ISSN NO. 2320-5407

*Journal homepage: http://www.journalijar.com*

*INTERNATIONAL JOURNAL OF ADVANCED RESEARCH*

**RESEARCH ARTICLE**

# A Review of Indian and Non-Indian Stemming: A focus on GujaratiStemming Algorithms

**Chandrakant Patel [(1)]; Dr. Jayesh M. Patel [(2)]**
Acharya Motibhai Patel Institute of Computer Studies, Ganpat University, Kherwa

## *Manuscript Info*

***Manuscript History:***

Received: 14 October 2015
Final Accepted: 16 November 2015
Published Online: December 2015

***Key words:***

Gujarati language;information retrieval; stemming algorithms and errors;

***Corresponding Author***

**Chandrakant Patel**

## *Abstract*

Gujarati language is indigenous language in the Indian State of Gujarat, known for its rich morphology structure. Gujarati information mining processes have become a current area of research. Many methods and approaches have designed and introduced algorithms to solve the problem of morphology and stemming of Gujarati language. Each researcher proposed his/her own standards, testing methodology and accuracy measurements to test his/her algorithm. Therefore, we can't make an exact comparison between these algorithms.



## INTRODUCTION

The WWW is growing extensively, and it helps as a major information resource for the individuals. As a result, the web mining plays an important role which saves all the required information available on the WWW. The difficult part of the web is, browsing for precise information. Since the documents on the web grow massive every day, the handling of these documents is necessary. Different methodologies have been proposed to get the textual information as needed. The text mining, accomplished using information retrieval (IR) methods. For example, a user searches for "Who is Dr. Kalam?" In this example, the unnecessary words are being removed, which includes who, is and Dr. This process is useful to make the search simpler by removing all unwanted words from theuser 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 stemmer ultimately increases the number of the documents of an IR system. Stemming has been focus for several decades because of its use and importance.

### I. Background and Related work:

The most cited algorithm in English was introduced by Porter in 1980 and then recursively apply different rules one by one and until no need to apply rule.The Porter stemmer faces main two issues. The first, it is not close to produced stem words for instance "generals" become "gener". The second, the words become overzealous for instance "punish" become "pun".

Krovetz argued in 1993 that meaning is essential while stemming. This algorithm resolved some issue of the Lovins and Porter stemmers.This algorithm reconstructed the rule set where word first search into dictionary and then remove a suffixes. So, Krovetz algorithm heavily depends on integrity of dictionary.

Research community makes more efforts to making stemmer with and without morphology of the language. (Majumder et al., 2007; Hull, 1996; Shrivastava et al., 2005; Ramanathan and Rao, 2003; Pandey and Siddiqui, 2008).

Natural language is note completely regular construct and processing on such a word by stemmer makes some mistakes. There are two types of error such as over stemming and under stemming. In over stemming, a two words probably same but represented as distinct after stemming process. Exa. “adhere” and “adhesion”. While in under stemming, two words actually different variants but probable consider as same. Exa. “experiment” and “experience”.

## II. Gujarati Morphology:

Gujarati is descended from Old Gujarati, which is derived from Sanskrit. In India, Gujarati is an official language for Gujarat State including UT of India, Daman and Diu.

Gujarati is a free order language. So, a grammar of this language is quit critical to understand and represent than other Indian language, independent freestanding morphemes. In general nouns refer to person, object, property, nature and action and in sentences; it will replace by subject or object. Basically nouns categories by gender, person, collection, material and nature wise. A gender forms sub categories by masculine, feminine and neutral. So, a noun can be represented as noun stem + gender marker +number marker.

**TABLE 1EXAMPLE OF GENDER NOUN**

| Masculine | Feminine | Neuter |
|---|---|---|
| ગ્રંથ | ચોપડી | પુસ્તક |
| વાંદરો | વાંદરી | વાદરું |
| છોકરો | છોકરી | છોકરું |

As a general rule, words ending in non-radical ઓ, ઈ or ઉ are respectively masculine, feminine or neuter. It will notice that the majority of these words are mono syllabic and that the final ઓ has arisen “from some phonetic corruption of the root syllable”. Such as ગો (a cow), ઘો (kinds of lizards), ધરો (young grass), હોહો (a loud voice), મો (the mouth).

