Measuring the Similarity between the Sanskrit Documents using the Context of the Corpus

Jatinderkumar R. Saini[1], Prafulla B. Bafna[2]
Symbiosis Institute of Computer Studies and Research
Symbiosis International Deemed University
Pune, India

***Abstract*—Identifying the similarity between two documents is a challenging but important task. It benefits various applications like recommender systems, plagiarism detection and so on. To process any text document one of the popularly used approaches is document term matrix (DTM). The proposed approach processes the oldest, untouched, one of the morphologically critical languages, Sanskrit and builds a document term matrix for Sanskrit (DTMS) and Document synset matrix Sanskrit (DSMS). DTMS uses the frequency of the term whereas DSMS uses the frequency of synset instead of term and contributes to the dimension reduction. The proposed approach considers the semantics and context of the corpus to solve the problem of polysemy. More than 760 documents including Subhashitas and stories are processed together. F1 Score, precision, Matthews Correlation coefficient (MCC) which is the most balanced measure and accuracy are used to prove the betterment of the proposed approach.**

Keywords—Cosine; dimension reduction; sanskrit; synset; matthews correlation coefficient

I. Introduction

The degree of matching between two text pieces based on their statistics as well as semantics is termed as the similarity between text pieces [24]. Statistics of the document means the length of the document, tokens present in the document, etc. The semantics of the document means the understanding meaning of the words present in the document. These documents/ text pieces can be in the form of the word, pdf and so on. There are various measures to calculate the similarity between the two documents. Jaccard, cosine similarity and so on. Cosine similarity is independent of the statistics of the document. Cosine similarity calculates the cosine value of the angle between two vectors. These vectors comprise the frequency of words in a multi-dimensional plane. Each word present in the document represents the dimension/feature [10]. Thus the orientation of the text document gets captured by cosine similarity instead of the magnitude only. Cosine similarity [11] is better than other similarity measures eg. Euclidean distance. The value near to '1' indicates the documents are most similar. Cosine value is always between '0' and '1'. Calculating cosine similarity between English, Hindi [12-14], Marathi [15] text documents [5] is a common task but processing Sanskrit language [33,30,28] and its morphological analysis [35] are critical tasks, as a result finding out the mapping between Sanskrit language texts is challenging. Sanskrit is assumed to be the mother of every language. Panini has introduced this grammar rich language before 2500 years ago. The Sanskrit language has been the traditional means of communication in Hinduism, Jainism, Buddhism, and Sikhism, still, Sanskrit text mining is an untouched area. Several kinds of literature are available in Sanskrit for eg. stories, subhashits and so on. A subhashita (Sanskrit: 'सुभाषित') can be explained as a legendary kind of Sanskrit concise poems to communicate the message of advice, aphorism and so on. Generally, Sanskrit subhashit or stories are related to all aspects of life. Subhashitas are significant in Indian traditional education and are used to teach values like truthfulness, courage, etc. which are applicable for each phase of life righteousness.

To extract any information from Sanskrit text, various techniques are used. DTMS is one of the techniques using which different operations can be carried out on Sanskrit corpus. Sanskrit documents are placed in rows and significant terms are placed in columns. The entry in the matrix represents the number of times the particular Sanskrit term occurred in the document. The significance of the term is decided based on the frequency of the term. The semantics of the term is considered in DSMS. DSMS uses synset groups in which semantically similar tokens are grouped. Instead of considering term frequency, synset group frequency is considered. It facilitates to solve the polysemy problem means one word used with different senses.

Dimension reduction means the removal of unnecessary features. Several methods are available for dimension reduction like principal component analysis, latent semantic analysis, etc. In the text processing [1][2], Different NLP tasks [5-9] are carried out like removal of stop words [3][4] [32][29] results in dimension reduction. Before stopwords, removal tokenization needs to be carried out for example 'ततो मक्षिका उड्डिय गता.' meaning 'The fly flew away' In this Sanskrit statement 'ततो' meaning 'from there' is removed after separation of tokens. Tokens of the sentence are 'ततो', 'मक्षिका', 'उड्डिय', 'गता', '.'. Lemmatization coverts words into their meaningful root form [31]. On the formulated document synset matrix, several applications could be built like plagiarism detection, document clustering, etc. Till now no research is carried out to find Sanskrit document similarity using semantics and context.

