J. Appl. Environ. Biol. Sci., 4(7S)436-443, 2014

© 2014, TextRoad Publication

ISSN: 2090-4274
Journal of Applied Environmental
and Biological Sciences
www.textroad.com

A Novel Stemming Approach for Urdu Language

Mubashir Ali, Shehzad Khalid, Muhammad Haneef Saleemi

Department of Computer and Software Engineering Bahria University
Islamabad, 44000, Pakistan

Received: September 1, 2014
Accepted: November 13, 2014

ABSTRACT

Stemming is one of the most important pre-processing steps in the process of Text Mining which boosts the performance of information retrieval (IR) system. It is also equally important for many other interesting research areas like natural language processing (NLP), text categorization etc. The main objective of stemming is to bring many grammatical word forms, for example parts of speech, gender, tense etc. to their stem or root form. Due to the rich morphological structure of Urdu language, it is a challenging task to develop an Urdu stemmer for information retrieval system. In this paper, we have proposed an effective rule-based stemming method for Urdu language to cope with the challenges of Urdu morphological structure. Our proposed Urdu stemmer generate the stem of Urdu words as well as borrowed words (words from other languages such as Arabic, Persian, Turkish, etc). The proposed methodology is compared with the existing Urdu stemming technique such as Light Weight Stemmer for Urdu Language to demonstrate the dominance of proposed Urdu stemmer as compared to the competitor.

KEYWORDS: Urdu, Stemmer, Prefix, Postfix.

1. INTRODUCTION

Urdu is a national language of Pakistan and is also the state language of India. It is an Indo-Aryan language. Urdu language is made up of with the combination of different foreign languages such that Arabic, Persian, Turkish, etc. These borrowed languages themselves are complex morphological languages. Resultantly, Urdu is a morphologically rich language with complexity inherited from the fparent languages. There exists a large amount of unstructured Urdu textual data in the world; by applying data mining techniques useful information can be achieved [16]. This can further be used for automated news categorization, Urdu document analysis and student-centered learning in Urdu medium [17].

Urdu is robust in both inflectional and derivational morphology [1]. Morphology deals with inner structure of words [2]. The major units of morphology are morphemes. Term morpheme is a smallest word unit that has a semantic interpretation and cannot be decomposed further [3].Morphemes can be free morphemes or bound morphemes [4]. Morphemes that exist freely are called free morphemes such as flower is free morpheme. On the other hand, morphemes that are made as a result of combining different morphemes are called bound morphemes i.e. in flowers, ‘s’ is a bound morpheme. To boost the performance of IR system, morphological analysis of Urdu language is very important. This important is based on the fact that IR system works on the root/stem form of a word rather than its inflected and derived form. The improvement in the performance of IR system is possible with the use of stemmer. Stemmer is an algorithm that produced the root form of the word. For example, an English stemmer should reduce the English words like, liking, liked, and likes to their stem “Like”. Likewise Urdu stemmer should restrict the Urdu words خبروں(news), خبریں(news), to Urdu stem word خبر(news).

In this paper, we introduce a novel stemming approach for Urdu text. Generic Urdu stemming rules are proposed, which have the ability to generate stem of any Urdu word. The rest of the paper is organized as follows: section 2 describes a brief overview of the related work. In section 3, proposed Urdu stemming method is described. Experiments are given in section 4 to demonstrate the effectiveness of proposed approach. Finally, section 5 gives the conclusion of the paper.

*** Corresponding Author:** Haneef Saleemi, Department of Computer and Software Engineering Bahria University, Islamabad, 44000, Pakistan. haneef_saleemi@yahoo.com

2. Background and Related Work

There are three approaches [5] that are commonly used for stemming such as affix stripping, table lookup, and statistical methods. Affix stripping approach [6] is used to obtain the stem of the word by removing the attached prefix and postfix from the word. In table lookup approach [6], each word and its associated stem is stored in structured table. This approach requires a lot of storage space for its implementation and its table needs to be updated manually for each new word. On the other side, in statistical approach [7] statistical analysis are performed based on corpus size.

