Sindh Univ. Res. Jour. (Sci. Ser.) Vol.49 (004) 813-816 (2017)

http://doi.org/10.26692/sujo/2017.12.0063

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Stemming Software Application of Shahmukhi Script Using Porter's Algorithm

M. IRFAN, J. A. MAHAR[++], H. SHAIKH, F. A. SURAHIO

Department of Computer Science, Shah Abdul Latif University, Khairpur Mir's, Pakistan

Received 16th May 2017 and Revised 24th September 2017

Abstract: Today, linguistics problems enforce to computer industry to develop and design such application that could give strength to eliminate complexity of languages keeping more than one meaning of the same words. A Couple of approaches originated and provided better mechanism and accuracy for English language. In, Pakistan, various languages exist and spoken by People living different provinces of Pakistan. Sindhi, Urdu, Balochi and Punjabi are the most common languages. Each language represents different meaning of the same word besides diacritical complexities and morphemes problems too. The morphological exceptions increases day by day in script natural languages and plenty algorithms have been introduced, stemmer is one of them. In this paper, Punjabi language is selected to identify its morphological issues on prefix, suffix and prefix-suffix words. For this, porter stemming algorithm has been chosen for getting results on developed corpus of 23962 words. Moreover, prefix words are calculated along 5.64% stemmer error rate (SER) and 1.47% with suffix words. Thus, from prefix-suffix words 3.92% are calculated. However, the entire accumulative 11.03% of SER is recorded on Punjabi language. The developed stemmer could be fruitful for programmers and another step forward to field of natural language processing projects.

Keywords: Stemming System, Punjabi Language, Shahmukhi Script, Porter's Algorithm

1. INTRODUCTION

Panjabi language presents on paper with the help of two different scripts i.e. "Shahmukhi" used by Pakistani people and "Gurmukhi" used by Indian people (Gulzar, 2010). Vast amount of information via Panjabi Shahmukhi exist on web network in the form of e-data which increases gradually on daily basis due to availability of numerous tools of genre on Internet.

Natural Language Processing (NLP) applications of Panjabi and also those languages which are rich in orthographic, derivational morphology and use inflectional words require stemming systems as vital (Mateen, 2017). Two exclusive types of words are found in Panjabi language; Primary and Secondary words. Primary words are not separable for instance نس while secondary are those type of words which can be divided in to multiple words like بے وفائی. Here stem word is وفا , بے is prefix and ئی is a suffix morpheme. The Panjabi dictionary is presented with a lot of compound words like تہہ خانہ،چِٹّا دُدّھ ،ودّھ گھٹّ.

Stemming intend to decrease inflectional and derivational type of words in to stem or root (Aitao, 2003). For automatic stemming of Panjabi, massive research work has been proposed (Kumar, 2011) (Dhawan, 2013) (Joshi, 2014). The increasing ratio of information founded on Panjabi Shahmukhi script necessitates effective and efficient techniques and mechanism for stemming the Panjabi words particularly using Shahmukhi script.

In recent times, (Mateen, 2017) suggested an algorithm along with flow chart for stemming the Panjabi words by using Shahmukhi script but unfortunately results are not good enough. Thus, in this research study, by following algorithm proposed by (Porter, 1980), a stemmer for Panjabi language is developed based on Shahmukhi script.

2. MATERIAL AND METHODS

Many languages have numerous words with multiple meanings and morphology which make their usage complex. Panjabi language is old and has a prehistoric background come up difficult with digital system applications. It is highly intricate to break and understand them.

Corpus Collection

Stemmers of different types have been found on web. Corpora and lexicons of preferred language are must for all stemmers, no matter in less or more. That's why for Panjabi shahmukhi script-based stemmer, corpus was developed with cautious manners. Books of Panjabi like Tamahi Sanjh By Sabir Nazar, Punj Ganj By Mian Muhammad Bux and Madho Lal Hussain Novel By Nain Skh were useful for the collection of corpus.

In the linguistics region, morphemes are known as grammatical of least amount. Prefix, stem and postfix is a chain to compose a Punjabi word so such words can further be divided in to prefix, postfix and stem.

