PASHTO LANGUAGE STEMMING ALGORITHM

SEBGHATULLAH ASLAMZAI
SAIDAH SAAD

ABSTRACT

This paper presents a stemming algorithm for morphological analysis for less popular or minor language like Pashto language. There is lack of resources and tools that can be applied in different applications such as in document indexing, clustering, language processing, text analysis, database search systems, information retrieval, and linguistic applications. The review of literature shows that only a few morphological studies have been conducted in the Pashto language, and research which focused on automatic stemming has not yet been fully analysed. In addition, no stemming algorithm has been proposed for extracting Pashto root words from the Pashto corpus, which is applicable for the above mentioned functions. Therefore, the objective of the current thesis is to develop a rule-based stemming algorithm for the Pashto language. The Pashto corpus is directly used as the input and the stemming algorithm uses both inflectional and derivational morphemes. The output is in the form of meaningful root word without affixes. Furthermore, the accuracy and strength of the proposed algorithm is evaluated using word count method. To validate the function of the developed algorithm, two native speakers of Pashto were recruited to evaluate the algorithm in terms of its accuracy and strength. The result of the study shows that the proposed algorithm has the accuracy of 87%. This study can have a great contribution to Pashto language in terms of extracting the root words useful for different purposes including data indexing, information retrieval, linguistic application, etc. This research also lays the ground for further studies on Pashto language analysis.

Keywords: Morphology, Conflation, Stemming algorithm, Root word, Pashto rules.

INTRODUCTION

Resource development normally focuses primarily on well-described and popular languages with a large number of speakers including English. Minor languages may also benefit from language resources that can be employed in applications such as in electronic dictionaries, topic detection of materials, document indexing, clustering, language processing, text analysis, database search systems, information retrieval, linguistic applications, computer assisted language learning, and so forth. However, the development of such resources has rarely been conducted, and are challenging. Often, the developed resources and tools for major and popular languages cannot be applied to minor or less popular ones. Hence, Pashto, as a less popular language, lacks the resources and tools that can be applied in different applications such as in document indexing, clustering, language processing, text analysis, database search systems, information retrieval and linguistic applications.

The review of literature shows that only a few morphological studies have been conducted in Pashto language, and research which focused on automatic stemming has not yet been fully analysed. In addition, no stemming algorithm has been proposed for extracting Pashto root words from the Pashto corpus, which is applicable for the above mentioned functions. Therefore, the objective of the current thesis is to develop a rule-based stemming algorithm for the Pashto language. The Pashto corpus is directly used as the input and the stemming algorithm uses both inflectional and derivational morphemes. The output is in the

form of meaningful root word without affixes. Furthermore, the accuracy and strength of the proposed algorithm is evaluated using word count method.

Morphological analysis technique is considered as computational processes analysing natural words via taking into account their internal morphological structures. On the other hand, stemming algorithms are regarded as processes that congregate all words with the same stem which have some semantic relations. Indeed, both processes are thought to be very constructive and valuable in different natural language applications including information retrieval, text compression, data encryption, text classification and categorization, automatic translation, and vowelization and spelling. The current paper focuses on stemming in Pashto language.

Stemming is the process that determines the stem of a certain word. To put it in another way, the purpose of a stemming algorithm is reducing variant word forms to a common and regular morphological root, known as "stem" (Bacchin et al., 2002). Hence, the term 'stemming' is widely employed by researchers and the main goal of the stemming algorithms and morphological analysis methods is removing all possible affixes, and as a result, reducing the word to its root word or stem (Frakes & Yates, 1992; Tagva et al., 2005).

There are different types of stemming algorithms such as affix removal, lookup table and statistical methods (Bento et al., 2005; Tordai, 2006). The most common stemming algorithms are affix removal algorithms. Affix removal stemming algorithms usually remove affixes (i.e., suffixes or prefixes) from words resulting in a root word known as a stem which is often approximately a word's root morpheme. Stemming algorithms are widely used in many kinds of language processing as well as text analysis systems. Furthermore, they are widely employed in database search systems, information retrieval, and linguistic applications.

A high-quality stemmer will convert all variations of words to the accurate root word. The stemming algorithms are mostly rule-based and employ the logical approach to remove or sometimes replace inflectional and derivational suffixes. Stemming algorithms stem the input word in order that all the alternative forms of a word can be conflated to the root word. Several rule based algorithms only transform the variant forms of word to a stem instead of root word which is useful in reducing the data size and complexity in the document collections and consequently, is helpful to improve the function of information retrieval. Nevertheless, many text mining applications, like topic detection, document indexing, and clustering, etc. require much more accuracy in index terms instead of index compression (Soori et al., 2013). Hence, in this paper, a Pashto algorithm is developed that is able to use the Pashto script as input data and extract the root words which are valuable for the above mentioned purposes.

