International Journal of Applied Engineering Research ISSN 0973-4562 Volume 19, Number 1 (2024) pp. 99-102
© Research India Publications.

A Comparative Study between Dhundhari and Shekhawati Stemming: Case of Very Closely Related Languages

Varda Pareek[1*], Nisheeth Joshi[2]

[1] Research scholar, Banasthali Vidyapith

Assistant Professor, Computer Science and Engineering,

Manipal University Jaipur, Raj-India,

[2]Speech and Language Processing Lab, Centre for Artificial Intelligence,

Banasthali Vidyapith, Rajasthan-India

[*] Corresponding Author

Abstract

Dhundhari and Shekhawati are part of a Rajasthani language group. Shekhawati is a neighbour region of a Dhundhar. Dhundhari is a richer morphological language than Shekhwati. Both are low-resources languages. In this paper, morphology of Shekhawati and dhundhari are analysed by developing rule-based stemmer. For that,103 suffixes, 32 prefixes of Dhundhari and 89 suffixes, 32 prefixes of Shekhawati are developed. 124 rules of Dhundhari and 99 rules for Shekhawati are created. The inflectional accuracy for Dhundhari was greater than Shekhawati whereas derivational accuracy for Shekhawati was greater than Dhundhari.

Keywords: Dhundhari Language, Shekhawati Language, Stemmer, Inflectional, Derivational, Rule Based.

INTRODUCTION

It is well known that NLP is a very fast-growing area of Artificial intelligence but most of the research in NLP till now focused on very few languages whereas 7000 languages exist in this world. These other languages are known as low resource languages as these languages are less searched or studied and less resources are available for these languages. These languages are less computerized and minimum privileged and very little taught. A group of Rajasthani languages are also low-resourced languages. Dhundhari and Shekhawati are part of the Rajasthani languages. Dhundhari is spoken in Jaipur, Dausa, Ajmer and, Tonk districts of Rajasthan whereas Shekhawati is spoken in Jhunjhunu, Sikar and Churu districts of Rajasthan. Dhundhari is the second most popular language of Rajasthan whereas Shekhawati is the third most popular language of Rajasthan. To analyze the morphology of these two languages, we developed a rule-based stemmer for both languages. In this paper, comparison of morphology of both languages by developing stemmer is shown.. Both languages have many Hindi words. So, mixed rules for both the languages separately with Hindi are tried to create. This task has been very much challenging for development of stemmers of both languages.

Example: word in Dhandhari is ‘लुगायां’- ‘- (i.e.noun) which has ‘लुगा’ - – (i.e. Noun) as a root word. Both come into a similar category. This is an example of inflectional stemming. ‘खायोड़ो’ – (i.e.noun) is an actual word where root word is ‘खा’ – -(i.e. verb). Here ‘खायोड़ो’ is a noun form and ‘खा’ is a verb form. This is an example of derivational stemming.

Example: word in Shekhawati is ‘करसी’- ‘- (i.e. verb) which has ‘कर’ - – (i.e. verb) as a root word. Both come into a similar category. This is an example of inflectional stemming. As a stemming example - ‘’ (करेड़ी) – (i.e. adjective) is an actual word where root word is ‘कर’ – - (i.e. verb). Here adjective form is converted in to form of noun. This is an example of derivational stemming.

To analyse morphology of both languages, Rule-based stemmers of Dhundhari and Shekhawati are developed. The number of rules developed for Dhundhari and Shekhawati, were 124 and 99 respectively. Number of prefixes were 32 for both while number of suffixes were 103 and 89 respectively. Further details of proposed methodology are discussed in proposed algorithm section.

Literature Review

Jaafar et al. (2017) introduced a program arrangement of benchmarking light. They also introduced a measurement called, “Gs-Score (for Global Stemming Score) that joins execution time with the precision of stemmers”. The precision of those frameworks was depending on the stemming system’s exactness. Furthermore it was notified that the “Holy Quran” is an astoundingly phenomenal text.

Barakhnin et al. (2017) expressed the algorithm for verbs stemming which depends on Porter's algorithm. Algorithms of stemming and verbs are portrayed in the article. They tried words having a place with various grammatical forms and didn't track down any blunders. This permitted them to pass judgment on the rightness of the proposed calculations.

