A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language

Sajjad Ahmad Khan[1], *Waqas Anwar*[1], *Usama Ijaz Bajwa*[1], *Xuan Wang*[2]
(1) COMSATS Institute of Information Technology, Abbottabad, Pakistan
(2) Harbin Institute of Technology, Shenzhen Graduate School, P.R.China
Sajjad_ak78@yahoo.com,waqas@ciit.net.pk,usama@ciit.net.pk,
wangxuan@insun.hit.edu.cn

ABSTRACT

Stemming is a procedure that conflates morphologically related terms into a single term without doing complete morphological analysis. Urdu language raises several challenges to Natural Language Processing (NLP) largely due to its rich morphology. The core tool of information retrieval (IR) is a Stemmer which reduces a word to its stem form. Due to the diverse nature of Urdu, developing its stemmer for an IR system is a challenging task. This paper presents a light weight stemmer for Urdu text, which uses rule based approach. Exceptional lists are developed to enhance the accuracy of the stemmer. The result of the stemmer is quite enough and can be effective in IR system.

KEYWORDS : Information Retrieval, Light weight stemmer, Exceptional Lists, Suffix and prefix list

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 varies in a word because of the position of that character in the word (start, end, medial, independent).

According to (Al-Khuli, M. 1991), Morphology deals with the internal structure of words. Hundreds thousands of words are contained in every human language and constantly fresh words are incorporated. These words are formed from a group of smaller components. In morphology, its main building blocks are morphemes. Morpheme is the smallest component in a language having some meaning[1]. There are two types of morphemes i.e. free and bound morphemes.

Morphemes which exists freely (alone) are called free morphemes whereas bound morphemes are made as a result of combination with another morpheme. The affixes are bound morphemes. Those morphemes that produce the grammatical formation of a word are called Inflectional morphemes[2.] Deriving new words from the existing ones is called derivational morphemes. In derivation, a different part-of-speech class is created by adding a bound morph to a stem.

In Urdu language, morphological processing becomes particularly important for IR. IR system is used to ensure easy access to stored information. 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 IR is the use of

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

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 69–78,
COLING 2012, Mumbai, December 2012.

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. jumping, jumped and jumps to the stem “jump”. Similarly the Urdu stemmer should stem the words بدنصیبی (unlucky), نصیب دار (lucky), بدنصیب (luckless) to Urdu stem word نصیب (luck).

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 organization is as follows: In section 2, different stemming approaches and rule based stemming algorithms are discussed, in section 3, Orthographic features of Urdu is discussed, in section 4, the Urdu morphology is discussed, section 5 discusses light weight stemmer, section 6 discusses the Proposed Urdu Stemmer and finally the evaluation of the stemmer is discussed in section 7.

2 Stemming approaches and algorithms

There are four kinds of stemming approaches (Frakes, et al.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 a disadvantage i.e. large memory space required for words and their stems and the difficulties in creating such tables. 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 study (J.B. Lovins, 1968) discussed the first English stemmer and used about 260 rules for stemming the English language. The study, suggested a stemmer consisting of two-phases. The first stage removes the maximum possible match from the suffix list sorted on the base of word lengths. The spelling exceptions are covered in the second stage.

The Porter stemmer (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 handled 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 study (S. Khoja, et al. 1999) explains an Arabic stemmer called a superior root-based stemmer. This stemming algorithm truncates prefixes, suffixes and infixes and then uses patterns for matching to pull out the roots. The study (Thabet, N. 2004) described 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 study (E.T. Al-Shammari, et al. 2008) proposed the Educated Text Stemmer (ETS). It is a simple, dictionary free and efficient stemmer that decreases stemming errors and needs lesser storage and time.

Bon was the first stemmer developed for Persian language (M. Tashakori, et al. 2002). 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 study (Mokhtaripour, et al. 2006) proposed a Persian 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*, has been developed for Urdu language (Q. Akram, et al. 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 used in Urdu. This algorithm removes the prefix and suffix from a word and returns the stem word.