RELATED WORK

Pashto language is one of the East Iranian classes of languages widely spoken in Afghanistan and Pakistan. Pashto is also spoken in India, Iran, Tajikistan, United Arab Emirates and the U.K. (Method, 2010; Zyar, 2003). Pashto has two major dialects such as Eastern Pashto and Western Pashto. Eastern Pashto is spoken in northeastern Pakistan, while Western Pashto is spoken in Afghanistan capital city, Kabul. Pashto also has two minor dialects such as Southern Pashto and Central Pashto. Southern Pashto is spoken in Baluchistan and in Kandahar, Afghanistan, whereas Central Pashto is spoken in northern Pakistan (Wazirstan). Pashto is a variation of the Arabic Script and is written using the Arabic Script (Zuhra & Nauman, 2005). Pashto has seven vowel sounds and its alphabet has a similar set of consonants like English. It has a series of retroflex consonants sounds (*t*, *d*, *r*, *n*.) which are made by curling the tongue backwards. This language has also complex grammatical rules (Zyar, 2003). There are some studies have been conducted on Pashto (Penzl, 1955; Khattak, 1988; Tegey & Robson, 1996; Babrakzai, 1999; Zuhra and Khan 2009; Bing, et al., 2010). The work of these linguists form the basis for the research work presented in this paper.

Khattak (1988) identifies different facets, for which a Pashto verb inflects and asserts that ,
> *"The formal distinctions of the Pashto verb reflect a variety of categories: tense, aspect, mood and voice. Referring to the Khattak (1988) in the subject or object position, the verb also inflects for person, number and gender."*

Khattak (1988) further says that the morphology of the Pashto verb shows only two simple tenses: present and past. The future is expressed with the help of a model clitic*ba*. He provides the basic structure of a Pashto verb, given below, where # indicates the potential positions for clitics. Verb= [aspect # negative # stem + agreement #]
Babrakzai (1999) provides the definition of agreement as follows:
> *"System of inflection that records a nominal's inherent features (usually person, number, gender/ or case) on another category, generally a verb, adjective or a determiner".*

According to Tegey and Robson (1996), agreement is indicated with personal endings, i.e. suffixes following the verb stem which show person and number. For the category of gender is restricted to the third person form of simple verbs and to the third person singular forms of the auxiliary (Penzl, 1955) called copula verbs of 'to be' (Zuhra and Nauman, 2005). However, the category of gender is found in third person plural form of this auxiliary in Yousafzai dialect (Khan and Zuhra, 2007).

A Pashto noun inflects for gender, number and case (Penzel, 1955). Different Pashto grammarians (Reshteen, 1994; Zyar, 2003] categorize the Pashto nouns into different masculine and feminine classes according to their final phonemes. Bellew et al. (1986) have also contributed significantly to the investigation about Pashto nouns. The Pashto adjectives have more or less the same inflectional properties and similar morphological behavior as those of Pashto nouns.

Correspondingly, Zuhra and Khan (2009) develop an inflectional morphological analyzer that can analyze different inflections of a Pashto verb, noun or adjective. The system is corpus-based. The developed system is capable of accepting input in the form of a transliterated Pashto verbal, nominal or adjectival inflection. Then, convert it to an Arabic-scripted Pashto equivalent. Finally, morphologically analyze the word and search and display all the sentences in the corpus, in which the word is used.

Bing et al. (2010) present a novel method to improve word alignment quality and eventually the translation performance by producing and combining complementary word alignments for low-resource languages. Instead of focusing on the improvement of a single set of word alignments, they generate multiple sets of diversified alignments based on different motivations, such as linguistic knowledge, morphology and heuristics. They also demonstrate this approach on an English-to-Pashto translation task by combining the alignments obtained from syntactic reordering, stemming, and partial words. The combined alignment outperforms the baseline alignment, with significantly higher F-scores and better translation performance.

However, most of the work nowadays, on deriving an automatic stemming construction has been conducted only in English language and Arabic language, leaving a vast research field for studies in other languages, including Pashto, especially as the literature demonstrates, and with the researcher's best knowledge, no research has been conducted on stemming in Pashto to extract the roots/stems. Therefore, this research study is undertaken to address the problem and fill the gap. Hence, the purpose of the current study is to extract the root words from Pashto language.

27

METHOD

Information retrieval algorithm is still lacking for Pashto in comparison with other languages such as English. Despite some researches on Pashto language in terms of information retrieval system and algorithm development, to date, no efficient stemming algorithm has been developed for Pashto language and yet to be built. This may be attributed to lack of Pashto development systems of the complicatedness of Pashto morphology. In addition, there is no stemming algorithm for extracting the Pashto roots word, even though there are plenty of stemming algorithms for other languages. One of the challenges of developing an efficient Pashto stemming algorithm is the impossibility of arriving at a comprehensive set of rules of Pashto word formation and affixation. Another issue can be associated with the demand for understanding complete Pashto language system and its analysis as well as its huge morphological rules to develop an efficient and powerful algorithm.

