Analysis and Development of Resources for Urdu Text Stemming

Abdul Jabbar
Department of Computer Science
Institute of Southern Punjab, Multan, Pakistan
a.jabbar73@hotmail.com

Sajid Iqbal
Department of Computer Science
Bahauddin Zakariya University, Multan, Pakistan
Sajid.iqbal@bzu.edu.pk

Muhammad Usman Ghani Khan
usman.ghani@kics.edu.pk
Department of Computer Science and Engineering
University of Engineering and Technology, Lahore

Abstract

Urdu has been facing various challenges in Natural Language Processing (NLP) due to its morphological richness. Stemming, as a preprocessing technique used in different applications of natural language processing, is one of the basic morphological operation applied in written text. The technique is used differently in its various applications like machine translation, query processing, question answering, text summarization and information retrieval. This paper presents the Urdu resources for Urdu text stemming such as affixes list, stop word list, stem word list and stemming rules to remove the infixes letter(s) and recoding to extract correct stems. Here, we collect 1211 affixes, 1124 stop words, 40904 stem word list and 35 rules with their various variations to remove the infixes.

Keywords: *Stemming, Urdu stem words, Urdu affixes, Urdu stop words, Natural Language Processing*

1. Introduction

Urdu language is different from other languages like English in terms of its linguistics and phonetic rules. It was developed in the 12th century from the regional Apabhramsha of northwestern India[1]. In the following paragraph, we list the prominent differences as compared to English language with respect to text stemming process.

Urdu script writing orientation is from right to left. In Urdu word alphabets are connected and do not start with capital letter as in English. Furthermore, most of the characters change their shape based on their position in the word and adjunct letters. Table 1 below demonstrates some of the Urdu characters morphology.

Table 1 Different shape of Urdu Character

آخری شکل Final	درمیانی شکل Medial	ابتدائی شکل Initial	حرف Letter
ع،ـع	ـعـ	عـ	ع
مرتفع، محلوقوع	تعارف، تعظیم	عابد، عندلیب، عروس بہار	مثالیں

In English each word is separated by a hard space, but in Urdu words are not always separated by hard

[1] http://www.britannica.com/topic/Urdu-language

space, e.g. اولے کا بدلا. Further, two or more words could be written as a single word like کیساتھ، ہمنواز. This is a challenging issue in Urdu text segmentation In English proper nouns always start with capital letter but in Urdu proper nouns do not start with a capital letter because Urdu has no letter casing. Due to this, it is a challenging task to extract proper nouns from Urdu text. English language researchers have used rules and tools like named entity recognition (NER) to mark proper nouns in given text. On the other hand, such tools for Urdu are either rare or have low accuracy rate [32],[33].

In English, inflectional or derivational forms are created by attaching affixes to either or both sides of the root. For example, words *readable* and *unreadable* are created from the root word *read.* In this example, "*un*" is a prefix and "*able*" is a suffix. However, in Urdu language, affix letter(s) could be found anywhere in the middle of the word. For instance, رسوم derived from رسم and اکابر derived from اکبر. This is a challenging issue in Urdu text stemming because, it is difficult to distinguish between affix letters and actual part of the root letters. Conversion from singular to plural is also different in Urdu language. There are multiple rules to do this. For example, a singular could be converted to plural by adding some suffix such as کتاب سے کتابیں, by adding some co-fix with suffix -Further there may be multiple rules to convert a word into its plural کتاب سے کتابیں /کتابوں/کتب. The case of broken plural words is also different, because they do not follow the normal morphological rules. Urdu broken plural words are somewhat like irregular English plurals. The difference is that English singular and plural words resemble to each other, such as man to men, but in the case of Urdu both may be non-identical مشہور سے مشاہیر . Another difference between two languages is that English has only uni-gram words even after derivation. In Urdu there are uni-gram, bi-gram and tri-gram words that are obtained after derivation for example غیر تربیت یافتہ ، شادی شدہ and کتابیں.

Resource development is basic and fore most step of Natural Language Processing (NLP) field. Urdu resource development starts with Urdu Zabta Takhti (UZT) that is standard code for Urdu characters, approved by Government of Pakistan (GOP) [8].