J.B. Lovin's [8] proposed first English stemmer that is based on rule-based strategy. In this stemmer, Lovin's defined 260 rules for stemming English word. This stemmer produces the stem of English words in two phases. In the first phase, it removes the maximum matching suffix defined in suffix table and recodes the word to produce valid stem. Spelling exclusions are handled in second stage. In 1980, Porter [9, 10] introduced another stemming method that is also based on rule-based strategy. This stemming technique removes the suffixes form words with the help of suffix list and some conditions are enforced to determine suffix to be separated. Porter stemmer has five steps and within each step, rules are applied until one of them passes the conditions. If a rule is recognized, the suffix is removed consequently, and the next step is executed . At the fifth step, recoding is performed and resultant stem is returned. Porter reduced the Lovin's rules upto 60.

Variety of effective stemming methods for Arabic language has been proposed. Khoja et al [11] proposed a rule based stemmer for Arabic language which is known as superior root-based stemmer. This stemmer truncates prefix, suffix and infix and then uses pattern for matching to generate root. It makes use of several linguistic data files such as punctuation characters, definite articles, list of all diacritic characters and 168 stop words in order to improve stemming accuracy results of proposed Arabic stemming technique. To stem Arabic text, Thabet, [12] introduced a light stemming approach. It is applied on classical Arabic in Quran to generate stem of Arabic words. This method reads each surah from text files as an input and after replacing all the uppercase letters with the lowercase letters, it generates a list of words for each Surah. This stemmer produced 96.6% accuracy for prefix and 97% for postfix stemming.

Regarding to Persian stemming, M. Tashakori [13] proposed first Persian stemmer called Bon, based on rule-based strategy. It is an iterative longest matching algorithm. Bon truncates longest possible morpheme from the word and this process is repeated until no more character left to truncate. After removing the matched prefix and suffix from Persian words, the achieved stem may be incorrect. Bon uses a re-coding technique to produce the correct stem of the processed Persian word. By using this Persian stemming method, the recall is improved by 40%. Another Persian stemmer [14] is developed by Mokhtaripour which is also based on rule-based approach. This stemmer works without the help of dictionary. In order to handle the borrowed words such as Arabic, English etc., this stemming method proposed certain rules. By enforcing these rules, stemming performance of this work is considerably improved. This stemmer is used in a query system and 46% accuracy of the query system was improved by using this proposed Persian stemming technique.

As for as Urdu language is concerned only two methods [1, 15] for Urdu stemming have been proposed i.e. Assas-band and Light weight Urdu stemmer. These stemmers can only handle prefix and postfix present in Urdu words. To remove these prefix and postfix from Urdu words, these techniques [1, 15] use very large lists of rules and exception lists. These are also highly dependent on these large lists. The large size of rules list and exception lists considerably affect the efficiency of existing Urdu stemming methods. As Urdu language is a union of other foreign languages i.e. Arabic, Persian, Hindi, Turkish, etc. proposed Urdu stemmers are not competent to produce the stem of loan words i.e. Arabic, Persian, Hindi, Turkish, etc.

3. Proposed Urdu Stemmer

In this section, we describe our proposed Urdu stemming approach to stem Urdu text. This stemming method is based on rule based strategy and used affix stripping technique to generate stem of word. The overview of proposed Urdu stemmer is presented in figure 1. To support the proposed Urdu stemming technique, we have developed various rules and exception lists which are as follow:

3.1 Prefix Rules List

Prefix is a smallest language unit that is attached to the start of the word. The prefix may be single or two characters long and sometimes it is a morpheme. In order to produce prefix rule, various grammar books and Urdu literature are consulted to get a list of 60 prefixes rules. The size of this list is much smaller as compared to that generated in earlier work [15]. Examples of prefix rules are بد, بر, نا, ال.

3.2 Postfix Rules List

Postfix is a smallest language unit that is attached to the end of the word. The postfix is normally one to two characters long and sometimes it is a morpheme. After consulting various grammar books and Urdu literature, we presented a list of 140 suffixes rules. The size of this list is significantly smaller than the size of list as presented by [15]. Samples of these suffixes are وں, وے, اتے, ویں.