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 i.e. Aspirated consonants and non 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 , مھ , نھ , لھ .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 Diacritical marks

The diacritical 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 accent mark or diacritic.[3]

There are 15 diacritical accent marks in Urdu (Malik, M. G. Abbas, et al. 2008). Diacritical 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 diacritical 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. For example the word **(بیل)** has two meanings i.e. a "creeping plant" as well as it means "bull". To remove the doubt between these two words, there should be Zabar after Beh **(ب)** for deducing the meaning as "bull".

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

4 Urdu morphological structure

Urdu verbs, nouns and adjectives are discussed in the following sections as in (Sabzwari, S, 2002):

4.1 Urdu verb

Verb (فعل) corresponds to occurrence or performing of something. That verb which does not take object is called intransitive verb (فعل لازم). When a verb needs a direct object then it is called transitive verb and we can called it (فعل متعدی) in Urdu. A root form is that morpheme of Urdu verb which is not changed among different morphological forms and it is also called base form. When a suffix (نا) is removed from the verb's lexicon form (infinitive form) then the left over part of an infinitive form will be a root, which is also called (مادہ) in Urdu. Causative Stem Form of verb can be achieved through adding of suffixes to root form. The Causative verb forms /transitivitized verb forms can be obtained through the roots of lower valency verb by adding Urdu suffixes: –aa (ا), –waa (وا) to the root form of verb. The causative verb types are known as stem forms. The infinitive (مصدر) is that kind of verb which has the suffix "نا". This form of the verbs also can be used in place of noun. Usually this form of the verb has a masculine suffix "نا". For feminine infinitive form the suffix "نے" and the suffix "نی" for oblique infinitive form are used. The repetitive form (استمراری) also known as imperfect or habitual form that is produced by appending suffixes to the root as: تا, تے, تی, تیں.

4.2 Urdu noun

A word which is the name of anything is called Noun, i.e. person's name, a place, an animal, thing, a concept, time, a situation etc.

Initially nouns are classified into proper and common nouns. Proper noun (اسم معرفہ) is the name of particular person, thing or place, e.g. Hafsa, Del Laptop , Karachi etc. The Common noun (اسم نکرہ) is the common name for any person, thing or place e.g. woman, pen, village, etc. The Common nouns are further divided into state, spatial, group, instrumental and temporal nouns. There are four fundamental properties related with Urdu nouns i.e. Gender, Number, Form and Case. The masculine and feminine gender is accepted by the Urdu's Nouns. For inanimate nouns, to get its gender categorization in Urdu, there is no common rule. Generally powerful, huge, dominant, heavy and larger items are masculine, whereas lighter, weak and smaller are feminine. Generally, masculine are shown through bigger nouns (اسم مکبر), whereas feminine are shown through smaller nouns (اسم مصغر). The Urdu nouns have two possibilities for number. One is called singular and second is called plural. In Normal form, when Urdu noun is listed in dictionary is called nominative form. When Nouns are followed by a postposition then it can be viewed in oblique form. Those Urdu nouns which belong to human being and sometimes other animate nouns contain a different form used to call/ address person, this type is called vocative.

4.3 Urdu adjective

An adjective express the status or action, quality that a noun refers to e.g. نیک آدمی(Pious man) , (Fresh banana) تازہ کیلا .Descriptive Adjective is the most common and significant type of adjective. It express attributes of the noun they qualify in terms of its color, size, dimensions, shape, sound, shade, personal trait, time, and quality. When the descriptive adjectives directly lead a nominal head as modifiers, in that case they are called attributive adjectives e.g. (Cruel king) ظالم بادشاہ ,سفید گیند (White ball).

The Possessive adjectives are used to show the ownership and the ownership relation is understand in two ways; whether, in noun phrases, adjectives lead the head noun as modifiers or they may be lead by a proper form of the genitive postposition **(**کا،کے،کی**)** e.g. حفصہ کا نیلا دوپٹہ (Hafsa's blue veil).