Maylawati et al. (2018) performed study for Indonesian text and developed stemmer for it. The precision of their algorithm was about 88.65%.

Swain and Nayak (2018) discussed a rule-based stemming approach and hybrid stemming approach in their study. The study revealed that the e preferable outcomes were obtained with hybrid stemming as compared to the rule-based stemming. Mulki et al. (2018) found that majority of tweets and comments reflect Tunisian perceptions of major social, economic, and political incidents. The tweet opinions followed, examined, and assessed through sentiment analysis and their results revealed that pre-processing procedures positively influenced the evaluation of both supervised and Lexicon-based classifiers.

Alhaj et al. (2019) examined the impact of "Information Science Research Institute (ISRI)", "Tashaphyne", and "ARLStem" stemmers toward the presentation of classification of Arabic Document. "Term Frequency-Inverse Document Frequency (TF-IDF)" was used for extricating element part from messages. Three algorithms of machine learning were analyzed for taking care of the classification of the documents. The outcomes showed that the SVM classifier is best among them.

Oo and Soe (2019) assessed various sorts of "Recurrent Neural Networks (RNNs)" for succession marking errands. Their study also proposed three variations of "RNNs-('LSTM,' 'Bi-LSTM' and 'GRU')" for division and procedure of stemming for joint words. Their outcomes showed that Bi-LSTM' model was more viable as compared to the rest of two models. To assemble neural grouping naming design for joint course of Myanmar word, the' NCRF++ toolbox' was utilized.

Kanan et al. (2019) analysed the gathered dataset with the help of three notable stemmers: the 'P', 'Khoja' and the 'Light' stemmers. 'SVM' and 'NB' classifiers were used with all three stemmed datasets. In both SVM and NB classifiers, the P-Stemmer' achieved a better result than the other two stemmers. The precision for 'F1-Measure' was 0.92 in SVM and 0.90 in NB classifier.

Nzeyimana (2020) handled morphological disambiguation as a grouping issue with a variable number of classes for each example. The author processed two kinds of highlights from each morphological division and fed those elements to a feed-forward neural network, lastly producing probabilities with a softmax function. Author trained the organization to limit cross-entropy misfortune work. Author utilized the baseline classifier to then anticipate the stem for the whole unlabelled vocabulary of Kinyarwanda verbal structures.

Bakar et al. (2023) also worked on rule-based stemmer which focused on 'e-khutbah texts' for getting words which contain affix of letter 'p'.

Morphology

a) Noun

Singular and masculine nouns in Dhundhari end with ({o}, ओ). Plural and masculine nouns in Dhundhari end with ({aa}, आ). Cases of Shekhwati in terms of masculine noun is also same. Singular and Feminine nouns fin case of Dhundhari and Shekhawati both end with ({ee}, ई). While for Plural and feminine nouns in both end with ({yan}, यां). Examples of masculine and feminine nouns in Dhundhari and Shekhawati are following.

Table 1- Nouns

	Feminine Noun	Masculine Noun
Singular	छ्याळी (chhyaali)-goat	नाडो (Naado)-pond
Plural	छ्याळयां (chhyaalian)-goats	नाडा(naada)-ponds

b) Adjective

Cases of Shekhwati and in Dhundhari in terms of Masculine as well as feminine adjectives are same. Singular and Masculine Adjectives end with ({o}, ओ). Plural and Masculine adjectives end with ({aa}, आ). Singular and Feminine Adjectives in Dhundhari end with ({ee}, ई). Plural and feminine Adjectives in Dhundhari also end with ({ee}, ई). Following table

Table 2 –Adjectives

	Feminine Adjective	Masculine Adjective
Singular	चोखी (chokhi)-good	चोखो (chokho)-good
Plural	चोखी(chokhi)-good	चोखा(chokha)-good

c) Verb

Verb in Dhundhari and Shekhawati are somewhat different. "है" is used as auxiliary verb with the main verb in Shekhawati whereas "छै" is used as auxiliary verb with the main verb in Dhundhari. Verbs of Present and past tense are same in Dhundhari and Shekhawati but verbs of future tense are totally different.

