# Challenges in Developing a Rule based Urdu Stemmer

**Sajjad Ahmad Khan, Waqas Anwar, Usama Ijaz Bajwa**
Department of Computer Science
COMSATS Institute of Information Technology, Abbottabad, Pakistan
sajjadkhan25@hotmail.com,waqas@ciit.net.pk,usama@ciit.net.pk

## Abstract

Urdu language raises several challenges to Natural Language Processing (NLP) largely due to its rich morphology. In this language, morphological processing becomes particularly important for Information Retrieval (IR). The core tool of IR is a Stemmer which reduces a word to its stem form. Due to the diverse nature of Urdu, developing stemmer is a challenging task. In Urdu, there are large numbers of variant forms (derivational and inflectional forms) for a single word form. The aim of this paper is to present issues pertaining to the development of Urdu stemmer (rule based stemmer).

## 1. Introduction

Urdu is an Indo-Aryan language. It is the national language of Pakistan and is one of the twenty-three official languages of India. It is written in Perso-Arabic script. The Urdu vocabulary consists of several languages including Arabic, English, Turkish, Sanskrit and Farsi (Persian) etc.
Urdu's script is right-to-left and form of a word's character is context sensitive, means the form of a character is dissimilar in a word because of the position of that character in the word (beginning, centre, on the ending) (Waqas et al., 2006).
In Urdu language, morphological processing becomes particularly important for Information Retrieval (IR). Information retrieval system is used to ensure easy access to stored information. It also deals with saving, representation and organization of information objects. Modules of an IR system consist of a group of information objects, a group of requests and a method to decide which information items are most possibly helping to meet the requirements of the requests. Inside IR, the information data which is stored and receives search calls usually corresponds to the lists of identifiers recognized as key terms, keywords. One of the attempts to make the search engines more efficient in information retrieval is the use of stemmer. Stem is the base or root form of a word. Stemmer is an algorithm that reduces the word to their stem/root form e.g. tested, testing, pretest and tester have the stem "test". Similarly the Urdu stemmer should stem the words **کم عقل** (senseless), **عقل مند** (sensible), **عقل مندی** (sagacity) to Urdu stem word **عقل** (sense). Stemming is part of the complex process of taking out the words from text and turning them into index terms in an IR system. Indexing is the process of selecting keywords for representing a document.
The smallest units of word which cannot be decomposed further into smaller meaningful units are called Morphemes.[1] They are of two kinds: free morphemes and bound morphemes. Morphemes which exist freely (alone) are called free morphemes whereas bound morphemes are made as a result of combination with another morpheme. For instance "flower" is a free morpheme, while "s" is the example of a bound morpheme.

The study of internal structure of words is called Morphology.[2] Deriving new words from the existing ones is called derivational morphemes e.g. Honour, Honourable, Honourably. Examples in Urdu: The words **چاہت** (love), **چاہنا** (to love) and **چہیتا** (lovely) are the derivatives of word **چاہ** (love). Those morphemes that produce the grammatical formation of a word is called Inflectional morphemes e.g. Boys. Examples in Urdu: The words **سخت تر** (harder) and **سخت ترین** (hardest) are the inflected forms of word **سخت** (hard).

The stemmer is also applicable to other natural language processing applications needing morphological analysis for example spell checkers, word frequency count studies, word parsing etc. The rest of the paper is organized as follows: In section 2, different rule based stemming algorithms are discussed. Section 3 gives an introduction regarding orthographic features. In section 4, several issues pertaining to Urdu stemmer are

[1] http://www.ielanguages.com/linguist.html
[2] http://introling.ynada.com/session-6-types-of-morphemes



*Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011,* , pages 46–51,
Chiang Mai, Thailand, November 8, 2011.

discussed in detail. Conclusion of the study and the future work is discussed in section 5.

## 2. Stemming Algorithms

There are four kinds of stemming approaches (Frakes, R.Baeza-Yates, 1992): table lookup, affix removal, successor variety and n-grams. Table lookup method is also known as brute force method, where every word and its respective stem are stored in table. The stemmer finds the stem of the input word in the respective stem table. This process is very fast, but it has severe disadvantage i.e. large memory space required for words and their stems and the difficulties in creating such tables. This kind of stemming algorithm might not be practical. The affix removal stemmer eliminates affixes from words leaving a stem. The successor variety stemmer is based on the determination of morpheme borders, i.e., it needs information from linguistics, and is more complex than affix removal stemmer. The N-grams stemmer is based on the detection of bigrams and trigrams.

