IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

Rule based Stemmer using Marathi WordNet for Marathi Language

Pooja Pandey[1], Dhiraj Amin[2], Sharvari Govilkar[3]

Assistant Prof, Dept of Information Technology, Matoshri College of Engineering & Research Center, Nasik, India [1]

Assistant Professor, Department of Computer Engineering, Pillai College of Engineering, New Panvel, India [2]

Associate Professor, Department of Computer Engineering, Pillai College of Engineering, New Panvel, India [3]

Abstract: Stemming is integral part of many natural language processing and information retrieval application. Stemmer for a given language basically extracts root or base word for the input word. Marathi word being rich in morphological variation requires an efficient stemmer which can deal with various morphological structures associated with words. Marathi WordNet consists mainly of Marathi root and base words with their Part of Speech information which is useful to reduce over and under stemming issues. Proposed system augments rule based approach with WordNet to perform stemming of the Marathi words with the help of Name entity and stem exception dataset.

Keywords: Stemmer, Root words, Marathi WordNet, Stem Word, Suffix, Inflection and Marathi

I. INTRODUCTION

The process of transforming multiple morphological forms of words or terms into a common related word is called stemming. Simple example of stemming process is where words "stemming", "stemmed", "stemmer" are transformed into common term "stem". Stemming is a crude heuristic process which cuts the ends of words and often further includes the procedure of removal of derivational affixes and inflection associated with the word [1]. Stemming process doesn't consider the context of words while it is stemmed as compare to lemmatizer which also considers the context and maps words forms into logical related root or base word.

A particular kind of very simple algorithm for performing stemming includes elimination of endings of terms by using suffix list which are used frequently. On the other hand a complex stemmer can apply knowledge related to morphology of that term for obtaining stem terms from derived words.

Rule-based and Machine learning are the two most common approaches for developing a stemmer for various language available in world [2]. In Rule based approach rule are generated based on the linguistic information available for the given language. Rules are mostly manually generated by linguistic experts, which are then used to remove suffix associated with the input word to stem. Machine Learning approach consists of supervised or unsupervised approach.

Supervised approach again requires linguistic knowledge in advance which can be further used to train system to perform stemming. In unsupervised approach specific knowledge of the language is not required and thus becomes a language independent stemmer.

Rule based approach are time consuming task but when finished can lead to more accurate results if rule set is of specific range and all the rules are generated by linguists whereas machine learning approach also requires predefined database to train the system but it can yield much better results for unknown or new terms which didn't exist in the dataset.

The main importance of stemmer is in information retrieval application where a stemmed query proportionally leads to increase in recall for the given application. In fact stemmer is an important pre-processing stem for different applications like text mining, text summarization and sentiment analysis. Most of the search engines are gradually upgrading from keywords based to semantics and context based answer retrieval for the search question. Natural language question answering systems are developed using ontology as representation of knowledge. The onto terms store in the ontology are in there root or base forms so an accurate stemmer is very useful of matching user query term with onto terms stored in the ontology.

WordNet are available for most of the languages available across the globe. Many natural language processing applications are using WordNet in different use cases. Marathi WordNet [3] is provided by IITB which provides different relations between synsets or synonym sets which represent unique concepts. WordNet is a large lexical database of Marathi. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept; WordNet really consists of four sub-nets, one each for nouns, verbs, adjectives and adverbs, with few cross-POS pointers. Cross-POS relations include the "morphosemantic" links

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

that hold among semantically similar words sharing a stem with the same meaning: observe (verb), observant (adjective) observation, observatory (nouns) [4].

While building ontology for Marathi question answering system one may consider the word present in the Marathi WordNet as the base or root form of the word. Most of the rule bases stemmer simply removes suffixes or prefixes without any knowledge of what will be root form of the word and may lead to incorrect stemming, same thing can be observed for any basic Marathi Lightweight stemmer which doesn't augment Marathi WordNet in its processing steps may lead to over or under stemming of words. So for such type of Marathi question answering system a stemmer is needed which doesn't over or under stems a word which is already flagged as base or root word in the Marathi WordNet.

Stemmers are available for Marathi language but yet there is no stemmer which augments WordNet with it reduces the problem of over stemming of the term. We have developed a rule based Marathi stemmer which contains big chunks of suffix and inflection removal rule with support of stem exception dataset and Marathi WordNet which reduces the problem associated with over stemming of word when WordNet is used in different natural language processing application.

