*Sindh Univ. Res. Jour. (Sci. Ser.) Vol.48 (4) 891-898 (2016)*

**SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)**

# Sindhi Stemmer for Information Retrieval System Using Rule-Based Stripping Approach

M. R. SHAH, H. SHAIKH, J. A. MAHAR, S. A. MAHAR

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

Received 11th January 2016 and Revised 29th September 2016

**Abstract:** For last few years, huge amount of information in Sindhi language has made available online in the form of e-data. The information retrieval system is used to ensure easy and efficient access to the stored information. Stemmer is the tool, which information retrieval system uses to decrease morphological variants of a word to its root or stem. As yet no any information retrieval system and stemmer for Sindhi language is available, hence, the access to data resources is not possible. In this paper, an algorithm is proposed applying rule based approach. The proposed algorithm depends upon our developed lexicon and linguistic rules. The 5327 words are incorporated in the lexicon among them 2142 words are those having prefix or suffix word morphemes. A number of 38 rules included the repository. The performance of the prefix, suffix and combined prefix-suffix words of Sindhi language is separately calculated. The cumulative performance accuracy of 84.85% is calculated using developed stemmer. The outcome of this stemmer will be beneficial for the developers of automatic Sindhi information retrieval system.

**Keywords:** Information Retrieval, Stemming, Rule-based Approach, Lexicon, Sindhi.

## 1. INTRODUCTION

Most of the information is nowadays made available online. English and European languages have proved their significant dominance on the web since its inception. With the passage of time, the web is now becoming multi-lingual. Specially, for the last few years, there has been significant growth seen in amount of information in Pakistani and other Asian languages. A large number of documents are made available on Internet in various Pakistani languages like Urdu and Sindhi.

Information Retrieval (IR) systems are being used to ensure easy access to stored information. Work carried out to IR systems for Pakistani languages is the matter of recent interests. Shortage of linguistic resources and tools seems to be a big problem against the development of IR. One of the efforts executed to make the search engines more efficient and effective in IR is the use of Stemmer, which almost every IR system use to decrease morphological variants of a word to its root of stem (Chen, 2003).

Scientifically and linguistically, morphology is a area related with the study of linguistics and mainly pertaining to words and their formation into grammatical units called morphemes (Al Ameed, 2005). Stemming is a common and widely used method for morphological normalization of natural language texts. Stemming is a process narrates morphological variants of a word and reducing inflected or sometimes resulting words to their root or stem shape.

The algorithms of stemming are classified into three groups: truncating, statistical and mixed methods. Each one of these has a typical way of finding the stems of the word variants. There are two types of stemming: (1) Light stemming which is restricted to removal of limited number of prefixes and suffixes to make appear a stem. (2) Heavy stemming which is also known as root-based stemming includes implicitly the removal of prefixes and suffixes without extracting the particular root (Al-Omarim, 2013).

Variety of stemming systems are available in English (Porter, 1980), Urdu (Kansal, 2012), Punjabi (Joshi, 2014), Arabic (Sajjad, 2015), Persian (Mehrad, 2011) and other languages. The performance accuracy of these developed stemmers is debatable. However, no such stemming systems are available for Sindhi Language. Sindhi language indicates several issues to Natural Language Processing (NLP) largely due to its rich morphology. But still research contributions are not up to the requirement, due to which no any IR System is yet available for Sindhi Language.

## 2. MATERIAL AND METHODS

A rule based stemmer for Sindhi language is developed and presented in this paper, which use the stripping approach. Indeed there are many challenges to

be faced in this study for Sindhi language and found some linguistic rules for removing the prefixes and suffixes from the stem.

## Data Collection

For computational automatic processing, the corpora are necessary requirement for a language. Since five decades, appearance of transform has been empirically examined in speaking and writing Sindhi language. As the corpus, 86,733 number of Sindhi words is used for the implementation of developed stemmer. The corpus is developed by (Mahar, 2014).

Initially, the complete text of Sindhi is divided into sentences. Then after, such sentences made appeared with the required statistical information. The calculated information of sentences and words was utilized for further computational processing. Consequently selected words insert to the developed application. The application had shown 50,327 word tokens. The symbolize word tokens are not equal to total words because homographic and same words are selected only once. **(Table-1)** displays the comprehensive and detailed statistics of Sindhi corpus.