5 Light weight stemmer

Light weight stemming is to find the representative indexing form of a word by the application of truncation of affixes (Imed Al-Sughaiyer, et al. 2004). The core objective of light weight stemmer is to preserve the word meaning intact and so increases the retrieval performance of an IR system.

In this paper, a light stemmer for Urdu text is proposed. Various lists are developed that help in finding a stem of an Urdu word. This stemmer is rule based and works by truncating all possible affixes from a word. The problem in this light weight stemming is that in some cases there is ambiguity e.g. a particular string of letters may or may not play a function of affix. A method is introduced for detecting such type of ambiguity that finds if a specific sequence in an affix is part of the original word. For this purpose Global Prefix Exceptional and Global Suffix Exceptional Lists are developed which are discussed in the coming sections.

6 Proposed Urdu Stemmer

During morphological analysis of Urdu text, it is noticed that there could be upto sixty different forms of a single verb (Rizvi, S, et al. 2005). Therefore reducing these forms to their stem forms is very important during IR tasks. Affixes perform an important role in making inflection and derivation of words. When these affixes are removed from a word, it gives a stem word. We have developed different lists for our stemmer. The details of these lists are given below.

6.1 Stop word list

Stop words are those words that occur frequently. For English stemmers, a stop word list is already maintained, similarly, it is necessary for Urdu stemmer, that there must be a stop word list. Therefore, to accomplish this task, various Urdu books and literature were studied and finally 150 stop words are generated for Urdu. Some of the stop words in Urdu are نے،کا،کے،سے،میں.

6.2 Prefix list

Prefix is that morpheme which is attached to the start of a word. The prefix may consist of only a single character, two or more than two characters and sometimes a complete word. After consulting many grammar books 180 prefixes were collected. A sample of these prefixes are بد،زبر،با،خار.

6.3 Suffix list

The suffix is that morpheme that is added to the end of a word. The suffix may consist of a character, more than a single character or a complete word. A list of suffixes, consisting of 750 suffixes for Urdu text is collected after consulting relevant literature. Samples of such suffixes are ات،باش،بندی،پسند،خانہ.

6.4 Global prefix exceptional list

There are some words that contain a prefix, in fact it is not a prefix but it is the part of that word e.g. in the word “نایاب”, this word contain a prefix “نا” , if it is removed then it produces a word “یاب”, which is incorrect from stemming point of view. Therefore, such type of words must need to be identified in advance and to be treated as an exceptional case. For our stemmer an exceptional list for prefixes (about 13000) is created called Global Prefix Exceptional List (GPEL). Samples of such type of words are اعتماد،اتفاق،ازل،درج .

6.5 Global suffix exceptional list

In some cases, when a suffixe is removed from a word and actually that suffix was the part of a stem word that should not be detached. An erroneous result will be produced due to irrelevant truncation of suffix e.g. in the word “نانا”, when the suffix “نا” is removed then it produces stem “نا”, which is incorrect. Therefore such type of words should be treated as an exceptional case. Samples of such type of words are فوت،دستیاب،لیاقت،مرد. For our stemmer an exceptional list for suffix (of length about 16000) is created.

6.6 Characters add-list (normalization of stem)

When affix stripping algorithm is applied on Urdu words, then some time we get an incomplete word, e.g. after stripping affixes from a word “خوبصورتی” (Beautifulness), we get stem “صور”, which is incorrect stem. Therefore, this word should be added with the character “ت” to form correct stem form i.e. “صورت” (appearance). For this reason five types of list are maintained for five characters; **(**ا،ه،ت،ی،ن**)**.

6.7 Stem 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, we developed a stem dictionary for Urdu words having 3500 words.

6.8 Proposed algorithm of Urdu stemmer

The *longest-match* theory regarding affix removal states that when more than one affixes result as a match, the one which is longest should be removed. This method is applied by first sorting the affixes in any class in order of decreasing length. In case of suffix removal, if *-ional* is removed when there is another match *-ational,* then more work has to be done to eliminate *-at,* that is, for another order-class. To avoid this extra order-class, *-ational* should precede *-ional* in the prefix/suffix exception list.