II. LITERATURE SURVEY

Most of the natural language application which considers semantic for various operations is very much dependent on WordNet available for languages. In the recent past many stemmers are developed for Indian languages including Marathi. Different approaches like rule based, machine learning are used for developing stemmer which is an integral part of pre-processing module in many natural language applications. More work has been done for development of stemmer using rule based approach over statistical approach. Due to existence of multiple morphological variants of single word in Marathi language stemming process is very critical as compared to other languages like Hindi which also belong to Devanagari script. There is no work done where Marathi WordNet is augmented within the stemming process for Marathi terms.

Authors Monika Dogra, Abhishek Tyagi and Upendra Mishra has developed stemmer for devanagari script where both prefix and suffix are removed from the word using hybrid approach where three different algorithm (lookup approach, prefix removal and suffix stripping algorithm) where used to perform stemming [5]. Unsupervised approach was used by Shahid Husain [2] to perform stemming of Marathi words where suffix rules are generated using frequency based stripping and length based stripping. Through length based stripping it is observed that over stemming cases are increased. Their proposed method is language independent and can be extended for other languages. Author has compared performance of rule based with unsupervised stemmer where suffix stripping rules are generated manually for rule based stemmer and unsupervised stemmer learns suffixes automatically from a set of Marathi words. Rule based approach for stemming lead to over stemming problems. It is observed that unsupervised approach provided better results over rule based stemmer [6]. Authors have extracted root word from given word of devanagari script using rule based approach [7] where they have devised rules for suffix and inflection removal from given words. Stop word list is created which are exempted from stemming which helped them to improve their overall results. Authors developed a set of suffix stripping and infection removal rules.

III.CHALLENGES FOR STEMMING MARATHI WORD

Marathi language exhibits high level of morphological variations, in Marathi language a single root word like महाराष्ट्रा can have various morphological variants like महाराष्ट्राचा, महाराष्ट्राची, महाराष्ट्रामध्ये, महाराष्ट्रासाठी, महाराष्ट्रावर, महाराष्ट्राकडे.

Basic word formation of Marathi consists of word = [(root/base word)] + [inflections/suffix]. For example word शिवाजींचा consists of inflection जीं and suffix observed here चा and the main root word is शिवाजी.

Suffixes in Marathi language can be plain suffixes like ी, ा, ै, ौ, ो, ि, ॉ or joined word suffix like ल्या, त्या, न्या, च्या or complex or standalone word suffix like साठी, वर, कडून, मध्ये, कडे

Due to Marathi language exhibiting high morphological structure many stemmers developed faced, over and under stemming related challenges.

In over stemming words get stemmed to an extent where meaning of word is lost (where stem words doesn't belong to language vocabulary) after stemming. In stemmer it is also observed that words after stemming changes its meaning from one context to another, this is also due to over stemming of words.

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. Over stemming often leads to scrambling of words like proper nouns and adjectives.

Under stemming is observed in stemmer results when words are partially stemmed. Here words get under stemmed and the stemmed word doesn't belong to the root

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

class. Also when two words belonging to same context are stemmed to different root then this problem is also called as under stemming of a word. For example both the word better and best must be reduced to the good which is the root of both these words, but if they are not all stemmed to good it indicates an error.

Most of the Marathi word contains inflection before word suffixes. When stemmer only remove suffix and leaves the inflection attached with word then the problem of under stemming is observed.

For example in word शिवाजींचा if only suffix चा is removed then observed root word will be शिवाजीं but actual root word is शिवाजी.

IV. PROPOSED RULE BASED MARATHI STEMMER

We propose a rule based stemmer for Marathi language which uses Marathi WordNet to reduce the problem associated with over and under stemming. Developed stemmer contains two modules: pre-processing and stemming modules. Figure 1 represents our proposed WordNet based Marathi stemmer.