Table 3– Future tense of verb-"खा" in Dhundhari

	Masculine singular	Masculine plural	Feminine singular	Feminine plural
First person	खावूंलो	खावांलां	खावूंली	खावांलां
Second person	खावैलो	खावोला	खावैली	खावोला
Third person	खावैलो	खावैला	खावैली	खावैली

Table 4– Future tense of verb-“खा” in Shekhawati

	Masculine singular	Masculine plural	Feminine singular	Feminine plural
First person	खास्यूं	खास्यां	खास्यूं	खास्यां
Second person	खासी	खास्यो	खासी	खास्यो
Third person	खासी	खासी	खासी	खासी

d) Adverb

There is no specific format of adverbs in Dhundhari and shekhawati.

Those can be divided into following categories:

Adverb of time: आज (today), काल (yesterday)

Adverb of place: ‘अंडै (there) in Dhundhari and अठै (there) in shekhawati’

Adverb of manner: ‘खाती-खाती (fast) in Dhundhari and बेगी (fast) in shekhawati’

Adverb of degree: पतळो (Thin)

Adverb of frequency: बर-बर (बार-बार)

Proposed Algorithm

In this article Shekhawati and Dhundhari morphologies were studied deeply and by rule-based stemmers for both were developed. The number of suffixes and prefixes were analysed as given in following table.

Table 5–Prefixes and Suffixes

	Number of Prefixes	Number of Suffixes
Dhundhari Language	32	132
Shekhawati Language	32	86

In the light of above prefixes and suffixes 124 rules of Dhundhari and 99 rules for Shekhawati were created.

Algorithm of Stemmer:

Step 1: Begin

Step 2: Mentioned Lists of Suffixes and Prefixes

Step 3: Calculated Length of all Suffixes

Step 4: Check suffixes in given word according to the length of suffixes.

Step 5: Combined all words after suffix removal and remaining input words.

Step 6: Check Prefixes in Combined words list according to length of prefixes.

Step 7: Declare Output.

Table 6- Examples of Dhundhari Stemming.

Suffix	Word	Stem
यां	छोर्यां	छोर
◌ा◌ं	मोट्यारां	मोट्यार
वैला	खावैला	खा
◌ी	आपणी	आपण

Table 7- Examples of Shekhawati Stemming.

Suffix	Word	Stem
यां	बायां	बा
◌ा◌ं	छोरां	छोर
सी	करसी	कर
◌ी	आपणी	आपण

Results and Discussions

The system was evaluated using the Standard accuracy of the developed systems were calculated by equation (1). For the purpose 4000 words of Dhundhari and 4000 words of Shkhawati were tested. Among 4000 words of each language, 2000 were inflectional and the 2000 words were derivational Table 8 and 9 summarize the results of this study.

$$Accuracy = \frac{Correct\ output}{words} \tag{1}$$

On looking at the results it was found that the system had certain limitations. Since Dhundhari and Shekhawati are dialects of Hindi, in certain cases words of Dhundhari, Shekhawati and Hindi are similar which caused conflicts in stemming rules. This is an area that needs refinement and will be addressed in future versions.

Table 8-Result of Dhundhari

	Total words	Correct Stem	Incorrect Stem	Accuracy
Inflectional	2000	1994	6	99.7%
Derivational	2000	995	10	99.5%
Combined	4000	1992	16	99.6%

Table 9-Result of Shekhawati

	Total words	Correct Stem	Incorrect Stem	Accuracy
Inflectional	2000	1984	16	99.2%
Derivational	2000	1988	12	99.4%
Combined	4000	1972	28	99.3%

Comparative Analysis

Comparative accuracy of inflectional stemming between Shekhawati and Dhundhari is shown in figure 1.