**Table-1 Statistical Information of Corpus**

| Corpus Type | Sentences | Word Tokens |
|---|---|---|
| Arts | 1897 | 6884 |
| Sports | 1656 | 7582 |
| Politics | 2590 | 13351 |
| Environment | 1819 | 7098 |
| Music | 3822 | 15412 |
| **Total** | **11,784** | **50,327** |

The process of tokenization is the division of selected Sindhi text into the possible tokens (Mahar, 2012). The white space, symbols, digits and punctuation marks are used as word limits that become constructive enough for the process of tokenization. The tokenization model of Mahar's is used in this research project for creation of word tokens. Hence, the process of stemming of Sindhi words is successfully analyzed, design and implemented.

The lexicon is a mandatory component of rule-based linguistic applications because it contains words and their morphemes. For Sindhi stemmer, a lexicon of 50,327 words is developed for the implementation of algorithm. Different types of morphemes are required to store in the lexicon, hence, five tables are created for each one. The created tables remain with names of root, compound, prefix, suffix and prefix-suffix words. The sample figure of the lexicon is depicted into **(Fig-1)**. The table is consisting of four columns; each represents separate kind of morpheme. The system automatically assigns IDs to each word. Due to assigned ID, the word can easily search and appropriate rule can be applied on the selected word.




**Fig.1 Layout of Developed Lexicon**

Moreover, word entries affect the performance of the created system specifically those NLP systems that depend on the lexicon. The maximum number of words will gain the maximum performance therefore, a chance is provided to the user for word entries. The sample layout picture of the developed application for the word entries in the lexicon is displayed in **(Fig.2)**.

## Proposed Algorithm

The automatic system is always based on the algorithm therefore, an algorithm is proposed that segment the Sindhi secondary word into possible morphemes. The text of Sindhi language is complex because Sindhi words are homographic in nature and writing system uses cursiveness. Therefore, developers face many difficulties at the time of stemmer development. Furthermore, the affixes morphemes are more difficult to segment as compare to prefix morphemes. The application is developed carefully that segment all types of Sindhi words and extract the stem successfully using linguistic rules and morphemes.

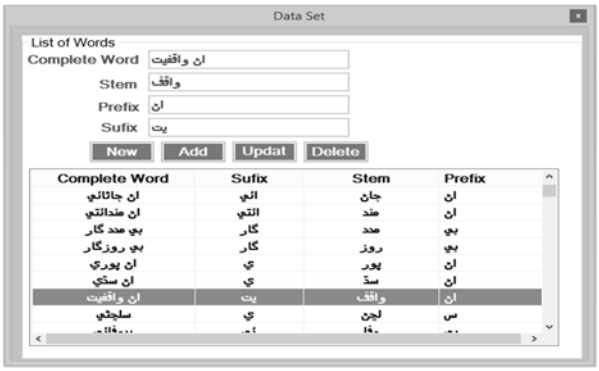



**Fig.2 Layout of Word Entries into the Lexicon**

The proposed algorithm is specially divided into three portions i.e. Tokenization, Rules selection and matching and Morphemes frequencies. The flow chart of the proposed algorithm is made available in (**Fig.3**). The Tokenization model of (Mahar, 2012) has been used in this system for creating the word token. In this process the complete Sindhi words are segmented from the Sindhi corpus. After tokenization word tokens are temporary stored in array named as 'WordTemp' for further process.

The developed application then accepts each token from the array and process of searching the match rule will begin for each input word. If the seek becomes successful then system applies the selected rule on the word and split the word into possible morphemes. After that control transfers to step where generated morphemes are compared with lexicon otherwise system display the selected word as the 'root word'.

For getting confirmation the system calculates the frequencies of morphemes and stores them separately into the database. This process is earlier performed by the system. After the preprocessing of the morphemes frequency system selects the morphemes having highest frequency. After that system selects the root word having maximum frequency and then system will display it. This whole process will continue till the last word of the inputted or typed text is encountered.

## 3. RULE BASED APPROACH

During the literature survey, it has been found that rule-based automatic systems are usually used to operate and interpret information effectively. Such type of systems can simply be developed using a set of statements and rules indicating the actions of the statements. It represents knowledge of the area through specific rules that show what to do in whatever the case may be. The algorithm is based on developed lexicon and linguistic rules. A repository is also developed for the created rules which are designed with the rule based approach. These rules are taken from Sindhi literature and from the Sindhi linguistics. Basically, there are set of rules in this approach, these rules are implement to inflected tokens to segment the prefixes and suffixes from the stem. Linguistic rules are the basics for the stemmer therefore; a rule-based approach in applied for the designing and development of the stemming rules. A total of 38 rules are included into the repository and used for the implementation and execution of the proposed algorithm as well software application.