3.3 Prefix Global Exception List (PrGEL).

The correct identification of prefixes is very important because a wrong interpretation of prefix leads to poor stemming resulting in a loss of significant information. In Urdu morphology, there exist some words having prefixes as they matched with one of the rules. But in reality, they are an integral part of some word. Removing such prefix will result in destruction of such words. For instance, when we remove prefix "با" from the word "بازو" (arm), then it returns stem "زو", which is incorrect. As these rules handle vast majority of valid prefixes, it is not logical to remove such rules to avoid destruction of some words. Such words are therefore handled as an exceptional case by putting them in exception lists. In our proposed work, we have developed a prefix global exception list of about 5000 words. This exception list of prefixes is significantly smaller in size as compared to that generated by [15].

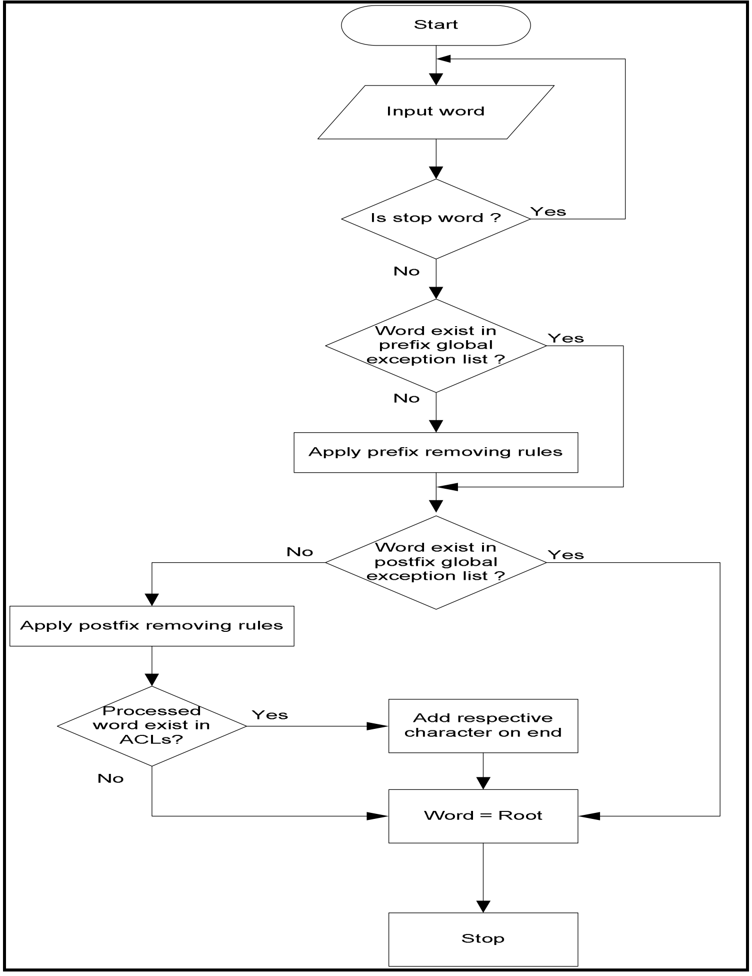
Start
Input word
Is stop word ?
Yes
No
Word exist in prefix global exception list ?
Yes
No
Apply prefix removing rules
Word exist in postfix global exception list ?
No
Yes
Apply postfix removing rules
Processed word exist in ACLs?
Yes
No
Add respective character on end
Word = Root
Stop

Figure 1: Overview of Proposed Urdu Stemmer

3.4 Postfix Global Exception List (PoGEL).

Exception list of postfix stemming is critical to avoid destruction of certain words due to application of postfix rules. There are many words in Urdu morphology that appears to have a postfix. If we truncate this postfix from the word, its incorrect form will be produced. For example, in the word "کرسی" (chair) when suffix "ی" is stripped then it generates stem کرس, which is invalid. Therefore to make sure the originality of these words, they must be treated as an exceptional case. A postfix global exception list of about 6000 words has been created to support this stemming work. The size of this list is considerably smaller than the size of list used by [15].

