International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

An Affix Removal Stemmer for Afaraf Text

Kelil Ali Ebrahim#* and Rahul Joshi^

#Department of Computer Science Symbiosis International University, Pune, India
^Department of Computer Science and IT, Symbiosis International University, Pune, India

Received 20 Sept 2017, Accepted 23 Nov 2017, Available online 30 Nov 2017, **Vol.7, No.6 (Nov/Dec 2017)**

Abstract

*Stemming is the action of removing affix from inflected words to their base stem word. Eg, in the set {speaker, speaking, speaks} the stem is 'speak'. Affix is contains of **suffix** and **prefix**. Suffix attached at end of root word where as prefix is attached begining of root word (Bijal Dalwadi ,2016). In general, word-stemming algorithm is developed for various AI (artificial intelligence) applications based on the morphological rules of specific natural languages. It is used in IR systems, indexing, document mining, document classifiers, document summarization, dictionary search etc. (Bijal Dalwadi et al, 2016; Belal Abuata et al,2015). In this survey the researcher refers articles, journal, and books those are directly related to research domain. In this paper the researcher try to review the related articles of previous Afaraf morphology works done, to guide us to design stemmer and understand Afaraf language morphological rules. And obtained good understanding on fundamental and state of the arts of the domain and tried to access latest approaches, methods and techniques.*

***Keywords**: Afaraf, Stemming, Affix, Prefix, Suffix, Rule based technique, over stemming, under stemming.*

1. Introduction

Stemming is the conflation of the different forms of a word into a single root form, i.e., the stem. Affix consists of **suffix** and **prefix**. Suffix attached ends of root word where as prefix is attached begining of root (stem) word (Bijal Dalwadi ,2016).

In general, word-stemming algorithm is developed for various artificial intelligence applications based on the morphological rules of specific natural languages. It is used in IR systems, indexing, document mining, document classifiers, document summarization, dictionary search etc. (Bijal Dalwadi *et al*, 2016; Belal Abuata *et al*, 2015). This paper suggests a stemmer for **Afar af text** ,which conflates term by **affix removal** and will use rule based techniques based on the morphological rules of afar lalngauge and prefix and suffix rules will apply. In Rule based approach rule are generated based on the morphological 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 form. Afaraf has two different affixes that are the prefix and suffix. Unlike English language stemmers that work good by removing only suffixes to get the stems form (Bijal Dalwadi *et al*, 2016). But an effective Afaraf stemmer must be able to remove both the **suffixes** and prefix. The researcher planned to accomplish in these survey good understands on fundamental and state of the arts of the area and tried to access latest approaches, methods, techniques and attempt reviewed literature articles and works done on Afaraf morphology. And have deep understanding on topics based on the existing researches.

2. Objectives of the literature survey

The main objective of the Literature survey is:

- To knowing who writes, what and where about text stemming development and design process
- To review previous work done of Afaraf morphology
- To distinguish the implements and sources for text/word stemming development process
- Access to latest approaches, methods and techniques
- Gaining an understanding on the fundamentals and state-of the art of the area
- Learning the definitions of the basic concepts

3. Literature survey (Related works)

The study that developed stemming algorithm for Gujarati language with experimental result of accuracy is 96.63 and used dictionary look up (offline) and rule based techniques and evaluate by over stemming and under stemming. The developed stemmer does not stem loan words (borrowed, Scientific, etc.) (Bijal dalwadi *et al* 2016).

*Corresponding author's ORCID ID: 0000-0001-8133-2573

A Rule based stemmer (*GUJSTER*) using Dictionary Approach developed *an* algorithm for reducing affixes from the Gujarati words and get optimal output as compare to the previous hybrid algorithms. And this stemmer checks stem word into online Gujarati dictionary (online), with the accuracy of 97.09%. The developer *Used rule based technique* and the developed algorithm can Stem loan words ((borrowed, Scientific, etc.). (Chandrakant D. Patel *et al*, 2017).

Stemming challenges in the **existing word stemming algorithms** for the Malay language that have a significant impact on the accuracy of word stemming algorithm to remove affix effectively. This research proves that till now stemming Algorithm for Malay text is in challenge (ineffective) because of morphological rules complexity.

Finally this study suggested that: These word stemming challenges could be addressed by having a *word stemming algorithm* that: has Stemming features of document preprocessing for removing special characters,

Has Word stemming rules based on morphological rules of Malay language to *stem standard derived words* Has Word stemming rules based on word patterns to stem non- standard derived words (loaned, scientific, informal (slang words)).