Therefore, the idea of developing a simple, efficient and powerful stemming algorithm with high proficiency and accuracy, for Pashto language is reinforced. This approach will take into account Pashto morphological rules and structures and also it will take advantage of new approach or a combination of stemming algorithms to obtain better results. Thus, the current research study develops a rule based stemming algorithm for Pashto language to remove the affixes and extract root words. In the following section, the rules of Pashto language are explained.

For the research method, the step can be decomposed into four main phases which are outlined in Figure 1
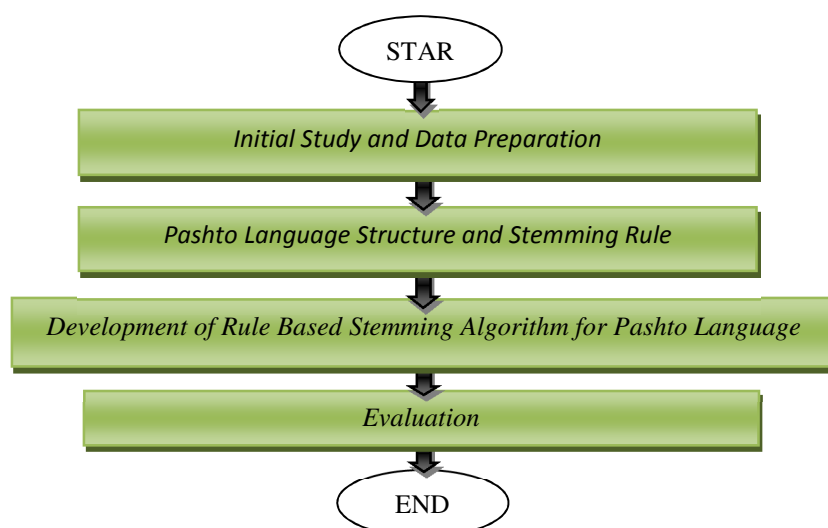

FIGURE 1. Research Methodology of Pashto Language Stemming

In phase one, we have searched the literature thoroughly and reviewed the related literature in order to understand stemming and morphology process as well as the approaches to morphology analysis (analytic & synthetic) and Pashto Language structure. In phase two, structure and stemming rules in Pashto language are identified and presented. In phase three, the development of rule based stemming algorithm for Pashto Language is described. Then, the corpus used for the current study is explained. Finally, the proposed algorithm is evaluated in terms of accuracy and speed by two native speakers of Pashto. To do so, many exemplifications of stemming algorithm using Pashto corpus are analyzed and discussed. Accordingly, the implication of the work is provided.

28

PASHTO LANGUAGE STEMMING ALGORITHM

In this study, a rule based algorithm is developed for Pashto language using the Pashto language rules. To this end, an algorithm is developed which is able to remove affixes from Pashto words and reduce them to root words. Then, the proposed algorithm is evaluated through two tests using common corpus of Pashto extracted from the internet. In addition, two native speakers of Pashto are recruited to evaluate the algorithm in terms of accuracy and speed. Below some of the main Pashto affixation rules are presented.

Rule: 1- Pashto affixation rule: words which are composed of five or more characters and end with suffixes: "ستان"،"تون"، "ونو"،"ان"، "و"،"ور"، "ناک"،"جن"، "وان"، "وال"، "تابه"Just remove the suffixes .For example:

Word with suffix	Suffix	Root word	English meaning
پوهنتون	تون	پوهن	University
موتروان	وان	موتر	Driver

Rule: 2- Words which are comprised of 4 or more characters and end with "ي" Remove the suffix and add"ه".For example:

Word with suffix	Suffix	Root word	English meaning
پيغلي	ي	پيغله	girls
شايستي	ي	شايسته	beautifully

Rule: 3- If the word has more than 5 characters and ends with "من"just remove the suffix " من ", For example:

Word with suffix	Suffix	Root word	English meaning
دردمن	من	درد	Sick, patient
واكمن	من	واک	Owner

Rule: 4- If the word is composed of 5 or more characters and ends with "يز" remove the suffix "يز" and add "ه" to the end. For example:

Word with suffix	Suffix	Root word	English meaning
تولنيز	يز	تولنه	Humanity
دوديز	يز	دود	Traditional

Rule: 5- Words that have more than 3 characters and start with the following prefixes (نه ، نا ، م ، بد ، نيک ، تش ، تل ، تور ، تک ، ډير) just remove the prefixes. For example:

Word with suffix	prefix	Root word	English meaning
ناپوه	نا	پوه	Uneducated
تل پاتی	تل	پاتی	Permanent

Rule: 6-   Words which have 5 or more characters and start with "لا" just remove the "لا". For example:

Word with suffix	prefix	Root word	English meaning
لازيات	لا	زيات	Mach more
لاښه	لا	ښه	Very good

Rule: 7- Words which have 5 or more than five characters and start with "وران" Just Remove "وران" . For example:

Word with suffix	prefix	Root word	English meaning
وران خولی	وران	خولی	Joker
وران کاره	وران	کاره	Bad worker

Rule: 8- When the same word is repeated as suffixes, it is removed as a root word. For example:

Word with suffix	prefix	Root word	English meaning
خوږی خوږی	خوږی	خوږی	sweetly
ترخی ترخی	ترخی	ترخی	smut

Rule 9-   Words that have more than 3 characters and start with the following prefixes (نه ، نا ، نیک ،) just remove the prefixes. For example:

Word with suffix	prefix	Root word	English meaning
ناپوه	نا	پوه	uneducated
ناروغ	نا	روغ	Sick

DEVELOPMENT OF CORPUS

In this study, a corpus of 30000 common Pashto words is used. The corpus is extracted from different sources such as a book by Wafa (1995) and three popular and mostly visited web sites such as (BBC Pashto news, Tolafghan and Benawa) which have more users throughout the world. In addition, the text is extracted from up to date and common sources understandable for highly educated people as well as less educated ones, in fact a balanced corpus is used (Dandapat et al., 2004). Balanced corpus is needed to process natural language processing tasks like stemming. Balanced corpus is a corpus that represents the words that are used in a language. As indicated in (Dandapat et al., 2004), texts collected from a unique source, for example, from scientific magazines, will probably be biased toward some specific words that do not appear in everyday life. Such types of corpuses are not balanced corpus so that they are not appropriate for many natural languages processing in general and stemming in particular except for special purposes. However, developing a balanced corpus is one of the difficult tasks in natural language processing research because it requires collecting data from a wide range of sources: fiction, newspapers, technical, and popular literatures (Tordai, 2006). As a result it requires much time and human effort. For this particular study, corpus was collected from different popular Pashto online newspapers and bulletins to balance the corpus.

Online newspapers, bulletins and public magazines are considered as consisting different issues of the community: social, economical, technological and political issues. The corpus data was selected from the (www.bbc.co.uk/pashto/afghanistan) and

(www.tolafghanistan.com) web sites with a lot of users all over the world. So, they are a potential source for collecting balanced corpus for natural language processing tasks. This corpus is used for evaluating the performance of the stemmer (Tesfaye & Abebe, 2010; Pingali et al., 2007).

THE ARCHITECTURE OF THE PASHTO STEMMING ALGORITHM

In this section, the developed stemming algorithm of Pashto language is explained in details (Figure 2). As the architecture of the algorithm shows, the process of the algorithm which is tokenization, normalization , affixes (prefix and suffix) removing, searching for prefixes and suffixes, remove all prefixes and suffixes, renormalization and listing out root words comprise the main stages of this algorithm.
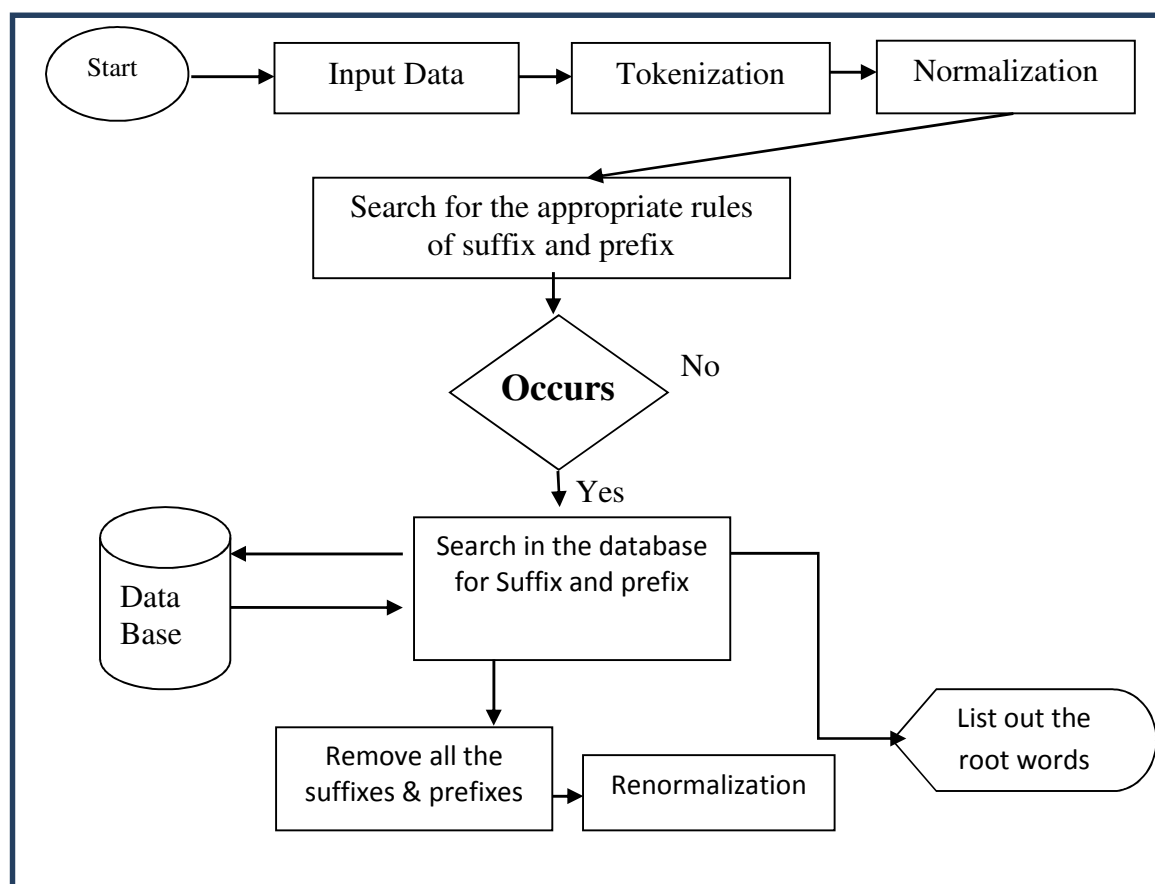
FIGURE 2. The process of Pashto Language Stemming Algorithm

