ au	ं am	ः aha

Fig.2. Modifiers used in Devanagari Script

When one wants to work on information extraction, text summarization, text mining, retrieval based on ontology, the common problem arises for all these applications is to find exact information for the user depending on the query given. If the query word is present in the document then the information / sentences are retrieved related to the query term. When the system compares query word with the input Devanagari script document then even though the term present in the document, mismatch may occur because of suffixes and inflections. These suffixes and inflections in Marathi makes the task of information retrieval, text mining very complicated. The solution to this is to extract root words by removing suffixes and inflections of the word. This process is called as Morphological Analysis. Performing stemming and removing inflections of the word is very important task to retrieve relevant information from the collection of document. There are variety of languages which uses Devanagari script such as Marathi, Hindi, Sanskrit and Prakrit etc. It is easy to extract root words by applying stemming algorithm for language like English. As Devanagari script is morphologically rich, stemming, Suffix removal and inflection striping is complex. For keyword based information retrieval, root words plays vital role in improving the performance of retrieval system. As Devanagari script is morphologically rich script which contains many suffixes and inflections in the words, retrieving accurate information becomes complex task. For effective information extraction there is a need to extract root words from input document. This problem demonstrates necessity of morphological analyzer.

We have performed the experiment on Marathi language in which input is randomly selected Marathi document and output is series of root words. We have designed CORPUS containing more than 3000 possible stop words and suffixes for Marathi language which are frequently used in Devanagari script. The CORPUS plays very important role for filtration of input document. Performing stemming and removing inflections of the word is very important task to retrieve relevant and detailed information from the document.

Example: In information retrieval system if we want to search information about "भारत", then query will not result all the sentences from document containing words like " भारताची, भारतासाठी, भारतामध्ये, भारताने" due to suffixes "ची, साठी, मध्ये, ने" and inflection of 'आ' in 'ता' in above words. This problem can be overcome by using Morphological operation on the each term to get actual root word. The root word for all the above words is "**भारत**" Once this root word is found the information retrieval becomes easy because the query term will match with all the above words and hence results all the sentences containing this term. Thus there is need to find out the root words from the document so that retrieval system will provide relevant information.

There has been a significant improvement in the research related to Devanagari script document. In recent year's research towards Indian languages is getting increasing attention. Our proposed architecture can be used to design Morphological analyzer for any language.

II. Literature Review

The morphological analysis for Devanagari script document requires many pre-processing stages such as tokenization, keyword recognition, stop word removal, stemming, removing inflections from the word etc. In India there is very less work has been reported in literature on Devanagari script. Extraction of root words is the preliminary task for any natural language processing activity. The Lemmatization and stop-word elimination are well studied for English and a few European Languages. Also there is no work done on the validation of script using UTF-8 as in [8]. Even, lexical analysis such as stemming for Marathi is not used in the modern and popular search engines such as Yahoo and Google. The stop word removal is very important task as it doesn't contribute much in information retrieval process. This stop word removal system has already been implemented for English language. As discussed by Manish Shrivastava, Pushpak Bhattacharya in [1] and Ashish Almeida, Pushpak Bhattacharya in [2], the inclusion of suffixes in indexing and stop-words elimination effect on the retrieval performance. An important observation is that the suffixes in Marathi language can also contribute to the semantics of the document and hence improves the retrieval performance by removing all suffixes from the document. The removal of inflectional suffixes are not possible by normal stemming operation so there is need of stemmer which is used to remove all the possible suffixes from the keyword and gives word stem. According to Upendra Mishra, Chandra Prakash as given in [3] the Maulik stemmer is purely based on Devanagari script (Hindi) and it uses the Hybrid approach (combination of brute force and suffix removal approach). In [4] the author evaluates a rule-based and an unsupervised Marathi stemmer. The rule-based stemmer uses a set of manually extracted suffix stripping rules whereas the unsupervised approach learns suffixes automatically from a set of words extracted from raw Marathi text. To detect suffixes automatically using unsupervised approach, Marathi WordNet is required which is not available for public use. Character recognition is very important for validation of script. The author in [7] uses UTF-8 provided by Unicode organization as in [8] for character recognition for Hindi script. The paper represents light stemmer which removes all of these suffixes, the longest suffix first. The list of 27 common suffixes is used in this paper. For Devanagari there are too many suffixes possible that may occur in the document. The CORPUS for all possible Stop words and Suffixes in Devanagari is not available for public use. We have designed rules for inflectional suffix stripping operation to achieve desired output as a root words.