To evaluate machine learning algorithms different parameters are available. Eg. Precision, accuracy, Matthews's Correlation coefficient. Matthews Correlation coefficient (MCC) is a quality measure for binary classification. It is a balanced measure because it considers all false and true positives and negatives. Precision is defined as the ratio of relevant documents to the retrieved documents. Accuracy is the ratio of a number of correctly classified documents to the total number of input documents.

The arrangement of the paper is as mentioned. The existing work carried out by other researchers is written as a literature review in the second section. The research methodology is stated in the third section. The fourth section depicts results and discussions and conclusions are presented as the fifth section to end the paper.

The proposed approach is unique because

1) It constructs synset for Sanskrit corpus.
2) It extracts the context of the corpus using semantics.
3) It builds concept space for the corpus.

II. Literature Review

Sanskrit is inflectionally strong language and the correct morphological analyzer is required to process Sanskrit text. In spite of being identified as a good analyzed language, morphological analysis of Sanskrit is a challenging task. A morphological analyzer is built which covers wider aspects of language and covers the complexity of words. The applications of the analyzer are used for Sandhi splitting, search engine, etc. [16] Modularity is used while building the morphological analyzer. Other modules like spell check etc. can be added easily. Some modules in the proposed approach are not working, but these modules are based on the grammar which can be easily handled manually. Accuracy of the analyzer is calculated and improvements are discussed [17]. One of the important and unique features of the Sanskrit language is focused known as a dual case. Confusion between plural and dual can be easily removed if grammar rules are followed. Indian treasure is explored to state its different common applications. Various challenges related to NLP along with features of NLP are explored. There are two aspects of similarity implicit and explicit, covering both of these aspects are necessary but critical for finding out the similarity between two texts. This is also termed as the NLP challenge. Different levels that are document or paragraph or sentence or word affect differently and that is why need to be considered to calculate the similarity between two text documents [18]. Sanskrit needs to be handled differently as its morphology is strong. Short texts of Sanskrit [34] are processed semantically using the morphological based approach. Words sematic membership in the sentence is considered assuming that each word has a different significance in the sentence. Ranking of the tokens is carried out using an adaptive measuring algorithm. Sanskrit complex text was processed [19] WordNet is a lexical database that includes vital components like Glosses etc. necessary to identify different NLP features. Domain ontologies could be built using WordNet also semantic relations, polysemy problems can be handled using Wordnet. Sanskrit Wordnet can be used to solve issues like word sense disambiguation (WSD) using gloss. Different techniques that have developed Sanskrit gloss are surveyed to depict different pattern types to solve WSD [20]. Text summarization is carried out to avoid the efforts and time required to read the document. Natural language processing tools are easily available for the English language. Very less work is explored for Indian languages because of the availability of fewer resources. Existing summarization techniques for Indian languages are surveyed and opportunities for research are discussed [21]. Sentiment analysis is useful for researchers to identify the views of individuals for various services, products, etc. The Internet has allowed exploring NLP tasks by providing a huge amount of data. Machine learning techniques facilitate to provide analysis of this data. The need for domain experts is been reduced for verifying the results due to deep learning algorithms. The sentiment analysis can be done using deep learning techniques and effectively than traditional machine learning techniques for resource-scarce languages. Major challenges like word embedding, ontology building are surveyed which can act as a stepping stone for NLP research [22]. Identifying the language present in the given text is called Language Identification (LI). This could be the important initial preprocessing step to carry out NLP tasks Automatic language recognition is the challenging process. India is a country of multiple languages and there is good scope for language identification problem, It will bridge the digital gap Indian and other languages. Hindi and Sanskrit text is separated using the N-gram approach. The languages which share the same scripts, the technique can be applied [23].

III. Research Methodology

The proposed approach is already being tested for Marathi and termed as DSMM [26] and are now experimenting with Sanskrit. Sanskrit parsers are not developed fully and still are in research phase. Same way the udmodel which is being used in the proposed approach a few times gives incorrect results for a few NLP tasks it impacts other sequential NLP processes. Knowledge of Sanskrit language is necessary and manual intervention, expertise is involved to get correct results unlike Document Synset Matrix Marathi [26]. Also, MCC is used to validate the results which were not used to asses DSMM. Also, the literature available for Sanskrit text similarity is very limited as compared to Marathi text similarity.