An algorithm based on the principle of longest-match uses only one order-class. The entire feasible combinations of affixes are compiled and then ordered based on their length, so the longest word comes first. When a match is not found on longer suffixes / prefixes, shorter ones are scanned.

```
A-    INPUT Word for stemming
      Initially Word = (Prefix)-stem-(Suffix)
B-    Search in Stop word list
      IF Word exists in Stop word list
          Return to step A (for next word)
      ELSE go to step C
C-    Search in Global Prefix Exceptional List    (GPEL)
      IF Word exists in GPEL
          THEN go to step E
      ELSE go to step D
D-    OPEN Urdu prefixes file
      READ prefix one by one from the file until EOF is
      reached
      IF there is a match
          THEN remove prefix from word
          So Word = stem-(Suffix)

E-    Search in  Global Suffix Exceptional List
      (GSEL)
      IF Word exist in GSEL
          THEN output Word = stem
          go to step H
      ELSE go to step F

F-    OPEN Urdu suffixes file
      READ suffix one by one from the file Until EOF
      reached
      IF there is a match
          THEN remove suffix from word
          So Word = stem
          go to step G
      ELSE go to step H
G-    Search in Characters-Add List (CAL) file
      IF Word exists in CAL file
          THEN Add the respective
              character to the word
          So Word = stem (normalized)
      ELSE
          Word=stem
H-    End
```

The proposed stemmer extracts the stem by removing the prefixes and suffixes having maximum length from a word. Thus for this purpose, the prefix and suffix lists are sorted in descending order before any prior operation.

To understand the proposed algorithm, please refer to Figure 1. Our proposed algorithm first checks the entered word that whether it is a stop word? If so then no processing will be done on that word and next word will be entered. When a word is not a stop word then the word is checked in global prefix exceptional list, if it exists then it means that the word has prefix(s) but should not be removed from the word because they are the part of stem. On the other hand if it does not exist in global prefix exceptional list then it means it has some prefix (s), thus prefix rules are applied to remove the maximum prefix from the word.

The word is now checked in global suffix exceptional list, if it exists then it is marked as the stem. But if it does not exist in the list then it means this word has some suffix (s). Therefore suffix rules are applied to remove the maximum suffix from the word.

To normalize the word form, the word is checked in five different lists maintained for **(**ا،ہ،ت،ی،ن**)** characters i.e. Characters-Add list. When a word is found in any of the five lists then respective character is added to produce a normal word form. Thus marks this resultant word as the stem.

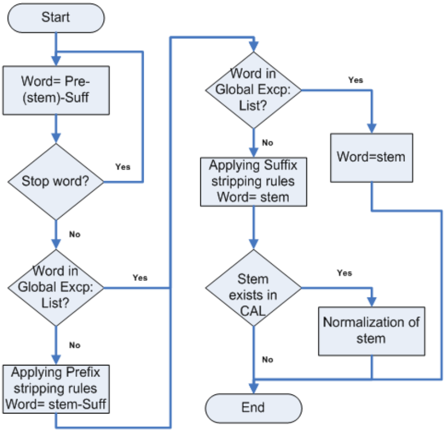
Start
Word= Pre-(stem)-Suff
Stop word?
Yes
No
Word in Global Excp: List?
Yes
No
Applying Prefix stripping rules Word= stem-Suff
Word in Global Excp: List?
Yes
No
Word=stem
Applying Suffix stripping rules Word= stem
Stem exists in CAL
Yes
No
Normalization of stem
End

Figure 1: Flow chart of Proposed Urdu Stemmer

7 Evaluation

We evaluated the proposed Urdu stemmer on three corpora[4] i.e. corpus-1 (9200 words), corpus-2 (27000 words) and corpus-3 (30000 words). These corpora include data in the form of verbs, nouns, adjectives, punctuations, numbers, special symbols etc. When the corpus-1 is fed to the stemmer, then in pre-processing step, the stemmer removed all stop words, numbers and punctuation marks. Thus after pre-processing steps, there were 4268 words left. Stemming is performed on 4268 words that produce 39.4% precision, 71.1% recall and 50.70% F1-Measure.

