Improving a Lightweight Stemmer for Gujarati Language

Chandrakant D. Patel[1] and Jayeshkumar M. Patel[2]
Acharya Motibhai Patel Institute of Computer Studies[12], Ganpat University, Kherwa

ABSTRACT

The origin of route of text mining is the process of stemming. It is usually used in several types of applications such as Natural Language Processing (NLP), Information Retrieval (IR) and Text Mining (TM) including Text Categorization (TC), Text Summarization (TS). Establish a stemmer effective for the language of Gujarati has been always a search domain hot since the Gujarati has a very different structure and difficult that the other language due to the rich morphology.

KEYWORDS

Gujarati Language, Gujarati Stemming, Light Stemmer, Linguistic resources, Morphological Analysis

1. INTRODUCTION

Now a day, an internet becomes primary tool to get resourceful information. Using the Web mining, we get only required data which we need to access and save our time and efforts as well from surfing on World Wide Web. Everyday different kinds of documents added massively on internet which needs to optimize. Text mining which a part of Web mining has been proposed with several methodology including Information Retrieval methods. For example, a user searches for "Who is Dr. Kalam?" In this example, the stop words are being removed, which includes who, is and Dr. This process is useful to make the search simpler by removing all stop words from the user input. The left out words are searched in each document and a matrix with the number of words, and their frequencies of occurrence.

In domain of Information Retrieval System (IRS), Stemmers covered almost modern indexing and search systems. Stemming is used to reduce the morphological modifications of a word into their stem, root or base word. Before indexing the stem word, Stemmers removing affixes (suffixes or prefixes) from words, makes as Stem words. The stemming process ultimately increases the number of retrieved documents in an IR system. Stemming has been focus for several decades because of its use and importance.

Stemming from one of the major tasks in automatic processing of rich language, such as Gujarati [1][2]. However, emerging is one of primary processing which reveal directly the result of any other morphological processing. Gujarati language a word highly reflected in language; it is one of the most challenging that the extracted algorithms do not always provide a best solution.

2. RELATED WORK

The algorithm of stemming can be classified into three groups: trimming methods, statistical methods hybrid methods. Each method has its own merits and demerits as per typical method to find a stem.

The first group is connected to a withdrawal from the word affixes. Lovins [3] proposed in 1968 that the first Lovins stemmer which removes the suffix had long been a word. In the end, removed the word right words change the wording which makes various adjustments has been recorded using different rules. The advantages of this algorithm is-it is very quick and can manage removal of double letter in words such as 'getting' being transformed to obtain 'get' and also deal with many irregular plural as - mouse and mice etc. The disadvantages of the Lovins approach are that it is time and data consumes. In addition, many of the suffixes are unavailable later in the repository.

Suffixes in English, which is mainly caused by a combination of small and simple suffixes ideas based on the most famous one in 1980 arising from the proposed methods to today's stemming i.e Porter algorithm [4]. It consists of five steps, and each step of the way until they were in a position, use several terms. If accepted as a rule, and the next step is to remove the suffix accordingly. The result appeared at the end of the fifth step was reviewed. Porter "snowball" [5] in the name of known and different languages to create their own stemmer that enables a comprehensive framework, which is used to design.

Stemmer Paice [6] is an interactive algorithm with a table containing 120 rules indexed by the last letter of a suffix. At each iteration, it tries to find a rule by the last character of the word. Each rule specifies a cut or replace purpose. If there is no such rule, it stops. The advantage is its simple form and each iteration taking care of the both cut or replace in accordance with the rule applied. The advantage is that it is very heavy and the resulting algorithm may occur.

Dawson stemmer is an extension of Lovins approach except that it includes a much more comprehensive list of about 1,200 suffixes. As Lovins it is too one pass stemmer and therefore it is very fast. The advantage is that it covers more suffixes that Lovins and is fast in execution. The disadvantage is that it is very complex and lacks a reusable standard setting work.

