International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8 Issue-8S3, June 2019

Preprocessing for Parts of Speech (POS) Tagging in Dogri Language

Shivangi Dutta, Bhavna Arora

Abstract*:*Natural language processing (NLP) is viewed among the most crucial fields of computer science, information retrieval and artificial intelligence. One such challenging feature in NLP is Parts of speech (POS) tagging. It is the process of labelling the words present in the corpus as the parts of speech. According to English grammar there are eight major parts of speech which are: noun, pronoun, verb, adjective, adverb, preposition, conjunction, interjection. Over the past few years, various researchers have compassed considerable amount of work using various pursues to closely supervised tagging and unmonitored ta gging. These methods of labelling are further divided into rules-based, stochastic and hybrid approaches. The language that has been taken for research work is Dogri Language which is based on Devanagari script. The paper presents the related work in the languages having same script as Dogri. The study helps in the selection of appropriate technique to be used for POS tagging for Dogri language. The paper also presents grammatical and inflectional analysis of Dogri language along with few rules for designing POS tagger. A section of the paper also demonstrates the results of preprocessing i.e. tokenization and stemming of Dogri text, which are considered as the initial steps in POS tagging.

Index Terms*:*Dogri language, Parts of speech tagging, stemming, tokenization.

I. INTRODUCTION

Language contains words and words are categorized into several types or parts of speech. Parts of speech tagging is the entire process of labelling the words present in the corpus as the parts of speech. As in the English, there are eight distinctive parts of speech viz Noun, Pronoun, Adjective, Verb, Adverb, Conjunction, Preposition, Interjection. Frequently used approaches that are used to actually enforce part-of-speech tagger are Rule Based approach, Statistical approach and Hybrid approach. Rule-based tagging system uses various rules associated with languages to tag the words in the corpus. In another type of approach i.e. Statistical approach or statistical parts of speech taggers the probabilities are calculated. Statistical taggers calculate the frequency or existence of words for a specific tag. Hybrid approach make use of both the rule-based approach as well as the statistical approach. The analysis of languages is a complex task. Numerous Indian tongues like Hindi, Malayalam, Tamil, Punjabi and Marathi have several part-of-speech taggers but less work is done in Dogri comparatively which is the regional language of Jammu and Kashmir. Dogri is a contemporary Indo-Aryan linguistic primarily vocalized inthe states of Jammu and Kashmir and the connecting ranges including Punjab, Himachal Pradesh and in some tehsils in Pakistan. Like any other language, Dogri is a language which is rich in morphology. Dogri Language has the pre-dominant word order such as Subject-Object-Verb (SOV) and has the flexibility to rearrange the constituents. Usually nouns are modulated for number, sex and case. Numbers are –singular and plural; two genders-masculine and feminine; and three cases- simple case, oblique case and vocative case [1]. In the first section, different techniques of POS tagging are discussed which is followed by the related work in the Devanagari script. The following section deals with grammatical and inflectional analysis of the Dogri language and some rules to identify different tags in Dogri Language. In the terminal section preprocessing of Dogri corpus will be performed.

.

Revised Manuscript Received on May 22, 2019.

Shivangi Dutta[1], Department of Computer Science &IT, Central University of Jammu, Jammu, India, shivangidutta94@gmail.com

Bhavna Arora[2], Department of Computer Science & IT, Central University of Jammu,Jammu, India.bhavna.aroramakin@gmail.com.