R programming is used which provides different packages like tm, quanteda, libraries like udpipe which provides a morphological analysis of the text, for example, identifying the part of speech (PoS), tense, gender and so on. Also functions like documentTermMatrix_tfidf(), udpipe_annotate (udmodel_ Sanskrit) and so on are provided by R programming. Wordnet provided by CFILT, IIT Bombay [27] is used to identifying the semantic relationship between different tokens. Fig. 1 shows the research methodology steps.

A. Collection of a Dataset, Creation of Corpus and Preprocessing

Sanskrit corpus is not available. The first step is to collect Sanskrit subhashits and stories [25]. The corpus belongs to 340 subhashits and 421 stories. The data is about 15 MB. Different preprocessing steps are carried out. In the first step, tokenization is carried out. Total tokens are 1, 11,123. Stop words are removed, total stop words counted are 34,287. Lemmatization is carried out on the remaining 76836 words.

Unique terms/ tokens are generated which are 56,198. The tokens having similar meanings are grouped and termed as synset groups. Synsets groups are formed and a total number of terms in the form of synset groups are 46,345.

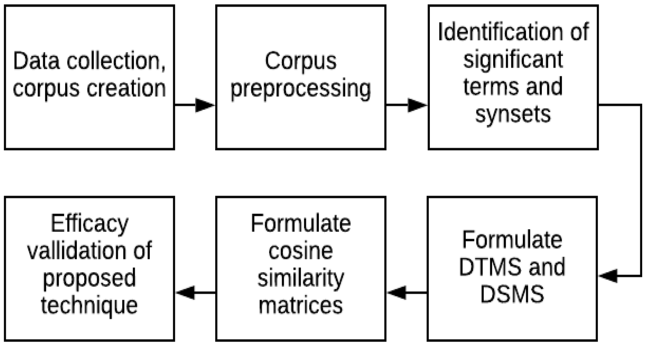
Data collection, corpus creation
Corpus preprocessing
Identification of significant terms and synsets
Formulate DTMS and DSMS
Formulate cosine similarity matrices
Efficacy vallidation of proposed technique

Fig. 1. Steps in Research Methodology.

B. Identification of Significant Terms and Formulation of DSMS

The frequency of each term is calculated and then the group frequency of each synset group is calculated. Group frequency means, the addition of the frequencies of all terms present in one synset group. Significant synset groups are identified based on a threshold. The synset groups having more than 75% threshold are considered while building DSMS. It means that, if maximum synset frequency is 100 then all those synsets having frequency more than 75, are considered as significant synsets/terms. Documents (subhashits and stories) occupy the rows of the matrix and synsets act as columns in the matrix. The entry in the table/matrix shows the frequency of that particular term/synset in the corresponding document. Due to synset formation, dimensions are reduced, also the context of the term gets considered. For example ‘पक्षे’ means in favor of as well as on the other hand, if ‘कृपा’ (kripa) and ‘पक्षे’ (pakshe) come together, it means its meaning is ‘in favor of’ and thus the semantics of the words gets identified. This problem of having multiple meanings of a single word is also known as word sense disambiguation or polysemy which is solved by identifying the sense in which it is used in the corpus. Also with the help of other identified significant terms / synsets, the context of the corpus is understood.

C. Calculate Document Similarity using Cosine Measure

Thus the proposed technique is different than the traditional one which considers only significant terms as column heads and formulates DTMS. DTMS not only lacks to identify semantics and context of the corpus but also dimensions are more than DSMS. On DSMS cosine measure is applied and the similarity of the document is calculated. Two documents are said to be similar if they have cosine measure greater than 0. 65 that is the threshold of cosine measure considered for document similarity is 65%. The value of the threshold is considered after evaluating thresholds ranging from 10% to 90%, precision calculated WAS MAXIMUM FOR 65% THRESHOLD.

IV. RESULTS AND DISCUSSIONS

Table I presents the morphological analysis of the entire corpus. It includes all the documents with their index with respect to the corpus. The paragraph code is assigned for every document. It represents the position of paragraph in the document. The sentence rank of each sentence of a paragraph is represented by sentence code, for example in the first documents, in the first paragraph, the second sentence is shown at the 4th Sr. No. The table also depicts lemma, part of speech and other information like gender, the tense of verbs, singular, plural and so on. For example for term ‘असौ’, it’s lemma is ‘अदस्’, its part of speech is a pronoun and its gender is masculine whereas ‘प्राह’ is a verb which is singular, representing the first person, having past tense and active voice.