### Rules for Prefix Morphemes

Many words are present in Sindhi dictionaries which are concatenated with the prefix morphemes such as لاجواب، بداخلاق، وڏڙا. In these Sindhi words, بد, لا, are prefix morphemes. There are 31 different prefix morphemes are used before Sindhi root words. The samples of prefix morphemes are displayed in **(Table 2)**.

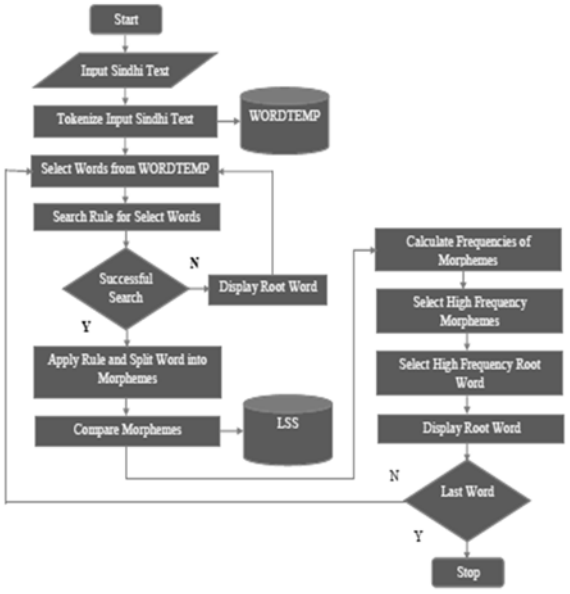



**Fig. 3. Flow Chart of the Proposed Algorithm**

**Table-2 Prefix Morphemes**

| | | | |
|---|---|---|---|
| نا | خوش | اڻ | نه |
| بد | سر | آ | م |
| بر | دست | مها | مهه |
| بي | هم | آ | وڏ |
| خود | لا | نر | پر |

Among the developed 38 rules, 16 were developed for prefix words. Few applied rules for finding the stem are as under.

**Rule 1-** If word starts with نا (Noon+Alif) then remove نا (Noon+Alif) from beginning.

For example- ناميد, نامناسب, نامراد

**Rule 2-** If word starts with پر (Pey+Rey) then remove پر (Pey+Rey) from beginning.
For example- پرزور, پرواهه, پرديس

**Rule 3-** If word starts with لا (Laamu+Alif) then remove لا (Laamu+Alif) from beginning.
For example- لاتعلق, لاعلم, لاجواب

**Rules for Suffix Morphemes**

Many words are present in Sindhi dictionaries which are concatenated with the suffix morphemes such as سمهڻ، اٽي، چوڪرو. In these Sindhi words, و, ڻ, are suffix morphemes. There are 85 different suffix morphemes are used after the Sindhi root words. The samples of suffix morphemes are displayed in **(Table 3)**.

**Table-3 Suffix Morphemes**

| | | | |
|---|---|---|---|
| ا | و | ائو | ڻ |
| ي | وارو | ئو | ٽي |
| يان | وال | آئو | ٽو |
| يو | واري | ياٽو | ن |
| يارو | ولو | آنتو | ان |

Among the developed 38 number of rules, 15 were developed for suffix words. Some of the suffixes rules are applied to find the rood word as under.

**Rule 1-** If word ends with ي (Ye) then remove ي (Ye) from end.
For example- وٹي, ہڏي, آزادي

**Rule 2**- If word ends with ا (Alif) then remove ا (Alif) from end.
For example- ڪنا, وڏا, پوچا

**Rule 3**- If word ends with ڻ (Nounn) then remove ڻ (Nounn) from end.
For example- ڪٽڻ, پنڻ, وڏڻ

**Rules for Prefix-Suffix Morphemes**

Many words are available in Sindhi dictionaries which are concatenated with the suffix morphemes. Some of the Prefix-Suffix rules are applied to find out the root word as under.

**Rule 1-** If word starts with نا (Noun+Alif) and ends with ي (Ye) then remove نا (Noun+Alif) from beginning and ي (Ye) from end.

