IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Rule based Stemmer for Marathi Language

N. Pise[1*], V. Gupta[2]

[1*]Computer Science Department, IES, IPS Academy, Indore, Madhya Pradesh, India
[2] Computer Science Department, Banasthali University, Rajasthan, India.

*Corresponding: nikitapise97@gmail.com, Tel.: +91-9669929694

Available online at: www.ijcseonline.org

Accepted: 24/May/2018, Published: 31/May/2018

***Abstract*—** Natural Language Processing (NLP) is a branch of artificial intelligence which deals with the analysis and synthesis of natural languages in the form of text and speech. NLP requires stemming algorithms to remove derivational and inflectional affixes without performing morphological analysis of the inputs. These algorithms are essential to extract root or stem words. The goal of stemming is to reduce word forms/grammatical forms to their root forms. To accomplish, specific knowledge of language is required. In NLP, the stemmer can be used to improve the efficiency of text summarization, text mining, information retrieval and sentiment analysis. In this paper, we proposed a rule based stemming approach for Marathi language using Marathi corpus, stopword list and suffix stripping rules.

***Keywords*—**natural language processing, stemming, corpus, marathi, suffix stripping and stopwords.

I. INTRODUCTION

Stemming is an important feature supported by Natural Language Processing (NLP) systems, Information Retrieval (IR) and Text mining [1]. Stemming is usually done by removing attached affixes from the root word. The main purpose of stemming is to reduce inflectional forms and derivational forms of a word to a base form. To do this various stemming algorithms have been developed. Each algorithm attempts to convert the morphological variants of a word to get mapped to the root or stem word. Marathi language exhibits high level of morphological variations. For example, a single root word like महाराजा can have various morphological variants like महाराजाचा, महाराजाची, महाराजासाठी, महाराजावर, महाराजाकडे [9].So the key terms of sentences or document are represented by their stem words rather than their original words to reduce the total number of distinct words in sentences or documents which results in reducing processing time of the final output. Stemming plays an important role in Information Retrieval System (IRS) for improving the performance of all languages. The goal of stemming is to diminish inflectional and derivational variant forms of a word to a common base form. A stemmer can execute operation of transforming morphologically identical words to root word without performing morphological analysis of that term[15]. This makes stemming an attractive preference to raise the ability of matching query and document vocabulary in information retrieval system. Many Natural languages like Dravidian languages (Tamil, Telugu, Malayalam and Kannada), Indo Aryan languages (Hindi, Bengali, Marathi, Gujarati) searching quality is increased because of using stemming algorithm. Thus various stemming are developed for various languages, but each one has its own advantages as well as limitations. Most of the stemming algorithms are language dependent. So there is an urgent need to develop language independent stemming algorithm to increase the searching efficiency [5][11].

Stemming is an important approach which is used in various applications of Natural Language Processing. Stemming requires syntactic and semantic knowledge about languages to extract root words or stem words by removing affixes from the given word. A lot of research has been done for many languages but there are very less number of approaches for Marathi Language. So here is a rule based approach proposed for Marathi. Some suffix stripping rules has been given to describe the separation affixes from the input words. There are three flow charts which describes the overall stemming algorithm. We have evaluated the results and found the efficiency of the proposed algorithm was quite high. Lemmatizer with more contextual information can also be used to increase the performance of root word extracting systems.

II. RELATED WORK

Stemming is a process to convert a word to its root word. It is also can be said as a process of operation in Natural Language processing usage. Stemming is very useful in information retrieval. Stemming is used as a convertor from differential word to root word. In this process, a single word means its different type of grammatical approaches. There are five stemming operation methods-

1) A combination of matching the string and the most frequent differential suffices model; also called as edited distance on dictionary algorithm.
2) A process focused on finite state automata also called as analyzer of morphology.
3) For the retrieval of a differential form of a lemma in possible format which focus on "radix trie" data structure.
4) The combination of rule-based and supervised training approach is known as Affix stemmer.
5) Fixed length truncation approach, NLP is use to give meaning of human languages to computers. It can make a drastic change in linguistic activities. By applying various rules and set skill different forms can be converted to their stem which can be called as Stemming. Stemming converts the word into stem without changing POS. Its usage is in text mining. The stem is not compulsory to be a dictionary word. The root word can only be derived from the rule based approach but its efficiency can be varied because it needs a lot of space for storing the rules. So by combining all the rules an approach statically can give more accurate results. When a variety of languages different grammar and different rules is used then edit distance is the perfect method for usage. Sometimes data structures can also becomes the perfect solution like radix tree. It can match prefix to find the most correct lemma of the given word. [8].