II. POS TAGGING TECHNIQUES

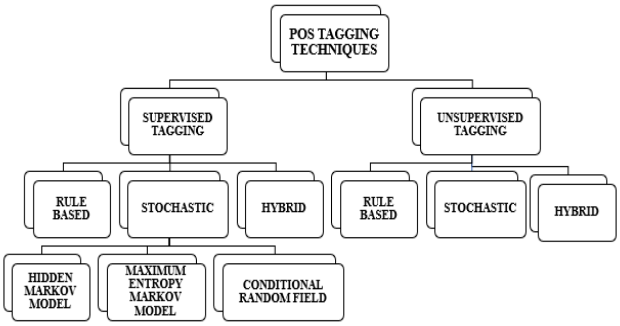


Fig. 1. POS Tagging Techniques [2]

The two basic POS taggers are the Supervised & Unsupervised Taggers. The previously interpreted corpus is used in supervised system tagging. The corpus is used to train the system and hence to acquire the occurrence frequencies of word-tags, rules and tag-set, sets etc. On the contrary, the pre-tagged corpora are not the prerequisite in case of unsupervised POS tagging. The unsupervised taggers make use of the methods according to which automatically tag allocation is done to the words present in the corpus [3]. Supervised and unsupervised techniques fall into three subcategories as described below:

A. *Rule Based*

The rule-based Parts of speech tagging system may be constructed using a set of hand-made rules as defined by the researcherwith the help of linguistics. Relative information is essential for allocation of POS tags to

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

the words present in the context. The executed system doesn't show effective results when the textual information is unknown [3].

B. Stochastic Based POS Tagger

To observe the adequate probable tag sequence, when we are given the observation sequence of n words w1$_n$, that is, find maximum P (t1$_n$ |w1$_n$). To compute P (t1$_n$ |w1$_n$) Bayesian classification rule is used which is given in equation 1.

$$P(x\,|\,y) = P(x).P(y|x)/P(y)\ (1)$$

- ***Hidden Markov Model (HMM) Based Tagger:*** A hidden Markov model (HMM) based tagger explore the adequate probable tag intended for respective term in a sentence. An HMM constructed tagger works by finding a label(tag) arrangement for entire sentence, rather than discovering a label for respective word discretely. For a given sentence w_1..., wn, an HMM based Parts of speech tagger works by finding a tag/label sequence t_1,....,t_n that maximizes the joint probability[4] which is given in the following equation 2.

$$P(t_1 \dots t_n, w_1 \dots w_n) = P(t_1 \dots t_n)P(w_1 \dots w_n|t_1 \dots t_n)(2)$$

- ***Maximum Entropy Based Tagger*:** The Hidden Markov Model based taggers are relatively easy to construct but it is really problematic to include additional complicated structures into such models. The maximum entropy (ME) based tagger however offers an ethical technique of integrating additional complicated structures into (HMM) probabilistic models. For a certain sentence say w_1....,w_n, a Maximum Entropy based tagger produces theconditional possibility of a labelled tag sequence: t_1….,t_n as:

$$(w_1 \dots\dots w_n|t_1 \dots \dots t_n) \approx \prod_{i=1}^{N} P(t_i|C_i)^{(3)}$$

Where C_1, …Cn are mainly described as the context used for respective word in the assumed sentence. The Maximum Entropy based taggers take the features to calculate $(t_i\,|C_i)$ [5].

- ***Conditional Random Field Model (CRF):***CRF is mainlya discriminatory model based on probabilism. CRF model has the advantages of ME model and overcomes the disadvantage of ME model i.e. CRF model is without the label bias problem [6].

C. Hybrid POS Tagger

Hybrid models combine both i.e. rules-based models and statistical models. Hybrid models use strong features from both the approaches i.e. stochastic approaches and the rules- based approaches which make it more efficient. The working of the system includes making a probabilistic model and hence applying certain rules to remove the ambiguity or errors and vice versa.