The second group is called the statistical methods; it contains stemmers that are based on statistical techniques. Most methods remove affixes but after the implementation of certain statistical procedures. In this group we can find the following text stemmers:

N-grams [7] stemmer regardless of the language in which the approach of the string-similarity is used to convert the word inflation in its root. An N-gram is a set of consecutive characters of n in a word. The main idea behind this approach is that, similar words will have a high proportion of n-grams in common. The advantage that it is independent of language and therefore very useful in many applications. The disadvantage is that it requires a considerable amount of memory and storage to create and store the grams of n.

HMM stemmer was proposed by Melucci and Orio and based on the concept of the Hidden Markov Model (them) that is finite-state machines. At each transition, the state sends a new symbol with a given probability from the table. This method does not require prior knowledge of

languages and the most likely path is found by using the Viterbi encoding in the automatic graphic.

The third group is called the mixed methods that include: the inflectional and derivational methods: the analysis of the inflectional and derivational morphology implies so much. Here the corpus must be very great to develop this type of stemmers and therefore basic corpora are part of the linguistic analyzers also. If the inflectional of the alternative words is related to the variations of the specific plural language like syntactic, sex, case, etc. whereas in the derivational variants they are connected to the part of the speech (POS) of an oration in which it appears the word.

Krovetz Stemmer [8] was presented in 1993 by Robert Krovetz. It indeed eliminates inflected suffixes in three stages. Since it is not to stemmer stems all the variants of words, it can be used like pre-stemmer

before applying the algorithm starting off. This would increase to the speed and the effectiveness of stemmer key. This stemmer not always produces a good recall and precision performance.

The Xerox has developed a lexical data base of English and other languages can analyze and also generate derivational and inflected morphology. The data base of inflectional reduces to each superficial form word that can be found in the dictionary, of the following way: the singular names (Eg children), the infinitive verbs (e.g. understand), etc. The advantages of this are that to stemmer it works well with a great document and also eliminates are also applied area codes where it wants all stems are valid words. The disadvantage is that the exit depends on the lexical data base that cannot be exhaustive. Since this method is based on a lexicon, that cannot contain words that are not part of the lexicon. Stemmer this has not been applied successfully in many other languages.

The Corpus based Stemmer has been proposed by Xu and Fencing. They have an approach that attempts to overcome some of the disadvantages of the Porter suggested. This refers to the automatic modification of confusion (classes - words) which has led to a common core, depending on the characteristics of a collection of text data using statistical methods. The basic assumption is that the forms of speech which must be combined to a certain collection co-produce in the compilation documents. By using this concept over and under stemming disadvantages are solved. For example, the way in which it works is Porter stemmer uses the first to identify the conflicted words and then the next step is to use the collection of statistics to redefine those words. The advantage of this method is that it can potentially avoid the fillings which are not appropriate for a given situation and the result is a true word, and not a collection of incomplete log. The drawback is that you must develop the statistical measure for each increase of collection and processing time separate as the first step, and algorithms are used in first before using this method.

Context-sensitive Stemmer was proposed by Funchun Peng. Basically to the words of the question of entry, morphological variants that would be useful in the search before the question was submitted to the search engine are predicted. This reduces the number of bad extensions, which in turn reduces the cost of additional computation and improves accuracy simultaneously. The advantage of this is it stemmer improves selective extension of the word next to the question and conservative event matching word in the document side. The disadvantage is the processing

time and the complex nature of stemmer. There may be errors in finding nominal sentences in question and the words of proximity.

Hybrid stemmer for Gujarati proposed by Pratik Patel et. al. [2] with completely unsupervised approach.

This algorithm based on Goldsmith's which was purely unsupervised with hand crafted suffixed. The advantage of this algorithm is hand crafted Gujarati suffix boosts the accuracy with light weight stemmer. The disadvantage of this algorithm is it can removes only suffixes.