A text corpus consisting of 18 million Urdu words is collected by the Center for Research in Urdu Language Processing (CRULP) at National University of Computer and Emerging Sciences in Pakistan [10]. CLE at University of Engineering and Technology, Lahore has also produced different resources for Urdu NLP. This paper focuses on construction of Urdu resources which can be used especially for stemmer development for Urdu text and evaluation, and generally for research in computational linguistics. This paper describes the lists of resources produced during my MS thesis and are available online[2].

The remainder of this paper is organized as follows: Section 2 gives the overview of stemming algorithms. Section 3 describes the development of stop words list. Section 4 introduces the Urdu affix list. The Development of the stem word list is described in section 5 and development of infix rules are described in section 6. In section 7 conclusion and discussion of future work can be found.

2. Stemming algorithms

In stemming, affixes are chopped off from derivational and inflectional forms of Urdu words to extract stem. For example, Urdu words خوشگوار، ناخوش in which نا and گوار are affixes and its common stem is خوش. According to Lovins [15] "*a stemming algorithm is a computational procedure that reduces all the words with same root by stripping each word of its derivational and inflectional suffixes*". Khoja and Garside [13] define the stemming algorithm in Arabic language context as "*Stemming is the process of removing all of a word's prefixes, suffixes and infixes to produce the stem or root*".

Anjali Ganesh [23] analyzed the English stemmer and classified them into three: truncating, Statistical and mixed. Moral [24] grouped stemming algorithms into algorithmic-based approaches and linguistic-based approaches. Moghadam [25] classify Persian stemmer into three classes: structural stemmers, table lookup stemmers and statistical stemmers. Basically, there are two types of stemmers and third is a combination of these two, these are described in detail below.

a) Rule base stemmer

This is most commonly used stemming technique. First stemmer [5] that is found in literature was a rule base stemmer. This stemmer is suitable for those

[2] https://sourceforge.net/projects/resource-for-urdu-stemmer/

languages that have well defined linguistic structure. Rule base approach is also a better choice for infixes removal. This stemmer can give promising results if excellent rules are developed. On the other hand, too much rules make it more complex with inclusion of contrary and recursive set of rules. To apply one rule you have to remove other (opposite) rule. This situation can easily lead toward deteriorated performance.

b) Statistical stemmer

Another popular approach is the use of statistical techniques. In such approaches, the system learns from available corpus and decides on unseen data using knowledge gained through its experience, i.e. statistical measures obtained through experience are used to remove prefixes and affixes. However, the use of statistical methods is popular among the language processing community. N-gram based statistics are used by W. B. Frakes [26]. Melucci [27] used HMMs and YASS: Yet another suffix stripper by Majumder [28] is a few of the examples. It can easily be seen that such approaches are limited to gain experience. In other words, such approaches rarely cover the complete grammar of the language [29]. For example, N-gram approach is better to remove suffixes only however, Urdu language has suffixes, co-suffixes, prefixes, infixes and circumfixes (prefixes and suffixes simultaneously) [30] and use of HMM requires that every word must start with the prefix [27]. This review shows that statistical approaches are insufficient in stemming process. A comprehensive review of stemming technique applied in Urdu, Persian and Arabic is present in [39].

c) Hybrid stemmers

Hybrid stemmer is combination of the rule base stemmer and statistical stemmer [40].

3. Development of stop words list

Usually a sentence contains stop words and content words as shown in figure 1. Stop words have no linguistic meaning, so useless in queries and can generally be safely ignored, e.g. in Urdu "کا","پر","سے" and in English "on", "in", "to", and "the". On the other hand, content words are keywords of any sentence and have lexical meaning. Content words list usually contains nouns, verbs, or adjectives. Urdu stop word list usually contains postpositions, determiners, pronouns, and conjunctions [7]. [2] Used 400 translated Urdu stop words from English stop words. In these translated Urdu stop words, some are not actually stop words e.g. طریقوں،لمحات وغیرہ. 1200 closed words list is provided in [22] and all the close words are not stop words, for instance بےشک, تسلیمات are not stop words but closed words, and these words are stem able words.