The low accuracy obtained is due to the occurrence of various issues in stemming Urdu words e.g. Compounding, Tokenization, Transliteration and Infixation. It is very difficult to classify the compound words as a single or multiple words e.g. مرہم پٹی (bandaging), خط وکتابت (correspondence). Some times the reduplication also produces ambiguity; whether it is treated as single or double word e.g. ساتھ ساتھ،آہستہ آہستہ،جگہ جگہ(together, slowly, at every place)

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. Tokenization process should be error free, hence producing correct tokens before applying an Urdu stemmer. It

[4] http://www.crulp.org/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm

is also observed that these corpora include Urdu transliteration of English words e.g. فیملیز (families), پروگرامنگ (programming). There is no prefix and suffix morpheme available in our developed lists for such type of words. 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. Thus light weight stemmer cannot handle words having infixes. Due to insufficient words in Character Add lists, it leads to error in stemming. Proper nouns and abbreviations also contribute in the error of stemming e.g. یو-ایس-اے (USA), لندن (London).

The same stemmer is applied on corpus-2 that produced 18378 words after pre-processing. It gives 49% precision, 78.6% recall and 60.36% F1-Measure. When corpuse-3 is fed to the stemmer it produces 19351 words after pre-processing. Our light weight stemmer produced precision 73.55%, recall 90.53% and 81.16% F1-Measure. The summary of the three corpora and evaluation of the stemmer after applying stemmer is given in table 1.

Characteristics	Corpus (1)	Corpus (2)	Corpus (3)
Words after Pre-processing	4268	18378	19351
Already stemmed Words	2233	10674	12428
Words to be Stemmed	2035	7704	6923
Correct Stemmed Words	802	3775	5092
Results			
Precision	39.4%	49%	73.55%
Recall	71.1%	78.6%	90.53%
F1-Measure	50.70%	60.36%	81.16%

Table-1: Description of corpora after pre-processing and evaluation of stemmer

Conclusion

Morphologically Urdu is a complex language. There exist a number of variants in this language for a single word. Urdu language is rich in both inflectional and derivational morphologies.

In this paper, a light weight stemmer for Urdu text is proposed. The proposed stemmer handles inflectional morphology. This stemmer removes prefixes and suffixes from a word to get stem word but before removing the affixes, the stemmer checks the exceptional cases. The stemmer gives 73.55% precision, 90.53% recall and 81.16% F1-Measure and it was also compared with *Assas-band*, the only available Urdu stemmer.

Generally stemmer increases recall at the cost of decreased precision. Our study proves that maximum matching affix approach is more suitable for developing the stemmer for Urdu language. Other regional languages of Pakistan (Punjabi, Pashtu, Sindi, and Kashmiri etc) are similar to Urdu in morphology. It would be interesting to observe whether similar techniques can be used to develop stemmers for these languages.

References

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

E.T. Al-Shammari, Jessica Lin, (2008). Towards an error-free Arabic stemming. *17th Conference on Information and Knowledge Management, iNEWS'08*, pages 1–6,Napa Valley, California, USA.

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

Imed Al-Sughaiyer, Ibrahim Al-Kharashi. (2004). Arabic morphological analysis techniques: a comprehensive survey. *Journal of the American Society for Information Science and Technology,* 55(3):189 – 213.

J.B. Lovins, (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1 and 2):22–31.

Malik, M. G. Abbas,B.C. Bhattcharyya, P. (2008). Hindi Urdu machine transliteration using finite-state transducers. *proceedings of COLING* 2008, pages 537–544, Manchester, UK.

M.F. Porter. (1980). An algorithm for suffix stripping. Program, 14(3): 130–137.

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

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

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

Rizvi, S. & Hussain, M. (2005), "Analysis, Design and implementation of Urdu morphological analyzer. *Engineering Sciences and Technology, SCONEST*, pages 1-7.

Sabzwari, S. (2002). *Urdu Quwaid*. Sang-e-Meel Publication.

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

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