[++]Correspondence Author: Email Javed Mahar: mahar.javed@gmail.com

Words in number 23962 are developed in total from which 17861 can be classified in to prefix, suffix and prefix-suffix while remaining is known as root words. Words with at least one morpheme are selected attentively. 5079 words are found with prefix morphemes, composed words having suffix morphemes are 12729 in number and words with prefix-suffix both are 53 in total. Lexicon like corpus developed on Porter's algorithm also been developed for implementation.

Words in a large number are necessary to evaluate Porter's algorithm. For computing process, a lexicon also has been developed by assembling morphemes like prefix, suffix and stem to carry out the practice of algorithm for Panjabi stemmer. A developed corpus of Panjabi is used for vocab. Stemming system has stratified in to four sections for likely words.

Developed corpus worked with 23962 individual token words, discussed before now. 17861 from them are prefix while remaining all along with suffix or connective are root words. Words that are not obligatory are in use for lexicon. Auxiliary Punjabi words in figure are shown in **(Table-1)**.

Table-1 Information of Secondary Words

Word Types	No. of Words
Prefix	5079
Suffix	12729
Prefix-Suffix	53
Total	17861

User can interact by entering his/her most likely word, the designed framework obtain the majority of great executions.

According to lexicon words with stem, suffix and prefix are some entries in to the database. Application is provided with the features like adding, updating and deleting properties. A new file for database can also be created.

Martin Porter's Algorithm

Ubiquitous stemming algorithm in fact categorized underlying Rule Based, Hybrid and statistical ways. In Rule Based the approaches put oneself that transforms rules to swerve words for the sake to remove prefixes and suffixes and enable potential to programmers in NLP. Likewise, a Martin porter algorithm is focuses on the morphemes specially.

In this research, Punjabi language words are selected instead of English set of words due to its huge availability via porter. So, this algorithm has been further altered according to Shahmukhi script of Punjabi language and it is known that researcher had taken such step to modify due to needs and requirement whenever they felt (Gupta, 2011).

It has no doubt still porter is being considered as well-known strategy and algorithm to resolve these lingua problems and found. The complete steps of the selected algorithm is given in (Porter, 1980).Upgrades and several unique ideas has been shared and explored time to time along alterations given by the research contributors. From the English dialects corner, quite a few thoughts are exist for suffixes fundamentals globally also slightly mixed addition.

It covers five phases (stages) and individual progression standards are jointed until the condition passes among one. Once accredited suffix wiped accurately following second stage is then performed event. Consequent arising stem move toward to complete the progression that individually returns. The algorithm of porter is divided into convinced numbers of vector steps that are given below to make the final root or stem along prosperity.

3. RESULTS

In this results section implemented system that we tried to provide facility to reduce ambiguousness of the words and make proper sense for Punjabi. An effort is made with selected porter algorithm for Punjabi language stemmer. Evaluation of the intact performance ratio and stemmed Error Rate (SER) use preceded in (Lee, 2003), also exercise underlying experiment.

$$\text{SER}=\frac{\text{Number of Incorrectly Stemmed Words}}{\text{Total Number of Words}}\ \text{X}\ 100$$

Thus, 17861 words amount has given to developed system and those are also classified into three major collections i.e. Prefix Words, Suffix Words and Prefix-Suffix words.

Prefix Word Stemming

Evaluating the execution of the prefix words different 1016 words were self-assertively taken for the testing from the dataset containing 5079 words which is around 20% of the arrangement dataset. Testing choose rate is also taken for the prefix and prefix-suffix. In prefix morphemes, around 31 distinctive morphemes are found. The words having prefixes are requested into three special groupings for the appraisal of the system execution and the connection of the prefix stemmed word with each other. The gathering of the prefix words are as under:

(1) Prefix words showing the sense of negation
(2) Prefix words showing the sense of adjective
(3) Prefix words showing the sense of antonym

The system performance is computed through execution and this execution is made with developed system also calculated isolate on above prefix word depiction for the effective manner. As earlier discussed in introduction section that error rate of the stemmer make possible to identify the system performance and Punjabi language is un-existed from the SER perspective and exist work does not match the required criteria that has done with this script language. The detail of the words along classes that has been used in this paper is given in **(Table-2)**.