Promising stemming Accuracy (correctly stemmed words), Efficiency (process speed), Compactness (reduce rules) (mohammed N.K *et al* ,2016;)

Assessing the impact of Stemming Accuracy on Information Retrieval – A multilingual perspective, was developed to evaluate stemmers for several languages (English, Portuguese, Spanish and French) that found in literature in term of **accuracy and impact on achievement of information** retrieval. And concluded that most accurate stemmer was not the one to achieve greatest improvement in *Information Retrieval* (**IR)**, in **none** of the language (Felipe N. Flores *et al*, 2016).

A rule based stemmer for ge'ez text is developed with accuracy of **82.42% and 17.58%** of error that is due to over stemming and under stemming error. In this research researcher used **rule-based technique** to generate rules based on morphological rule of Ge'ez language and used **python** for developing tool (Abebe Belay *et al*, 2017).

The stemmer for Marathi text is developed by using **Marathi Word Net, which** used for prevent **over stemming and under stemming**. Marathi Word Net consists mainly of Marathi root and base words. Also in addition to affix removal they use **root verification** by using **Word Net, Marathi name entity dataset** and **exceptional word dataset** that prepared by researcher. And used rule-based techniques (Pooja Pandey *et al*, 2016)

The primary goal of Rule-based stemmer for Arabic Gulf dialect was to derive an efficient algorithm to extract the stem of dialect words used in Arabian Gulf countries (Kuwait, Bahrain, Qatar, UAE, Saudi Eastern Area, and South of Iraq). Used rule based techniques. The paper proves that the existing MSA (Modern Standard Arabic) stemmer (khoja and darwish, 1999 and 2000 respectively) are not applicable to Gulf Arabic dialects and their performance was low. Result for the three stemmers for Arabic gulf dialects with total number of words 5436 were: - *for new rule based algorithms = 88%, khoja = 39 % and darwish =28%.* And the new algorithm can handle all *known Arabic dialects* by defining new rules and integrating these rules with the current used rules (Belal Abuata *et al*, 2015).

A novel root based Arabic stemmer developed in this study and better in accuracy when compared with two well-known Arabic stemmers (Khoja and Garside *et al*, 1999; Ghwanmeh 2009) and with the accuracy of 75.03% by using more additional rules based on morphological rules of Arabic language and it is based on light and heavy (root-based (root verify) stemming methods and use C# language for coding (Mohammed N. Al-Kabi *et al*, 2015).

A Hybrid Approach used to Stem Punjabi Words working for Punjabi text is developed with average accuracy of 81.27%. Used naïve techniques that emphasis on the word to be found in the database (root word and stripping suffix databases) and uses suffix stripping techniques to remove suffixes from words (Puneet Thapar *et al*, 2014).

4. Afaraf Morphology

Afaraf language is speaking in the horn of Africa by ethnic group of Afar. Afaraf language is part of the Cushitic branch of the Afro-Asiatic family. More 2 Million peoples speak it and most of native speakers are people living in Ethiopia, Djibouti and Eretria Loren (F.Bliese *et al*, 1981; jamal qabdaulkadir, *et al*, 2007,Usman taha *et al* 2015).

Morphology in Afaraf deals with all combinations that form words or parts of words. Two main classes of morphemes, stems and affixes: Stem is the root of the word, supplying the main meaning, and Affixes add additional‖ meanings in words.
Eg. Oggol-**e** root oggol (accept)
T-abbe root abbe (hears)

Affixes

Affixes: Affixes are combination of suffix and prefix and well-known methods that are used to identify morphological variants which is common to all languages.
Prefixes: A Prefix is an affix that attached in front of a stem.
Eg**. t-**ablee (do you see)
Suffixes: A Suffix is an affix that attached after the stem. Eg. gex-**xee (**did you go**)**
2.2.2 **Word Formation**
Word formation: The word can form one or more morphemes. There are three broad classes of ways to form words from morphemes and Afaraf use the three forms in word formation, they are: **inflection**, **derivation** and **compounding (**jamal qabdaulkadir, *et al*, 2007,Usman taha *et al* 2015)

Inflection is the combination of a word stem with a grammatical morpheme, usually resulting in a word of the same class as the original stem, and usually filling some syntactic function and is productive in Afaraf (Jamal qabdaulkadir, *et al*, 2007,Usman Taha *et al* 2015). The word **inflectional** morphemes modify a word's tense, number, aspect, and so on.
Eg. Gex (go), Gex-**e** (he went)