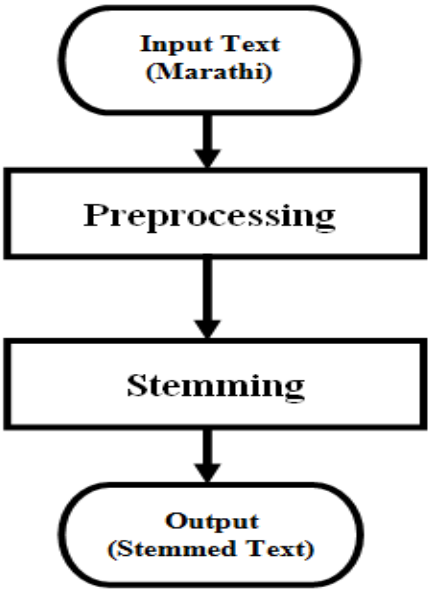

Fig. 1. Proposed rule based Marathi stemmer

Sample input and final output for rule based Marathi stemmer.

Input Text: भारताची राजधानी नवी_दिल्ली (New Delhi) आहे.

Stemmed text: भारत राजधानी नवी_दिल्ली आहे

Pre-Processing module: This module deals with pre processing of input text to and is divided into two parts filtration and tokenization. Figure 2 describes pre processing module.

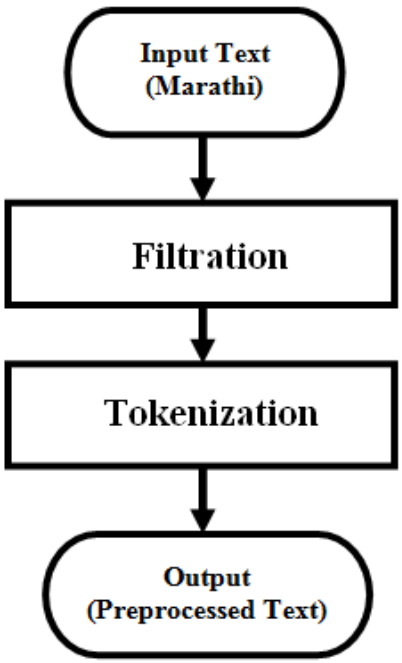

Fig. 2. Preprocessing Module

In filtration input text is filtered out to remove any non Devanagari Unicode but it is ensured that some punctuation marks like "_" and "-" are not excluded as they are also used in Marathi language word formation. Tokenization is the basic and important module of any NLP application.

At top most level of tokenization process document are split into paragraphs. Paragraphs are then split into individual sentences and lowest level sentences are broken into individual words.

In some scenarios word may be further tokenized into ngrams. For tokenization of a sentence into a set of words, space between two words is considered as parameter to split them.

Input Text: भारताची राजधानी नवी_दिल्ली (New Delhi) आहे.

Filtered Text: भारताची राजधानी नवी_दिल्ली आहे

Tokenized Text:

Token 0: भारताची

Token 1: राजधानी

Token 2: नवी_दिल्ली

Token 3: आहे

Stemming module: The output of this module is stemmed word if word matches the stemming criteria. It is further subdivided into three parts: root verification, suffix removal and inflection removal. Figure 3 describes stemming module.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

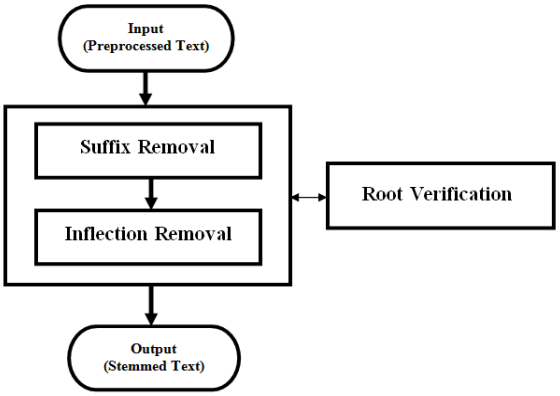

Fig. 3. Stemming Module

Sample input output for stemming module
Input:

Token 0: भारताची

Token 1: राजधानी

Token 2: नवी_दिल्ली

Token 3: आहे

Output:

Token 0: भारत

Token 1: राजधानी

Token 2: नवी_दिल्ली

Token 3: आहे

Stemmed text: भारत राजधानी नवी_दिल्ली आहे.

In root verification word (stemmed word or original word) is checked whether it is root word or not. It is verified whether word is root or not when word meets one of the three conditions. Figure 4 represents three conditions which are executed for root verification process. First Condition, here word is matched with Marathi WordNet; if word is present in the WordNet then it is flagged as root or base word. Second Condition, here word is matched with Marathi name entity dataset created by us; if word is present in Marathi name entity dataset then it is flagged as root word. Name Entity dataset is a simple Marathi Name Entity set which consist common names of person, places and organizations.