TABLE I. MORPHOLOGICAL ANALYSIS OF SANSKRIT TEXT

Sr. No	Document code	Paragraph code	Sentence code	Sentence	Token code	token	lemma	Part of speech	Other information
1	doc1	1	1	ततः असौ प्राह	1	ततः	ततस्	ADV	*NA*
2	doc1	1	1	ततः असौ प्राह	2	असौ	अदस्	PRON	Gender=Masc
3	doc1	1	1	ततः असौ प्राह	3	प्राह	प्राह	VERB	|Number=Sing|Person=1|Tense=Past|Voice=Act
4	doc1	1	2	क्षत्रियस्य तिस्रः भार्या धर्मम् भवन्	1	क्षत्रियस्य	क्षत्रिय	NOUN	Case=Gen|Gender=Masc|Number=Sing
5	doc1	1	2	क्षत्रियस्य तिस्रः भार्या धर्मम् भवन्	2	तिस्रः	त्रि	NUM	Gender=Fem|Number=Plur
9	doc2	1	1	ि	1	ि	ि	SP	*NA*

Instead of using any stop word removal technique, only adjectives, adverbs, nouns, and verbs are selected. It resulted in the removal of stop words in the form of discarding numbers, special characters, and pronouns, etc. Also instead of using stemming, lemmatization is used and unique lemmas/terms are identified. The frequency of each term is identified. For example the frequency of 'ध्यानम्' is 30. Group frequency is calculated by adding up the frequency of similar terms 'वृत्त', 'वार्ता' are similar terms and are grouped called as a synset group. Table II shows the frequency of 56198 unique terms in the corpus. Table III shows the frequency of 46, 345 synset groups. In DTMS 'वृत्त' having frequency 4 is not considered as significant terms because its frequency is less than the threshold frequency. But it is an important word. This importance is considered in DSMS as it adds up the frequency and the particular words in synset groups are treated as significant words.

TABLE II. TOKENS AND FREQUENCY FOR DTMS

Sr.No	Token	Frequency
1	ध्यानम्	30
2	वार्ता	26
3	स्तोत्रम्	23
4	फाल्गुन	21
5	वरदः	15
6	सत्यः	4
7	वृत्त	4
		.

TABLE III. TOKENS AND FREQUENCY USING DSMS

Sr. No.	Synset	Frequency
1	वृत्त, वार्ता	30
2	क्षत्रिय, नृप	28
3	ध्यानम्	10
		..
4	स्तोत्रम्	10

TABLE IV. DOCUMENT TERM MATRIX SANSKRIT

Document	क्षत्रिय	बार्या	धर्म	वैश्	सुत	नृप
doc1	1	2	1	1	1	1
doc 2	0	1	1	0	0	0
doc 3						
doc 4						
doc 5	0	0	1	0	0	0

Table IV shows DTMS for a sample of four documents. The frequent terms are placed as column heads and documents are placed in rows. The entry in the matrix shows the frequency of that term in the document. For example, 'धर्म' has occurred 1 time in document 1. Table V shows the DSMS for the sample of five documents in the corpus. Synset groups are placed in a column, unlike DTMS. It's clear that not only dimensions are reduced but also semantics is being considered that is 'क्षत्रिय, नृप' having the same meaning are grouped while formulating DSMS. In fact, by looking into other significant terms/synset groups context of the corpus is also.

Table VI presents the document similarity matrix using the cosine measure using DSMS. According to the applied 65% threshold, all the documents pairs having a measure of more than 65% are termed as similar. In case of documents, D1 and D2 having measure 0.75 are observed to be similar and D1 and D3 having measure 0.43 is termed to be non-similar.

A confusion matrix is constructed for the DSMS technique and is presented in Table VI. Different evaluation parameters such as F1 score, precision, Matthews Correlation Coefficient, and accuracy are calculated for both the techniques that are DSMS and DTMS and shown in Table VII. All the parameters of the proposed technique (DSMS) produce better results than the existing technique that is DTMS (Table VIII). F1 score, precision, and accuracy are improved by 0.2 measures and Matthews Correlation Coefficient is improved by 0.1 measure.

