Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 5, June 2021: 5209-5214

Development of Stemmer for Afar-af text: A Hybrid Approach

Kelil Ali Ebrahim

Department of Computer Science,College of Engineering and Technology Samara University, Samara, Ethiopia,PoBox: 132, Email: kelilaliebrhm@gmail.com

Abstract: Utmost natural language processing systems practices stemmer as a distinct module in their architecture. Specially, it is crucial for developing, machine translator, speech recognizer and search engines. In linguistic morphology, stemming is the process for reducing inflected (or sometimes derived) words to their root, stem or base form.

In this article, a stemming system for Afar-af is presented. This system takes as input a word/terms and removes its affixes (suffix, prefix) rendering to a **rule based** algorithm. This stemmer is not adequate to describe every rule applied in Afar-af word formation. Consequently, N-**gram** is combined with the rule to handle cases that are not covered by rule in the **hybrid** approach of this stemmer. The algorithm follows the well-known Porter algorithm for the English language and it is advanced according to the grammatical rules of the Afar, language.

Afar-af morphology was studied and defined in order to model the language and develop an automatic procedure for conflation. The inflectional and derivational morphologies of the language are discussed

Afar-af words are very rich in morphology and requires an operative stemming algorithm, which can regulate diverse morphological arrangements that are associated with words.

An evaluation of the system indicates that the algorithms accuracy works with better performance than other earlier stemming algorithms for Afar-af giving accuracy of 98.73 percent. Furthermore, Possible extensions of the planned work and advance evaluation approaches are briefly reviewed

Keywords: Afar-af, Stemmer, Hybrid, Affix, Rule based.

1. INTRODUCTION

One of the efforts to make the search engines further operative in information retrieval was the practice of word stemming. A stemming algorithm is a process that diminishes all words with the similar stem to a common form by stripping of its inflectional and derivational suffixes [23]. The core objective of the stemming procedure is to remove all probable affixes and therefore diminish the word to its stem [23,12]. Applying stemming, several modern-day search engines associate words with prefixes and suffixes to their root word, to make the search wider in the meaning that it can confirm that the highest number of significant matches is involved in search outcomes. In this article, discussed the two critical phase of preprocessing and Stemming [14,15].

This article used a hybrid approach that combines rule based and n-gram composed to develop stemmer for Afar-af text. Porter's algorithm which uses rule based approach to design the stemmer of English language [2,11.13]. Herein stemmer each rule that works for each morpheme formation is not comprehensively known due to the nature of the language. Therefor statistical algorithm takes over when there are no rules that applies for a given word.

Stemming is a process for reduce derived or deflected words to their base or root forms. For example, in the set {stemmer, stemming, stemmed and stems} in to root form '**Stem' [1]**.

Stemming improves Information Retrieval performance generally by bringing different forms of a word which share a common meaning under one heading [22].

The study also covers text process like tokenization, stop-words removal, and normalization as pre-processing phase. Example suffix: - {Abeh, Abeyyo, Abetto, Abele,Abneh, Abeenih,Abaanah….} the root or stem form is '**Ab'** English meaning (he did, I will do, you will do, he will do, we did, they did, they are doing…). example for both prefix and suffix: - in the word 'Maabiyyo' () ma is a prefix that used as negation and iyyo is suffix that that shows activity, and '**ab'** is the root form for the word 'maabiyyo'.

2. RELATED WORKS

In Afar af language the stemmer algorithm was designed by Osman Taha and Kelil Ali in which both of them have used Rule based technique to find the stem word of a derived word.

However, the key problem in these stemmers were over-stemming and under-stemming because of they used only ruled based approach for Afar-af language which is morphologically rich [1].

Over-stemming occurred when the words that are not morphological variants are conflated. Under-stemming occurred when words that are truly morphological variants are not conflated.

The study that developed stemming algorithm based on rule based techniques for Gujarati language with algorithm claimed experimental accuracy of 97.09% [4]. MarS algorithm for Marathi language which is a morphologically rich language. The algorithm MarS practiced on their particular dataset and using rule based method. The approach they used brings high over stemming & under stemming errors. In this tasks the developed algorithm claimed an accuracy of 79.97% [5]. The study for Malay text using three rule based stemmer. These algorithms used to remove suffixes, prefixes, infixes, enclitics, particles and proclitic [6]. Study for Punjabi language stemming algorithm to establish the stem word of Panjabi language. The study used a hybrid approach, to develop the algorithm by removing the derivational forms to find out the root word. They also had used rule-based techniques and lookup table to performs this algorithm. The experimental accuracy of this work shows 86% [16]. A light weight stemmer of hybrid approach for Gujarati language was designed [17] to develop stem word using IR system and get rid of derivational words by using EMILLE corpus to execute this hybrid approach for Gujarati language. The result of this approach gives 68% accuracy.