Derivation is the combination of the word stem with a grammatical morpheme, typically resulting in a word of a different class. In case of derivation Afaraf morphology is unproductive (Jamal qabdaulkadir, *et al*, 2007,Usman taha *et al* ,2015).
Eg. 'Maqub' is a noun; 'Aqub' is a root of verb
Compounding is the joining of two or more words to form a new word. Afaraf use a word compounding like other languages.
Eg. Tibba-**exceh** (I keep silent)
Tibba-**inneh** (we keep silent)
Tibba-**inteh** (she keep silent)
Tibba-**inxic** (you, keep silent)

4.1 Alphabets and Sounds

Afaraf uses Latin character (Roman alphabet), and the basic sound system of Afaraf has some modifications on sound of consonant and vowels. Afaraf has the 17 consonants and ten vowels Afaraf has ten vowels: [a], [i], [e], [u], [o] and their long counterparts [aa], [ii], [ee], [uu] and [oo] **(**Usman taha *et al* 2015). Furthermore, in 20th century and in the 2nd millennium the two necessities introduced in Afaraf's written words. One has something to do with the consonants; the consonants have increased from 17 to 21 that use all the Roman alphabet and the other with the double consonants, which is two-letter and three composed sounds like **sh, ch, kh, ts and tch and** they used mostly for nouns (Usman taha *et al*, 2015)

In this Afaraf morphology literature review tried to reviewed includes Afaraf word formation (inflection, derivation and compounding), Affix (suffix and prefix), Afaraf grammar, dialect and varieties, alphabet and sounds, number morphology (singular and plural), Personal pronoun, Adjectives, Adverbs, Verbal-noun, Strong and weak verb, Indefinite pronoun, Conditional and subjunctive mood, Linkage and Afaraf tenses to remove suffixes and prefixes from the word and produce stem word.

5. Proposed Stemmer

The main objective of this study is to develop Affix removal stemmer for Afaraf text. Stemming is the process or normalization that reduces the morphological variants of words like inflected or derived words to a common form usually called a stem by the removal of affixes, in this study tokenization, normalization, stop word removals and stemming would be covered.

5.1 Rule based techniques

Rule based techniques: It is used to generate rules based on morphological rules of each languages and the most commonly accepted technique for its high precision and recall.

Normalization: Normalization Afaraf Normalization involves process of normalizing a terms in the document in to similar case format. For instance Xaagu' to xaagu' Qari' to qari' are all normalized to understandable as lower case.

Stop-word removal: Stop-words are words, which occur frequently in every document so in this study those stop words, will remove.

Tokenization*:* Tokenization is the task of chopping it up into pieces, called *tokens*, perhaps at the same time throwing away certain characters, such as punctuation. **Eg**: input: Friends, Romans, Countrymen, lend me your ears; output: Friends Roman Countrymen lend me your ears (Bijal Dalwadi, 2016).

Stemming*:* It is the process of removing affix from derived words to their base stem form for instance stemmer, stemming, stems into stem (Bijal Dalwadi, 2016). Proposed stemming algorithm for Afaraf text based on morphological rule of Afaraf