Hybrid stemmer for Gujarati proposed by Kartik Suba et. al. with inflectional and rule based stemmer. The inflectional stemmer performed POS based stemming which has been inspired by Goldsmith approach. The derivational stemmer deals with derivation of the words either by affixes or morpheme bounry which has been inspired by Jurafsky and Martin approach.

Many algorithms have been developed to come from a wide range of languages including Indian and Non-Indian languages. For Gujarati language, there are three different approaches to come: light weight approach; and statistical approach [9] and hybrid stemmer. Yet no complete stemmer for this language available.

In addition to the introduction, the paper consists of four sections. In the second section, we present the Gujarati morphology. In third section, we describe our suggested stemming algorithm. Then we present and discuss the results. Finally, we draw our conclusion.

3. GUJARATI LANGUAGE

Gujarati is official language of the State of Gujarat, India and one of the 22 official languages recognized by the Government of India. It is spoken by almost 46 million people in the world, giving it the rank of 23rd most spoken language in the world. Gujarati is written from left to right and there is no capitalization in it.

3.1 Gujarati Morphology

The morphological nature of the Gujarati language is bit complex. The Gujarati has three genders such as masculine, neuter and feminine, numbers such as singular and plural, cases such as nominative, oblique and locative. The gender of a noun is determined by either its meaning or ending. The Gujarati adjective has two categories such as declinable and indeclinable. This adjective can determine by its ending based on cases. The Gujarati verbs are inflected based on combination of gender, number, person etc.

4. STEMMING ALGORITHM

4.1 Definition

A mother is a word without affixes, of an approach of elimination called light stemmer. Relate to a process to be detached a small unit from the prefixes and/or suffixes. In this algorithm, we use supervised methodology, to make handcrafted rules for prefix and suffixes after the processing and taking help from expert for the Gujarati Morphology.

4.2 Linguistic resources

To improve the output of our proposed algorithm and to solve the word inaccuracy, we assumed certain properties of the language in the form of:

15 prefixes with a length ranged from 1 to 2
characters 152 suffixes with a length ranged from
1 to 4 characters
95 stop words with a length ranged from 1 to 3
characters 736 dictionary words as root words

1099 Gujarati words from e-newsPaper (**Source: epaper.divyabhaskar.co.in**)

4.3 Pre-processing (in Figure 1)

It is crucial part in any Information Retrieval Processing, we should keep in mind what we are going to input and that's bifurcate by two basic steps.

Truncation: to eliminate punctuation symbols, special characters, numbers, stop words (a special file where we have listed out words which is less value while we considering for index terms) and non-Gujarati characters/words. For that purpose, we use many machine learning algorithm to process the same step by using Super Vector Machine [10], Radial Basis Function, Naïve Bayes Classifier [11] and K-Nearest Neighbor [10] algorithm or hybrid algorithm. After applied algorithm on it, it will return each document with a vector of terms, calculation of terms frequencies (TF) and inverse document frequencies (TF-IDF).

Separation: to breakdown text into distinct word using as delimiters, whitespace and newline.

