2017 International Conference on Computer, Communications and Electronics (Comptelix)
Manipal University Jaipur, Malaviya National Institute of Technology Jaipur & IRISWORLD, July 01-02, 2017

MarS: A Rule-based Stemmer for Morphologically Rich Language Marathi

Harshali B. Patil, Ajay S. Patil
School of Computer Sciences,
North Maharashtra University,
Jalgaon, Maharashtra, India
patilharshalib@gmail.com, aspatil@nmu.ac.in

***Abstract*—Stemming is a technique that transforms morphologically similar terms into a unique term without doing a complete morphological analysis. Stemming is used as a preprocessing step in many Natural Language Processing (NLP) applications like Information retrieval (IR), Machine Translation, Parsing, Summarization, etc. The present work explores the application of stemming to the task of information retrieval. In IR, stemming is generally used for two main purposes: decreasing index size and for increasing system performance. This paper presents a stemmer for Marathi language which uses rule-based technique. The average accuracy achieved by the proposed stemmer is 79.97% when tested on a collection of 4500 unique words from the news corpus among nine runs. Since the accuracy of the proposed stemmer is satisfactory it can be effectively useful in several NLP systems for Marathi language.**

Keywords—Natural language processing, stemming, Marathi, rule-based

I. INTRODUCTION

To form a meaningful sentence sometimes the change in the form of words is required due to the change in gender, number, tense, etc. e.g. "रामाने (राम) आंबा खाल्ला (खा)" (Ram had eaten the mango). These suffixes are attached to the words for showing the relation of that word with other words like verb. As seen in the above example in order to show the relation of subject with verb the suffix is attached to the subject and hence its original form gets changed. The inflected form of words make the automated processing of the documents a difficult task. For instance in absence of word normalization techniques "रामाने" (Ramane) and "राम" (Ram) are treated as two different terms. Use of word normalization techniques is the solution for these types of problems. Stemming is one of the important word normalization techniques. Stemming refers to the process of removing affixes from words. The stem is a part of word that is common to all its inflected variants. Table 1 depicts the example of stemming for Marathi language.

TABLE I. STEMMING EXAMPLE

Stem	Inflected forms
पुराव	पुरावा, पुरावे, पुरावेसुद्धा, पुरावेही, पुराव्याखाली, पुराव्याच्या, पुराव्यानिशी, पुराव्याला, पुराव्याशिवाय, पुराव्यासह, पुराव्यांच्या, पुराव्यांनिशी, पुराव्यांसाठी, etc.

Stemming is used in IR for increasing the recall and reducing the size of index files. There are techniques like simple rule-based, affix stripping, suffix substitution, dictionary-based, n-gram, HMM, clustering –based, corpus-based, hybrid, etc. This paper presents the rule-based stemmer for Marathi language. Marathi is an Indo-Aryan language. It is predominantly spoken by the native people of Maharashtra state, parts of neighboring states like Gujarat, Madhya Pradesh, Goa, Karnataka, Chhattisgarh and Andhra Pradesh, union-territories of Daman and Diu and Dadra and Nagar Haveli. Marathi is also spoken by Maharashtrian emigrants worldwide, especially in the United States, United Kingdom, Israel, Mauritius and Canada. It is one of the 22 official languages of India. It is the official language of state of Maharashtra and co-official language of Goa. Many official websites of Maharashtra Government are hosted in Marathi language. NLP systems are required to access this e-data available in the Marathi language. Stemmer is a building block for almost all NLP systems. IR is one of the important NLP applications. In IR due to use of inflected forms of terms e.g. "गंगा नदीच्या काठी असलेली गावे " (1700 hits) and "गंगेच्या काठाशी असणारी गावे" (1150 hits) both of these two queries generated different number of hits. However since the above two queries are semantically same the result should return same set of documents. This paper proposes stemmer for Marathi which will be useful in resolving such types of problems. The rest of the paper is organized as : Section 2 presents the related work to the topic, followed by section 3 describes the proposed stemmer, while section 4 consists of the results followed by conclusion and future work.