3.5 Add Character Lists (ACLs).

Sometime, the truncation of postfix from Urdu words results in incomplete stem. For example, after applying the suffix rules the word جگہوںwill become جگwhich is incorrect. Therefore, a character Hey (ہ) will be added at the end of word جگto make it a meaningful word place (جگہ). For our proposed stemmer, we have developed 8 separate lists w.r.t. characters (الف, ت, ر, س, ن, و, ہ, ی) to attach at the end of such incomplete stems.

3.6 Non Informative Word / Stop Word List.

Non informative words are those that occur frequently and do not provide valuable information to understand the sentence and its type. In order to clean the dataset form non informative words, a static list of 200 words is generated by consulting Urdu language experts, grammar books and Urdu literature. Some example words are کا, کی, کے, نے..

3.7 Stem Word Dictionary.

Stem dictionary contains a list of words followed by their actual stem. This dictionary is essential to validate the accuracy of stemming algorithm. After studying various grammar books and Urdu literature, we developed a stem word dictionary of about 10000 words to verify the accuracy of proposed stemming method. Some examples of stem words are نظر, جذب, جبر, حکم.

3.8 Proposed Urdu Stemmer Algorithm

The proposed algorithm is based on longest-match theory which states that when more than one stemming rule is matched for a given word, then apply that rule which removes maximum number of characters from the word to reduce it to its potential stem. To achieve this, we need to find all possible rule matches rather than applying the rule immediately matched. Our proposed algorithm compiles all possible affixes once and arranged them based on their length. Affix with maximum length is removed from the word. The algorithm is comprised of following steps:

1) Input a word to get its stem.

2) Search the word in stop word list.

a) Filter out the word if it is a stop word such as if its match is found from the non-informative word list. Ignore that word and select the next one from the word sequence.

b) If word does not exist in non-informative word list, then go to step 3.

3) Search the word in Prefix Global Exception (PrGEL) List.

a) If word exists in PrGEL then go to step 4.

b) If word is not found in PrGEL, then apply prefix removing rules and remove the maximum matched prefix from the word and go to step 4.

4) Search the word in Postfix Global Exception (PoGEL) List.

a) If word found in PoGEL, mark the processed word as stem and go to step 5.

b) If word does not exist in PoGEL, then apply the postfix removing rules.

c) If any one of the postfix removing rule is matched, then remove the maximum matched suffix from the word and search the processed word in Add Character Lists (ACLs).

d) If processed word found in any ACLs, then attach the respective character to the end of processed word. Mark the processed word as stem and go to step 5.

e) If processed word does not found in any ACLs, mark the processed word as stem and go to step 5.

f) If none of the postfix rule is applied then mark the word as stem and go to step 5.

5) Repeat steps 1-4 for all words.

4. Experimental Studies

To evaluate the performance of our proposed Urdu stemming methodology, four self-generated Urdu headline news corpora have been used. Brief overview of these Urdu headline news corpora is given in TABLE 1.

4.1 Experiment 1: Evaluation of Proposed Urdu Stemmer.

The purpose of this experiment is to evaluate the stemming accuracy of proposed Urdu stemmer on variety of Urdu datasets. We evaluated the proposed Urdu stemmer on the unique words of Urdu headline news corpora. After removing the less informative words in a pre-processing step 32000 unique words are extracted. Proposed prefix and postfix rules as discussed in section 3.1 and section 3.2 are applied on 32000 unique words. The performance of the proposed prefix and postfix rules is measured using the number of words that matched prefix and postfix rules. We also report the number of True Positives (correctly stemmed words) and False Positives (incorrectly stemmed words) achieved using application of these rules on different corpora. Accuracy of our proposed stemming rules is then computed as the ratio of the True Positives and the number of words that matched stemming rules. The prefix stemming accuracy of proposed Urdu stemmer is presented in TABLE 2. It is observed that the proposed prefixes rules are showing good accuracy results i.e. 85.64%, 87.91%, 83.59%, 85.28% respectively using all the corpora.