The (J.B. Lovins, 1968) published the first English stemmer and used about 260 rules for stemming the English language. She suggested a stemmer consisting of two-phases. The first stage removes the maximum possible ending which matches one on a redefined suffix list. The spelling exceptions are covered in the 2nd stage.

The (M.F. Porter, 1980) developed the stemmer on the truncation of suffixes, by means of list of suffixes and some restrictions/conditions are placed to recognize the suffix to be detached and generating a valid stem. Porter Stemmer performs stemming process in five steps. The Inflectional suffixes are handled in the first step, derivational suffixes are handling through the next three steps and the final step is the recoding step. Porter simplified the Lovin's rules upto 60 rules.

Different stemmers have also been developed for Arabic language. The (S. Khoja and R. Garside, 1999) developed an Arabic stemmer called a superior root-based stemmer, developed by Khoja and Garside. This stemming algorithm truncates prefixes, suffixes and infixes and then uses patterns for matching to pull out the roots. The algorithm has to face many problems particularly with nouns. The (Thabet. N., 2004) created a stemmer, which performs on classical Arabic in Quran to produce stem. For each *Surah*, this stemmer generates list of words. These words are checked in stop word list, if they don't exist in this list then corresponding prefixes and suffixes are removed from these words.

The (Eiman Tamah Al-Shammari, Jessica Lin, 2008) proposed the Educated Text Stemmer (ETS). It is a simple, dictionary free and efficient stemmer that decreases stemming errors and has lesser storage and time required.

Bon was the first stemmer developed for Persian language (M. Tashakori, M. Meybodi & F. Oroumchian, 2003). Bon is an iterative longest matching stemmer. The iterative longest matching stemmer truncates the longest possible morpheme from a word according to a set of rules. This procedure is repeated until no more characters can be eliminated. The (A. Mokhtaripour and S. Jahanpour, 2006) proposed a Farsi stemmer that works without dictionary. This stemmer first removes the verb and noun suffixes from a word. After that it starts truncation of prefixes from that word.

Till date only one stemmer i.e. Assas-Band, developed for Urdu language (Q. Akram, A. Naseer and S. Hussain, 2009). This stemmer extracts the stem/root word of only Urdu words and not of borrowed words i.e. words from Arabic, Persian and English words. This algorithm removes the prefix and suffix from a word and returns the stem word. This stemmer does not handle words having infixes.

## 3. Orthographic Features of Urdu

According to (Malik M G Abbas et al., 2008), Urdu alphabet consists of 35 simple consonants, 15 aspired consonants, 10 vowels, 15 diacritical marks, 10 digits and other symbols.

### 3.1 Consonants

Consonants are divided into two groups:

**a. Aspirated Consonants**

There are 15 aspirated consonants in Urdu language. These consonants are shown by a grouping of a simple consonant to be aspirated. A special letter called Heh Doachashmee (ھ) is used to mark the aspiration. Aspired Consonants are بھ, پھ, تھ , ٹھ, جھ, چھ, دھ , ڈھ , کھ , گھ , رھ, ڑھ , لھ , نھ, مھ

**b. Non Aspirated Consonants**

Urdu language consists of 35 non aspirated consonant signs that represent 27 consonant sounds. Various scripts are employed to show the similar sound in Urdu, For example: Sad **(**ص**)**, Seen **(**س**)** and Seh **(**ث**)** represent the sound [s].

### 3.2 Vowels

Urdu has ten vowels. Seven of them contain nasalized forms. Out of these seven, four long vowels are represented by Alef Madda (آ), Alef (ا), Choti Yeh (ی) and Vav (و) and three short vowels are represented by Arabic Kasra (Zer), Arabic Fatha (Zabar) and Arabic Damma (Pesh). In Urdu language, the Vowel demonstration is context sensitive. For example, the Urdu Choti Yeh (ی) and Vav (و) can also be used as a consonant (Malik M G Abbas et al., 2008).