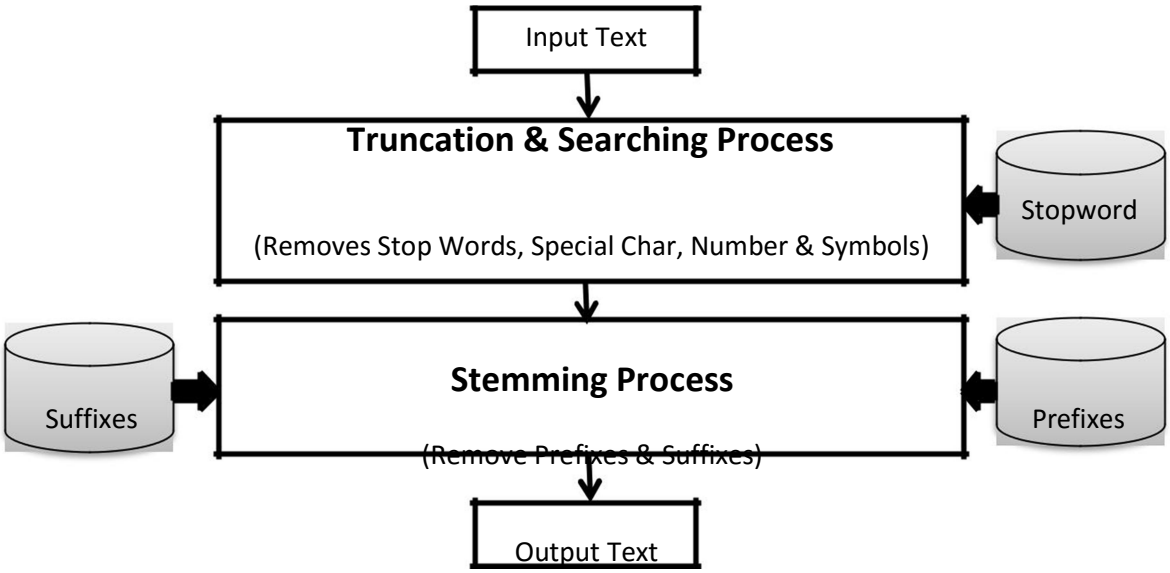

Figure 1 Stemming Algorithm Process

Searching: each word which getting after the processing from separation, first we have to check in our local root dictionary, If the word is available then we skipping the processing of Stemming Process because in this dictionary, we listed out those words which is already processed before. So, it will save our processing timing for stemming word.

4.4 Stemming Process. (in Figure 1)

The stemming algorithm, we have two special files with listed words which with range of 1 to 4 characters. This process consists two important sections:

Removing all possible prefixes from a word to have the possible outputs.

Removing all possible suffixes with mid-outputs to determine the correct stems according to the Gujarati language.

Add the stem word in our local dictionary for future purpose.

5. ANALYSIS OF RESULT

For the testing this algorithm, authors developed an application on Processor of 2.10 GHZ, 4 GB RAM and Netbeans 8.0.2 with several textual files to maintain different era such as Prefix list, Suffix List, Dictionary List and Input file and Output file.

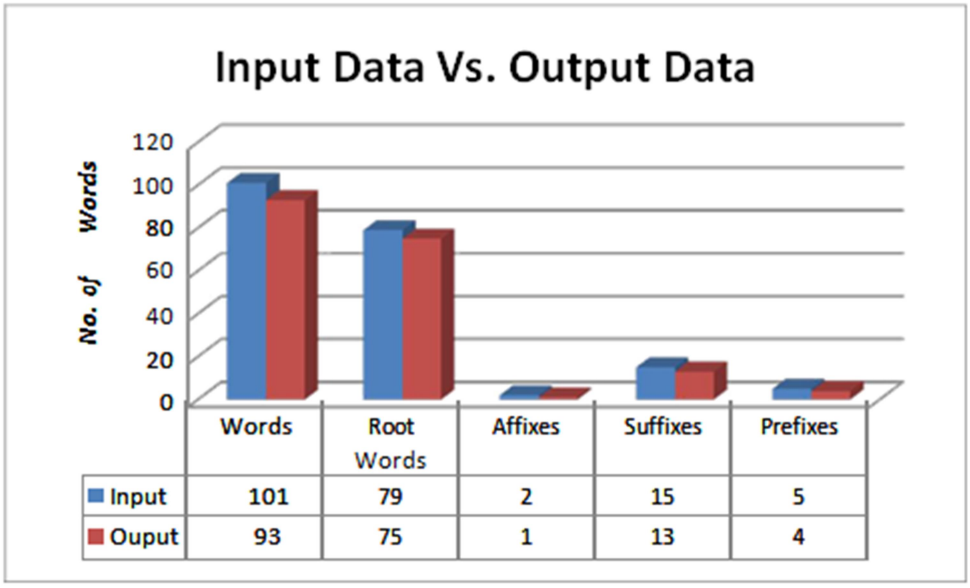

Figure 2 Sample of Input versus Output Data

After the input the words, we found the above results and I found some words cant read due to whitespace character while processing in the application. Here, I got some surprised stem words which are

meaningless and it automatically added to root dictionary such as □□□□□□□ (a sub caste in local region) but the result was □□□ (a colour).