II. RELATED WORK

The work related to stemmer development has started in the late 60s. The first stemmer was developed by Lovins (1968) for English language using context sensitive longest match basis [1]. The early work related to stemmer development was carried out for English language only; some most notable stemmers for English includes the stemmers developed by Porter (1980) [2], Dawson (1974) [3] and Paice (1990) [4]. As there is a dramatic growth in the data appearing on World Wide Web in regional languages, the research for development of resources in regional languages has started. Several research efforts in the area of tools development for Marathi language have been reported in recent past by Patil (2016) [5], Mhaske (2016) [6], and Patil (2015) [7]. The detail analysis of stemming algorithms available for Indic languages is done by Patil et al. (2016) [8]. Many Stemmers are available for Indic languages but the work for Marathi stemmer

development has been started recently. The currently available stemmers for Marathi language include the stemmer developed by Almeida et al.(2010) that uses n-gram technique [9]. They obtained the mean average precision (MAP) of 35.79% against baseline MAP of 23.94%. Majgaonkar(2010) reported their efforts for Marathi stemmer development. They used rule-based and unsupervised techniques, and concluded that with statistical suffix stripping approach they obtained maximum accuracy for a test dataset of 1500 unique words [10]. Dolamic et al. developed light and aggressive stemmers for Marathi and obtained a change in MAP of 41.6% with aggressive stemmer and 13.9% with light stemmer [11]. Husain (2012) used an unsupervised approach for stemmer development using frequency based method and length based method. The author tested this stemmer for Urdu and Marathi language and found that length based method works better as compare to frequency based method when tested on 1200 words [12]. It would be interesting to investigate the performance of a stemmer for Marathi that hybridizes rule-based approach with other stemming techniques. To begin with we propose a rule-based stemmer for Marathi in section III.

A. Approaches for stemming

Several approaches for stemmer development of various languages are explored by researchers that includes the simple approach like affix stripping methods, statistical approach like clustering based, table lookup approach and hybrid approach. Fig. 1 summarizes various approaches used for stemming.

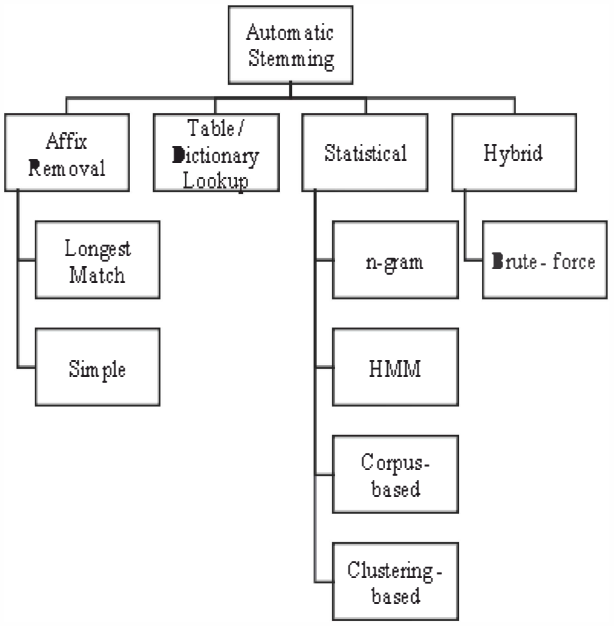

Fig. 1. Automatic stemming techniques

The rule based approach uses the list of suffixes along with the criteria under which it can be removed. The removal of affixes is either done based on the longest match found or in an iterative manner. These types of stemmers generate more under-stemming and over-stemming errors. In table / dictionary lookup technique a table of corresponding terms along with their stem is used. The stemming is done by searching the corresponding term in the table and retrieving the stem related to that term. These types of stemmers have the capability to assign correct stem for exceptional cases like mice – mouse. Statistical technique based stemmers apply some statistical measures for stemmer development like n-gram, HMM, clustering-based method, corpus-based method, etc. First the system is trained with a large corpus and then inflected forms are submitted to the trained module for stemming. When any two or more approaches are combined for stemmer development then the stemmer becomes hybrid. Most hybrid stemmers were developed by combining lookup table with affix removal. They overcome the disadvantages of one technique by combining it with the other; for example the lookup table based stemmer cannot stem a term if it is not present in the table. Hence, when the lookup table based stemmer is combined with affix removal then the terms that are not present in dictionary are also stemmed.