TABLE V. DOCUMENT SYNSET MATRIX SANSKRIT

Document	क्षत्रिय, नृप	बार्या	धर्म	वैश्	सुत
doc1	2	2	1	1	1
doc 2	0	1	1	0	0
doc 3					
doc 4					
doc 5	0	0	1	0	0

TABLE VI. COSINE SIMILARITY MATRIX FOR THE CORPUS

Documents	D1	D2	D3		D761
D1	1	0.75	0.45		0.23
D2	0.75	1	0.35		0.67
D3	0.45	0.35	1		0.88
D761	0.23	0.67	0.88		1

TABLE VII. CONFUSION MATRIX

Actual/Predicted	Similar documents	Non-similar documents	Total
Similar documents	431	94	525
Non-similar documents	45	191	236
Total	476	285	761

TABLE VIII. EVALUATION PARAMETERS FOR BOTH TECHNIQUES

Sr. No	Parameter	DSMS	DTMS
1	F1 Score	0.86	0.63
2	Precision	0.82	0.61
3	Matthews Correlation Coefficient	0.60	0.51
4	Accuracy	0.81	0.61

V. CONCLUSIONS AND FUTURE WORK

DSMS is proposed which identifies the semantics of the term. Synset groups are formed to achieve dimension reduction. Total terms present in the DSMS contribute to identify the context of the corpus. Thus polysemy problem gets solved for the Sanskrit corpus processing which is strongly inflectional language. F1 Score, precision, accuracy and Matthews Correlation Coefficient (MCC) are used to prove the betterment of the technique. F1 score and precision are improved by 0.2 measure and Accuracy and MCC are improved by 0.1 measure. A total of 761 documents are processed including Subhashits and stories. The same approach can be extended for other regional languages.

REFERENCES

[1] Bafna P.B., Saini J.R., 2020, “An Application of Zipf's Law for Prose and Verse Corpora Neutrality for Hindi and Marathi Languages”, International Journal of Advanced Computer Science and Applications, 11(3).

[2] Bafna P.B., Saini J.R., 2020, “Marathi Text Analysis using Unsupervised Learning and Word Cloud”, International Journal of Engineering and Advanced Technology,9(3).

[3] Rakholia, R. M., & Saini, J. R. (2015, March). The design and implementation of diacritic extraction technique for Gujarati written script using Unicode Transformation Format. In 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-6). IEEE.

[4] Kaur, J., & Saini, J. R. (2016). POS Word Class Based Categorization of Gurmukhi Language Stemmed Stop Words. In Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2 (pp. 3-10). Springer, Cham.

[5] Rakholia, R. M., & Saini, J. R. (2017). Automatic Language Identification and Content Separation from Indian Multilingual Documents Using Unicode Transformation Format. In Proceedings of the International Conference on Data Engineering and Communication Technology (pp. 369-378). Springer, Singapore.

[6] Saini, J. R., & Rakholia, R. M. (2016). On continent and script-wise divisions-based statistical measures for stop-words lists of international languages. Procedia Computer Science, 89, 313-319.

[7] Rakholia, R. M., & Saini, J. R. (2017). A rule-based approach to identify stop words for Gujarati language. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 797-806). Springer, Singapore.

[8] Rakholia, R. M., & Saini, J. R. (2016). Lexical classes based stop words categorization for Gujarati language. In 2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Fall) (pp. 1-5). IEEE.

[9] Raulji, J. K., & Saini, J. R. (2017, January). Generating Stopword List for Sanskrit Language. In 2017 IEEE 7th International Advance Computing Conference (IACC) (pp. 799-802). IEEE.

[10] Kaur, J., & Saini, J. R. (2020). Designing Punjabi Poetry Classifiers Using Machine Learning and Different Textual Features. International Arab Journal of Information Technology, 17(1), 38-44.

[11] Rakholia, R. M., & Saini, J. R. (2017). Information Retrieval for Gujarati Language Using Cosine Similarity Based Vector Space Model. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 1-9). Springer, Singapore.

[12] Venugopal-Wairagade, G., Saini, J. R., & Pramod, D. (2020). Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List. arXiv preprint arXiv:2002.00171.

[13] Bafna P.B., Saini J.R.,2019, “Hindi Multi-document Word Cloud based Summarization through Unsupervised Learning”, 9th International Conference on Emerging Trends in Engineering and Technology on Signal and Information Processing (ICETET-SIP-19), Nagpur, India.

[14] Bafna P.B., Saini J.R., 2019, “Scaled Document Clustering and Word Cloud based Summarization on Hindi Corpus”, 4th International Conference on Advanced Computing and Intelligent Engineering, Bhubaneshwar, India, in press with Springer.(December).

[15] Bafna P.B., Saini J.R., 2020, BaSa: A Context based Technique to Identify Common Tokens for Hindi Verses and Proses, in press.