The postfix stemming accuracy results are achieved by using proposed postfix rules are given in TABLE 3. As obvious from the stemming results given in Table 3 that proposed postfix rules give the best stemming results with the significant accuracies i.e. 91.05%, 90.54%, 88.22%, and 88.67% respectively.

4.2 Experiment 2: Comparison of Proposed Urdu stemmer with Competitor.

The aim to conduct this experiment is to compare the stemming accuracy of proposed Urdu stemmer with existing Light Weight Urdu Stemmer. The experiment is performed on our internally generated Urdu headline news datasets as described in TABLE 1. It also aims to demonstrate that our proposed Urdu stemming approach is generic for any kind of Urdu corpora. The competitor rules are applied on 32000 unique Urdu words. The accuracy results of competitor approach for prefix stemming and postfix stemming are presented in TABLE 4 and TABLE 5. It is observed that competitor performance is significantly affected by the incorrect identification of prefixes and postfixes. The word generated by competitive stemmer is not a valid word because their rules break down a lot of compound words. They have also generated erroneous prefix and postfix rules that are the part of words.

In Urdu vocabulary, there are large numbers of compound words e.g. گلدان(Flower pots), آبپاشی (Irrigation) etc. Compound words do not have any stem because these are formed with the combination of other words. These words have their own significant meanings. The breaking down of these compound words will definitely causes the wrong stemming and the loss of useful information. For example word آبپاشی (Irrigation) has a unique significant meaning if prefix آبis removed then the meaning of this word will totally destroy. There are lots of compound words i.e. تنگدست(poor), دلفروش(blindly follower), etc that have been destroyed by competitor rules. Competitor approach is also not able to handle borrowed words effectively. The comparison of proposed stemming approach with the competitor is given in TABLE 6 and TABLE 7. The comparison of stemming accuracies of proposed stemming approach with the competitor demonstrates that proposed stemming approach gives best results as compared to competitor.

Table 1.A brief overview of experimental corpora.

Sr. #	Corpora	Dataset Description	Total Words	Unique Words
1	Corpus 1	An Urdu headline news corpus. It contains the news of two different categories i.e. politics and weather	12500	5070
2	Corpus 2	It is also an Urdu headline news corpus. It comprises of two different news classes i.e. sports and terrorist.	7250	3080
3	Corpus 3	It consists of unique Urdu word. It has developed by using various grammar books and Urdu dictionaries.	24238	24238
4	Corpus 4	A comprehensive headline news corpus obtained by combining corpus 1, corpus 2 and corpus 3.	43988	32388

Table 2: Stemming accuracy results of proposed prefix rules.

Corpora	Total Words Tested	Number of Words that Matched Prefix Rules	True Positive	False Positive	Accuracy %
Corpus 1	4819	195	167	28	85.64%
Corpus 2	2943	182	160	22	87.91%
Corpus 3	24238	323	270	53	83.59%
Corpus 4	32000	700	597	103	85.28%

Table 3: Stemming accuracy results of proposed postfix rules.

Corpora	Total Words Tested	Number of Words that Matched Postfix Rules	True Positive	False Positive	Accuracy %
Corpus 1	4819	2280	2076	204	91.05%
Corpus 2	2943	1460	1322	138	90.54%
Corpus 3	24238	18023	15900	2123	88.22%
Corpus 4	32000	21763	19298	2465	88.67%

Table 4: Stemming accuracy results of competitor prefix rules

Corpora	Total Words Tested	Number of Words that Matched Prefix Rules	True Positive	False Positive	Accuracy %
Corpus 1	4819	920	154	766	16.73%
Corpus 2	2943	413	57	356	13.80%
Corpus 3	24238	2238	288	1950	12.86%
Corpus 4	32000	3571	499	3072	13.97%

Table 5: Stemming accuracy results of competitor postfix rules

Corpora	Total Words Tested	Number of Words that Matched Postfix Rules	True Positive	False Positive	Accuracy %
Corpus 1	4819	2760	1520	1240	55.07%
Corpus 2	2943	1835	840	995	45.77%
Corpus 3	24238	20023	7990	12033	39.90%
Corpus 4	32000	24618	10350	14268	42.04%

Table 6: Comparative accuracy results of competitor and proposed prefix rules

Corpora	Total Words Tested	Competitor Prefix Rules Accuracy %	Proposed Prefix Rules Accuracy %
Corpus 1	4819	16.73%	85.64%
Corpus 2	2943	13.80%	87.91%
Corpus 3	24238	12.86%	83.59%
Corpus 4	32000	13.97%	85.28%

Table 7: Comparative accuracy results of competitor and proposed postfix rules.