For example- نافرماني , ناانصافي, نااميدي

**Rule 2**- If word starts with لا(laamu+Alif) and ends with ي (Ye) then remove لا(laamu+Alif) from beginning and ي (Ye) from end.
For example- لاپرواهي , لاگرزي, لاتعلقي

**Rule 3**- If word starts with اڻ(Alif+Nounn) and ends with و (Wav) then remove اڻ(Alif+Nounn) from beginning and و (Wav) from end.
For example- اڻڄاڻو, اڻسڌو, اڻٻڌو

## 4. IMPLEMENTATION AND RESULTS

The application and its implementation process along with the obtained results are defined in this section. The morphemes are divided into three classes i.e. prefix, suffix and prefix-suffix. The developed software application assessed through each class. Hence, results of every morpheme are calculated separately.

During the stemming process the system initially realize the prefix morpheme (if any) of the word and then removes it on priority basis. Let consider the word **بد** بختي system search the prefix morpheme بد and then remove the searched morpheme from the complete word. The resulting word is now,بختي now system inspects the suffix morpheme ي and eliminates it from the word. After this process, only the root word **بخت** will get from the system.

The corpus and rule based is collectively used for the implementation of the Sindhi stemming system. The proposed algorithm is chiefly based on the searching and matching process of morphemes and developed rules. The dynamic matching is also adapted by different researchers to handle word entries of lexicon efficiently based on excellent matches. The database setting is considered as key component in the development of a lexicon. Hence, in view of appropriateness the SQL server 2005 was taken. The SQL setting is properly supportable and suitable enough for different writing styles of Sindhi. The VB.Net framework is used for the implementation of proposed and this framework is capable enough of connecting the database system of any Arabic Script-Based language including Sindhi.

From the input text, system automatically chooses every token respectively. After that the rules and morphemes are examined as suitable one can be selection and applied for the word encountered. In the situation of word mismatch the software will display such word as out of vocabulary. The accuracy of the developed application is tested by scoring the incorrectly stemmed words similar to (Lee, 2003). The Stemmed Error Rate (SER) is defined as:

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

For evaluation the performance of the developed Sindhi stemmer, a number of 2142 Sindhi words were prepared and stemmed error rate is separately calculated for the word morphemes: Prefix Words like نا مناسب,

Suffix Words like كاميابي, Prefix-Suffix words like اٽلڪيو

**Prefix Word Stemming**

Upon receiving prefix word the system automatically search the appropriate rule for the input word from the rule base system. If comparison is successful, then system calculates and selects the highest frequency of the prefix morphemes. For the performance evaluation of the prefix words a number of 194 words were selected randomly from the dataset having 773 tokens which is about 25% of the training dataset. The selected testing percentage is also taken for the suffix and prefix-suffix words. Many of these words are already available in dictionaries having prefix morphemes. After searching out the prefix morphemes, about 31 different morphemes are found.

The prefix words are divided into three different groups for the evaluation of the system performance and the comparison of the prefix stemmed word with each other. The words having prefix morphemes are basically classified by (Narejo, 2015). The classification of the prefix words are: (1) The prefix words showing the sense of negation like بد قسمت (2) The prefix words showing the sense of adjective like لا محدود (3) The prefix words showing the sense of antonym **پرديس** . It should be noted that a few words are also found in dictionaries that are not lying under these classification, these words are ignored as the performance of the system is not appeared to be up to the standard as required. The summary of the calculated results is given in **(Table 4)**.

**Table-4 Stemmed Error Rate of Prefix Words**

| Prefixes Classes | No. of Words | Correctly Stemmed | Incorrectly Stemmed | SER% |
|---|---|---|---|---|
| Negation | 71 | 65 | 6 | 8.45 |
| Adjective | 102 | 99 | 3 | 2.94 |
| Antonym | 21 | 18 | 3 | 14.29 |
| **Total** | **194** | **182** | **12** | **25.68** |

The calculated results are the evidences that the SER of Antonym is 14.19 which become the highest among the all. Normally, prefix morphemes are utilized in Sindhi that show a negative and antonym concept in the word meaning of the original meaning of the word like **بد** and پر. Such sort of prefix morphemes are also used and considered as separate words in Sindhi language. Hence, the developed stemmer makes segments of such kind of words separately. Due to this, SER is relatively increases. However, normal Adjective prefix words are successfully segmented with the SER of 2.9% that is inferior to the Negation and Antonym sense words.