Article developed rule-based & string matching stemmer over hybrid approach, for Ngoko Javanese language. The work used to two different algorithms that is string matching and Rule based approach to find out the stem word of Javanese language. After applying of these two technique give accuracy 66% [18]. Adel Rahimi [19] developed hybrid stemmer algorithm for Persian language. Herein work

he had designed hybrid stemmer used dictionary & rule-based stemmer to get root or stem word for Persian language. The accuracy of this work shows 97%.

Meryeme Hadni et al. [20] announced Arabic stemming algorithm, to find out the stem form of Arabic words, by using hybrid stemmer techniques. This algorithm gives better accuracy than others for Arabic text categorization.

3. STEMMING PROCESS

The figure in Fig.1 indicates the process of stemming. When a word is given for stemming, the algorithm gives the root word with the assistance of a lookup table. A lookup table holds the derived form of the words [7.21]. A stemming development gets the inflected form of a word from the look up table. The key cons of stemming are that it cannot work with fresh and unaware words. It can only receipts all varied words from the lookup table so varied words essentially required to be existing in this table. For simple morphology language like English, the table extent could be simple.

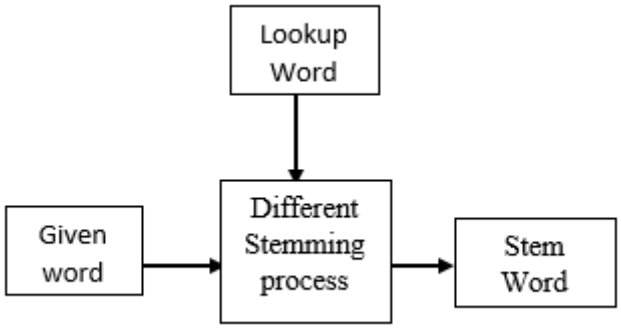

Fig.1. Stemming Process

3.1. Rule based Approach

Herein method, a set of rules are determined to discover the root word [9]. The rules may be designed using different methods as displayed in Fig.2 In suffix stripping more or less letters are eliminating from the end of the given word. Example: (Afar-af) **abiyya** into ‘ab’. In porter stemmer rules are there to find out the stem word. Instance: (English) infection into ‘infect’.

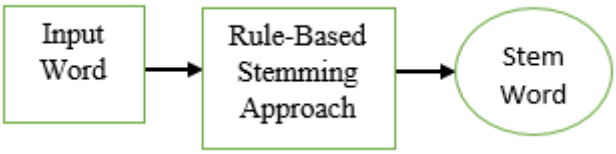

Fig.2. Rule based Approach

3.2. Hybrid Approach

Herein approach two or more methods are used to advance stemming performance as shown in Fig.3. For example, in this approach may mix a Brute-force algorithm for building lookup table, then it applies suffix tree algorithm to find out root word. In a particular language there are amount of words

and lots of interactions between the words. So it is moderately hard to store all these data in a certain list [8,10]. The lookup table is supportive for storing it in a small list entitled exception list. If any word is not existing in the exception list, then to find the root word Suffix stripping algorithm perhaps employ.

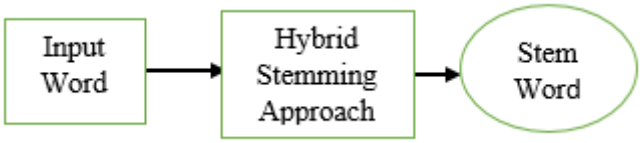

Fig.3. Hybrid Stemming Approach

Improved Hybrid stemming algorithm for Afar-af text

1. READ the word to be stemmed from file
2. OPEN stop word file
Read a word from the file until match happens or End of File reached
IF word exists in the stop word list
Go to 5
Else
Go to 3
3. If word matches with one of the rules
Remove the suffix and do the necessary adjustments
Go back to 3
ELSE
Go to 6
4. Return the word and RECORD it in stem dictionary
5. IF end of file not reached
Go to 1
ELSE
Stop processing
6. IF there is no applicable condition and action exist
Apply N-gram stemmer
Go to 4

First of all, this algorithm checks if a word is in the stop word list or not. If found in the list, the word is excluded from additional procedure and nothing returned to the calling

routine; end and process the next word if any. If the word is not in the stop word list, the word is tested for any match in the rule clusters. If a match is found, the respective act for that rule will be applied. If a match is not find n-gram stemmer is generated and returns the stem for the calling function. further the suffix –s in rule cluster is also held by n-gram as it has no universal rule.

After the developments, the fresh or improved stemmer is run on the similar set of experiment data that was used for the first version. Thus, the amount of under-stemmed and over-stemmed words were

reduced to 0.60% for (20 words) and 4.22% (95 words) correspondingly. The over-all mistakes account for 5.13% (110 words) and the performance of the new stemmer is improved to 97.83%.

4. Conclusion and recommendation

Stemming algorithm is crucial for extremely inflected languages like Afar-af for several applications that call for the stem of a word. In this article, a hybrid stemming technique was used that tries to decide the stem of a word according to morphological rules and n-gram. The technique fit in two distinct stemming methods to advance the general performance of the stemming procedure. Correspondingly to the evaluation of the investigates, it could be concluded that a general accuracy of almost 97.83% is inspiring outcome which displays stemming could be accomplished with little error amounts in extremely inflected languages like Afar-af language.