III. PROPOSED STEMMER

This section discusses the development process of proposed stemmer. Initially the orthographic structure of Marathi is discussed followed by the actual algorithm for proposed stemmer along with the generic rule-set. Marathi is written in the Devanagari script. It is written from left to right. The language comprise of 36 consonant letters and 16 initial-vowel letters. These consonants and vowels are used for constructing meaningful Marathi words. The vowels are not written as it is when they are combined with consonants. There are different symbols for each vowel (vowel sign), which are added to the consonant symbol to indicate combined syllables. When there is no symbol then it is understood to have short vowel अ in combination of that consonant. e.g. - Adding आ (A) to consonant "क" (k) is denoted by extra vertical line like "क + आ = का" (kA) this vertical line is called *kana.*

A. Suffixes in Marathi

Grammar is naturally divided into four parts viz., orthography, etymology, syntax and prosody. From these four parts part 2 i.e. etymology treats the inflection of words and their classification. To form a correct/meaningful sentence sometimes the change in the form of a word is required due to the change in gender, number and tense. Such change does not occur in all types of words and are called as non-inflectional words. The words in Marathi are broadly categorized into two types namely inflectional and non-inflectional. Noun, pronoun, adjectives and verbs are inflectional words while adverb, post-position equivalent to preposition in English, conjunction and interjection are categorized as non-inflectional words according to the Marathi grammar, but there are some exceptions. Words in Marathi follow specific patterns. Some common patterns found in Marathi words are given in the form of regular expressions along with examples in table II. Here T represents token, S is use to denote stem, R represents root of the word, and I denotes inflection present in that token.

The suffixes in Marathi are categorized into 3 different types according to Majgaonkar et al. (2010) as plain suffixes, joint word suffixes, and complex suffixes. Plain suffixes are same as dependent vowel signs like "ा,ु,े" etc. e.g.- "मुलगा".

TABLE II. PATTERNS FOUND IN MARATHI WORDS

Sr. No.	Regular expression for Marathi words	Example
1	T → S/R	घर = घर
2	T → (S/R).I	घरासमोर = घर (घरा) + समोर
3	T → (S/R).(I)+	घरासमोरचादेखील = घर (घरा)+ समोरचादेखील
4	I+→ I.(I)*	समोरचादेखील = समोर + चा + देखील

Joint word suffixes are those suffixes which are formed by merging two or more consonants and vowels. The joint words are formed by merging any of the consonants with the morphological variants "्या" of the consonant "य".Variants like "ल्या, च्या" are considered as joint word suffixes. e.g.-"देवासारख्या". Complex suffixes are formed by combining two or more consonants with the plain suffixes e.g.–"देवासाठी".

B. Inflections in Marathi

In Marathi there are 8 parts-of-speech (POS) as noun, pronoun, verb, adverb, adjective, post-position, conjunction and interjection. Among these 8 POS the words falling in noun, pronoun, adjective, and verb category suffer from inflections due to tense, numbers, genders, etc. In Marathi there are eight cases, two numbers, three genders and three tenses. Generally case markers, postpositions along with participles act as Marathi suffixes. Morphologically Marathi nouns are inflected for gender, number and case. These elements determine its agreement with the verb. Marathi pronouns undergo the inflections for gender, number, case and person. Case markers and postpositions are generally attached to Marathi nouns and pronouns. The various characteristic suffixes for Marathi cases are: "स, ला, ते, ने", etc. The postpositions are attached to words instead of case markers or along with case markers for showing the relation of words with other words. In Marathi language adjectives are inflected due to gender and number. Those adjectives whose masculine singular terminates in "आ", have different forms for three genders [13] e.g.- the adjective "चांगला" (good) whose masculine singular terminates with "आ" becomes "चांगली" for feminine case and "चांगले" in neutral case. Adjectives in Marathi are inflected only when case markers are attached to the words for which adjective is demonstrating the information **e.g.** - "पिवळ्या टोपीचे ". Verbs in Marathi may be divided into two classes, active and neuter. They may be further subdivided into active transitive, active intransitive, causal, regular, irregular, defective and auxiliary. There are three tenses, three voices and six moods in Marathi like indicative, conditional, potential, subjunctive, imperative, and infinitive. The Marathi verb is inflected on account of person, number, gender, as well as tense and mood. The inflectional termination is added to the root of the verb to show the tense of the verb. e.g.- root of the verb is like "लिहि" (write) and the inflected forms are like "लिहिणे", "लिहिले", (write, written) etc.