III. RELATED WORK

The section describes the associated work in the languages having script as Devanagari i.e. Hindi, Marathi, Sanskrit. Various POS tagging techniques used by different researchers to tag the words is presented in the Tables below. The table highlights the techniques used by various authors. The proposed work by the researchers and the results are also described in the followingtables:

Table I: POS Tagger for Hindi Language

Technique	Author & reference no.	Proposed work	Results
RULE BASED APPROACH	**Smriti Singh et.al** [7]	Used corpora of 15,562 words. comprehensive morphological study, high-coverage lexicon including CN2 (algorithm based on decision tree) are some of the techniques used by the author.	Accuracy= 93.45%
	Navneet Garg et.al [9]	have done parts of speech tagging for Hindi language by Rule Based tagger. The data set taken was from news:17233, essay:5039, Stories:3877	News: precision= 89.94%, Recall: 92.84% F-measure: 91.37%, Essay: Precision:81.36% Recall:87.32% F-measure:84.23% Stories: Precision:85.11% Recall:88.32% F-measure:87.06%.
	Aniket Dalal et.al [10]	The paper gives an overview of maximum entropy models, feature functions used in Hindi POS tagging and chunking	POS=89.346% and Chunking=87.399%

STATISTICAL APPROACH	**Nisheeth Joshi et.al [4]**	To train the system, 3,58,288 words from tourism field were used. Not much effort was put in developing morphological analyser instead 5 annotators for creating POS tagged corpora was used. A clear description of HMM is given to resolve the ambiguity of the words	Accuracy = 92.13%.
	Rajesh Kumar et.al[11]	have done parts of speech tagging for Hindi language using a probability-based model called Hidden Markov Model (HMM)	Precision=96.46% Recall= 90.13% f-measure=93.17%
HYBRID APPROACH	**Kanak Mohnot et.al [12]**	Hindi text is tokenized into singular form and POS category is applied by POS rules with the help of Hindi Database. If no rule is formed, then apply rules considering previous and next token and display the results	Accuracy=89.9%.
	Praveshkumar Dwivedi et.al[13]	designed an algorithm which after normalizing the text, checks the word properties and analysis of word root/stem. Tagging is done word root/stem exist otherwise morphological analysis is done in which prefix and suffix is applied to root/stem and morphological synthesizer is applied and hence words are tagged.	----
	Vijeta Khicha et.al [14]	built a combination of hidden markov model and rule based model using java language. Firstly, the Devanagari format is checked and segmentation is done. After that tokenization is done and hence tagging is done using HMM	Precision = 92.56% and Accuracy = 87.55%.

Table II: POS Tagger for Marathi Language

Technique	Author & reference no.	Proposed work	Results
RULE BASED APPROACH	**Pallavi Bagul et.al [15]**	discussed rule-based technique for POS tagger in Marathi text and have shown number of rules which work well. The ambiguity is resolved using Marathi grammar rules as described in the paper and assignment of particular tags are done.	----
	ShubhangiRathode et.al [16]	developed a part of marker for Marathi Language in particular. The approach used by them is rule	Accuracy=95.05%

		based and according to the researcher the effectiveness of the proposed POS tagger in Marathi is relatively more than that of Shallow- Parser and NLTK.	
STATISTICAL APPROACH	**Jyoti Singh et.al [17]**	have developed the part of speech tagger primarily for Marathi language using a probabilistic approach. Unigram, Bigram, Trigram and HMM are approaches used by the researchers for the development of the tagger	Unigram accuracy =77.38%, Bigram accuracy= 90.30%, Trigram accuracy =91.46% and HMM accuracy=y 93.82%.
	Nita V. Patil [18]	used HMM technique to train and test POS tagger for Marathi Language. Unigram, bigram and trigram language models are used for the prediction of the most probable sequence of labels for the specified sequence of words. Viterbi decoding algorithm is used.	Accuracy = 86.61%

Table III: POS Tagger for Sanskrit and Nepali language

Technique	Author & reference no.	Proposed work	Results
RULE BASED APPROACH	**Namrata Tapaswi et.al [19]**	developed a simple POS tagger to tag each word of the sentence automatically. The approach is tested for 15 tags and 100 words of the language to acquire the desired results.	----
	ArchitYajnikn [20]	presents POS tagging method for Nepali script using both HMM and the Viterbi algorithm. Viterbi algorithm is found to be faster than HMM and the system	Accuracy = 95.43%.