Availability of Resources:

The stop words are most frequently occurring words in Devanagari script document and carry no meaning for information retrieval system. The suffixes are always attached with the root words. The presence of suffixes may degrade the retrieval performance. For effective morphological analysis there is need of Stop words and Suffix corpus. Unfortunately these resources are not available for public use.

III. Proposed Architecture

The objective of this paper is to extract root words from Devanagari script document using Morphological Analyzer. Very less work has been done in India on Devanagari script due to unavailability of resources such as WordNet, Ontology and Corpus etc. Unless we extract root words from the input document, we cannot achieve effective result because retrieval system requires exact match of query word with words in the input document. The presence of such words degrades the matching query term efficiency.

We have developed Morphological Analyser for Devanagari script which gives root word by removing suffixes and inflections of the words. After extracting root words the query term can be easily compared with root words from the input document so as to generate relevant result in information retrieval system. This system can acts as preliminary stage for all information retrieval system such as Text mining, Text summarization, Categorization etc.

The input document is in Marathi language. First of all the input document is filtered for the purpose of removing all the special characters. The input document is then scanned to validate whether the input contains any other script or not. If so we will eliminate those irrelevant characters to maintain pure Devanagari script document. This can be done by using UTF-8. The pure Devanagari script document is then tokenized so that we can deal with each keyword separately. The customized corpus is used for removing stop words and suffixes from the document. This step is important as stop words are most frequently occurring words in Devanagari script and carry no meaning for keyword based information retrieval system. Suffix removal lead to achieve stem of word which will be further used for extracting actual Root word. The rules are created to remove inflections of the word so that user can get accurate root word.

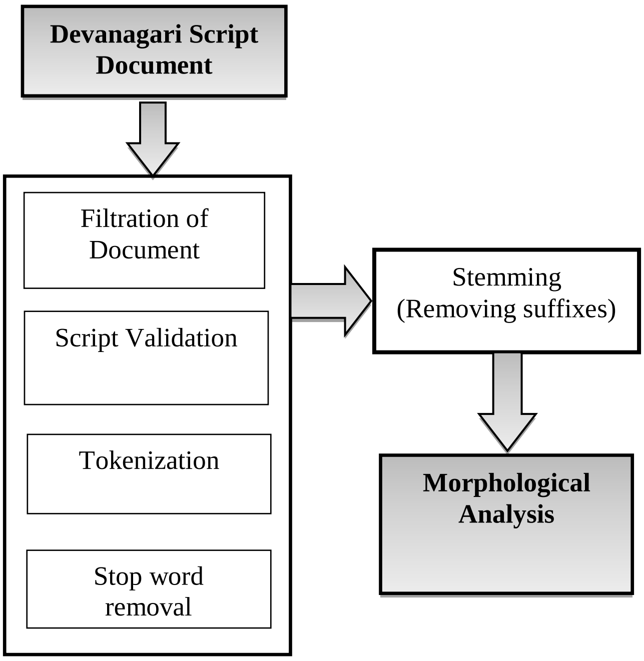


Fig.3. Morphological Analyzer for Devanagari Script

The input to the system contains randomly selected Devanagari script documents especially in Marathi language. The following steps are used for extracting root words from Devanagari script documents.

Step 1: Filtration of Devanagari Script

As presence of special characters in Devnagari documents degrades the performance, it needs to be removed. This removal of special characters from Devanagari script is called as filtration of document. Token creation of special characters and its recognition with UTF-8 is time consuming which leads to memory wastage. The special characters such as “ ” ‘ ’, . / ? [ ] { }: ; \ | ~ ! @ # $ % ^ & * ( ) _ - = + < > are frequently used in many language scripts. These characters will not contribute towards final result.

Step 2: Script validation

Keyword recognition is very important stage in text mining because the resultant information is totally depends on the language and nature of query supplied to the system. The input document may contain some words or sentences in other script or language. Here we are analyzing whether the input document is in Devanagari script or not. The words which are not valid to Devanagari script are simply removed from further processing. To perform this operation we have used Unicode values called UTF-8[8] for Devanagari script document. We compared UTF-8 list with each character of each token, if match found the character is valid and allowed otherwise removed from the document. The aim of this phase is to maintain pure Devanagari script document as input to Morphological Analyzer.

Algorithm for Validation of Input document:

1. Apply filtration algorithm. If already applied then ignore this step.
2. Use the character set as UTF-8
3. Scan the input document.
4. Compare each character from scanned input document with UTF-8.
5. If character is present in the UTF-8, then it is valid to Devanagari script otherwise not.
6. Ignore all the invalid Devanagari script characters.
7. Repeat step 3 till all characters from input script document get verified.
8. Store all the valid Devenagari character, words in file to process further.