C. Compilation of suffix list

A one-class list of suffixes was initially compiled based on the linguistic knowledge that comes from the Marathi grammar books by Burgess (1854). It was observed that still there exist several suffixes that need to be added in the resulting list; so more suffixes were learned by digging deeper into the corpus. For learning new suffixes 1,02,639 unique words were extracted from nearly 3000 text documents. The extracted words are then inspected manually for finding more suffixes.

D. The proposed stemmer

In Marathi language when the words undergo inflection, first it gets converted into an oblique form and then the inflection suffixes are attached; so three different sets of suffixes are created. Following algorithm describes the proposed stemming process is detail.

MarS: Marathi Stemmer

Input : File f : File consists of Marathi words ,
Suffix set S: Set of all suffixes,
PS: Set of all plain suffixes (vowel sign),
JW: Set of joint word suffixes
Output : Stemmed file f'

Begin
sv = ø
Read the file f to be stemmed
Tokenize the file f and populate token vector TV
For each token t in TV do
  If t ends with one of the suffix s ϵ S then
    remove s from token t and store result into stem
  Else
    stem = t
  Endif
  If stem ends with one of the suffix s ϵ PS then
    remove s from stem and store result into stem
  Endif
  If stem ends with one of the suffix s ϵ JW
    remove s from stem and store result into stem
  Endif
  add stem in stem vector sv
Endfor
Write sv in output file f'
return f'
End

The stemming process is carried out sequentially in three steps: in the first step, the inflection suffixes from the words are removed based on longest common match principle. Then the plain suffixes are removed and finally the joint word suffixes are removed. For implementation purpose the complied list of suffixes was converted into if-then else types of rules to extract those suffixes from the words. Due to space constraint we are unable to present the entire rules here, so four generic rules are given in table III that represents all the rules used by the proposed stemmer.

TABLE III. GENERIC STRUCTURE OF MARATHI STEMMING RULE

Rule Id	Observation	Action taken
R1	tϵTV and sϵS and length(t) > length(s) and t ends with s	Strip s and return stem
R2	tϵTV and sϵS and length(t) > length(s) and t ends with s but t do not starts with sy	Strip s and return stem
R3	Stem obtained in R1/R2 ends with PS and length(stem)>2	Strip PS and return stem
R4	Stem obtained in R2 ends with JW but stem do not start with 'त'	Strip JW and return stem

In Marathi when case markers are attached to the words, then the word form gets changed while attaching the case marker. This phenomenon is called as *Samanyaroop* in Marathi language. For handling such cases R3 and R4 are use. e. g. "देश - देशा, शाळा - शाळे, कळी - कळ्या, आंबा - आंब्या, घोडा - घोड्या".

Though the proposed approach is simple and easy it requires extensive efforts (manual) for developing the rule-set. The advantage of this technique is that it is corpus independent, as well as requires less computations as compared to other statistical techniques.

IV. RESULTS

This section presents the results obtained by the proposed stemmer first we discuss the dataset used, then the stemmer accuracy, its strength followed by error analysis.

