Comparative Analysis of Urdu Based Stemming Techniques

Muhammad Hassaan Rafiq¹, Shiza Gul Niazi², Subaika Ali³
hassaan.rafiq@lgu.edu.pk¹, shizaniazi013@gmail.com², subaika.ali@gmail.com³

Abstract

Stemming reduces many variant forms of a word into its base, stem or root, which is necessary for many different language processing application including Urdu. Urdu is a morphologically rich and resourceful language. Multilingual Urdu words are very challenging to process due to complexity of morphology. The Research of Urdu stemming has an age of a decade. The present work introduces a research on Urdu stemmers with better performance as compare to the existing Urdu stemmer.

1. INTRODUCTION

A stemmer provides base, stem or root form from different modified forms of a words, for instance word “mixed”, and “mixing” will be reduce to stem i.e “mix” in English stemming. Stemmers are important for a number of application which requires the use of base form instead of derived form of words. Word Sense Disambiguation, Information Retrieval systems, word count studies, text mining, automatic discretization are few of the applications that make the use of stemming. Stemming of morphologically rich languages (MRLs) provides a greater impact as compared to less inflectional languages. For example an Urdu word has more than 60 inflectional forms. Urdu is spoken by approximately 200 million people and if we include Hindi speakers who speak Urdu in their Hindi words in their vocabulary then it will become second largest language with 588m speakers in the World. This sort of Urdu language will be known by linguists as Hindi-Urdu. The diversity and Morphological richness of Urdu language introduce a number of challenges in stemming. Vocabulary of Urdu contains many words taken from different languages that result into unavoidable complexity because these borrowed words have a tendency to be handle according to the grammatical principle of its source language. Some of the most prominent source language of Urdu vocabulary are Persian, Turkish, Sanskrit, Arabic, Hindi, and English. Here are some of

few examples explaining this characteristic of Urdu vocabulary. The words selected for these examples are normally used in daily routine Urdu language.

- 1))baag, garden(
- 2) Sanskrit, (umang, aspiration)
- 3) English, (television, television)
- 4))khatoon, woman(
- 5) Arabic, (manzil, destination)
- 6) Hindi, (aag, fire)

These examples taken from “six different languages” based on “six Urdu words”. The majority of Urdu vocabulary words belong to different foreign languages. Owing to these facts it is essential to adopt a customized approach for Urdu stemming in order to improve performance and accuracy. Hence, this approach is selected to handle the challenging nature of Urdu stemming. In Urdu stemming, every Urdu vocabulary word is stemmed manually according to the type of word. Before this selection, a thorough analysis of various Urdu stemming approaches has been performed. The possibility of further enhancement in its performance is also described. Rest of the paper is organized as follows: Section 2 presents the related work, in section 3 design approach of dictionary based stemmer is given and section 4 draws a conclusion.

2. LITERATURE REVIEW

Stemmers can be developed by using a number of approaches such as rule-based, hybrid and statistical. Statistical approaches are usually not fully dependent on the prior linguistic knowledge of the concerned language; they make use of corpus analysis to calculate the occurrences of stems and affixes. On the other hand, rule based and hybrid approaches process the text according to framed rules derived from the grammar of the concerned language. Current state-of-the-art Urdu stemmer Assas-Band, developed by Akram et al., claims 91.20% accuracy which is the lowest among all stemmers as comparison is given in the Table. On the other hand Assas-Band is standing at the first position among all the reported Urdu stemmers since 2009 till 2016. This fact denotes the challenging nature of multilingual Urdu vocabulary. Table 1 provides a comparative view of various proposed solutions deploying statistical, rule-based and hybrid approaches regarding Urdu stemming. Statistical approaches presented in is, in fact, two studies on the basis of one work so, there is only one statistical approach proposed by the author. Authors in table 1 used word pattern matching schemes to handle infixes and they first time addressed the problem of Urdu words having infixes to some extent. However, the average accuracy of their results is just 77.39%. In the most recent work presented in 2016 authors used a rule based approach. Their approach is capable of handling Urdu loanwords and Urdu compound words. The overall accuracy of their stemmer is 88.91%. They declared the “Urdu Light Weight Stemmer” as current state-of-the-art, while it offers the lowest performance regarding accuracy. However, the discussed approaches do not provide remarkable results for Urdu stemming as shown in figure. The prime concern of this issue is the flexibility of Urdu language that introduces numerous exceptions against the framed rules (for rule-based approaches) and derived patterns (for statistical approaches). Following description could be helpful to understand the morphological richness of Urdu language. Urdu verbs have various inflections for habitual, infinitive, past, non-past and imperative forms and these verbs also inflect to show agreement for the case, respect, gender, and number. These twenty inflected forms of a regular verb are

further duplicated for transitive and causative forms and all these inflected forms make a total of more than sixty variant forms. A language with this much morphological variations and with a vocabulary that contains the majority of its words borrowed from more than half dozen languages cannot be handled successfully without a precise and more Urdu specific approach.

Table 1: Analysis of year wise related work.