IV. GRAMMATICAL AND INFLECTIONAL ANALYSIS

This section presents the study of modification in different grammatical aspects like number, gender etc. in Dogri Language [21].

In grammar, inflection is defined as the modification of a word to denote different categories of grammar such as number, tense,person, case, voice, aspect, gender, and mood. An inflection denotes one or more grammatical categories with an infix, prefix or suffix.

- ***Noun*:**

It refers to people, animal, object, idea, concept, feeling etc. In Dogri a noun hosts the attributes like gender, number and case.

Inflection of nouns by gender

In Dogri language, the genders- masculine and feminine are mostly same as in Hindi language.

The Dogri consonants which are ending with noun in the masculine form have certain different feminine forms which can be formed by using certain suffixes like अनी, आनी, ई and एआनी or ऐनी

Example:

MASCULINE	DOGRI FEMININE
शेर	शेरनी
जेठ	जेठानी

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

दादा	दादी
डॉक्टर	डॉक्टरनी

Inflection of nouns for number

In Dogri language, the conversion of masculine nouns having suffix (आ) into their corresponding plural forms can be done by replacing the suffix with (ए).

Examples:

SINGULAR FORM	PLURAL FORM
बोतल	बोतले

In feminine noun आं (āṃ) is added at the suffix, e.g.,

कताब	कताबां

- ***Pronoun:***

The words used in place of Noun and substitute a noun or Noun phrase are called pronoun. Some Dogri pronouns are: आपूं, असें, इक- दुए, कोहदा

- ***Adjectives:***

Words that describe nouns. Adjectives are used before or after the noun.

Inflection of adjectives by gender

In Dogri language, adjectives in the masculine form mainly end with आ and in order to change the masculine adjective to its corresponding feminine form, we need to replace "आ" with "ई"

Examples:

SINGULAR MASCULINE	SINGULAR FEMININE
सयानाजागत	सयानीकुड़ी

Inflection of adjectives by number

In Dogri language, the singular masculine adjective ends by आ/ā. To alter it into the plural form, we need to replace it with ए/e

Example:

SINGULAR MASCULINE	PLURAL MASCULINE
कालाघोड़ा	कालेघोड़े

In Dogri Language, the feminine adjectives in the singular form that end with ई/ī is converted to its plural form by adding यां(yāṃ)

Example:

SINGULAR FEMININE	PLURAL FEMININE
कालीघोड़ी	कालीघोड़ियां

- ***Verbs:***

Verb inflection for gender

In verbs in Dogri language, the gender in the masculine form have "आ"assuffix, and the verbs in feminine form have "ई"as suffix which is shown below:

Example:

MASCULINE	FEMININE
जागतरोआदा	कुड़ीरोआदी

Verb inflection for number

In verb of number, singular is denoted by "दा" whereas plural is denoted by "दे" in suffix.

Example:

SINGULAR	PLURAL
फुलखिलदाऐ	फुलखिलदेन

- ***Compound verbs***

Compound verbs are the union of the main verb in addition to an auxiliary verb. Typically, the auxiliary verb is a word that complements sense to the main verb and drops its individual self-determining meaning.

Example:

कम्मकरनापौना– the word "पौना" is the compound verb. "कर" is mainly the root word of the main verb and पौना are the auxiliary verbs

- ***Adverb**:*

Adverbs are mainly associated with the verb and describe the verbs.

Example तौलेतौलेचलो – adverb here in the sentence is तौलेतौले

FOLLOWING DOGRI RULES ARE USEFUL TO RECOGNIZE VARIOUS TAGS:

1. There is a high probability that noun follows an adjective.

Example: ओइकसच्चादेसवासीऐ।

सच्चा is an adjective and देसवासी is a noun.