Stop words

پاکستان میرا وطن ہے

Content words

Figure 1: Example of stop word and content word

Up to now, no significant work has been carried out to find the stop words from Urdu text. After studying the various Urdu grammar books and literature [3], [4], [18], [22], [34], [35], [36], [37]. We developed stop word list consisting of 1124 Urdu words

4. Development of Affixes lists

The affixes are a morpheme or set of morphemes/words that are frequently attached to other words and create new words. Internal structure of Urdu word is shown in figure 2.

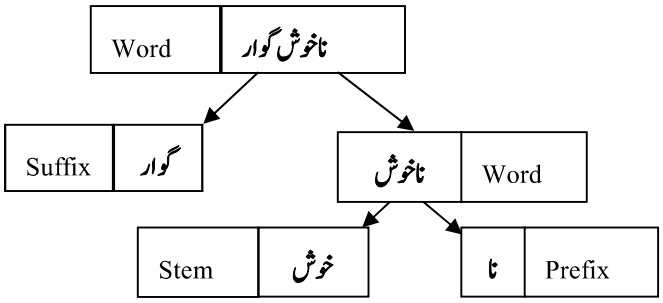

Figure 2 Internal structure of Urdu words Figure

A prefix is a morpheme or word that attaches with the start of the words and changes its meaning, e.g. ناپاک in which نا is a prefix morpheme and words خوش اخلاق in which خوش is a prefix word. On the other hand, suffix is such morpheme or word that comes at the end of the words and it does not change the meaning of the word, e.g. لڑکیاں in which ان are a suffix morpheme and the word دل پسند in which پسند is suffix word. Infixes are letters that can be anywhere in the middle of words, e.g. اکابر with which letter ا is an infix and root word is اکبر.

Urdu not only borrows the words from other languages, but also borrows affixes. Sometimes loan words are also used as affixes to make a hybrid word by Hybridization [31]. In table 2, some examples of affixes/words from the Hindi, Persian, Arabic and English are listed. In Urdu language, loan affixes word/letters from one language can be attached to other language loan affixes, words/letters to create a new single or compound word.

Table 2 Sample affixes list

Language	Prefixes	Words	suffixes	Words
Persian affixes in Urdu	تہ	تہ بند	انہ	مردانہ
Hindi affixes in Urdu	ک	کراہ	ک	بیٹھک
Arabic affixes in Urdu	ال	القران	فی	فی صد
English affixes in Urdu	ڈبل	ڈبل روٹی	سٹور	کریانہ سٹور

Qurat-ul-Ain [1] identified 174 prefixes and 712 postfixes. Khan [11], [12] collected 180 prefixes and 750 suffixes for Urdu text. Mubashir [16] mentions 60 prefixes and 140 suffixes. 122 suffixes and 15 prefix are identified by [7]. After studying the various Urdu grammar books and literature [3], [4], [9], [18], [22], [34], [35], [36], [37], we constructed prefixes 643, suffix 568; then arrange them according to their length.

5. Development of Stem word list

Stem words list is essential to validate the extracted stem. Khan [12] construct a stem words list of 3500 words for Urdu. Mubashir [16] developed a stem words list of about 10000 words. After studying the various Urdu grammar books and literature [3], [4], [9], [18], [22], [34], [35], [36], [37] we construct a root word list containing 40904 words. Examples of stem words are تدبیر، حکم.

6. Development of infixes rules

After studying the various Urdu grammar books and literature [9], [18], [34], [35], [36], [37]. We chop off the infixes letters from the Urdu words using orthographic pattern. We construct 10 rules for Urdu word length 4 letters with variations of rules, 12 rules for Urdu word length 5 letters with variations of rules and 13 rules for Urdu word length 6 letters with variations of rules. A sample list of infixes removal rules are given in table 3 and complete rules are available online.[3]

7. Conclusion

Generally, Information Retrieval (IR) system used variant forms of the query word by stemming process. We have pointed out the differences between Urdu and English language with respect to stemming process. We have also described different types of stemming approaches and borrowed affixes from Arabic, Persian, Hindi and English languages. In this paper, we presented required linguistic resources for Urdu text stemming. Evaluation of these language resources are given in [38].

[3] https://sourceforge.net/projects/resource-for-urdu-stemmer/

.