In sentence, a token which represent action known as verb. A verb will make meaning of sentence. In other words, a verb means an action will take place by subject or person. The verb may be inflected or non-inflected which is depend on tense. The category of verb depends on person as well as tense. So, ainflected verb can be represented as verb stem + infinitive and non-inflected verb can be presented as verb stem + inflectional materials.

## III. Review of Indian and Non-indian Stemming algorithms

### A. English Stemmer

Julie Beth Lovins et. al., proposed a first publication with two main principles used in the construction of stemming algorithm. i.e. iteration and longest match.

Wahiba et. al. modified version of the original Porter stemming algorithm for the English language and calculated the number of errors and represented new version of Porter stemmer compared with Paice and Lovins stemmer.

### B. Spanish Stemmer

Asuncion Honrado et. al., a Spanish stemming algorithm which is based dictionary lookup with 300 rules to reduce suffixes from words.

### C. Uyghur Stemmer

Aishan Wumaier et. al., described CRF combined FSM stemming method to generate Uyghur noun inflectional suffixes using morphotactic rules in reverse order.

Aishan Wumaier et. al., proposed maximum entropy combined with FSM to resolve ambiguity between ending part of the some words.

### D. Arabic Stemmer

May Y et. al., proved that rule based algorithm provide highest accuracy among other algorithmsand got the result in rules based about 14% and positional letter rank algorithm by 7% to 10%.

Anas Boubas et. al., developed general verb patterns and then applied these pattern to derive morphological rules.

### E. Gujarati Stemmer

No work of derivational stemming for Gujarati language. Kartik Suba et. al., presented two stemmers for Gujarati language – The first stemmer based on a hybrid approach which basically work for inflected words and second stemmer baser on rule based which basically work for derived words.Authors aim is to ensure that the entire related stem may or may not be a meaningful word in the vocabulary of the language. The inflectional and derivational stemmer achieved accuracy of 90.7% and 70.70%, respectively.

Hardik Joshi et. al., experimented with less significant meaning words and eliminates from Gujarati document and improved the MAP.

Miral Patel et. al., proposed a clustering algorithm for Gujarati language which is preprocess for stemming with 50,000 tagged words set.

Juhi Ameta et. al. proposed rule based stemmer of Gujarati language and evaluated their results by supervised by human experts. The authors focus on the usage of open class such as noun, verb, adjective, adverbs in Gujarati language with list of 167 suffixes for extraction of root words. By experiments authors found that adding more suffixes, increased over-stemming errors and worked with limited suffixes such as સફાયો->સફ.

Pratik Patel et. al. developed unsupervised algorithm based on Goldsmith's (2001) take all split method with list of hand crafted 59 Gujarati suffixes. In this algorithm, there were two phase; In training phase – any word by taking all possible splits which maximizes the function as the optimal split position i.e. the entire word is considered as a stem with no suffix. In second phase – a list of stems and suffixes along with their frequencies and achieved an accuracy of 67.86%.

Juhi Ameta et. al., described the use of stemming and POS tagging for transliteration for Gujarati – Hindi Machine Translation.

Jikitsha Sheth et. al., suggested DHIYA a stemmer for Gujarati language based on the morphology of Gujarati and accuracy of the stemmer is 92.41%.

Niraj Aswami et. al., developed stemmer and analyzer for Gujarati after making rules from dictionary and corpus for suffix replacement where inflected as well as derived words would present. Authors consider prefixes and suffixesfor stemming process.

Majumder et al. (2007) present a system called YASS (Yet Another Suffix Stripper) that uses a corpus to learn suffix stripping rules. In this proposed study, string distance algorithm is used, for clustering the related words and intuition is to return long matching prefixes and to correct early discrepancy.

### F. Bengali Stemmer

There are several stemming algorithms proposed for Bengali language. Bengali language is the morphologically rich in nature. Sandipan Sarkaret. al., evaluated Bengali stemmer and identified few factors such as rules based stemmer, POS tagger, Lexicon and Standard performance techniques required for Bengali language.

### G. Nepali Stemmer

Chiranjibi Sitaula presented hybrid algorithm and the processing of words based on two methods i.e. traditional rules and string match approach and authors got accuracy of 70.10%.