Step 3: Tokenization

The pure Devanagari script document is passed through tokenization to get valid tokens which can be also called as Lexicons. With the help of lexical analyzer one can tokenize the input document as one token per line. Here space is used to generate tokens.

Step 4: Stop word removal

Stop words are the most frequently occurring words within the collection of document and thus they have very little discriminatory value. Stop words represent noise, and may take more time on processing and reduce overall retrieval performance. They tend to create huge posting lists which take up lots of disk space and degrade the performance of retrieval .Thus, it is usual practice to identify and eliminate stop-words in the process of searching. While searching for particular keyword using query, the system may include all the records/tokens of the input document for the process of searching regardless of their relevance.

The stop word makes up a large portion of the text in the document. Thus it is required to remove such stop words from index to save the searching time, and also to enhance searching performance. We have designed Corpus of all the possible stop words that may occur frequently in the Devanagari script especially in Marathi language. Table1 shows few examples of stop words that are normally used in Marathi language.

Table 1. Examples of Stop words used in Devanagari Script (Marathi)

असं	अथवा	अशा	असा
आणखी	आणि	आत्ता	अरेच्चा
अरेपण	इतर	इथे	इथं
इतके	इतक्यात	इत्यादी	एवढ
किंवा	कारण	की	केंव्हा
जे	जेथे	ज्याच्या	जेंव्हा
जी	जिथले	जिथे	जिच्यामुळे
तरी	तर	तरच	त्याच्या
त्यांना	तेंव्हा	नेमकी	परंतु
त्याचा	बरं	बापरे	म्हणजे
पण	मध्ये	सध्या	व
मधे	वारंवार	ह्यावर	ह्यांच्यावर
वरचेवर	याच्यासाठी	याच्यात	याच्यातच
यात	यांचे	यांसाठी	यांहून
यावर	या	त्यामुळे	त्यामुळेच

Step 5: Stemming

Suffix stripping is an important step required in a number of natural language processing applications such as information retrieval, text summarization, document clustering etc. The widely used method for this processing is Stemmer which uses a suffix list to remove suffixes from words. The stem is not necessarily the linguistic root of the word. We have designed Corpus of all possible suffixes that occur frequently in the Devanagari script. The corpus is used to remove suffixes from input document.

The result of stemming is stem of word that can be given as input to Morphological Analyzer for further processing. The observation is made on the result of Stemming and it is found that stem of word normally contains inflections. The inflections in the stem word cannot be removed using simple stemming operation. To do this we must have some standards which will easily deal with inflections of the word. Table2 shows few examples of suffixes that are normally used in Marathi language.

Table 2. Examples of Suffixes in Devanagari Script (Marathi)

च्या	न्या	ऱ्या	ऱ्या
ख्या	वर	साठी	पासून
तून	कडून	कडून	मुळे
साठी	प्रमाणे	वरून	वर
च	चे	तील	पर्यंत
चा	कडेही	नी	नीही
मधले	तही	तपण	तसुद्धा
कडे	पाशी	पर्यंत	पूर्वक
तच	हून	प्रमाणे	लेल्या

Step 6: Morphological Analyzer

The aim of morphological analysis is to recognize the inner structure of the word. A morphological analyzer is expected to produce root words for a given input document. Devanagari script is morphologically rich language in which the case markers and postposition markers are usually manifested as suffixes. The root and stem of word may differ in their forms. The words after stemming are analyzed to check whether they are inflected or not. This can be done by creating and comparing rules with the words. If stem word is inflected then the root word is formed by addition of replacement characters with stem word.

The rules are formed to find actual root word. The system searches the perfect match from the set of rules and if the keyword after stemming called stem of word contains any inflection then those inflections are removed from the keyword. There is a need to design some standard set of rules which will enable the system to process the stem of words and find the actual root word.

Rule Format:

List of Characters → Replacement Character. e.g. मा मी मु मे → *म्*. The meaning of this rule is whenever the word ends with “ मा मी मु मे” or has inflection “ आ ई उ ए” are replaced by the character “ म” with inflection “ अ” .

We can use following data to get the root for inflected word:

1. List of all the possible suffixes.
2. Rules for inflected words to be replaced by another character.
3. The replacements characters to be made after removal of suffix so that valid root can be formed.

IV. Experiments and Results

Input to System:

The input for the system contains Devanagari script document in Marathi language.

“सागर मराठीमध्ये शब्दांचा मूळ शब्द शोधण्याचे काम करत आहे. मराठीतील मूळ शब्दाचा उपयोग करून कोणतीही मराठी माहिती सुलभपणे मिळवता येईल. आजही मराठीमधला शब्दकोश संशोधनासाठी उपलब्ध होणे कठीण आहे. त्यामुळेच, भारतामध्ये देवनागरी लीपीवर खूपच थोड्या प्रमाणात संशोधन झालेले आहे. (**End of Input Text**).”