Table 3 Sample infixes removal rules with variation

Set of Rules: Words Length 4 and Stem Words Length 3 Characters					
Rule No. 1					
Index		**3**	**2**	**1**	**0**
Orthographic pattern		ـ	و	ـ	ـ
Input word	امور	ر	و	م	ا
Stem Word	امر	ر		م	ا
Rule No. 1 Variations A					
Index		**3**	**2**	**1**	**0**
Orthographic pattern		ـ	و	ـ	ـ
Input word	خطوط	ط	و	ط	خ
Invalid Stem	خطط	ط		ط	خ
Deletion	ط			ط	خ
Stem Word	خط			ط	خ
Rule No. 1 Variations B					
Index		**3**	**2**	**1**	**0**
Orthographic pattern		ـ	و	ـ	ـ
Input word	حصول	ل	و	ص	ح
Invalid Stem	حصل	ل		ص	ح
Insertion	ا	ل	ص	ا	ح
Stem Word	حاصل	ل	ص	ا	ح
Rule No. 1 Variations C					
Index		**3**	**2**	**1**	**0**
Orthographic pattern		ـ	و	ـ	ـ
Input word	سجود	د	و	ج	س
Invalid Stem	سجد	د		ج	س
Insertion	ه	ه	د	ج	س
Stem Word	سجده	ه	د	ج	س

References

[1] Akram, Qurat-ul-Ain, Asma Naseer, and Sarmad Hussain. "Assas-Band, an affix-exception-list based Urdu stemmer." In Proceedings of the 7th Workshop on Asian Language Resources, pp. 40-46. Association for Computational Linguistics, 2009.

[2] Burney, Aqil, Badar Sami, Nadeem Mahmood, Zain Abbas, and Kashif Rizwan. "Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors." International Journal of Computer Applications 46, no. 19 (2012).

[3] BBC Urdu (2016): News and research articles retrieved from http://www.bbc.com/urdu

[4] DAWN News(2016): News and research articles retrieved from http://www.dawnnews.tv/

[5] Ethnologue Languages of the World (2015). "Urdu." Retrieved 29 November, 2015, from https://www.ethnologue.com/language/urd.

[6] ENCYCLOPAEDIA BRITANNICA (2015). "Urdu language." Retrieved 01 DECEMBER, 2015, from http://www.britannica.com/topic/Urdu-language.

[7] Gupta, Vaishali, Nisheeth Joshi, and Iti Mathur. "Design & development of rule based inflectional and derivational Urdu stemmer 'Usal'." In Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on, pp. 7-12. IEEE, 2015.

[8] Hussain, Sarmad, and Muhammad Afzal. "Urdu computing standards: Urdu zabta takhti (uzt) 1.01." In Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International, pp. 223-228. IEEE, 2001.

[9] Hussain, Sara. "Finite-state morphological analyzer for urdu." PhD diss., National University of Computer & Emerging Sciences, 2004.

[10] Hussain, Sarmad. "Resources for Urdu Language Processing." In IJCNLP, pp. 99-100. 2008.

[11] Khan, Sajjad, Waqas Anwar, Usama Bajwa, and Xuan Wang. "Template Based Affix Stemmer for aMorphologically Rich Language." International Arab Journal of Information Technology (IAJIT)12, no. 2 (2015).

[12] Khan, Sajjad Ahmad, Waqas Anwar, Usama IjazBajwa, and Xuan Wang. "A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language." In 24th International Conference on Computational Linguistics, p. 69. 2012.

[13] Khoja and Garside, "Stemming Arabic Text" 1999. Available online at URL: http://zeus.cs.pacificu.edu/shereen/research.htm#stemming [accessed 27/12/2015].

[14] Kwintessential (2015). "The Urdu Language." Retrieved 02 December, 2015, from http://www.kwintessential.co.uk/language/about/urdu.html.

[15] Lovins, Julie B. Development of a stemming algorithm. MIT Information Processing Group, ElectronicSystems Laboratory, 1968.

[16] Mubashir Ali, Shehzad Khalid, Muhammad Haneef Saleemi "A Novel Stemming Approach for UrduLanguage" ISSN: 2090-4274, Journal of Applied Environmental and Biological Sciences, J. Appl.Environ. Biol. Sci., 4(7S)436-443, 2014, www.textroad.com

[17] Qureshi, Anwar & Awan" Morphology of the Urdu Language", International Journal of Research in Linguistics and Lexicography, INTJR-Volume 1-Issue 3, September 2012,

[18] Ruth Lail Schmidt (1999). URDU: AN ESSENTIAL GRAMMER.