2. There is a probability that a post position follows a noun.

Example: उन्पानीचबट्टासुटेया।

पानी -noun and च -postposition.

3. There is a high probability of verb following a noun.

Example: ओरुट्टीखादा।

रुट्टी- noun, खा- verb

4. There is a probability that the word preceding 'दा','दी', 'न', 'ने', 'दे' is a verb.
5. In Dogri language, the main verb is followed by an auxiliary verb.

Example: कब्जेकरियै।

कब्जे- main verb, करियै- auxiliary verb.

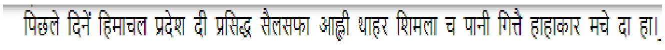
V. PROPOSED WORK

The proposed work in this research is designing and implementation of part of speech tagging system that closely tagsDogri words that are input to the system.The paper covers the initial stage of POS taggingwhich include acceptance of the relevant (Dogri) text, tokenization and stemming of the input text. The input data is the Dogri (Devanagari script) text collected from social media.

The Dogri text data which is saved in the form of a text file is an input to the system. The result of which is shown in the following Fig. 2.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

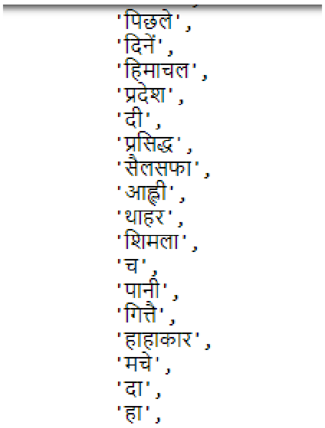
A) Input Dogri text:



पिछले दिनें हिमाचल प्रदेश दी प्रसिद्ध सैलसफा आह्ली थाहर शिमला च पानी गित्तै हाहाकार मचे दा हा।

Fig. 2. Dogri text

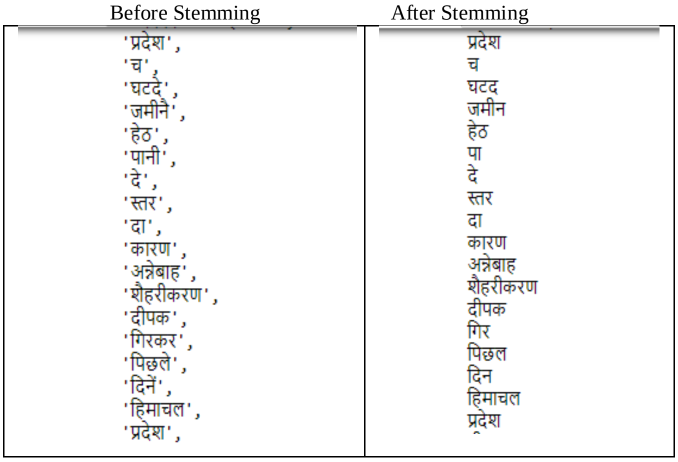
B) Tokenization of text: The accepted Dogri corpus is converted into tokens as per delimiter. The result is shown as in the Fig. 3.



'पिछले',
'दिनें',
'हिमाचल',
'प्रदेश',
'दी',
'प्रसिद्ध',
'सैलसफा',
'आह्ली',
'थाहर',
'शिमला',
'च',
'पानी',
'गित्तै',
'हाहाकार',
'मचे',
'दा',
'हा',

Fig. 3. Tokenization of Dogri Text

C) Stemming of text: Stemming is defined as the process of converting the words similar in morphology to their respective root words by removing the ending or suffixes from the words. Like any other language,Dogri language is rich in morphology. A set of rules are made considering the language to accomplish the process of stemming. The resulting words after performing stemming is not necessarily a root word. To overcome this problem a database of the root words is needed to be maintained to provide the relevant output[22]-[24]. The result of the Dogri text before and after stemming is shown below in Fig. 4.