Stemming is a pre-processing step of text mining and also a major requirement of natural language processing functions for information retrieval. Many other different algorithms of stemming are also there, first is Truncating method, in which suffixes and prefixes of a word are removed whose rules are described as Lovins, Porters, Paice/Husk and Dawson. Second is Statistical method, in which stemming is based on statistical analysis and techniques whose procedures are classified according to their rules as N-Gram, HMM and YASS. Third is Inflectional and Derivational method, in which the word variants are related to language specific syntactic variations and Parts of Speech of a sentence whose rules are classified as Krovetz and Xerox Inflectional and Derivational Analyzer. Fourth is Corpus Based method, which is modified version of Porters rule. Fifth is Context Sensitive method, in which context sensitive analysis is done using statistical modeling on the query side. It concludes that none of them give 100% output but are good enough to be applied to the text mining, NLP or IR applications. Over-stemming is when two words with different stems are stemmed to the same root. This is also known as a false positive. Under-stemming is when two words that should be stemmed to the same root are not. This is also known as a false negative. Light-stemming reduces the over-stemming errors but increases the under- stemming errors. On the other hand, heavy stemmers reduce the under-stemming errors while increasing the over-stemming errors. The different stemming for both Indian and non -Indian language accuracy and errors was found by Bijal et.al [9]. A new improved light stemming algorithm proposed by Thangarasu et.al [6] for less computational steps which is used to get good stemmed Tamil words. Also uses K-means clustering for the performance of Tamil language. A new stemmer "MAULIK" was proposed for Hindi language by Mishra et.al [13] using Devnagari script and hybrid approach. Anjali Ganesh Jivani [8] has discussed various methods of stemming and their comparisons in terms of usage, advantages as well as limitations. The fundamental difference between stemming and lemmatization is also discussed. Perhaps developing good lemmatizer could help in achieving the goal. But this paper does not deal with recently developed stemming algorithms. A new light stemming technique was introduced and compared this with other stemmers to show the improvement of search effectiveness in Arabic language by Mohamad et.al. [11] An approach for finding out the stems from the text in Bengali was presented by Das et.al. In this paper, they maintained two different hash tables, first one containing all possible nominal inflections and the second one containing all possible verbal inflections for Bengali language. An unsupervised approach for the development of stemmer in Urdu and Marathi language had been presented by Husain. And frequency based and length based stripping are proposed for rule generation. But author was using N gram method. N gram approach requires large memory size. In this research the approach is based on different stemming algorithms for different language [16].Stemming and Lemmatization: A Comparison of Retrieval proposes to compare document retrieval precision performances based on language modeling techniques, particularly stemming and lemmatization.

III. PROBLEM IDENTICATION

There are mainly two errors in stemming are – over stemming and under stemming[12].

A. Over-stemming:

Over-stemming is when two words with different stems are stemmed to the same root. This is also known as a false positive. For example words कहावत and कहानी both are reduced to the word कहा after stemming. Although these two words exhibit different meaning, still they are reduced to same root word कहा which in English means to speak.

B. Under-stemming:

Under-stemming is when two words that should be stemmed to the same root are not. This is also known as a false negative. For example in word शिवाजीीचा, if only suffix चा is removed then observed root word will be शिवाजीी but actual root word is शिवाजी. Paice has proved that light-stemming reduces the over-stemming errors but increases the under- stemming errors. On the other hand, heavy stemmers reduce the under-stemming errors while increasing the over-stemming errors.

IV. COMPONENTS

A. Some suffix stripping rules:

There are 50 suffix stripping rules. Some of them are given below-

- If word having length greater than 5, has suffix 'साठी' then remove the suffix which gives stemmed output.
- If word havi9ng length greater than 4, has suffix 'ाच्या' then remove the suffix which gives stemmed output.
- If word having length greater than 3, has suffix 'ात' then remove the suffix and add 'णे' which gives stemmed output.
- If word having length greater than 2, has suffix 'ावी' then remove the suffix and add 'णे' which gives stemmed output.
- If word having length greater than 1, has suffix 'ी' then remove the suffix which gives stemmed output.