Corpora	Total Words Tested	Competitor Postfix Rules Accuracy %	Proposed Postfix Rules Accuracy %
Corpus 1	4819	55.07%	91.05%
Corpus 2	2943	45.77%	90.54%
Corpus 3	24238	39.90%	88.22%
Corpus 4	32000	42.04%	88.67%

5. Conclusion

This paper presents a novel stemming approach for Urdu text. In proposed Urdu stemmer, we have developed generic prefix and postfix rules that can b applied on any kind of Urdu datasets. These rules are significantly smaller in size as compared to competitor. For experimental analysis, it is observed that our proposed stemming approach gives superior accuracy results as compared to competitor i.e. A Light Weight Urdu Stemmer. Our approach is also capable to handle compound words and loan words (words borrowed from other languages i.e. Arabic, Turkish, Persian, Hindi, etc).

REFERENCES

[1] Q. Akram, A. Naseer and S. Hussain. Assas-band, an affix- exception-list based Urdu stemmer. Proceedings of the 7th Workshop on Asian Language Resources. Singapore. pages 40–47. (2009)

[2] M. Al-Khuli. A dictionary of theoretical linguistics: English-Arabic with an Arabic- English glossary. Published by Library of Lebanon. (1991).

[3] K. Riaz. Challenges in Urdu Stemming (A Progress Report). BCS IRSG Symposium: Future Directions in Information Access (FDIA). (2007)

[4] S. Ahmad, W. Anwar, U.I. Bajwa. Challenges in Developing a Rule based Urdu Stemmer. Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP). Chiang Mai, Thailand. pages 46–51. (2011)

[5] Bento, Cardoso and Dias. Progress in Artificial Intellegence, 12th Portuguese Conference on Artificial Intelligence, pages 693 –701. (2005)

[6] A. A. Sharifloo, M. Shamsfard. A Bottom up Approach to Persian Stemming. IJCNLP. Pages 583-588. (2008)

[7] A. G. Jivani et al. A Comparative Study of Stemming Algorithms. IJCTA. vol 2, no. 6. Pages 1930-1938. (2011).

[8] J. B. Lovin's. Development of a stemming algorithm. Mechanical Translation and Computer Linguistic. vol.11, no.1/2, pp. 22-31, (1968).

[9] M.F. Porter. An algorithm for suffix stripping. Program. 14: 130-137. (1980)

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

[11] S. Khoja and R. Garside. Stemming Arabic Text. Lancaster, UK. Computing Department, Lancaster University. (1999)

[12] N. Thabet, Stemming the Qur'an. In the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. pages 85-88. (2004)

[13] M. Tashakori, M. Meybodi & F. Oroumchian. Bon: first Persian stemmer. Lecture Notes on Information and Communication Technology. pages 487-494. (2002)

[14] Mokhtaripour and S. Jahanpour. Introduction to a new Farsi stemmer. CIKM Proceedings of the 15th ACM international conference on Information and knowledge management.Arlington, Virginia, USA. pages 826-827. (2006).

[15] S. Ahmad, W. Anwar, U.I. Bajwa, X. Wang. A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language. Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP). Mumbai. pages 69–78. (2012)

[16] M. Waqas. Local Government Management And Performance. VFAST Transactions on Education and Social Sciences, vol 2, no. 2. (2014)

[17] S. Jawad, F. Wahab. A Student-Centered Effective Learning Framework For Quality Education. VFAST Transactions on Education and Social Sciences. Vol 3, no 1. (2014)