A. Dataset used

The dataset used for this experiment is a Marathi corpus obtained from FIRE[1]. FIRE is the Forum for Information Retrieval for Indian languages. The corpus consists of 99,275 documents of Marathi newspapers, Maharashtra Times and Daily Sakal spanning from April 2004 to September 2007. All the documents and topics are encoded in UTF-8. The document format is adopted from TREC. Since a gold standard set of data is not available, we have prepared a gold standard manually for our experiment; it consist of 4500 unique words extracted from 50 documents of FIRE corpus. We compare the result produced by our system with actual stem available in the gold standard. Analysis of the result is based on factors like number of correct and incorrect stems produced by the proposed stemmer.

B. Stemmer Accuracy

While evaluating stemmers either the performance of stemmer is evaluated or the impact of stemming in underlying application is evaluated. Various factors are considered in the evaluation of stemmers independent of IR. We evaluated the proposed stemmers by measuring the % accuracy, the strength of the stemmer and the errors produced by it. The % accuracy provided by the stemmer based on the formula mentioned by Husain(2012) is calculated as follows:

$$Accuracy = \frac{Numberofcorrectlystemmedwords}{Totalnumberofwords} X100 \quad (1)$$

[1] http://www.isical.ac.in/~fire/

Table IV presents the % accuracy obtained in all 9 runs. From table IV it is observed that the maximum accuracy of 83% is obtained by the proposed stemmer for run no. 8 and for run 4 we obtained the minimum accuracy of 75%.

TABLE IV. STEMMER ACCURACY

Sr. No.	Total terms	Total Correct	Total Incorrect	% Accuracy
1	500	403	97	80.6
2	500	414	86	82.8
3	500	402	98	80.4
4	500	375	125	75.0
5	500	391	109	78.2
6	500	397	103	79.4
7	500	403	97	80.6
8	500	415	85	83.0
9	500	399	101	79.8

C. Stemmer strength

Frakes and Fox (2003) have mentioned several ways to measure the stemmer strengths: number of words per conflation class, index compression, the word change factor, number of characters removed, hamming distance, inter stemmer similarity, and a similarity matrix [14]. Among these 7 factors we used first 4 factors to measure the strength of proposed stemmers. The mean number of words per conflation class (MWC) is computed by following formula:

$$MWC = {}^{N}/_{S} \quad (2)$$

Where N denotes number of unique words before stemming and S denotes number of unique stems after stemming. The index compression factor (ICF) is calculated as following:

$$ICF = {}^{N-S}/_{N} \quad (3)$$

where N and S are same as above. The word change factor indicates the proportion of the words in a sample that have been changed in any way by the stemming process. The mean number of characters removed is the average number of characters removed when stemming is applied to a text collection.

$$MNCR = \frac{\sum Number\ Characters\ removed\ from\ each\ term}{Total\ number\ of\ terms} \quad (4)$$

The various factors suggested by Frakes and Fox have also been considered for evaluating the proposed stemmer and table V summarizes the results.

TABLE V. STEMMER STRENGTH

Sr. no.	Factor	Result
1	Mean number of words per conflation class	1.64
2	Index compression	0.39
3	Word change factor	0.75
4	Mean number of characters removed	2.41

D. Stemming errors

According to Paice the stemmers can be evaluated independent of IR by calculating two types of errors generated by the stemming process under-stemming errors and over-stemming errors [15].

Under-stemming error: This refers to the words that should be grouped together by stemming, but that are not grouped. This type of errors causes a single concept to be spread over various different stems. So it will tend to decrease the recall in an IR system. These errors are generated when words with same meaning reduced to different stem.

Over-stemming error: This refers to the words that should not be grouped together by stemming but they are grouped. These errors cause the meaning of stems to be diluted, which will ultimately affect the precision of IR. These errors are generated when semantically distinct words are merged to the same stem.

TABLE VI. STEMMING ERRORS

Sr. No.	Total terms	% over-stemming	% under-stemming
1	500	4.0	24.2
2	500	5.0	20.8
3	500	5.6	20.2
4	500	8.4	33.0
5	500	4.2	23.0
6	500	6.4	19.4
7	500	7.6	39.2
8	500	4.8	22.8
9	500	7.8	14.0

The proposed stemmer makes an average 5.97% over-stemming errors and 24.06% under-stemming errors.