[16] Murali, N., Ramasreee, D. R., & Acharyulu, D. K. (2014). Kridanta Analysis for Sanskrit. Int. Journal on Natural Language Computing, 3(3), 33-49.

[17] Jha, Deeptanshu, Rashmi Jha, and Varun Varshney. "Natural Language Processing and Sanskrit." International Journal of Computer Engineering & Technology 5, no. 10 (2014): 57-63.

[18] Bharati, A., Kulkarni, A. P., & Sheeba, V. (2006, April). Building a wide coverage Sanskrit Morphological Analyser: A practical approach. In The First National Symposium on Modelling and Shallow Parsing of Indian Languages, IIT-Bombay.].

[19] Shevgoor, S. K. (2017, June). A Morphological Approach for Measuring Pair-Wise Semantic Similarity of Sanskrit Sentences. In Natural Language Processing and Information Systems: 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings (Vol. 10260, p. 162). Springer.

[20] Kulkarni, Malhar, Irawati Kulkarni, Chaitali Dangarikar, and Pushpak Bhattacharyya. "Gloss in sanskrit wordnet." In International Sanskrit Computational Linguistics Symposium, pp. 190-197. Springer, Berlin, Heidelberg, 2010.

[21] Verma, P., & Verma, A. (2020). Accountability of NLP Tools in Text Summarization for Indian Languages. Journal of Scientific Research, 64(1).

[22] Nankani, Hitesh, Hritwik Dutta, Harsh Shrivastava, PVNS Rama Krishna, Debanjan Mahata, and Rajiv Ratn Shah. "Multilingual Sentiment Analysis." In Deep Learning-Based Approaches for Sentiment Analysis, pp. 193-236. Springer, Singapore, 2020.

[23] Sreejith, C., Indu, M., & Raj, P. R. (2013, July). N-gram based algorithm for distinguishing between Hindi and Sanskrit texts. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-4). IEEE.

[24] Naik, R. R., & Landge, M. B. (2019). Plagiarism Detection in Marathi Language Using Semantic Analysis. In Scholarly Ethics and Publishing: Breakthroughs in Research and Practice (pp. 473-482). IGI Global.

[25] www.sanskritebooks.org › category › sanskrit › stories available on 21-04-2020.

[26] Bafna P.B., Saini J.R., 2020,Marathi Document-Similarity Measurement using Semantics-based Dimension Reduction Technique, International Journal of Advanced Computer Science and Applications 11(4).

[27] ww.cfilt.iitb.ac.in/wordnet/webswn/wn.php available on 21-04-2020.

[28] Raulji J.K., Saini J.R. (2016) "Sanskrit Machine Translation Systems: A Comparative Analysis", International Journal of Computer Application, 136(1):1-4.

[29] Raulji J.K., Saini J.R. (2016) "Stop-Word Removal Algorithm and its Implementation for Sanskrit Language", International Journal of Computer Application, 150(2):15-17.

[30] Raulji J.K., Saini J.R. "A Rule Based Architecture for Sanskrit to Gujarati Machine Translation System", Proceedings of International Conference on Emerging Trends in Engineering, Science and Technology (ICRISET-2018), Changa, India, in press.

[31] Raulji J.K., Saini J.R. (2019) "Sanskrit Lemmatizer for Improvisation of Morphological Analyzer", Journal of Statistics & Management Systems, Taylor and Francis, 22(4):613-625.

[32] Raulji J.K., Saini J.R. (2020) "Sanskrit Stopword Analysis through Morphological Analyzer and its Gujarati Equivalent for MT System", Proceedings of International Conference on ICT for Sustainable Development (ICT4SD-2019), Panaji, India, 93:427-433.

[33] Raulji J.K., Saini J.R. (2020) "Bilingual Dictionary for Sanskrit – Gujarati MT Implementation", Proceedings of International Conference on ICT for Sustainable Development (ICT4SD-2019), Panaji, India, 93:463-470.

[34] Raulji J.K., Saini J.R. (2020) "Sanskrit-Gujarati Constituency Mapper for Machine Translation System", Proceedings of IEEE Bombay Section Signature Conference (IBSSC-2019), Mumbai, India, 1-8.

[35] Saini J.R., Raulji J.K. (2020) "Peer Analysis of “Sanguj” with Other Sanskrit Morphological Analyzers", Proceedings of 2nd International Conference on Computing Analytics and Networking (ICCAN-2019), Bhubaneshwar, India, 1119:65-73.