Table-2 Stemmed Error Rates of Prefix Words

Prefixes Classes	No. of Words	Correctly Stemmed	Incorrectly Stemmed	SER%
Negation	276	269	6	2.17
Adjective	641	639	2	0.31
Antonym	99	95	3	3.16
Total	**1016**	**1003**	**11**	**5.64**

Table-2 listed the classes that has been used their attributes mentioned in first 5 column of this table. Thus, 1016 amount of words are selected and given to the developed system along porter one algorithm. Between them, 1003 correct words (stemmed) found and 11 achieved incorrectly.

So, accumulative results of SER are 5.64 received. These results calculated in this manner that a negotiate class of SER found with 2.17 among the correct 269 words from the incorrect 6 words. But, with adjective class 641 words have been given to developed system and found the 639 words. Just 2 incorrect words we receive accompanied with 0.31 SER and it is feasible. Besides, 99 antonyms have been selected and successfully achieved 95 correct words, there are only 3 words recorded as incorrect and finally 3.16% of SER.

Suffix Word Stemming

Once the developed frameworks gets words having suffix morpheme on that time it scans the most legitimate sprint for choose word from base class when framework gets the word having suffix morpheme, at that time framework consequently scans the most legitimate execute for chosen word from the run base structure. The amount of the words taken randomly to test and evaluate from predefined developed dataset of 2546. The lexicon dataset is just 12729, thus, Punjabi linguists are agreed that there are 74 suffix morphemes exist.

It is primarily to setup the words that have suffix morphemes and in this way coherent exploratory targeted word with suffix is sorted into 5 distinguish classes. The ascending orders of the prefix with conceivable cases of Punjabi words are as under.

(1) Suffix words in singular sense
(2) Suffix words of plurality sense
(3) Suffix words showing adjectival sense
(4) Suffix words classed in masculine sense
(5) Suffix words of feminine sense

Furthermore, the classes contains prefixes experimented via some other languages spoken in Pakistan and this idea has taken from (Narejo, 2015) research contribution and results covered along average SER listed in **(Table-3)** even few illustration are also given for further clarification of compared recorded results. Thus, the table is sorted into five columns (fields) showing isolatable information.

Table-3 Stemmed Error rate of the Suffix Words

Suffixes Classes	No. of Words	Correctly Stemmed	Incorrectly Stemmed	SER %
Adjective	4832	4823	9	0.19
Singular	1599	1596	4	0.25
Plural	1924	1913	11	0.57
Masculine	955	952	3	0.31
Feminine	3419	3414	5	0.15
Total	**12729**	**12698**	**32**	**1.47**

Singular, plural, masculine and feminine are the concern classes of system that are given and these are estimated 12729 words. An amount of correct words are 12698 that are stemmed however found 32 incorrect words from them along 1.47% summation of SER. But, if we take individual class then the amount of the adjective class words are 4832 found and among only 32 are achieved incorrect with 0.19 SER received from this words amount corner. In the same way, 1599 words of singular class are used and among these correct words just 4 incorrect words be acquainted 0.25 SER with our system.

From the plural class angle, 1924 words that were given to system as corrected and delivered 1913 incorrect words among them but only 11 words have been found accompany by 0.57 SER. Later, like others we have given 955 corrected set words but, the system back to 3 incorrect words with 0.31 SER results ratio. Last class of our developed system is feminine having 3419 words given to system as corrected and found 5 words incorrect and returned 0.15 SER.

Prefix-Suffix Stemming

Right when system gets word having both prefix-suffix morphemes, by then the structure normally checks for the most appropriate regulate for picked word from the oversee base structure. If relationship is productive, by then structure figures frequencies of each related word and plays out the chose calculation finally it picks the word having most bewildering repeat of prefix-suffix morphemes.

For evaluating the execution of prefix-addition words, 53 words were discretionarily taken from the dataset. All words were tried through the created Punjabi stemming mechanism since constrained words are found from the distributed published and deployed Punjabi books. The discovered results are depicted in **(Table-4)**

Table-4 Stemmed Error rate of the Prefix-Suffix Words

No. of Words	Correct	Incorrect	SER
53	51	2	3.92

Due to majority of the prefixes and suffixes morphemes, 3.92% SER achieved through this class and individual root words are also used as pathway and got resource from books and magazines and life style of the Punjabi peoples. But fact is that it is quite and enough difficult to isolate the words using Punjabi because it primarily based on the lexicon and corpus based applications.