V. CONCLUSION AND FUTURE WORK

This paper describes the Marathi rule-based stemmer (MarS) based on longest common match principle. Our approach first tokenizes the document and then for each token performs the stemming based on the applicability of the rule. The average accuracy of 79.97% was obtained when we tested the proposed rule-based stemmer in 9 different runs consisting of 4500 unique terms. The over-stemming and under-stemming error rate is quite high because some single word suffixes in Marathi are the part of stem itself in some cases while sometimes they act as suffixes, and this is the reason for errors. Since a similar Marathi word represents different meaning they generate the under-stemming errors. If the POS category of the word is known prior to stemming then the errors can be reduced upto some extent. The word-stem dictionary can also be used to reduce the errors generated while stemming. In future we will develop a dictionary based stemmer for Marathi language. We shall then hybridize the rule based and dictionary based approach and study its performance.

ACKNOWLEDGMENT

The authors are very much thankful to the Forum for Information Retrieval Evaluation for providing the Marathi corpus. The work presented in this paper is supported by SAP DRS-II research grant of UGC, New Delhi.

REFERENCES

[1] J. B. Lovins, "Development of a Stemming Algorithm", Mechanical Translation and Computational Linguistics, 1968, vol. 11, pp. 22-31.

[2] M. F. Porter, "An Algorithm for Suffix Stripping", Bulletin of the Association for Literary and Linguistic Computing, 1980,vol. 14(3), pp. 130-137.

[3] J. L. Dawson, "Suffix Removal for Word Conflation", Bulletin of the Association for Literary and Linguistic Computing, 1974, vol. 2(3), pp 33-46.

[4] C. D. Paice, "Another Stemmer", ACM SIGIR 1990, vol. 24(3), pp.56-61.

[5] N. V. Patil, A. S. Patil and B. V. Pawar, "Issues and Challenges in Marathi Named Entity Recognition", International Journal on Natural Language Computing, 2016, vol. 5(1), pp. 15-30.

[6] N. T. Mhaske and A. S. Patil, "Issues and challenges in Analyzing Opinions in Marathi Text", International Journal of Computer Science Issues, 2016, vol. 13(2), pp. 19-25.

[7] V. B. Patil and B. V. Pawar, "Modeling Complex Sentences for Parsing through Marathi Link Grammar Parser", International Journal of Computer Science Issues, 2015, vol. 12(1), pp. 108-113.

[8] H. B. Patil, B. V. Pawar and A. S. Patil, "A Comprehensive Analysis of Stemmers Available for Indic languages", International Journal on Natural Language Computing 2016,vol. 5(1), pp. 45-55.

[9] A. Almeida, P. Bhattacharyya, "Experiments in N-Gram Based Indexing and Retrieval in Marathi", Forum for Information Retrieval Evaluation, 2010.

[10] M. M. Majgaonker and T. J. Siddiqui, "Discovering Suffixes: A Case Study for Marathi Language", International Journal on Computer Science and Engineering, 2010, vol. 2(8), pp. 2716-2720.

[11] L. Dolamic and J. Savoy, "Comparative Study of Indexing and Search Strategies for the Hindi, Marathi and Bengali Language", Special Issue of ACM Transaction on Asian Language Information Processing on IR for Indian Languages,2010, vol. 9(3).

[12] M. S. Husain, "An Unsupervised Approach to Develop Stemmer", International Journal on Natural Language Computing, 2012, vol. 1(2), pp. 15-23.

[13] E. Burgess, Grammar of the Marathi language, American Mission Press 1854.

[14] W. B. Frakes and C .J. Fox, "Strength and Similarity of Affix Removal Stemming Algorithms", ACM SIGIR, 2003 vol. 37(1), pp. 26-30.

[15] C. D. Paice, "Method for Evaluation of Stemming Algorithms Based on Error Counting", Journal of the American Society for Information Science, 1996, vol. 47(8), pp. 632-649.