Information System and Smart City 2023; 3(1): 481.
Original Research Article

NLP-reliant Neural Machine Translation techniques used in smart city applications

Ritesh Kumar Dwivedi[1,*], Parma Nand[1], Om Pal[2]

[1] *Department of Computer Science and Engineering, Sharda University, Greater Noida 201306, India*

[2] *Department of Computer Science, University of Delhi, New Delhi 201306, India*

*** Corresponding author:** Ritesh Kumar Dwivedi, ritesh.dwivedi@nic.in

ARTICLE INFO

Received: 13 January 2024
Accepted: 5 February 2024
Available online: 2 April 2024

doi: 10.59400/issc.v3i1.481

ABSTRACT: For smart city applications, Neural Machine Translation (NMT) methods based on Natural Language Processing (NLP) are crucial as they facilitate information sharing and communication among diverse populations. NLP techniques are used in many domains related to smart cities, such as development and research, business, industries, media, healthcare, and residences and communities. The majority of people in India communicate using their regional languages. The majority of applications used by users in smart cities will mostly accept English as input. These people will be able to interact with these smart city devices in their native tongues more effectively with the help of effective machine translation. Just 10% of Indians use English as their primary language of communication; there are 22 official regional languages in India. So, there is requirement of better machine translation using Natural language processing (NLP). Natural language processing for Indian regional languages has a very long way to go until it surpassing the abilities of existing rich NLP applications and techniques for English language. Machine Translation is technique of Natural Language Processing (NLP) which provides better inter-lingual communication. For low resourced Indian languages effective machine translation systems became important for establishing proper communication. Machine Transliteration is a technique to convert source language into target language using machine. The developed system takes English language as input and then applies machine translation techniques to translate the source language into multiple languages using trained RNN model and multi-lingual search model which search the input word across all the datasets and generate the output into other Indian languages such as Hindi, Tamil. Our approach achieves top performance for English-Hindi language pair and comparable results for other cases.

KEYWORDS: NLP; Recurrent Neural Network (RNN); Neural Machine Translation (NMT)

1. Introduction

Various ILNMT architectures for various Indian languages, such as Hindi, Kannada, Tamil (HRLs), Marathi (LRLs), Nepali (ZRLs), and Sinhala, were covered by many researchers. An MNMT system was presented by researchers to overcome the problems associated with low-resource language translation. This model consists of two MNMT systems: one-to-many for English-Indic and one-to-many for Indic-English. Each system has a shared encoder-decoder that has thirty translation directions and fifteen language pairs. In smart city, effective multilingual communication is required. In order to facilitate communication between multilingual locals, visitors, and city officials, NMT offers real-time translation services. Smoother interactions in domains like emergency response, transportation, and public services are made possible by this. Initially, the machine transliteration approaches mainly based on traditional and statistical methods. But, after emergence of deep learning techniques, Researchers are adopting these approaches for machine transliteration. Over the years, various rule-based approaches and Named Entity Recognition (NER) approaches are used for multi-lingual translation using phonetics of source and target. These multi-lingual translation approaches were mostly statistical and language specific. In the view of various Indian regional languages, there exists the immense need for machine multi-lingual transliteration.

The existing machine transliteration systems has various challenges such as pronunciation varies across multiple languages and different dialects of the same language, different structure of different languages, missing of common delimiters in few languages, translation ambiguity of regional language words, usage of homograph words for which the meaning of the word changes with the change in context, existence of multiple scripts in few languages such as Punjabi has Gurmukhi, etc., use of multiple languages in a sentence, automatic POS tagging based on sentence grammatical structure. These challenges make the multilingual translation very difficult. So, there is a huge demand of effective approaches to overcome the above discussed challenges and to improve of existing models. According to literature various machine learning techniques have been utilized for this task such as Direct MT system, which is based on Dictionary Lookup, Statistical MT system based on corpus and statistical models, interlingua based system based on universal natural language, knowledge based MT system which is based on deep-learning based and its core is neural networks.

The system proposed in many research papers works on some of these challenges and enhancing the performance accuracy of the machine translated text. There are various types of machine learning systems designed using traditional algorithms over the years which are also highlighted in the research work. Many different neural network models and architectures are developed since last few years.