As the figure 2 illustrates, after the data was entered, the functions of Tokenization and normalization are performed. In the process of tokenization, the words are chopped up into pieces, known as tokens, perhaps at the same time throwing away certain characters such as punctuation. For example: When the input is: "friends, romans, countrymen, lend me your ears", the output will be: friends/romans/countrymen/lend/me/your/ears. Hence, in the tokenization process, the input text is tokenized word by word by using delimiter as space. In the normalization process, the words are rearranged and prepared for the next process. Some words that do not need to go through the process of stemming, directly go to output.

31

Next, the words with affixes go through the process of stemming. In this process, the affixes of words are compared with the affixes in the database and as result the words with affixes are reduced to root words. Some of the algorithm used, are shown in figure 3 and 4 for suffix and prefix removal. Finally, the stemmed words are renormalized and listed out as root words.
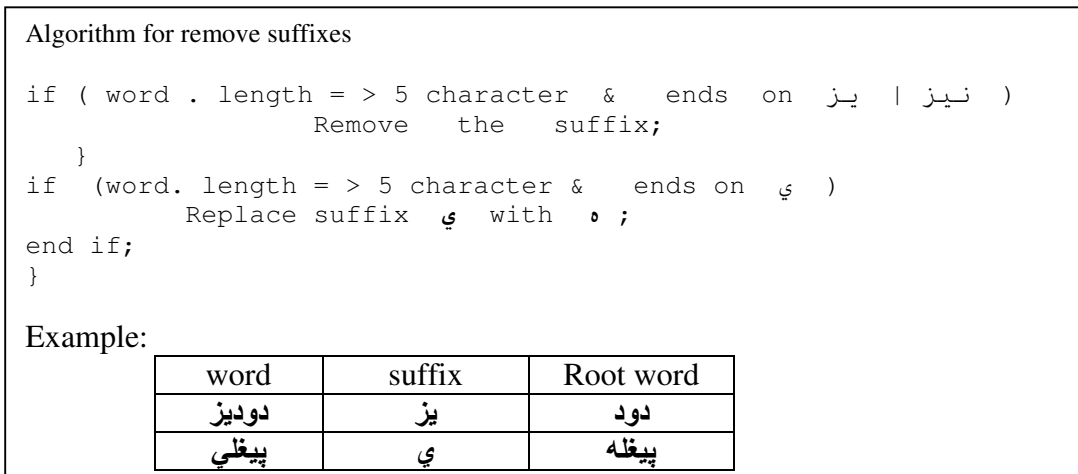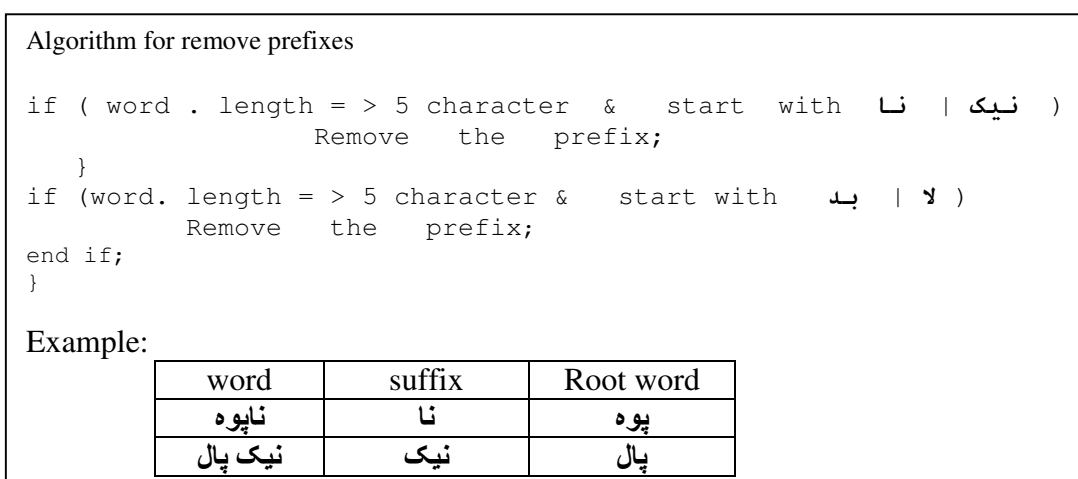
```
Algorithm for remove suffixes

if ( word . length = > 5 character  &   ends  on  نـيز | يـز  )
                Remove   the   suffix;
   }
if  (word. length = > 5 character &   ends on  ي  )
        Replace suffix  ي  with  ه ;
end if;
}
```