Filtration:

When this input is given to our system the filtration will remove all the special characters from the system as they are not part of further processing.

Script validation:

In this step, the characters which are not valid to Devanagari script are simply removed from the input. This task is done using UTF-8. Here the output we obtain is pure Devanagari script. The English characters in given example is removed from the input to maintain pure Devanagari script document.

Tokenization:

The output generated after keyword recognition is fed to tokenization process where the input Devanagari script is tokenized to ease the further processing. The tokenization is done by detecting spaces between the keyword i.e. when space is reached the token get formed. All the tokens are then given to stop word removal process.

Stop word Removal:

The system finds all the possible stop words from the input by comparing it with Corpus of Stop words we have designed. The matched stop words are ignored from further process as they don’t carry any meaningful information. Table 3 shows few stop words and their occurrences found in the given input script.

Table 3. Stop words Found in the Input Devanagari Script.

Stop words Found	Example of Occurrence
त्यामुळेच	त्यामुळेच भारत…….
कोणतीही	कोणतीही माहिती……

Stemming:

This step plays an important role as it removes all the possible suffixes from the script. Suffixes are those characters that normally appear at the end of words. The suffixes are of two types, plain suffixes and complex suffixes. The system identifies all those suffixes by comparing it with the Corpus of suffixes we designed. The Corpus contains all the possible suffixes used in Devanagari script document for Marathi language. Table4 shows possible suffixes found in the given input.

Table 4. Suffixes Found in the Input Devanagari Script.

Suffixes Found	Example of Occurrence
च	खूपच
ही, ले	कोणतीही, झालेले
मध्ये, मधला	*भारतामध्ये*, मराठीमधला
वर	लीपीवर
चा, चे	*शब्दांचा*, शोधण्याचे
ता, तील	मिळवता, मराठीतील
पणे	सुलभपणे
साठी	संशोधनासाठी

The output after stemming contains Stem of words which may have many inflections.

Morphological Analyzer:

The inflections present in the stem of words can be removed by using Morphological analyzer. The analyzer uses the rule based approach to remove inflections from the stem word. Table 5 shows some examples of Rules used for given input script.

Table 5. Examples of Rules used in Morphological Analyzer.

List of Characters	Replacement Character
ता ति ती तु तू ते तै तो तौ	*त*
ना नि नी नु नू ने नै नो नौ	*न*
णा णि णी णु णे णो ण्या	*ण*
ला ली लां ले	*ल*
षे षां	*षा*

The output of the morphological analyzer is divided into three columns such as Original word, Stem of word and Root word as shown in Table 6. The stop words found in the input script are indicated as S_W.

Table 6. Result of Morphological Analyzer

Original Word	Stem Word	Root Word
सागर	सागर	सागर
मराठीमध्ये	मराठी	मराठी
शब्दांचा	शब्दां	शब्द
मूळ	मूळ	मूळ
शब्द	शब्द	शब्द
शोधण्याचे	शोधण्या	शोधण
काम	काम	काम
करत	कर	कर
आहे	आहे	आहे
मराठीतील	मराठी	मराठी
मूळ	मूळ	मूळ
शब्दाचा	शब्दा	शब्द
उपयोग	उपयोग	उपयोग
करून	करून	करून
कोणतीही	S_W	S_W
मराठी	मराठी	मराठी
माहिती	माहिती	माहित
सुलभपणे	सुलभ	सुलभ
मिळवता	मिळवता	मिळवत
येईल	येईल	येईल
आजही	आज	आज
मराठीमधला	मराठी	मराठी
शब्दकोश	शब्दकोश	शब्दकोश
संशोधनासाठी	संशोधना	संशोधन
उपलब्ध	उपलब्ध	उपलब्ध
होणे	होणे	होणे
कठीण	कठीण	कठीण
आहे	आहे	आहे
त्यामुळेच	S_W	S_W
भारतामध्ये	भारता	भारत
देवनागरी	देवनागरी	देवनागरी
लीपीवर	लीपी	लीपी
खूपच	खूप	खूप
थोड्या	थोड्या	थोड
प्रमाणात	प्रमाणा	प्रमाण
संशोधन	संशोधन	संशोधन
झालेले	झालेले	झालेल
आहे	आहे	आहे

V. CONCLUSION

Information Retrieval from Devanagari script document needs extract root words to do further processing. An attempt is made to design the Morphological Analyzer for Devanagari script. There are many factors that may affect the performance of the system for Devanagari script. Stemming alone cannot find the relevant information if the words in the document have more inflections. Inflections in word may degrade overall performance of search. Accuracy of Morphological analyzer is totally depends on how effectively one can generate rules for eliminating inflections from the word.