[19] Daily Pakistan (2015). "Urdu declared second most popular language among 2301 others." Retrieved 01 December, 2015, from http://en.dailypakistan.com.pk/pakistan/urdu-declared-second-most-popular-language-among-2301-others/.

[20] García, María Isabel Maldonado. "Comparación del léxico básico del Español, el Inglés y el Urdu." Unpublished doctoral dissertation-UNED, Madrid 500 (2013).

[21] Urdu words list got from http://www.cle.org.pk/software/ling_resources/wordlist.htm Retrieved 02 DECEMBER, 2015,

[22] Urdu closed words list retrieved on 09 march 2016 from http://cle.org.pk/software/ling_resources/UrduClosedClassWordsList.htm

[23] Jivani, Anjali Ganesh. "A comparative study of stemming algorithms." Int. J. Comp. Tech. Appl 2, no. 6 (2011): 1930-1938.

[24] Moral, Cristian, Angélica de Antonio, Ricardo Imbert, and Jaime Ramírez. "A survey of stemming algorithms in information retrieval." Information Research: An International Electronic Journal 19, no. 1 (2014): n1.

[25] Moghadam, Fatemeh Momenipour. "Comparative Study of Various Persian Stemmers in the Field of Information Retrieval." Journal of information processing systems 11, no. 3 (2015): 450-464.

[26] W. B. Frakes. (1992). Information Retrieval: Data Structures & Algorithms, Chapter 8,Retrieved 01 October,2015, from http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap08.htm

[27] Melucci, Massimo, and Nicola Orio. "A novel method for stemmer generation based on hidden Markovmodels." In Proceedings of the twelfth international conference on Information and knowledge management, pp. 131-138. ACM, 2003.

[28] Majumder, Prasenjit, Mandar Mitra, Swapan K. Parui, Gobinda Kole, Pabitra Mitra, and KalyankumarDatta. "YASS: Yet another suffix stripper." ACM transactions on information systems (TOIS) 25, no. 4 (2007): 18.

[29] Anzai, Yuichiro. Pattern Recognition & Machine Learning. Elsevier, 2012.

[30] Husain, M. S., Ahamad, F., & Khalid, S. (2013). A language Independent Approach to develop Urdustemmer.

In Advances in Computing and Information Technology (pp. 45-53). Springer BerlinHeidelberg.

[31] Qureshi, Anwar & Awan" Morphology of the Urdu Language", International Journal of Research in Linguistics and Lexicography, INTJR-Volume 1-Issue 3, September 2012,

[32] Singh, U., Goyal, V. and Lehal, G.S., 2012. Named Entity Recognition System for Urdu. In COLING (pp. 2507-2518).

[33] Riaz, Kashif. "Rule-based named entity recognition in Urdu." In Proceedings of the 2010 named entities workshop, pp. 126-135. Association for Computational Linguistics, 2010.

[34] Board, P. T. (2010). "اردو قواعدوانشاء" for Class-10th. Lahore: Punjab Textbook Board.

[35] Bloch, Dr. Sohail Ahmad (2012), " بنیادی اردو قواعد " , Muqtadrah Qumi Zuban Pakistan, Islamabad.

[36] Haq, Molvi Abdul (1996), "قواعد اردو", Anjaman Tariqi e Urdu, New Dehli (Hind)

[37] UEP (2014)," تخلیق اردو گرائمر", for class 8th, Unique Education Publisher, Urdu bazar Lahore.

[38] A. Jabbar et al. " Effective Urdu Stemmer Based Hybrid Approach". Information Processing & Management (under major revision).

[39] A.Jabbar et al. "A survey on Urdu and Urdu like Language Stemmers and Stemming Techniques" Artificial intelligence review (under minor revision)

[40] Lehal, RohitKansal Vishal Goyal GS. "Rule Based Urdu Stemmer." In 24th International Conference on Computational Linguistics, p. 267. 2012.