Example:

word	suffix	Root word
دوديز	يز	دود
پيغلي	ي	پيغله

FIGURE 3. Algorithm for Suffixes Removing

```
Algorithm for remove prefixes

if ( word . length = > 5 character  &   start  with  نا | نيک  )
                Remove   the   prefix;
   }
if (word. length = > 5 character &    start with   بـد | لا  )
        Remove   the   prefix;
end if;
}
```

Example:

word	suffix	Root word
ناپوه	نا	پوه
نيک پال	نيک	پال

FIGURE 4. Algorithm for Prefixes Removing

EVALUATION AND RESULT

According to Sawalha and Atwell (2008), "many computational linguists have designed and developed algorithms to solve the problem of morphology and stemming. They add that each researcher proposed his own gold standard, testing methodology and accuracy measurements to test and compute the accuracy of his algorithm. Therefore, in this work, to evaluate the function of the developed algorithm, two tests were conducted using the Pashto corpus as input data. The corpus data is obtained from two sources namely, web site (www.bbc.co.uk/pashto/afghanistan) and a book by Wafa (1995), which have common Pashto words and texts as mentioned before. In addition, two Pashto native speakers were recruited to check the performance of the proposed algorithm.

32

TEST ONE OF THE PASHTO STEMMER

In this test, the corpus data was selected from the (www.bbc.co.uk/pashto/afghanistan) web site which has a lot of users all over the world. The corpus consists of over 30000 words which were stemmed successfully.

FIGURE 5. The original data of the Pashto language

As it was mentioned, the selected corpus is balance in terms of formality, that is, both formal and informal languages data have been used. Figure 5 shows the original data of Pashto language and Figure 6 shows the output of data after the process of stemming being done.

FIGURE 6. The output of the data after stemming

Hence, as the data shows the original corpus was successfully stemmed to root words. For example, the term افغانستان was stemmed to افغان. The term خپلواک was reduced to خپل. Comparing the stemmed corpus to the original one, it is revealed that almost the entire original corpus has successfully been conflated to root words.

The result of the stemming of the Pashto corpus shows that the developed algorithm has reduced all of the words with affixes (over 3000 words) to their root words accurately and successfully. This indicates that the accuracy of the algorithm is high. As the Table 1 shows,

33

many Pashto words were conflated to root words successfully and accurately. The table shows the inflected words, root words and the English meaning.

TABLE 1. Example of Stemmed Words

No	Inflected Word	Root Word	English Meaning
1	افغانستان	افغان	Afghanistan
2	بشری	بشر	Humanity
3	حقونو	حق	Rights
4	خپلواک	خپل	Private
5	وایی	وای	Says
6	ولسمشریزو	ولسمشر	Presidential
7	نوماندانو	نوماند	Candidates
8	پراختیا	پراخ	Development
9	پلانونو	پلانو	Plans
10	ورکری	ورکر	To give
11	جنگئنو	جنگ	Fights
12	مسیولیتونه	مسیولیت	Responsibilities
13	پلانونه	پلان	Plans
14	نوماندانو	نوماند	Candidates
15	انتقالی	انتقال	Transfer
16	جنجالی	جنجال	Controversial
17	عملی	عمل	Practically
18	تشفکری	فکری	Thinker

Table 2 demonstrates the result of the algorithm in two tests. I can be seen that in each test 300 words were used to be stemmed to their root words. In test one, 279 words from 300 words were correctly conflated, which is 92.66%. In test two, 275 words from 300 words were stemmed to their roots successfully that is 91.66%. This stemmer is run on the test set of 3000 words which is assumed to be balanced. The literature from which the rule of the stemmer developed is totally different from the test set. This was done deliberately in order to predict the performance of the stemmer in the real world data.

In test one, the output from the stemmer indicates, out of 300 words 17 words were under stemmed and 19 words were over stemmed. In test two, he output from the stemmer indicates, out of 300 words 14 words were under stemmed and 18 words were over stemmed. Totally, this stemmer generates 68 stemming error in two tests. As a result, the accuracy of the stemmer becomes 88.66%. The result of the study is consistent with the findings of a study by Estahbanati et al. (2011) who developed an algorithm for Persian language whose accuracy was around 90 %.