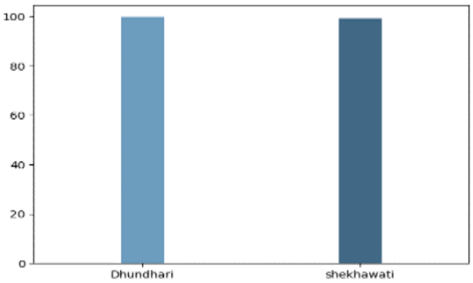

Figure 1: Comparative accuracy of inflectional Stemming between Dhundhari and shekhawati

Comparative Inflectional stemming accuracy between Dhundhari and Shekhawati is shown in figure1. According to figure 1, accuracy of inflectional stemming of Dhundhari is 99.7% is slightly greater than and accuracy of inflectional stemming of Shekhawati is 99.2%.

Now, comparative accuracy of derivational stemming between Shekhawati and Dhundhari is shown in figure2.

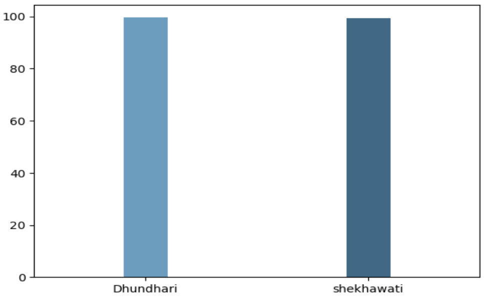

Figure 2: Comparative accuracy of derivational stemming between Dhundhari and Shekhawati

Comparative derivational stemming accuracy between Dhundhari and Shekhawati is shown in figure2. According to figure 2, accuracy of derivational stemming of Dhundhari is 99.5% and accuracy of derivational stemming of Shekhawati is 99.4%.

CONCLUSION

Shekhawati and Dhundhari are dialects of Rajasthani language. In this study, rule-based stemmers for both languages are developed and 124 rules for Dhundhari and 99 rules for Shekhawati are created. Study found that Dhundhari is morphological richer than Shekhawati as 103 suffixes, 32 prefixes are developed for Dhundhari and 89 suffixes, 32 prefixes are developed for Shekhawati. It has been observed in study that 99.6% accurate results for Dhundhari and 99.4% accurate results for Shekhawati were obtained.

Developed stemmers have limitations as there are many common words among Dhundhari, Shekhawati and Hindi. Hindi-Dhundhari joint stemming rules for Dhundhari stemmer and Hindi-Shekhawati joint stemming rules for Shekhawati stemmer were created. This caused some disruption during creating rules.

REFERENCES

[1] Alhaj, Y. A., Xiang, J., Zhao, D., Al-Qaness, M. A., Abd Elaziz, M., & Dahou, A. (2019). A study of the effects of stemming strategies on Arabic document classification. IEEE Access, 7, 32664-32671.

[2] Bakar, Z. A., Anuar, N., & Ismail, N. K. (2023). Extraction of Malay Root Word that Starts with Letter P in Malay e-Khutbah using Rule Based. *International Journal of Software Engineering and Computer Systems*, *9*(1), 39-45.

[3] Barakhnin, V., Bakiyeva, A., & Batura, T. (2017). Stemming and word forms generation in automatic text processing systems in the Kazakh language. Computational technologies, 22, 11-21.

[4] Jaafar, Y., Namly, D., Bouzoubaa, K., & Yousfi, A. (2017). Enhancing Arabic stemming process using resources and benchmarking tools. Journal of King Saud University-Computer and Information Sciences, 29(2), 164-170.

[5] Kanan, T., Sadaqa, O., Almhirat, A., & Kanan, E. (2019, October). Arabic light stemming: A comparative study between p-stemmer, khoja stemmer, and light10 stemmer. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 511-515). IEEE.

[6] Maylawati, D. S. A., Zulfikar, W. B., Slamet, C., Ramdhani, M. A., & Gerhana, Y. A. (2018, August). An improved of stemming algorithm for mining Indonesian text with slang on social media. In 2018 6th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-6). IEEE.

[7] Nzeyimana, A. (2020, December). Morphological disambiguation from stemming data. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 4649-4660).

[8] Oo, Y., & Soe, K. M. (2019, October). Applying RNNs Architecture by Jointly Learning Segmentation and Stemming for Myanmar Language. In 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE) (pp. 391-393). IEEE.

[9] Swain, K., & Nayak, A. K. (2018, September). A review on rule-based and hybrid stemming techniques. In 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA) (pp. 25-29). IEEE.