B. Some stopwords from the database:

Stop words are those words that occur frequently. For English stemmers, a stop word list is already maintained, similarly, it is necessary for Marathi stemmer, that there must be a stop word list. Therefore, to accomplish this task, various Marathi books and literature were studied to find out stopwords in Marathi language. There are total 205 stopwords in our database. Some of the stopwords in Marathi are given below-

Table 1.Stopwords

Stopwords (Database1)
काही(kahi)
ती(ti)
असे(ase)
म्हणून(mahnun)
याच्या(yachya)
मी(mi)
पण(pann)
नाही(naahi)
आहे(aahe)
होत(hot)
आपल्या(aaplya)
करून(karun)
ते(te)
आणि(aani)

C. Some words with their stem words from the database :

In the second database there are 40,000 words corresponding to their stem words. Some of them are given below:

Table 2.Words with their stem words

Words	Stem words (Database2)
आभाळाचे (*aabhalache*)	आभाळ (*aabhal*)
जरासे (*jarase*)	जरा (*jara*)
कथ्नाच्या (*kathnachya*)	कथन (*kathan*)
लोकच्(*lokach*)	लोक(*lok*)
महाभारतात(*mahabharataat*)	महाभारत(*mahabharat*)
नव्सात (*navsaat*)	नवस (*navas*)
नव्रदेवाप्रमाणे(*navrdevapramane*)	नवरदेव(*navardev*)
पिंडाची(*pindachi*)	पिंड(*pind*)
संरक्षणासाठी (*sangrakshnasathi*)	संरक्षण(*sangrakshn*)
समाजाला(*samajala*)	समाज(*samaj*)
तोंडे(*tonde*)	तोंड(*tond*)
दासबोधासारखा(*dasbodhasarkha*)	दासबोध(*dasbodh*)
द्वेषाने(*dveshane*)	द्वेष(*dvesh*)
युतीयोगाचे(*yutiyogache*)	युतियोग(*yutiyog*)

V. METHODOLOGY

We proposed a rule based stemmer for Marathi language which uses certain rules to reduce the under and over stemming problems and gives more efficient output. Developed stemmer contains following steps to get stemmed output. This process consists of two modules-Pre-processing module and stemming module. Figure given below represents our proposed rule based stemmer:

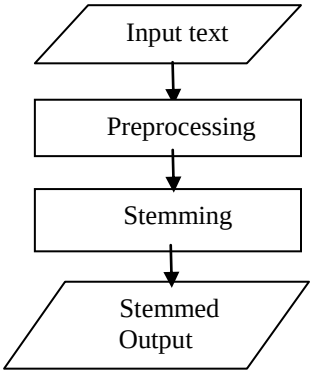


Figure1. Rule based approach for Marathi language

Example: Stemmed output text corresponding to given input text for rule based Marathi stemmer-

Input text- माझी शाळा (school) स्टेशनच्या जवळ आहे
(majhi shala (school) stationchya javad aahe)

Stemmed output text- माझी शाळा स्टेशन जवळ आहे
(majhi shala station javad aahe)

A. Pre-processing module: Pre-processing is converted into two parts–filtration and tokenization. Given figure is showing pre-processing of input text:

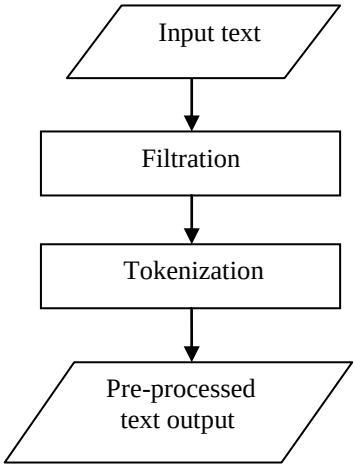


Figure2. Pre-processing module

Filtration is the process of removing non Devnagari Unicode characters like '@', '$', '%', '#', '^' etc. gets eliminated from the text by excluding some special characters like '-' and '_' those are frequently being used in Marathi sentences. This filtered output becomes input for the further processing. The next is tokenization which considered as crucial process in NLP. Tokenization is process of splitting the text into smaller parts called tokens. Paragraphs get split into sentences and then sentences get split into individual word. Spaces between the words in the sentences considered as parameters to split them.