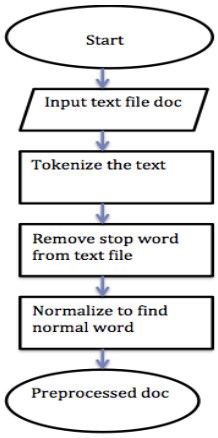

Fig. 1 Flow chart of preprocessing

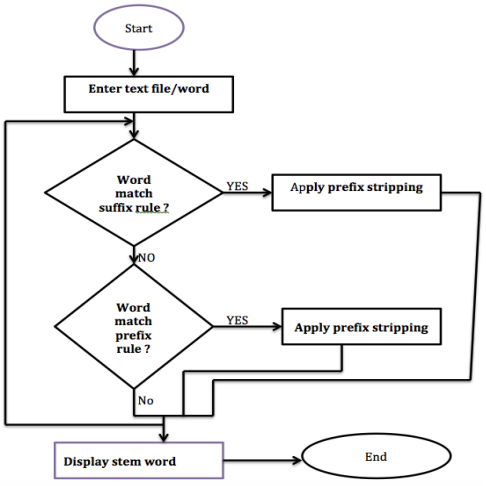

Fig 2. Proposed Afaraf text Stemming flow chart

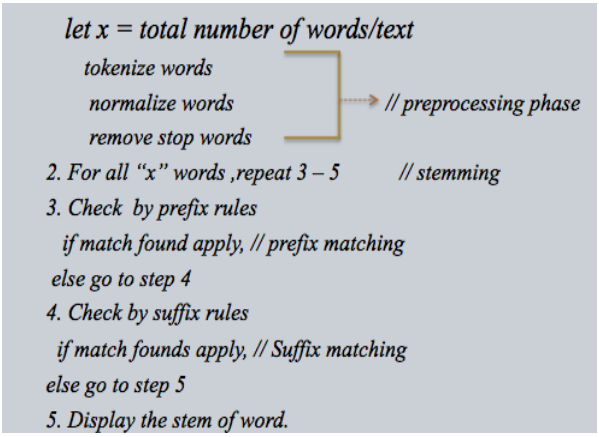

Fig.3 Proposed Afaraf texts words stemming algorithm

In this process of stemming the system takes text file as in put and it perform stemming after the preprocessing phase such as: - tokenize, normalize and remove stop words. Finally it performs stemming process by removing suffix and prefix from verbs and display-stemmed words as output.

Table 1 Sample output of Afaraf stemmer

Verb	Stem	Suffix	Prefix
Gaciyya	Gac	iyya	
Gexiyya	Gex	iyya	
Taaxige	aaxige	-	t
Table	Able	-	t

Conclusion

This study of literature survey was done to improve an understanding on the fundamentals and state-of the art of the area and identify the tools and sources of text stemming development process to develop stemming algorithm for Afaraf text. In this study also tried to access latest approaches, method and concepts about the topic.

The researcher referred articles, journal and books and so on those are directly related to the studying are and reviewed the Afaraf (Afar language) morphology literature work done before such as Afaraf word formation (inflection, derivation and compounding), Affix (suffix and prefix), and other morphological rules of Afaraf.

Reference

Bijal Dalwadi,(2016) An affix Removal stemmer for Gujarati text, *INDIACom*, pp 2296-2299.

Chandra Kant D. Pate,(2017),GUJSTER: a Rule based stemmer using Dictionary Approach ,"*ICICCT* , pp 496-499

Mohamad Nizam Kassim,(2016),Word Stemming Challenges in Malay Texts: A Literature Review,"*IcoICT* , pp 2326-2329.

Pooja Pandey1, Dhiraj Amin**,**(2016), Rule based Stemmer using Marathi Word Net for Marathi Language." *IJARCE,* pp 278-281,

Abebe Belay Adege, Yibeltal Chanie Manie,(2017), Designing a stemmer for ge'ez text-using rule based approach,*IJESR* ,(8), pp 1574-1578.

Sangita D.Makhija,(2016),Study of different Stemmer for Sindhi Language based on Devanagari Script", *INDIANCom,* pp 2326-2329.

Flores, F.N., Moreira, V.P,(2016),Assessing the impact of stemming accuracy on information retrieval,Inf. *Process, Manage*,(52) , 840–854.

Belal Abuata , Asma Al-Omari,(2015),A rule-based stemmer for Arabic Gulf dialect." *Com.Inf.Sci*, pp 104-112.

Puneet Thapar,(2014),A Hybrid Approach used to Stem Punjabi Words. "IJCSMC, (3), pp 1-9.

Tomáš Brychcín, Miloslav Konopík,(2015) HPS: High precision stemmer, *Info. Process,Manage*, (51), pp 68–91.

Osman Taha(2015)Development Of A Stemmer For Afaraf Text Retriev, *Addis ababa university* pp 10-32

Mohammed N. Al-Kabi , Saif A. Kazakzeh , Belal M. Abu Ata , Saif A. Al-Rababah , Izzat M.

Alsmadi,(2015), A novel root based Arabic stemmer, *Computer and Inf. Scie* ,(27), pp 94–103.

Jamaal Qabdulkadir Reedo,(2007), Qafar Afak YAbti Rakiibo, *Qafar Afih Cubbuso kee Gaddaloysiyyi Fanteena*, ,pp 3-134.

Loren F.Bliese,(1981), A Generative Grammar of Afar. *The Summer Institute of Linguistik and University of Texas at Arlington*, pp 29-90.