The proposed morphological analyzer acts as a preliminary step to achieve relevant output for the applications like text mining, text summarization, semantic Information retrieval based on ontology etc. by removing suffixes and inflections of the string. Our proposed approach will minimize inflections of words so that the further task will become easy for retrieving desired information. The research towards regional languages is increasing day by day. There is a large scope to design the complete resources for Devanagari script such as WordNet, Ontology and Corpus etc. to achieve better result in information retrieval applications.

REFERENCES

[1] Pushpak Bhattacharya, Manish Shrivastava, Nitin Agrawal, Bibhuti Mohapatra, Smriti Singh, IIT Bombay "Morphology Based Natural Language Processing tools for Indian Languages" 2012.

[2] Ashish Almeida, Pushpak Bhattacharyya IIT Bombay "Using Morphology to Improve Marathi Monolingual Information Retrieval" IEEE 2012.

[3] Upendra Mishra, Chandra Prakash, "MAULIK: An Effective Stemmer for Hindi Language" International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 Vol. 4 No. 05 May 2012.

[4] Mudassar M. Majgaonker, Tanveer J Siddiqui, Discovering suffixes: A Case Study for Marathi Language, International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2716-272.

[5] Deepak Kumar, Manjeet Singh, and Seema Shukla "FST Based Morphological Analyzer for Hindi Language", JSS Academy of Technical Education Noida, Uttar Pradesh, India, 2010.

[6] Dr. Riyad Al-Shalabi, Dr. Ghassan Kanaan, Dr. Ahmad Hasnah "Stop word removal algorithm for Arabic language", IEEE 7803-8482-2/2004.

[7] Leah S. Larkey, Margaret E. Connell, Nasreen Abduljaleel, "Hindi CLIR in Thirty Days", University of Massachusetts, Amherst. ACM Transactions on Asian Language Information Processing, 2003, 2(2), pp. 130-142.

[8] http://www.unicode.org/charts/PDF/U0900.pdf for UTF-8 Unicode's used in Devanagari.

[9] http://www.unicode.org/Public/6.1.0/charts/CodeCharts.pdf_contains more than 200 scripts Unicode's and their ranges used throughout the world.

[10] http://www.cfilt.iitb.ac.in/indowordnet/index.jsp___Center for Indian language technology (CFILT), by IIT Bombay.

[11] http://ltrc.iiit.ac.in/analyzer/marathi/all_out by IIIT, Hyderabad.

Authors' Profiles

Sharvari Govilkar is Associate professor in Computer Engineering Department, at PIIT, New Panvel, University of Mumbai, India. She has received her M.E in Computer Engineering from University of Mumbai. Currently She is pursuing her Ph.D. in Information Technology from University of Mumbai. She is having eighteen years of experience in teaching. Her areas of interest are Text Mining, Natural language processing, Information Retrieval, domain specific ontology construction etc.

Dr. J. W. Bakal received M.Tech in Electronics Engineering, from Marathwada University. Later, He has completed his Ph.D. in the field of Computer Engineering from Bharati University, Pune. He is a PhD supervisor in CSE at University of Mumbai. He is presently working as principal at the S.S. Jondhale College of Engineering, Thane, India. He was a chairman of board of studies in Information Technology in University of Mumbai. His research interests are Telecomm Networking, Mobile Computing and Information Security. He has publications in journals, conference proceedings, and books in his credits. During his academics tenure, he has attended, organized and conducted training programs in Computer and Electronics branches. He is life member of professional societies such as IETE, ISTE INDIA. He is also a member of IEEE. He has prominently worked for IETE as a chairman, Mumbai section.

Sagar Kulkarni is Assistant Professor in Computer Engineering department at PIIT, New Panvel, University of Mumbai, India. He has completed M.E. in Computer Engineering from University of Mumbai, India. Sagar has received BE in CSE from Shivaji University, Kolhapur. He is having Eight years of experience in teaching. His area of interest are Text Mining and Summarization, System Programming and Compiler construction, Natural Language Processing, Information Retrieval etc.

How to cite this paper: Sharvari S. Govilkar, J. W. Bakal, Sagar R. Kulkarni,"Extraction of Root Words using Morphological Analyzer for Devanagari Script", International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.1, pp.33-39, 2016. DOI: 10.5815/ijitcs.2016.01.04