Lastly, I believe that this paper contributes in the stemming study and proposal a retrieval instrument for Afar-af text that could be used on web, perhaps have its weaknesses and additional enhancements may be needed to advance the stemming algorithm and the effectiveness of the stemming methods.

5. REFERENCE

6. Kelil Ali Ebrahim and Rahul Joshi. An Affix Removal Stemmer for Afaraf Text (INPRESSCO). [2017].
7. Adel R.: Hybrid stemming for Persian. arXiv: 1507. 03077v2. (2015)
8. Designing Stemmer for Afaraf Text using Rule based approach Kelil Ali, Dr saidhbi. (Innovations in Computer Science and Engineering (pp.281-288) (2022)
9. Chandrakanta, D. P., Jayesh, M. P.: GUJSTER: A Rule based stemmer using Dictionary Approach. InternationalConference on Inventive Communication and Computational Technologies (ICICCT). (2017) 496-499.
10. Harsali, B. P., Ajay, S. P.: Mars: A Rule-Based Stemmer for Morphologically Rich Language Marathi. International Conference On Computer, Communications and Electronics. (2017) 580-584.
11. Abdul, M., M. Kamran M., Zubair, N., H. M. Danish, M. Hassan S., Qaiser A.: A Hybrid Stemmer of Punjabi Shahmukhi Script. IJCSNS International Journal of Computer Science and Network Security. 17(8), (2017).
12. Mohamad, N. K., Mohd, A. M., Anazida, Z., Amirudin, A. W.: Word Stemming Challenges in Malay Texts: A Literature Review. International Conference On Information and Communication Technologies (Icoict). (2016
13. Vaishali, G., Nisheeth, J., Iti, M.: Design & Development of Rule Based Inflectional and Derivational Urdu Stemmer ‘Usal’. International Conference On Futuristic Trend in Computational Analysis & Knowledge Management. (2015) 7-12.
14. .Md. Nesarul, H., Md. Hanif, S.: Bangla Parts-Of Speech Tagging Using Bangla Stemmer And Rule Based Analyzer. International Conference On Computer and Information Technologies. (2015) 440-444.

15. M.Thangarasu, Dr.R.Manavalan. : Stemmers For Tamil Language: Performance Analysis. International Journal of Computer Science & Engineering Technology (IJCSET). 4 (07 Jul 2013) 902-908.
16. Pratikkumar P., Kashyap P., Pushpak B.: Hybrid Stemmer for Gujarati. International Conference On Computational Linguistics (COLING), Beijing. (2010) 51-55.
17. F. Amin, W. Hadikurniawati, S. Wibisono, H. Februariyanti, J. S. Wibowo.: A Hybrid Method of Rule-Based And String Matching Stemmer For Javanese Language. Journal of Theoretical and Applied Information Technology. 95(19) (2017) 4973-4982.
18. S. Singh D., I. Shrestha.: A New Stemmer for Nepali Language. 2016.
19. N. Alami, M. Meknassi, S. A. Ouatik and N. E. E.: Impact of Stemming On Arabic Text Summarization. IEEE. (2016) 338- 343.
20. R.J. Prathibha, M.C. P.: Design of Rule Based Lemmatizer for Kannada Inflectional Words. International Conference On Emerging Research in Electronics, Computer Science and Technology. (2015) 264-269.
21. Adel R.: Hybrid Stemming for Persian. Arxiv: 1507. 03077v2. (2015)
22. .Wahiba Ben Abdessalem K.: A New Stemmer To Improve Information Retrieval. International Journal of Network Security & Its Applications (IJNSA). 5 (4) (2013) 143-154.
23. Meryeme H., Said Alaoui O., Abdelmonaime L.: Effective Arabic Stemmer Based Hybrid Approach for Arabic Text Categorization. Nternational Journal of Data Mining & Knowledge Management Process (IJDKP). 3 (4) (2013) 1-14.

A. Ghazvini, M. J. Ab Aziz.: Stemming Algorithm for Different Tenses to Improve Persian Dictionary. IEEE Symposium On Industrial Electronics and Applications (ISIEA) (2012) 50-53

24. Mubashir, A., Shehzad, K., Muhmmad, H. A.: Pattern Based Comprehensive Urdu Stemmer and Short Text Classification. IEEE. 6 (2017) 7374-7389.
25. Dhawan, Singh, Garg Hybrid Approach for Stemming in Punjabi. Chandni Dhawan et al, International Journal of Computer Science & Communication Networks,Vol 3(2), 101-104.
26. Debela Tesfaye, Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach (ADDIS ABABA UNIVERSITY). (2010).
27. Lovis, Development of a Stemming Algorithm. (Mechanical Translation and Computational Linguistics, vol.11, nos.1 and 2, (1968).