Before Stemming	After Stemming
'प्रदेश',	प्रदेश
'च',	च
'घटदे',	घटद
'जमीनै',	जमीन
'हेठ',	हेठ
'पानी',	पा
'दे',	दे
'स्तर',	स्तर
'दा',	दा
'कारण',	कारण
'अन्नेबाह',	अन्नेबाह
'शैहरीकरण',	शैहरीकरण
'दीपक',	दीपक
'गिरकर',	गिर
'पिछले',	पिछल
'दिनें',	दिन
'हिमाचल',	हिमाचल
'प्रदेश',	प्रदेश

Fig. 4. Stemming

VI. CONCLUSION AND FUTURE WORK

Parts of speech tagger assigns appropriate tags to the words in the sentences. The process involves study of the techniques used to assign appropriate tags to words. Related study carried out by various researchers in the languages having Devanagari script has been done in the paper. The study helps us in havinga better knowledge about the language and the techniques used for tagging.

The paper also includes the study of various grammatical and inflectional analysis of Dogri language. Rules for the identification of various POS tags for Dogri language is also discussed with the help of appropriate examples. Initial step of POS tagging is acceptance of relevant script and chunking of sentences into words and converting morphologically rich words into their root words which is performed and corresponding results are also shown.

Dogri language contains number of ambiguous words. Ambiguity of the Dogri words will be removed and hence POS tagging will be accomplished in the future.

REFERENCES

1. S. Kumar, “Developing POS Tagset for Dogri,”*Language in India www.languageinindia.com,* vol.18, no. 1, 2018.
2. S. Bhatta, K. Parmara and M. Patelb, “Sanskrit Tag-sets and Part-Of-Speech Tagging Methods- A Survey,” I*nternational Journal of Innovative and Emerging Research in Engineering (IJIERE),* vol. 2, 2015
3. S. Rathod and S. Govilkar, “Survey of various POS tagging techniques for Indian regional languages,” *International Journal of Computer Science and Information Technologies (IJCSIT),* vol. 6, pp. 2525-2529,2015
4. N. Joshi, H. Darbari and I. Mathur, “HMM BASED POS TAGGER FOR HINDI,”*The Second International conference on Parallel, Distributed Computing technologies and Applications (PDCTA),* pp. 341-349,2013
5. A. Ekbal, R. Haque, and S. Bandyopadhyay, “Maximum Entropy Based Bengali Part of Speech Tagging,” *Advances in Natural Language Processing and Application, Research in Computer Science (RCS)Journal (33),* pp.67-78,2008.
6. J. Lafferty, A. McCallum and F. C.N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001),* pp. 282-289, 2001.
7. S. Singh, K. Gupta, M. Shrivastava and P. Bhattacharyya, “Morphological Richness Offsets Resource Demand- Experiences in Constructing a POS Tagger for Hindi,” *Proceedings of the COLING/ACL -Association for Computational Linguistics,* pp. 779–786, July 2006.
8. N. Mishra and A. Mishra, “Part of Speech Tagging for Hindi Corpus,” *Int. Conf. Commun. Syst. Netw. Technol.*, pp. 554–558, 2011.
9. N. Garg, V. Goyal, and S. Preet, “Rule Based Hindi Part of Speech Tagger.,” *COLING (Demos)*, vol. 2, no. December, pp. 163–174, 2012.
10. A. Dalal, K. Nagaraj, U. Sawant, and S. Shelke, “Hindi Part-Of-Speech Tagging and Chunking: A Maximum Entropy Approach,” *Proceeding NLPAI Mach. Learn. Compet.*, pp. 1–4, 2006.
11. R. Kumar and S. S. Shekhawat, “PARTS OF SPEECH TAGGING FOR HINDI LANGUAGES USING HMM,” *Int. J. Sci. Res.*, vol. 7, no. 4, pp. 42–44, 2018.
12. K. Mohnot, N. Bansal, S. . Singh, and A. Kumar, “Hybrid approach for Part of Speech Tagger for Hindi language,” *Int. J. Comput. Technol. Electron. Eng.*, vol. 4, no. 1, pp. 25–30, 2014.
13. P. K. Dwivedi and P. K. Malakar, “Hybrid Approach Based POS Tagger for Hindi Language,” *Int. J. Res. Stud. Comput. Sci. Eng.*, vol. 4840, no. August, pp. 63–68, 2015.
14. V. Khicha and M. Manna, “Part-of-Speech Tagging of Hindi Language Using Hybrid Approach,” *Int. J. Eng. Technol. Sci. Res.*, vol. 4, no. 8, pp. 737–741, 2017.
15. P. Bagul, A. Mishra, P. Mahajan, M. Kulkarni, G. Dhopavkar, and M. T. Scholar, “Rule Based POS Tagger for Marathi Text,” vol. 5, no. 2, pp. 1322–1326, 2014.
16. S. Rathod, S. Govilkar, and S. Kulkarni, “Part of Speech TAGGER for MARATHI Language,” *Sixth International Conference on Computational Intelligence and Information*