### 3.3 Aerab Marks

The aerab marks are those marks that are added to a letter to change the pronunciation of a word or to differentiate among similar words. It is also called as diacritical mark or diacritic.[3]

There are 15 accent marks in Urdu (Malik M G Abbas et al., 2008). Accent marks (Zabar, Zer, Pesh, Ulta Pesh, Do-Zabar, Do-Zer, Do-Pesh etc) represent vowel sounds. These are placed above or below of an Urdu word. The accent marks are very rarely used by people in writing Urdu. When the diacritic of a character in a word is changed then it could entirely change its meaning. These accent marks play a significant role in the right pronunciation and recognition of meaning of a sentence, such as:

درخت پر انگور کی بیل ہے۔
(A vine is on the tree)

and بیل گھاس کھا رہا ہے۔
(The bull is eating grass)

In the first sentence, the word (بیل) means "a creeping plant" or a "vine" while in the second sentence it means a "bull". To remove the doubt between these two words, there should be Zabar after Beh (ب) in the second sentence.

### 3.4 Special Characters

There are two special characters used in Urdu which are discussed bellow:

**a. Hamza (ء)**

Hamza is used to separate two consecutive vowels sounds. For example, in آؤ (come), Hamza is separating two vowel sounds i.e. Alef Madda (آ) and Vav (و).

**b. Heh Doachashmee (ھ)**

Heh Doachashmee (ھ) changes the action of a simple consonant and makes it aspired consonant. For example, پھ = ھ + پ , جھ= ھ + ج

Examples in words: پھل ,جھنڈا
(Flag, Fruit)

## 4. Issues in developing an Urdu Stemmer

### 4.1 Morphological rich language

Urdu is morphologically rich language. It produces high number of derivational and inflectional words for a single word form. There are 57 different forms that can be generated from a single Urdu word (Rizvi, S. & Hussain, M., 2005). For Example, some different forms of Urdu word پڑھ (read) are:

پڑھنا،پڑھا،پڑھے،پڑھیں،پڑھی،پڑھنی،پڑھو،پڑھوں،
پڑھا،پڑھانا،پڑھاتے،پڑھاتا،پڑھوا،پڑھواتا،پڑھوں

Besides its own vocabulary, the Urdu vocabulary also consists of large number of Arabic, Persian, Hindi and English words etc. Thus Urdu language inherits the characteristics of the above mentioned languages too and as a result stemming process becomes a challenging task. We cannot achieve a good level of precision if a stemmer of any borrowed language is used as a stemmer on Urdu words. The reason is that, the Arabic stemmer will just stem Arabic words that are used in Urdu as borrowed words and a Persian stemmer will just stem borrowed Persian words etc.

By using traditional process of modeling every form of a word as a unique word generates a lot of problems for Natural Language Processing applications such as growth of vocabulary, inflectional gaps, larger out-of-vocabulary rates and poor language model probability estimation.

The relation among words in Urdu is found by using inflecting nouns, postposition and pronouns to state case information, number and gender. Inflecting verbs to reproduce number, gender and person information etc. Inflecting adjectives are to agree with the noun in number, gender and case. Thus, the standard stemmers which are developed for English words are not practically implementable for Urdu language.

### 4.2 Engineering issues

Urdu is bidirectional language and electronically we cannot represent it in ASCII form. Such type of language is represented by a special character

[3] http://www.the-comma.com/diacritics.php

set called Unicode. The Arabic Orthography Unicode Standards are used to process Urdu.
Unicode is not supported by many programming languages. The languages that support Unicode include C#, Python and Java etc. Some programming language support Unicode but the IDE may not support it fully.

### 4.3 Diacritical Marks

Special attention should be given to the diacritical marks while developing an Urdu stemmer. The stem of an Urdu word changes with the use of these marks. For example عالم is used in two senses, when Zabar is placed above the character ع and on ل, then its meaning is *people* and its stem is عالم (people). But when Zer is placed below ل, then its meaning is *scholar* and its stem is علم (knowledge).
Similarly رسل word has two meanings. One is *messengers* when Pesh is used on ر and س with stem رسول (messenger) and other is *access* when Zabar is used on ر and س with stem ارسال (sending). Another example is the word خاتم , which has two meanings (The last/ring), the first one has stem ختم (finish) and second has خاتم (ring).

### 4.4 Compound Words

For word formation, compounding is one of the morphological procedures. The grouping of two words which already exist is called a compound word (Payne, Thomas E., 2006). When two or more than two lexeme stems are merged together to produce another lexeme, then it is called compound word (Sproat. R., 1992). Examples are: Firefighter, Blackbird, Water-hose, Hardhat, Rubber-hose and Fire-hose in English.
It is very difficult to classify the compound words as a single or multiple words. The (Durrani N., 2007) discussed three schemes of compound words in Urdu i.e. AB, A-o-B and A-e-B.