Literature revealed that various neural network architectures have been developed to improve machine Translation accuracy. But, Neural Machine Translation (NMT) requires huge dataset to learn the patterns or language rules as well as context of the corpus. There two types of Neural Machine Translation (NMT) architectures. One is Recurrent Neural Network (RNN) along with LSTM and Neural Machine Translation (ConvS2S NMT) framework used for the machine translation along with experimental methodology and results. Many researchers described the process of how the Indian language can be processed by applying various techniques such as tokenization, pre-processing techniques, machine translation, Recurrent Neural Network (RNN), full-text search using CLIR model and final output generation in multiple Indian languages. Many researchers have performed machine translation on the English-Hindi parallel corpus which compiles all corpuses available in public domain including 1.40 million parallel segments consisting of sentences, phrases and dictionary entries. Machine Translation is one of the key areas of Natural Language Processing (NLP) and helps in breaking the language barrier and in facilitation of inter-lingual communication. With the evolution of information technology, many documents are available in multi-lingual languages so the need for effective machine translation systems became important for establishing proper communication. Machine Transliteration is a technique to convert source language into target language using machine.

This paper details the approach and methodology incorporated for the machine transliteration of the source language input text into other Indian languages accurately and unambiguously without changing the phonetics and the pronunciation of the source language text. Initially, the machine transliteration used to be done on traditional and statistical methods. With the emergence of deep learning techniques, few research attempts have been made using deep learning. Over the years, various rule-based approaches and Named Entity Recognition (NER) is used for multi-lingual translation based on phonetics of source and target languages using statistical and language specific methods. In the presence of various Indian regional languages, the immense need for machine multi-lingual transliteration has emerged.

The existing machine transliteration systems had various challenges such as pronunciation varies across multiple languages and different dialects of the same language, different structure of different languages, missing of common delimiters in few languages, translation ambiguity of regional language words, usage of homograph words for which the meaning of the word changes with the change in context, existence of multiple scripts in few languages such as Punjabi has Gurmukhi, etc., use of multiple languages in a sentence, automatic POS tagging based on sentence grammatical structure, straightway comparison not possible due to wide variation in language pairs, missing sounds etc. The system proposed in this research paper works on some of these challenges and enhancing the performance accuracy of the machine translated text. There are various types of machine learning systems designed over the years which are also highlighted in the research. Direct MT system, which is based on Dictionary Lookup, Statistical MT system based on corpus and statistical models, interlingua based system based on universal natural language, knowledge based MT system which is based on artificial intelligence and deep-learning based systems based on machine learning and neural networks.

Neural Machine Translation (NMT) is a new approach of machine translation with significant advantages over traditional approaches in terms of better translation performance and reduced model size. It translates as an end-to-end trainable supervised machine learning problem. The neural network consists of encoder and decoder network. The encoder extracts a fixed size length of vector representation from a variable length input sentence. The decoder then generates correct variable length target translation. The developed system tokenizes the source language text and then the text is pre-processed using various techniques such as stemming, lemmatization, stop-word removal, etc. and them the machine transliteration process is performed based on the trained NMT models and Seq-2Seq model using segmentation techniques such as character based or byte-pair based. After the machine translation of the input text, the text is further searched using multi-lingual search engine which searches the source language words in the target multi-lingual datasets. The output of the search engine provides the details of the translated word in the target dataset document along with the text position.