Input text: माझी शाळा (school) स्टेशनच्या जवळ आहे
(majhi shala (school) stationchya javad aahe)

Filtrated output: माझी शाळा स्टेशनच्या जवळ आहे |
(majhi shala stationchya javad aahe)

Tokenized text:

Token 0: माझी (*majhi*)

Token 1: शाळा (*shala*)

Token 2: स्टेशनच्या (*stationchya*)

Token 3: जवळ (*javad*)

Token 4: आहे (*aahe*)

B. Stemming module: The proposed rule based stemmer extract the root or the stem words by applying suffix stripping rules on the suffixes having maximum length. The overall process done in the stemming module is shown:

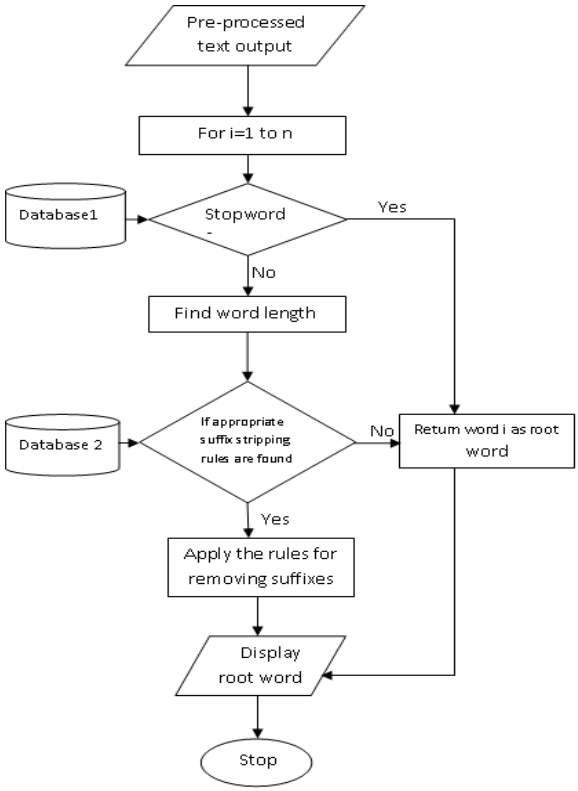


Figure3. Stemming module

Pre-processed input and stemmed output for rule based Marathi stemmer.

Input Text:

Token 0: माझी (*majhi*)

Token 1: शाळा (*shala*)

Token 2: स्टेशनच्या (*stationchya*)

Token 3: जवळ (*javad*)

Token 4: आहे (*aahe*)

Output Text:

Token 0: माझी (*majhi*)

Token 1: शाळा (*shala*)

Token 2: स्टेशन (*station*)

Token 3: जवळ (*javad*)

Token 4: आहे (*aahe*)

Stemmed output: माझी शाळा स्टेशन जवळ आहे

(majhi shala station javad aahe)

VI. RESULTS AND DISCUSSION

For evaluating the performance of the stemmer identification of correct stemmed outputs for given input is required.

Accuracy of the algorithms is calculated by the formula given below:

Accuracy = (Correct output ÷ Total Input) ×100%

We have performed evaluation on 350 words and then we got 86% accuracy.

VII. CONCLUSION AND FUTURE SCOPE

Stemming plays a dynamic role in information retrieval system and its effect is very huge, related to that analysis on various stemming algorithms. Stemming algorithm is also useful in dropping the size of index files as the number of words to be indexed are reduced to common forms or so called stems. Some stemming algorithms reaches better in some area, other reaches better in some other area. So that in future, researchers will go for more number of executions for the stemming algorithm methods and their benefits for various Indian and Non-Indian languages information retrieval system. A lot of stemming algorithms are exiting for many languages there still remain a lot to be done for improving the accuracy of the output. There is a need for an approach and a system which increases the efficiency by removing over and under stemming in a better way. Development of lemmatizer with better understanding of context of the word in the sentence can help in getting effective outputs.