**a. AB formation**

This scheme involves only joining of two free morphemes e.g. مرہم پٹی (Bandaging) , میاں بیوی (husband wife), couple literally, حال احوال (condition). AB form of compounds is further classified into Dvanda, Tatpurusa, Karmadharaya and Divigu (Sabzwari S, 2002).

**b. A-o-B formation**

This formation of Urdu compounds contains a linking morpheme "o" and is represented by a character **"و"** , e.g.عجزوانکساری (soberness and humility), خط وکتابت (correspondence), امن وامان (law and order).

**c. A-e-B formation**

In this formation constituent words are connected with the help of one of the enclitic short morphemes; zer-e-izafat or hamza-e-izafat e.g. صدرمملکت (president) is combined by a diacritical mark "Zer" below ر called as zer-e-izafat while in جذبۂ دل (heart's spirit) and خلفائےاسلام (Islamic caliphs), the diacritical mark hamza **(ۂ)** is used as a hamza-e-izafat.
Some times the reduplication also produces ambiguity; whether it is treated as single or double word e.g. جگہ جگہ،آہستہ آہستہ،ساتھ ساتھ
(together, slowly, at every place)
Therefore there should be some rule for the identification of compound words. Thus these points should be considered while developing an Urdu stemmer.

### 4.5 Tokenization

The natural language processing applications need that the entered text should be tokenized for further processing. English language generally uses white spaces or punctuation marks for the identification of word boundaries.
Although in Urdu, space character is not present but with increasing usage of computer, it is now being used, for generating right shaping and to break up words.
Example: صدرنےدورسےوزیرکوآوازدی
(The President called away the Minister)
In the above sentence there are eight words (tokens) but computer will consider the whole sentence as a single word because the computer will generate tokens on the basis of space occurrence. As due to non-joiner characters (here ر،ے،و،ز) in the words, no space occurs among words, so this whole sentence is considered as a single word.
Therefore, during stemming, these non-joiner characters wrongly generate tokens of input text, stemmer will generate wrong resultantly stem.
Tokenization process should be error free, hence producing correct tokens before applying an Urdu stemmer.

### 4.6 Affixes Removal

The word affix is used by the linguists for expressing that where a bound morpheme precisely be joined to a word. The Prefix, Suffix and infix are called affixes. Due to the use of affixes, a single word may contain a lot of variants and by removing these affixes (prefix and suffix) from a word will result into a stem word e.g. بدگمانی (mis presumption). After removing the Urdu prefix

and suffix from this word, produced a stem word گمان (presumption).
A lot of stemmers (except for Urdu) were developed for stripping off prefixes & suffixes from a word but there is little work done on infix stripping from a word. We cannot get stem word of an Urdu word by only stripping off prefixes and (or) suffixes e.g. اقوام (nations) , مساجد (mosques) , علوم (knowledge).
These words contain infixes and large amount of such type of words are present in Urdu. Special attention should be given to those Urdu words having infixes. After studying the morphology of Urdu words, it is noticed that if patterns for such type of words (having infixes) are made, then a correct stem could be achieved.

### 4.7 Exceptional Cases

#### a. Exceptional words

The removal of affixes (Prefixes and Suffixes) from a word produces a stem word but some times truncating these affixes leads to an erroneous stem e.g. نادار. Here نا is a prefix, where the stemmer eliminates it by producing دار , which is not a correct stem of the above stated word.
It means that in some words, the affixes play the role of stem characters and should not be removed. Such type of words should be treated as an exceptional case. In Urdu, there are a lot of words that can be treated as an exceptional case, thus for a stemmer, such word lists should be maintained in advance.

#### b. Urdu digits, Arithmetic Symbols and Punctuations

Urdu is read and written from right to left but when numbers are introduced, it is read and written from left to right.

حفصہ کی برتھ ڈے ۲فروری ۲۰۰۹ہے
(Hafsa's birthday is 2nd February 2009)

The Urdu digits (۰-۹), Arithmetic Symbols (+,-,*, /) and Punctuation marks (-,؟ , ، , ، , ',",؛ ,:) should be treated as an exceptional case during developing Urdu stemmer.

### 4.8 Stem-word Dictionary

To check the accuracy of any stemmer, there should be a stem word dictionary. After studying relevant literature, it is noted that there is no stem dictionary available for Urdu text. Therefore, development of an Urdu stem dictionary is necessary for testing the accuracy of a stemmer on huge corpus.