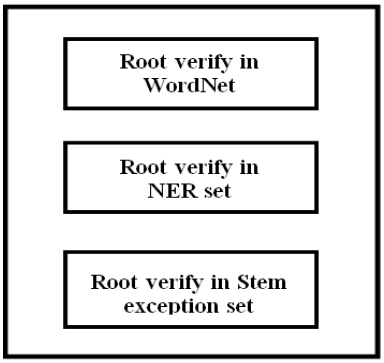

Fig. 4. Root verification Process

Words like धीरज, भारत are names of person and place and such words should not be stemmed as they are already root words. Third condition, here word is matched with Marathi stem exception dataset created by us; if word is found in Marathi stem exception dataset then it is flagged as root word.

Some words which are present in stem exception dataset राजधानी, मुला, मामा should not be stemmed as after stemming meaning of word will be completely lost.

Suffix removal: Most of the Marathi word has suffixes attached to it. Marathi language has a bigger list of suffixes as compared to Hindi language which also belong to Devanagari script. Marathi language is order free language like Hindi language where subject and verb can occur at any position in the sentence.

To identify the subject present in the sentence case markers and postposition attached too Marathi word are used. Both case markers and postpositions are suffixes attached to Marathi word.

Word in any language or script is a group of characters and suffix can be single or multiple continuous characters in a word. Basic way of identifying and splitting suffixes from base word or root word is to identify the portion from the end of word which matches with existing suffix list and then split the base word from suffix.

While splitting suffixes from base words it is first verified that length of suffix is not larger than length of word to reduce chances of over and wrong stemming. To remove suffix through rule based approach a predefined suffix list is created. Figure 5 shows some examples of suffix seen in Marathi language.

लेला	तात	मुळे
पुढलीच	पुढलीसुद्धा	मागचा
मागचीपण	मागच्याच	मागलाही
मागली	मागल्याही	मधूनच
मधलीच	मधचे	मधचीसुद्धा

Fig. 5. Suffix observed in Marathi Language

Inflection removal: Most of the Marathi words are inflected. Inflection relates to modulation in pitch of the voice during pronunciation of a word. Like word शिवाजींचा is having चा as suffix and the sound जीं is the inflection attached to base word Shivaji.

Mostly noun, pronoun and verb exhibit inflection in there formation. Infection removal rules are generated by us, which are mostly applied after suffixes are removed from the word.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 10, October 2016

Some inflection removal rules are as follows:

If word after removal of suffix contains डा, डी, डे then convert it to ड.

If word after removal of suffix contains ता, ति, ती, तु, ते then convert it to त for example in word भारताची after removing suffix ची word becomes भारता and here after applying inflection removal rules i.e. ता is converted into त and we get root word भारत

V. CONCLUSION

Stemmer is integral part of many natural language processing and information retrieval application. Most of the application which deals with semantic and ontology tends to use corresponding language WordNet frequently. Stemmer which augments WordNet in the stemming process provides much efficient functionality to such class of natural language processing applications. Through root verification process the common problem associated with over and under stemming in a stemmer is reduced proportionally. Further hybrid system can be developed using rule based and machine learning approaches to tune the system to handle more unknown words and generate more rules for the process of stemming.

REFERENCES

[1] http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html last accessed on October 22,2016
[2] Shahid Husain, An unsupervised approach to develop stemmer. International Journal on Natural Language Computing (IJNLC), 2012.
[3] http://www.cfilt.iitb.ac.in/wordnet/webmwn/english_version.php last accessed on October 20,2016
[4] https://wordnet.princeton.edu/ last accessed on October 20,2016
[5] Monika Dogra, Abhishek Tyagi and Upendra Mishra. An effective stemmer in Devanagari script, Intl. Conf. on Recent Trends In Computing and Communication Engineering, 2013.
[6] Mudassar, Tanveer J Siddiqui. Discovering suffixes: A Case Study for Marathi Language, (IJCSE) International Journal on Computer Science and Engineering, 2010.
[7] Sharvari S. Govilkar J. W. Bakal and Sagar R. Kulkarni, Extraction of Root Words using Morphological Analyzer for Devanagari Script, I.J. Information Technology and Computer Science, 2016.