ACKNOWLEDGMENT

I would like to express my sincere thanks to Dr. Archana Keerti Chowdhary, Principal of IPS Academy, Indore for providing me necessary guidance and support to carry out experiment also thanks to Dr. Namrata Tapasvi, H.O.D. of CS Department for providing research facilities and for constant technical and moral support.

REFERENCES

[1]Ciravegna F, Harabagiu S, "Recent Advances in Natural Language Processing".IEEE,2013.

[2] Garje, G. V., & Kharate, G. K. "Survey of machine translation systems in India." International Journal on Natural Language Computing (IJNLC) Vol, 2, 47-67, 2013.

[3] Hovy, E., & Lin,C.Y., "Automated text summarization and the SUMMARIST system". In Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998. Association for Computational Linguistics, (1998, October).

[4] Lin, C. Y. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop (Vol. 8), (2004, July).

[5] M. Kasthuri and S. B. R. Kumar "A Comprehensive Analyze Of Stemming Algorithms For Indian And Non-Indian Languages" International Journal of Computer Engineering and Applications, Volume VII, Issue III, September 14.

[6] M.Thangarasu., R.Manavalan, "A Literature Review: Stemming Algorithms for Indian Languages", International Journal of Computer Trends and Technology (IJCTT), volume 4 Issue 8, August 2013.

[7] Mihalcea, R., & Tarau, P.,"TextRank: Bringing order into texts. Association for Computational Linguistics", (2004, July).

[8] Ms. Anjali Ganesh Jivani, "A Comparative Study of Stemming Algorithms", International Journal of Computer Technology and Applications, Vol.2 (6), PP 1930-1938, NOV-DEC 2011.

[9] Mudassar, Tanveer J Siddiqui, "Discovering suffixes: A Case Study for Marathi Language", (IJCSE) International Journal on Computer Science and Engineering, 2010.

[10]Rohit Kansal Vishal Goyal G. S. Lehal, "Rule Based Urdu Stemmer". Proceedings of COLING 2012: Demonstration Papers, pages 267–276, COLING 2012, Mumbai, December 2012.

[11] Sajjad Ahmad Khan1, Waqas Anwar1, Usama Ijaz Bajwa1, Xuan Wang2, "A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language", Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP),, COLING 2012, Mumbai, December 2012.

[12]Snigdha Paul, Mini Tandon, Nisheeth Joshi and Iti Mathur, "Design of a rule based Hindi Lemmatizer", pp. 67–74, 2013.

[13]Upendra Mishra, Chandra Prakash, "MAULIK: An Effective Stemmer for Hindi Language", International Journal on Computer Science and Engineering (IJCSE) Vol. 4 No. 5, PP.711-717, May 2012.

[14]V.Gupta,N.Joshi,I.Mathur,"Design & Development of Rule Based Infectional and Derivational Urdu Stemmer 'Úsal'" ,INBUSH-ERA-2015,7-12,2015.

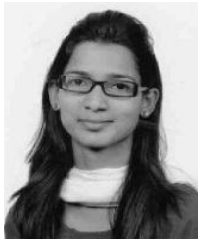
[15] V.Gupta,N.Joshi,I.Mathur, "Design & Development of Rule Based Urdu Lemmatizer",IEEE,2015.

[16]V.Gupta,N.Joshi,I.Mathur, "Rule based stemmer in Urdu",Computer and Communication Technology(ICCCT) 2013 4th International,2013.

[17]Virat V. Giri, Dr.M.M. Math & Dr.U.P. Kulkarni, "A Survey of Automatic Text Summarization System for Different Regional Language in India", In Bonfring International Journal of Software Engineering and Soft Computing, Vol. 6, Special Issue, October 2016

Authors Profile

Ms. Nikita Pise is pursuing her Bachelor of Engineering from IES IPS Academy, Indore, India.



Ms. Vaishali Gupta is pursuing her Ph.D in Computer Science & Engineering from Banasthali University, Rajasthan, India. She has interest in language processing specifically for Indian Languages. She has developed various NLP tools for Hindi and Urdu language. Her current research interest includes Natural language processing, Machine Translation and Information Retrieval.