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Technology(CIIT), 2016, pp. 131–138.

17. J. Singh, N. Joshi, and I. Mathur, "Development of Marathi part of speech tagger using statistical approach," in *Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2013*, 2013, pp. 1554–1559.
18. Nita V. Patil, "POS Tagging for Marathi Language using Hidden Markov Model,"*International Journal of Computer Sciences and Engineering IJCSE*, vol.6, no. 1, pp. 2347-2693, 2018.
19. N. Tapaswi, B. Santorini, and M. A. Marcinkiewicz, "Treebank Based Deep Grammar Acquisition and Part- Of-Speech Tagging for Sanskrit Sentences."
20. A. Yajnik, "Part of Speech Tagging Using Statistical Approach for Nepali Text," *Int. J. Cogn. Lang. Sci.*, vol. 11, no. 1, pp. 76–79, 2017.
21. P. Dubey, "The Hindi to Dogri machine translation system : grammatical perspective," *Int. J. Inf. Technol.*, 2018.
22. V. Gupta, "Hindi Rule Based Stemmer for Nouns," *Int. J. Adv. Res. Comput. Softw. Eng.*, vol. 4, no. 1, pp. 62–65, 2014.
23. B. P. Pande, P. Tamta, and H. S. Dhami, "A Devanagari Script based Stemmer," *Int. J. Comput. Linguist. Res.*, vol. 5, no. 4, pp. 119–130, 2014.
24. A. Pimpalshende and A.R. Mahajan, "Extraction of Root Words Using Morphological Analyzer for Hindi Text," *International Journal of Soft Computing*, vol.13, no. 5, pp 134-138, 2018.

AUTHORS PROFILE

Shivangi Duttareceived B.E. in Computer Science and Engineering from Mahant Bachittar Singh College of Engineering & Technology, Jammu, India and is currently pursuing her M.Tech. degree in Computer Science and Technology from Central University of Jammu, Jammu, India. Her research interests include Natural Language Processing and Big Data analytics.

Bhavna Arorareceived the Bachelor's degree from Kurukshetra University, India and Master's Degree from IMT Ghaziabad, India. She has done her Ph.D. in Computer Science & IT from University of Jammu in 2011. She is currently working as an Assistant Professor with Department of Computer Science & IT, Central University of Jammu, and has over 21 years of work experience as an academician and also successfully supervised Ph.D. and M. Tech students. She has also published three book chapters in Springer and thirty-seven research papers in national/international journals. She has also received UGC-Startup Grant for a Project in cyber-crime. She has attended International conference in Thailand which was funded by DST. She has also attended many national and international conferences. She is member of IEEE, ACM, ISTE, SIE and member of Board of reviewers, Technical Program committee in various International conferences. Her research interest's include Information Security, Data Mining and Natural Language Processing.