### H. Urdu and Marathi Stemmer

These both languages are highly inflected in nature. In recent, language independent algorithm was developed i.e frequency based stripping and length based stripping. This algorithm is based on N-gram spilling model and evaluated with 1200 words are extracted from the EMILE corpus.

### I. Hindi Stemmer

In Hindi language, extremely less work has been done – especially for stemming algorithm. Hindi is a national language of India but not too much linguistic tools were developed. So, it is in early stage of research in Hindi Language. There are three methods to developed rules for suffixes: Stripping, Statically and rule based approach. Nouns plays vital role in Hindi language. So, Vishal Gupta et. al. proposed Hindi stemming for nouns with 16 suffixes stripping based approach.

## IV. New Vision: Why do we need Gujarati Stemmer?

In rural areas, 43% of the non-users of the Internet said they would adopt the medium if the content was provided in local language. In urban areas, 13.5% of the non-users mentioned that they would use the Internet if content is provided in local languages.

Internet users in India could increase by 24% if local language content searching is provided on the Internet, said a 25 February report by the Internet and Mobile Association of India and IMRB International.

Due to rapid growth of internet, stemming algorithm plays vital role in Information Retrieval System ((IRS) for improving of all regional languages. Most of stemming algorithm works based on rule based approached and is not quite sufficient for IRS. There is immediately need of language independent stemming algorithm for all regional language. The stemmer can improve the impact over recall and precision in IRS. In stemming algorithm – recall and precision can improve through part of speech and explore the ambiguity words with suffixes. Another issue to eliminations of stop words in Gujarati text documents contributes to a significant amount of increase in precision values in IRS. The size of sample that is considered in statistical stemming is under debate, if the smaller size of sample then stemming will faster but larger sample will take long time for stemming. So, optimal sample must be considered that may cover optimal time for stemming in regional language.

The goal of stemming is to reduce inflectional and derivational variants forms of word to a common base form. NLP is an under research area and where IRS always need stemming algorithm. Gujarati language is free order and rich morphology language. To developing a stemming algorithm will impact to IRS. Gujarati is resource poor language and such an algorithm make significant level in regional language. An algorithm for Gujarati language will boost up searching quality and efficiency.

## Selected References:

[1] Al-shammari, E. T. (2010). {Improving Arabic Document Categorization : Introducing Local Stem}. 385-390.

[2] Anjali, Jivani, G., & Anjali, M. (2007). {A Comparative Study of Stemming Algorithms}. October, 2(6), 1930-1938.

[3] Aswani, N., & Gaizauskas, R. J. (2010). Developing Morphological Analysers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages. LREC.

[4] Aswani, N., Gaizauskas, R., & Court, R. (1980). {Developing Morphological Analysers for South Asian Languages : Experimenting with the Hindi and Gujarati Languages}. 811-815.

[5] Bijal, D., & Sanket, S. (2014). {Overview of Stemming Algorithms for Indian and Non-Indian Languages}. 5(2), 1144-1146.

[6] Boubas, A., Lulu, L., Belkhouche, B., & Harous, S. (2011). {GENESTEM : A Novel Approach for an Arabic Stemmer Using Genetic Algorithms}. 77-82.

[7] Clifton, A., & Sarkar, A. (2010). {Morphology Generation for Statistical Machine Translation}. Northwest Regional NLP Workshop.

[8] Elrajubi, O. M. (2013). {An Improved Arabic Light Stemmer}. 2013, 33-38.

[9] Eryigit, G., & Adali, E. (2004). {an Affix Stripping Morphological Analyzer for Turkish.Pdf}. IASTED international Conference, 299-304.

[10] Estahbanati, S. (2011). {A New Stemmer For Farsi Language}. 25-29.

[11] Frakes, W. B. (1984). Term Conflation for Information Retrieval. Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 383-389). Swinton, UK, UK,: British Computer Society.

[12] Frakes, W. B., & Fox, C. J. (2003, #apr#). Strength and Similarity of Affix Removal Stemming Algorithms. SIGIR Forum, 37(1), 26-30.

[13] Gupta, V., Joshi, N., & Mathur, I. (n.d.). {Design \& Development of Rule Based Inflectional and Derivational Urdu Stemmer â€˜ Usal â€™}. (i).