6. CONCLUSION

The proposed algorithm revised using any intelligence model with accuracy. We hope that our proposed algorithm promising result after the adopting new handcrafted rules which again prevent meaning less word. We also evaluate new algorithm with IRS with precision and recall, improved. Since implementation of this algorithm also testing using different regional language for further processing.

REFERENCES

[1] J. Sheth and B. Patel, "Dhiya: A stemmer for morphological level analysis of Gujarati language," 2014 Int.
Conf. Issues Challenges Intell. Comput. Tech., pp. 151–154, 2014.
[2] P. Patel, K. Popat, and P. Bhattacharyya, "Hybrid Stemmer for Gujarati," Comput. Linguist., no. August, pp. 51–55, 2010.
[3] J. B. Lovins, "Development of a stemming algorithm," Mech. Transl. Comput. Linguist., vol. 11, no. June, pp. 22–31, 1968.
[4] M. F. Porter, "An algorithm for suffix stripping," Progr. Electron. Libr. Inf. Syst., vol. 14, no. 3, pp. 130– 137, 1980.
[5] M. F. Porter, "Snowball : A language for stemming algorithms," pp. 1–15, 2001.
[6] C. D. Paice, "An Evaluation Method for Stemming Algorithms," Proc. 17th Annu. Int. ACM SIGIR Conf.
Res. Dev. Inf. Retr., pp. 42–50, 1994.
[7] B. P. Pande, P. Tamta, and H. S. Dhami, "Generation , Implementation and Appraisal of an N-gram based Stemming Algorithm," pp. 12–14.
[8] R. Krovetz, "Viewing Morphology as an Inference Process," Proc. 16th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. (SIGIR '93), 1993.
[9] P. Majumder, M. Mitra, S. Parui, G. Kole, C. Vision, P. R. Unit, P. Mitra, C. Science, and K. Datta, "r P Fo ie w r P Fo ee r R ie w," 2006.
[10] E. T. Al-shammari, "Improving Arabic Document Categorization : Introducing Local Stem," pp. 385–390, 2010.
[11] J. R. Sheth and B. C. Patel, "Stemming Techniques and Naïve Approach for Gujarati Stemmer," pp. 975–
8887, 2012.

Authors

Chandrakant D. Patel, Research Scholar, M. C. A., B. C. A.

Chandrakant D. Patel having 10+ years' experience in Academic with UG and PG Courses. Prior to joining Acharya Motibhai Patel Institute of Computer Studies (PG Department), he worked with Shri C. J. Patel College of Computer Studies (UG Department), Visnagar from 2005 to 2011. His areas of interest include Web Page Categorization and Stemmer in Indian languages. He has also written a book for Hemchandracharya North Gujrat University BCA students titled "Computer Organization" in 2011. He has completed OP sponsored by UGC at ASC, Rajkot. He has 2 national and 4 international papers in his credit.

Jayeshkumar M . Patel, Associate Professor, Ph. D., M. C. A., B. B. A. (Marketing & Finance)

Jayeshkumar M. Patel, having rich experience of 11 Years in Academics, Industry, Research and International exposure, is holding Doctorate in ERP (Computer Science) from North Gujarat University. Rewarding his research work, he has been awarded “Career Award For Young Teachers” from AICTE, Delhi. He is working as a recognised Ph.D. guide at G.T.U., H. N. G.U., Ganpat University and also with many other reputed universities. He has good number of research under his name and presented more than 47 research papers in International and National Journals and Conferences. He has delivered number of expert talk in SANDHAN Programme and UGC Sponsored Programme. He is also the member of board of studies and selection committee of different universities.