### 4.9 Different Urdu words having same stem

In Urdu, there are a lot of words that are different in meaning but their stem is same e.g. تاثیر (characteristic) and آثار (signs). As we mentioned that the meaning of these two words are different from each other but their stem is same i.e. اثر
Similarly the words ملوک (rulers) and ملایک (angels) are two different words having single script for their stem without diacritical marks i.e. ملک. The word ملک has two meanings i.e. ruler or angel. The word اصول (principles) and اصلیت (facts) have same stem i.e. اصل (principle/fact). Such type of words needs attention while developing a stemmer for Urdu language.

### 4.10 Code switching

Code switching, in linguistics, is the parallel use of more than one languages during conversation. The code switching in Urdu language is common and it accepts foreign words especially from English, e.g. یہ کیمرہ borrowed ہے (This Camera is borrowed).
In this example the Urdu text is from right to left-wards, while the English word "borrowed" is from left to right. The tokenization of the above sentence is performed in proper way electronically but Urdu stemmer will not stem the foreign word "borrowed", which is an issue.

## 5. Conclusion and Future Work

Stemmer is the core tool of any IR system. In this paper we have discussed some rule based English, Arabic, Persian and Urdu stemmers. Very less work has been done on Urdu stemmer due to its complex and rich morphology. Besides its own vocabulary, Urdu is also influenced by other morphology such as Arabic, Persian, Hindi, English etc. We have pointed out some challenges pertaining to the development of an Urdu stemmer. These issues should be considered while developing a rule based Urdu stemmer.
After studying different stemmers developed for Arabic, Persian and Urdu languages, we intend to develop an efficient rule based Urdu stemmer which will not only handle those Urdu words having prefixes and suffixes but also infixes. We will make patterns for handling infixes. For preprocessing of the proposed Urdu stemmer, Urdu stop word list will be maintained. An Urdu stem-word dictionary will also be prepared for evaluation purposes.

## References


A. Mokhtaripour and S. Jahanpour, 2006. *Introduction to a New Farsi Stemmer*, CIKM'06, November 5–11, Arlington, Virginia, USA.

Durrani N. 2007. *Typology of Word and Automatic Word Segmentation in Urdu Text Corpus*. *National University of Computer and Emerging Sciences,*Lahore, Pakistan.

Eiman Tamah Al-Shammari, Jessica Lin, October 30, 2008. *Towards an Error-Free Arabic Stemming*, iNEWS'08, Napa Valley, California, USA.

Frakes, R.Baeza-Yates, 1992. *Information Retrieval: Data Structures & Algorithms*, New Jersey: Prentice Hall PTR.

J.B. Lovins, 1968. *Development of a stemming algorithm. Mechanical Translation and Computational Linguistics*, 11, pp.22–31.

Javed I. 1985. *New Urdu Grammar.* Advance Urdu Buru, New Dehali

Malik, M. G. Abbas. Boitet, Christian. Bhattcharyya, Pushpak. 2008. *Hindi Urdu Machine Transliteration using Finite-state Transducers*, proceedings of COLING 2008, Manchester, UK.

M.F. Porter, 1980. *An algorithm for suffix stripping*, Program, 14(3) pp. 130-137.

M Tashakori, MR Meybodi, F Oroumchian, 2003. *Bon: The Persian stemmer*, in Proc. 1st EurAsian Conf. on Information.

Payne, Thomas E. 2006. *Exploring Language Structure, A Student's Guide*. Cambridge: Cambridge University Press.

Q. Akram, A. Naseer and S. Hussain, 6-7 August 2009. *Assas-Band, an Affix- Exception-List Based Urdu Stemmer*, Proceedings of the 7th Workshop on Asian Language Resources, pp. 40–47, Suntec, Singapore.

Rizvi, S. & Hussain, M. 2005, *Analysis, Design and Implementation of Urdu Morphological Analyzer*, Engineering Sciences and Technology, SCONEST 2005. Student Conference, pp. 1-7

Sabzwari, S. 2002, *Urdu Quwaid*. Lahore: Sang-e-Meel Publication

S. Khoja and R. Garside, 1999. *Stemming Arabic Text,* Lancaster, UK, Computing Department, Lancaster University.

Sproat, R. 1992. *Morphology and Computation*. The MIT Press

Thabet, N. 2004. *Stemming the Qur'an* In the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.

Waqas A., Xuan W., Lu Li, Xiao-long W. 2006. A Survey of Automatic Urdu Language Processing. International Conference on Machine Learning and Cybernetics, pp: 4489-4494