[14] Gupta, V., Joshi, N., & Mathur, I. (2013). {Rule Based Stemmer in Urdu}. 129-132.

[15] Harman, D. (1991). How effective is suffixing? Journal of the American Society for Information Science, 42(1), 7-15.

[16] Hegde, Y., Kadambe, S., & Naduthota, P. (2013). {Suffix Stripping Algorithm for Kannada Information Retrieval}. 527-533.

[17] Krovetz, R. (1993). Viewing Morphology As an Inference Process. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 191-202). New York, NY, USA: ACM.

[18] Lee, J., Othman, R. M., & Mohamad, N. Z. (2013). {Syllable-based Malay Word Stemmer}. 7-11.

[19] Lovins, J. B. (1968). {Development of a stemming algorithm}. Mechanical Translation and Computational Linguistics, 11(June), 22-31.

[20] Mahmud, R., Afrin, M., Miller, E., Iwashige, J., Ellis, E., Miller, D. O., et al. (2014). {A Rule Based Bengali Stemmer}. 2750-2756.

[21] Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., & Datta, K. (2007, #oct#). YASS: Yet Another Suffix Stripper. ACM Trans. Inf. Syst., 25(4).

[22] Mayfield, J., & McNamee, P. (2003). Single N-gram Stemming. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (pp. 415-416). New York, NY, USA,: ACM.

[23] Paice, C. D. (1990, #nov#). Another Stemmer. SIGIR Forum, 24(3), 56-61.

[24] Paice, C. D. (1994). An Evaluation Method for Stemming Algorithms. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 42-50). New York, NY, USA,: Springer-Verlag New York, Inc.

[25] Popat, P. P., & Bhattacharyya, P. (2010). Hybrid Stemmer for Gujarati. 23rd International Conference on Computational Linguistics, (p. 51).

[26] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.

[27] Porter, M. F. (2001). Snowball: A language for stemming algorithms. Snowball: A language for stemming algorithms.

[28] Pragisha, K., & Linguistics, M. T. (2013). {STHREE : Stemmer for Malayalam Using Three Pass Algorithm}. (Iccc), 149-152.

[29] Raj, P. C. (2013). {LALITHA : A Light Weight Malayalam Stemmer Using Suffix Stripping Method}. (Iccc), 244-248.

[30] Ransliteration, T. (2013). {I MPROVING T HE Q UALITY OF G UJARATI -H INDI M ACHINE T RANSLATION T HROUGH P ART - OF - S PEECH T AGGING AND S TEMMER ASSISTED}. 2(3), 49-54.

[31] Sembok, T. M., Ata, B. M., & Bakar, Z. a. (2011). {A Rule-Based Arabic Stemming Algorithm}. Proceedings of the 5th European Conference on European Computing Conference.World Scientific and Engineering Academy and Society (WSEAS)., 392-397.

[32] Sheth, M. J. (2014). {Dhiya : A Stemmer for morphological level analysis of Gujarati language}. 151-154.

[33] Sitaula, C. (2013). {A Hybrid Algorithm for Stemming of Nepali Text}. 2013(July), 136-139.

[34] Sunitha, K. V., & Kalyani, N. (2009). {A Novel approach to Improve rule based Telugu Morphological Analyzer}. Nature \& Biologically Inspired Computing (NaBIC 2009) IEEE, 1649-1652.

[35] Taghva, K., Beckley, R., & Sadeh, M. (2003). {A Stemming Algorithm for the Farsi Language}

[36] Taghva, K., Elkhoury, R., & Coombs, J. (n.d.). {Arabic Stemming Without A Root Dictionary}. 3-8.

[37] Thangarasu, M., & Manavalan, R. (2013). {A Literature Review : Stemming Algorithms for Indian Languages}. 4(8), 2582-2584.

[38] Wumaier, A. (2009). {Conditional Random Fields Combined FSM Stemming Method for Uyghur}.

[39] Wumaier, A., & Tursun, P. (2009). {Uyghur Noun Suffix Finite State Machine for Stemming}. 9-12.

[40] Wumaier, A., Kadeer, Z., Tursun, P., & Tian, S. (2009). {Maximum Entropy Combined FSM Stemming Method for Uyghur}. 51-55.