**Suffix Word Stemming**

When system receives the word having suffix morpheme, then system automatically searches the most proper rule for selected word from the rule base system. If the comparison goes successful, then the system calculated frequencies of all the related words and select the word having highest frequency of suffix morphemes. Let consider the input word is گهرواري, the system calculates the frequencies of all stored words having the prefix morpheme واري like پاهرواري اندرواري پاڳنواري on the basis of applied rule and the frequency of the word the system select the root word from our developed lexicon and then system displays the root گهر.

The 291 words were randomly selected from the training dataset for testing and evaluating the system. In training dataset, 1162 word tokens are included. There are 85 suffix morphemes found during the words searching and identification from Sindhi dictionaries. It is essential to classify the words having suffix morphemes. Therefore, for the comparison and experimental intention the words with suffix were categorized into 5 different classes: (1) The suffix words in singular sense like بکيو (2) The suffix words of plurality sense like سونارا (3) The suffix words showing adjectival meaning sense like پاڳيريو (4) The suffix words classed in masculine sense like چوڪرو (5) The suffix words of feminine sense like گهرواري. Summary of results with the standard of stemmed error rate is illustrated in **(Table 5)**.

**Table-5 Stemmed Error Rate of the Suffix Words**

| Suffixes Classes | No. of Words | Correctly Stemmed | Incorrectly Stemmed | SER % |
|---|---|---|---|---|
| Adjective | 438 | 432 | 6 | 1.37 |
| Singular | 147 | 145 | 2 | 1.36 |
| Plural | 121 | 114 | 7 | 5.79 |
| Masculine | 167 | 166 | 1 | 0.6 |
| Feminine | 289 | 286 | 3 | 1.04 |
| **Total** | **1162** | **1143** | **19** | **10.16** |

The calculated results demonstrated that the least SER is achieved with Masculine class among the all. Though, Adjective and Singular classes as well as Feminine class achieved good performance accuracies such as 1.37%, 1.36 and 1.04% correspondingly. The collective SER is 10.16%. It is because of the Plural class of suffix words which lonely stands with an SER of 5.79%. It can be seen that due to such calculations the overall performance of the developed system is affected and not at acceptable level. If the researchers reduce the SER in plural class then the performance of the system will automatically improved. On the other

hand, encouraging results are achieved with Masculine and Feminine classes.

**Prefix-Suffix Stemming**

Many words are found in Sindhi language having the prefix-suffix morphemes. These words are formed with the concatenation of both prefix and suffix morphemes. When system receives word having both prefix-suffix morphemes, then the system automatically searches for the most appropriate rule for selected word from the rule base system. If comparison is successful, then system calculates frequencies of all related words and selects the word having highest frequency of prefix-suffix morphemes. For example the input word is بداخلاقي the system calculates the frequencies of all stored words having the prefix-suffix morphemes بد and ياlike **بدنظمي**,بدمغزي بدڪرداري on the basis of applied rule and the frequency of the words. System selects the root word from our developed lexicon and then system displays the root اخلاق.

For evaluating the performance of prefix-suffix words, from the training datasets having 207 words, 52 tokens were selected randomly. This is about 25% of the training dataset. The calculated results are depicted in (**Table 6**).

**Table-6 Stemmed Error Rate of the Prefix-Suffix Words**

| No. of Words | Correct | Incorrect | SER |
|---|---|---|---|
| 52 | 47 | 5 | 9.61 |

**Cumulative Results**

The performance of the developed stemmer application based on proposed algorithm can be calculated when the performance of all the word types are collectively calculated. The designed stemming software has been gauged in testing 194 prefix words, 291 suffix words and 52 prefix-suffix words. The calculated results revealed segmentation error rate of 15.15%. The computed collective SER of special word classes is made available in (**Fig.4**)