Year	Author Name	Method Used	Accuracy
2009	Akram et.al [9]	Assas-Band an Affix-Exception-List Based Urdu Stemmer. (Rule-Based)	91.20 %
2013	Vishali Gupta et.al [10]	Rule Based Stemmer in Urdu (Rule-Based)	86.50 %
2013	Mohd. Shahid Hussain [11]	A Language Independent Approach to Develop Urdu Stemmer (Statistical)	82.56%
2014	Ali et. al [12]	A Novel Stemming Approach for the Urdu Language (Rule-Based)	87.61%
2015	Sajjad Khan et.al [13]	Template Based Affix Stemmer for a Morphologically Rich Language (Rules + Word Pattern Matching)	77.39%
2016	Ali et.al [14]	A Rule-based Stemming Method for Multilingual Urdu Text (Rule -Based)	88.91%

3. PROPOSED METHODOLOGY

After the careful analysis of Urdu vocabulary and the research in Urdu stemming etc., the dictionary-based approach is selected as the most appropriate one. It is also called the lexical lookup, table look-up or the brute force approach. In this approach, a dictionary of word-

stem pairs is prepared manually. Manual stemming is performed in a word-stem pair fashion according to the selected vocabulary.

Table 2: Urdu words with their Stems

Word	Stem	Word	Stem
ملتان	ملتان	بامعنی	معنی
ابواب	باب	لا حاصل	حاصل
آرام دہ	آرام	ماہرین	ماہر
نسخہ جات	نسخہ	باعث نجات	نجات

Along with the stemming, a stop word list is also developed. Stop words are non-content words or functional words that are usually ignored in search queries because no one uses them as query words and they can possibly index the whole corpus and hence make the search results entirely useless. Interestingly, there is no predefined criterion for stop words in Urdu. Different studies show different lists of stop words. A stop word in a language can be generally defined as a token that does not provide any linguistic meaning. Aqil Burney et al also mentioned that no proper work has been done on stop words in the Urdu language so they simply translated 421 English stop words in Urdu to use them as Urdu stop words. There are no guiding rules to identify stop words. Different studies proposes stop words lists of different numbers. In a study a list of 150 words declared as stop words list while mentioned a list of 200 stop words in their study. To avoid such vague approach regarding Urdu stop words in my study, I framed out some rules to decide either a word should consider as a stop word or not. The most important rule among all is that a stop word must neither be an inflected word nor a stem word because such type of stop word selection will create contradictions in any stemming approach. A sample list of collected stop words is presented in Table 3.

Table 3: Sample list of Urdu stop words

چونکہ	بلکہ	پر
ارے	یہی	لہذا
از	تاکہ	مگر
البتہ	نہ	بہ
اوپر	تو	چنانچہ
اوہ	کوئی	کہ
کو	کہ	حتیٰ

A list of special characters is also developed to filter out such characters from the input text in the text pre-processing stage.

Table 4: List of special creators

0 1 2 3 4 5 6 7 8 9	۰ ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹
><[] # ? = { }	^ , . - \ / _
\$ () + , * % @	& " ' ! ' ; :

4. CONCLUSION

Stemming is important and basic component in many natural language processing applications. The accuracy of the stemmer strongly affects the results of the system in which it is used. Various stemmer development techniques are being explored and studied for different languages across the world. This paper surveyed the stemmers available for Urdu language. This paper presents a rule based stemming method for Urdu language. Proposed method has the ability to generate the stem of Urdu words as well as loan words (words belong to borrowed languages i.e. Arabic, Persian, Turkish, Hindi, etc).

REFERENCES:

- [1] Akram QU, Naseer A, Hussain S. 2009. "Assas-Band, an affix-exception-list based Urdu stemmer" In Proceedings of the 7th Workshop on Asian Language Resources, Association for Computational Linguistics, 40-46.
- [2] Gupta, V., Joshi, N., Mathur, I. (2016) Design and Development of a Rule-Based Urdu Lemmatizer Proceedings of International Conference on ICT for Sustainable Development pp 161-169, doi. 10.1007/978981-10-0135-2_15
- [3] Husain, Mohd Shahid, Faiyaz Ahmad, Saba Khalid. 2013 "A language Independent Approach to develop Urdu stemmer." Advances in Computing and Information Technology. Springer Berlin Heidelberg, 4553.
- [4] Ali.M, K.S, S.H.M, 2014 "A Novel Stemming Approach for Urdu language" ISSN:20904274 Journal of Applied Envoironmetal and biological sciences, J. Appl Envoiron. Biol. Sci., 4(7S)436-443
- [5] Khan, Sajjad Ahmad, et al 2012. "A Light

Weight Stemmer for Urdu Language: A Scarce Resourced Language." 24th International Conference on Computational Linguistics.

[6] Ali M, Khalid S, Haneef M, Iqbal W, Ali A, Naqvi G. 2016 A Rule based Stemming Method for Multilingual Urdu Text. International Journal of Computer Applications. 2016 Jan.