Cumulative Results

With the porter algorithm develop linguistic application performance has calculated on prefix 5070 words and 12729 suffixes, similarly, prefix and suffix both words tested via this algorithm and develop framework of the software along possible and positive results. The entire system explored 11.03% of SER that is fine results we believe. Finally, the results that are collective during testing along SER are given in **(Fig.1)**. It displays the word types and their SER ratios as a pictorial form. To our best knowledge these results are much better than available accuracy and performance results.

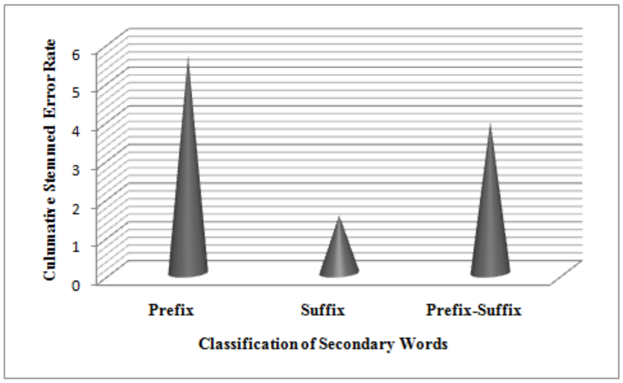

Fig.1 accumulative Words SER of Proposed Algorithm

4. CONCLUSION

Information retrieval system considers stemmer as an essential tool for retrieving information. This paper reflects corpus and morphemes based approaches of Porter's algorithm due to which Punjabi stemming system developed successfully. Experiments conducted by using Stemmer system for Panjabi language and acquire 88.97% accuracy results. Collaborative words of Panjabi with suffix, prefix and suffix-prefix are cover up in this execution. Every word in term of calculation and evaluation, calculated with high consideration. Case-by-case calculation has done on Panjabi words morphemes classes and SER. With prefix, SER gives 5.64%, 1.47% achieved of suffix words and with both Prefix-Suffix words 3.92% has been recorded. The SER gives 11.03% as a cumulative result and we find it very well. The Panjabi stemmer has been developed by following corpus of 23962 words.

REFERENCES:

Aitao, C. (2003). Building an Arabic Stemmer for Information Retrieval, In the Proceedings of the 11th Text Retrieval Conference, Berkeley, Pp. 631-639.

Dhawan, C., J. Singh, K.. Garg, (2013). Hybrid Approach for Stemming in Punjabi. International Journal of Computer Science & Communication Networks, 3(2), 101-104.

Gulzar, M. A. (2010). Issues of Language(s) Choice and Use: A Pakistani Perspective. Pakistan Journal of Social Sciences, 30(2), 413-424.

Gupta, V., G. S. Lehal, (2011). Automatic Keywords Extraction for Punjabi Language. International Journal of Computer Science Issues, 8(3), 1694-0814.

Joshi, G., K. D. Garg, (2014). Enhanced Version of Punjabi Stemmer Using Synset. International Journal of Advanced Research in Computer Science and Software Engineering, 4(5), 1060-1065.

Kumar, D., P. Rana, (2011). Stemming of Punjabi Words by using Brute Force Technique. In International Journal of Engineering Science and Technology, 3(2), 1351-1357.

Lee, Y. S., K. Papineni, S. Roukos, (2003). Language Model Based Arabic Word Segmentation. The 41st Annual Meeting of the Association for Computational Linguistics, Sappora, Japan, 399-406.

Mateen, A., M. K. Malik, Z. Nawaz, H. M. Danish, (2017). A Hybrid Stemmer of Punjabi Shahmukhi Script. International Journal of Computer Science and Network Security, 17(8), 90-97.

Porter, M. (1980). An Algorithm for Suffix Stripping. Program, 14(3), 130-137.

Narejo, W. A. (2015). An Algorithm for Sindhi Word Segmentation into Morphemes. MS Thesis, Department of Computer Science, Shah Abdul Latif University, Khairpur Mir's.