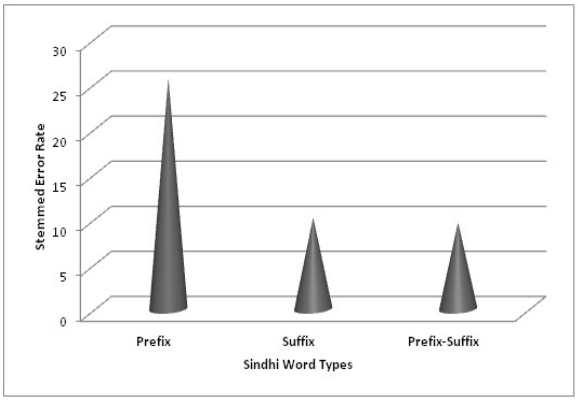



**Fig.4. Cumulative Words Stemmed Error Rate**

The results proved that inferior results are achieved with prefix-suffix words class that is considerable achievement. The SER of words having prefix-suffix is 9.61% collectively. Stemmed error rate of suffix words generates an SER of 9.16 %. The cause of the largest SER as compared to prefix-suffix word classes is described prior as the Adjectives with suffix occasionally present as individual words in the script of Sindhi. The calculated stemmed error rates of prefix words are vary as many prefixes are considered as a separate word in Sindhi language. Therefore the developed automatic system is not able enough to differentiate such prefix morphemes with the complete word because these word morphemes are also used as complete words.

## 5. AUTOMATIC SINDHI STEMMER

Two separate boxes are given in the developed interface. Both boxes are used for the process of the input Sindhi text. The box given in left side is fixed of showing the text that is inputted or typed in the software application. In this box, user can type the required text as well as select and upload any text file in DOC format for further processing.

When the text is given to the software application, it is the time to click the button named 'Process' that is shown in the left most side on the upper side. When the user clicks the button 'Apply' the software begins the actual function of stemming. The process will complete by displaying the outcome on the interface at in output box given in the right side. The availability of the both boxes is entirely dissimilar in terms of type of Sindhi text. The simple text can be inputted into the input box and if user needs to clear the inputted text for new text entry then user has to press 'Clear' Button and system automatically deletes the input texted from the input box and user may again insert new Sindhi text into input box for stemming the words.




**Fig.5 Text Input and Output Interface**

The output box represents the outcome in two separate columns i.e. Complete Word and Stem Word as displayed in **(Fig. 5)**. Both columns are entertained with stem and complete word. From the given Sindhi text, every word is automatically processed and then displayed into its specified column. The separate captions are given in the interface for understanding the input and output of the Sindhi stemmer.

## 6. CONCLUSION

Automatic stemmer is a mandatory component of an IR system without which IR system is nothing. An algorithm is proposed applying the rule based approach and then implemented for the Sindhi stemmer receiving the performance accuracy of 84.85%. The performance of the prefix, suffix and prefix-suffix words is individually calculated for comparison. The SER of 25.68% is received with Prefix words, 10.16% with Suffix words and 9.61% is recorded with Prefix-Suffix words. The collective SER of 15.15% is calculated and recorded. The accuracy of developed stemming software is not at acceptable level yet appending more words and rules in the repository will definitely optimize more the performance of developed Sindhi stemmer. The presented work is one step forward for the development of an automatic Sindhi language processing applications.

## REFERENCES:

Al Ameed, H. K., S. O. Al Ketbi, A. A. Al Kaabi, K. S. Al Shebli, (2005). Arabic Light Stemmer: A New Enhanced Approach. The 2nd International Conference on Innovations in Information Technology, Dubai, 1-9.

Al-Omari, A., B. Abuata, M. Al-Kabi, (2013). Building and Benchmarking New Heavy/Light Arabic Stemmer. The 4th International Conference on Information and Communication Systems, Irbid, Jordan, 1-6.

Chen, A., F. Gey, (2003). Building an Arabic Stemmer for Information Retrieval. In The Proceedings of the 11th Text Retrieval Conference, Berkeley, 631-639.

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.

Kansal, R., V. Goyal, G. S. Lehal, (2012). Rule Based Urdu Stemmer. In the proceedings of 24th International Conference on Computational Linguistics, Bombay, India, 267–276.

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.

Mahar, J. A., H. Shaikh, G. Q. Memon, (2012). A Model for Sindhi Text Segmentation into Word Tokens. Sindh University Research Journal (Science Series), 44(1), 43-48.

Mahar, S. A, (2014). Comparative Analysis of Vowel Restoration for Arabic Script Based Languages Using N-Gram Models. MS Thesis, Department of Computer Science, Shah Abdul Latif University, Khairpur, 31-32.

Mehrad, J., S. R. Berenjian, (2011). Providing a Persian Language Singular-Stemmer System (RICeST Stemmer). International Journal of Information Science 9(2), 13-22.

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

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

Sajjad, K.., W. Anwar, U. Bajwa, X. Wang, (2015). Template Based Affix Stemmer for a Morphologically Rich Language. The International Arab Journal of Information Technology, 12(2), 146-154.