TABLE 2. the Result of the Algorithm

Test No.	Total Words	Correct Results	Incorrect Results		Percentage of Correct Results	Average
			under-stemming	over-stemming		
1	300	264	17	19	88%	88.66%
2	300	268	14	18	89.33%	

TEST TWO OF PASHTO STEMMER

In the second test, a text of over 3000 words was extracted from a book by Wafa (1995) for the purpose of stemming. As discussed before, the selected corpus is also balanced; this is a famous historical book about the history of Afghanistan which has many Pashton readers. Figure 7 shows the example of Pashto language that taken from Wafa's book.

As the data shows, the words in the original text were successfully stemmed to root words. For example, the term افغانستان was stemmed to افغان. A comparison of the original and the stemmed corpus demonstrates that almost all words were stemmed to root words accurately and successfully (refer figure 8).
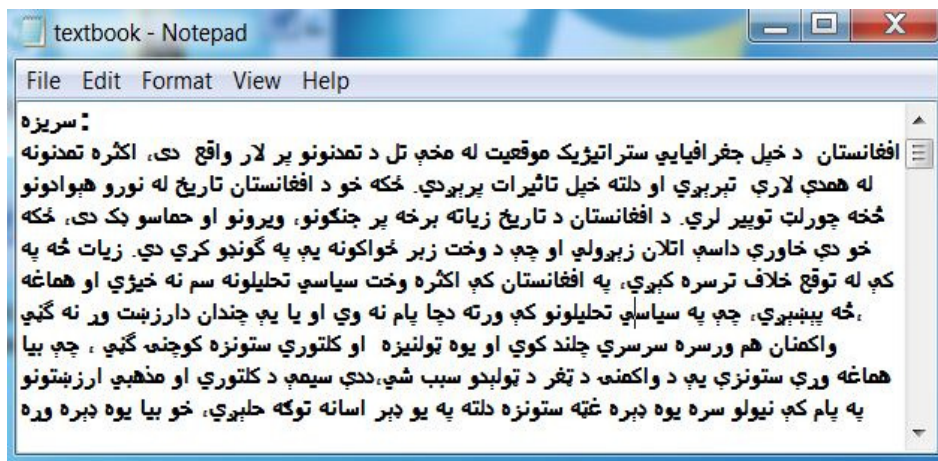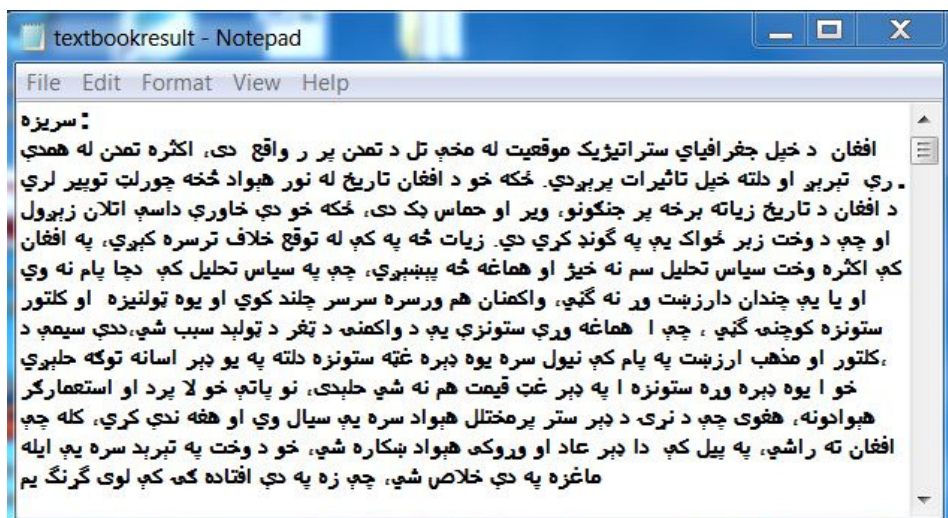

FIGURE 7. The original Pashto corpus.

FIGURE 8. The result after stemming

35

In addition to performing two tests, using Pashto corpus, to evaluate the function of the proposed algorithm, two academic native speakers of Pashto were recruited to evaluate it in terms of the accuracy and strength. Some texts were used as a corpus for the current study, were given to two native speakers of Pashto to conduct the stemming manually. The result of their stemming was compared with the result of the data obtained from the proposed algorithm using the same text. They stated that the result of the stemming algorithm is 87.66 percent correct. This indicates that the algorithm has the accuracy of 87 percent.

TABLE 3. A comparison of stemming result based on native speakers' evaluations

No. of words	Native Speaker One Evaluation	Percentage	Native Speaker Two Evaluation	Percentage	Stemming Algorithm correct result (researcher result evaluation)	Percentage	Final Result
600	522	87%	524	87.33%	532	88.66%	87.66%

Table 3 shows the comparison of native speakers' evaluation and algorithm stemming result used by the researcher. It is seen that the result of evaluation by native speaker one and native speaker two are 87 % and 87.33% respectively. This means that based on their evaluation the stemming algorithm has the accuracy of around 87 percent. It is also observed that according to researcher's evaluation, the algorithm's accuracy is 88.66 percent. Hence, on average the accuracy of the algorithm is 87.66 percent.

FINDINGS AND CONCLUSION

In this paper, error counting approach is adapted to evaluate the algorithm in terms of the number of accurately conflated results (Estahbanati et al., 2011). Accordingly, the numbers of correctly conflated words as well as incorrectly conflated ones are counted for analysis. The output from the stemmer was then checked against the respective expected valid stem. These errors were then described in terms of under stemming and over stemming. Over stemming occurs when too much of the term is removed, while under stemming occurs when too little of the term is removed. Evaluation of the effectiveness of stemming algorithms is necessary to reveal specific error patterns. This information can subsequently be used to improve the algorithm where possible. Some error types, however, are inherent to the affix removal method and without the additional information provided by, for instance, a dictionary, these errors cannot be avoided.

As discussed, in this paper, a rule based algorithm was developed for Pashto language. Based on the two tests performed for evaluation of the proposed algorithm, and the evaluation conducted by native speakers of Pashto, it was shown that the algorithm can perform the function of stemming with high accuracy. According to the evaluation by native speakers of Pashto, the accuracy of the developed algorithm is over 87 percent, which indicates great difference compared with that proposed by Zuhra and Khan (2009) in which their algorithm accuracy was 71 percent. In addition, this study has addressed some of the other issues of the past study. For example, here, Pashto words are directly used as input and the stemming algorithm deals with both inflectional and derivational morphemes. Moreover, the output is meaningful root word without affixes. Hence, the proposed algorithm in this paper shows great superiority in terms of the accuracy and strength compared to that developed by Zuhra and Khan (2009).

REFERENCES

Bento, C., Cardoso, A. & Dias G. 2005. Progress in Artificial Intelligence. 12th Portuguese Conference on Artificial Intelligence, (EPIA 2005). *Lecture Notes in Artificial Intelligence*, vol. 3808: 693-701.

Dandapat, S., Sarkar,,S., & Basu, A. 2004. A Hybrid Model for Part-f-Speech Tagging and Its Application to Bengali. *Journal of World Information Society*, 43(6):384-390.

Estahbanati, Reza Javidan, & Mehdi Nikkhah .2011. A new multi-phase algorithm for stemming in Farsi language based on morphology. *International Journal of Computer Theory and Engineering,* 3(5): 623-627.

Khan, M.A. & Zuhra. F.T. 2007. Morphological Analyzer for the Past Tense Verbs in Pashto. Paper presented at the Conference on Language and Technology, August 7-11, Bara Gali Summer Campus, University of Peshawar, Pakistan.

Method, P. 2010. Where in the world do they speak Pashto? http://www.pimsleurmethod.com/blog/2010/11/01/where-in-the-world-do-they-speakpashto [March 17, 2011].

Sawalha, M. & Atwell, E. 2008. Comparative evaluation of Arabic language morphological analysers and stemmers. Paper presented at the *COLING 2008*, 22nd International Conference on Computational Linguistics, 18-22 August 2008, Manchester, UK.

Soori, H., Platos, J., & Snásel, V. 2013. A Linguistic Method into Stemming of Arabic for Data Compression. *DATESO 2013*: 119-128

Tesfaye, D. & Abebe, E. 2010. Designing a Rule Based Stemmer for Afaan Oromo Text. *International Journal of Computational Linguistics (IJCL),* 1(2):1-11.

Tordai, A. 2006. Stem, Stemming, Stemmer on the Benets of Stemming in Hungarian, Master thesis, Faculty of Science, University of Amsterdam.

Pingali, P., Varma, V., & Tune, K.K. 2007. Evaluation of Oromo-English Cross-Language Information Retrieval. In: *IJCAI 2007 Workshop on CLIA*, Hyderabad, India.

Zuhra, F.T. & Nauman. H. 2005. *The computational morphology of Pashto*, Master thesis, Department of Computer Science, University of Peshawar, Peshawar.

Zuhra,,F.T. & Khan, M.A. 2009. A Corpus-Based Finite State Morphological Analyzer for Pashto. Paper presented at the *Conference on Language and Technology 2009 (CLT09),* January 22-24, National University of Computer and Emerging Sciences, Lahore Campus, Pakistan.

Zyar, M.A. 2003.*Pashto Grammar*. Peshawar: Danish Publishing Association.

Sebghatullah Aslamzai,
Information Technology Department,
Faculty of ICT, Kabul University
Kabul, Afghanistan
sebghat_aslamzai@yahoo.com

Saidah Saad
Faculty of Technology & Information Science
National University of Malaysia
43600, Bangi, Selangor
Malaysia
saidah@ukm.edu.my