2. Literature review

The deep learning-based approach to machine transliteration described by authors Soumyadeep et al.[1] details the two types of Neural Machine Translation (NMT) architectures namely Recurrent Neural Network (RNN) based NMT framework and Convolutional Sequence-to-Sequence Neural Machine Translation (ConvS2S NMT) framework used for the machine translation along with experimental methodology and results. The authors Harish and Rangan[2] details the process of how the Indian language can be processed by applying various techniques such as tokenization, pre-processing techniques, machine translation, Recurrent Neural Network (RNN) , full-text search using CLIR model and final output generation in multiple Indian Languages. The authors Kunchukuttan[3] have performed machine translation on the English-Hindi parallel corpus which compiles all corpus available in public domain including 1.40 million parallel segments consisting of sentences, phrases and dictionary entries. The corpus was tested on statistical as well as Neural Machine Translation systems. The source text is normalized using 'Moses' tokenizer for English and IndicNLP for Hindi and then processed. NMT setup described in their research work included RNN based encoder-decoder-architecture containing 512 GRU units each. The NMT system was trained using 'Adam optimizer' with learning rate 0.0001. The system designed by this author uses direct MT technique which can convert web-based documents language along with Babylon translation software and online translation tool PROMT and Google Translator. Also, offline translation tools are briefed in this research work including Systran, METAL, English to Bangla phrase-based machine translation, Anglabharati, Anuvadaksh, UNL Based Encovertor-decovertor, Anusaaraka, Sampark, etc. The authors Vidya et al.[4] have described the approach for cross-lingual information retrieval amongst various languages along with emphasizing on the pre- and post- processing strategies for the queries entered in source language. The system designed uses Google Translator for translating into languages supported by the Google Search Engine. The have described various MT systems in field testing or as web service such as 'Anglabharti' project was launched for machine translation from English to Hindi, 'Mantra' MT system which translate from English to Hindi in specified domain of personal administration, office orders etc., 'Anusaaraka MAT system' which translates Kannada, Bengali, Marathi, Punjabi to Hindi, 'Shiva & Shakti" MT system which translate from English to Hindi, Universal Networking Language (UNL) based English-Hindi MT system. The authors Sanjay and Pramod have also briefed on various types of Machine Translation systems such as Anusaaraka, Mantra, Matra, AnglaBharti, AnuBharti, Shiva and Shakti, Anubaad, Sampark used for translation of Indian languages. The paper described the literature survey done for transliteration from one language to another using various approaches. The authors Narayan[5] has described the machine translation using a combination of rule based and quantum neural network approaches. The paper describes the quantum neural architecture-based algorithm used for machine translation from English to Hindi language. The author Sheshadri[6] has performed experiments on neural machine translation on Hindi to English parallel corpus. The author Islam et al.[7] has performed study on the various applications on NLP developed for North-East languages. Bhattacharyya.[8] highlighted Indian languages has diversity, So for Indic languages, proposed solutions must be applicable to multiple languages. Godase and Govilkar[9] focuses on different Machine translation projects and highlighted the approach and observations in detail. Khan[10] conducted a very minute research work toward Urdu transliteration. Mallick et al.[11] verifies the efficacy of the proposed approach from the higher BLEU scores achieved as compared to the state of the art for translation task on the German-English dataset. Manogaran et al.[12] uses deep learning techniques to extract opinions from large datasets. Philip[13] provides and analyses an automated framework to obtain such a corpus for Indian. Ramesh et al.[14] demonstrates MT systems produced via a social media-based human evaluation scheme. Singh and Kumar[15] inspected on the word vectors of 66 ambiguous Punjabi nouns for an explicit WSD system of Punjabi language. Srivastava and Govilkar[16] presents a survey of paraphrase detection techniques for Indian regional languages. Vathsala and Holi[17] analyses the social media data for code-switching and transliterated to English language using a special kind of RNN. Yu et al.[18] proposes a "reread" mechanism to transfer the outputs of the first-pass encoder to the second-pass encoder. Zhou et al.[19] proposes a deep neural network--based system combination framework leveraging both minimum Bayes-risk decoding and multi-source NMT.

3. Types of MT systems

There are various types of MT systems that have evolved over the years as described below:

1) Rule based Approach MT System—It consists of collection of grammar rules, bilingual or multi-lingual lexicon, dictionary and software programs to process the rules.
2) Direct translation approach—It translates direct word to word with the help of bilingual dictionary.
3) Interlingua based translation approach—It presents source text into an intermediary (semantic form) called Interlingua and then further it is translated to target text. The advantage is the analyzer and parser of source text is independent of the target text generator.
4) Transfer based translation approach—It performs analysis, transfer and generation. Firstly, the source language is parsed to produce syntactic representation of the sentence. Then the results are converted into equivalent target language-oriented representations. Further, morphological analyzer is used to generate final translated text.
5) Statistical based approach—This approach is based on the statistical and knowledge models which are extracted from bilingual or multi-lingual corpora. The training is done using supervised/ unsupervised statistical machine learning algorithm which builds statistical tables with information about characteristics of well-formed sentences, and correlation between sentences and between languages. Further decoding is done to find the best translation for the input sentences. There are 3 types of statistical based approach namely word-based translation, phrase-based translation, and hierarchical phrase-based model. The best translation is chosen based on highest probability of the probability distribution function represented as $p(e|f)$.
$$e = \text{argmax}\, p(e|f) = \text{argmax}\, p(f|e)p(e) \tag{1}$$
6) Hybrid based translation—This combines both rule-based and statistical based approach. This approach utilizes some rules for pre-processing the input data as well as post-processing the output generated.
7) Knowledge based translation—This required complete understanding of the source text prior to translation into target text. It is implemented on interlingua architecture. The system needs to be supported on lingual semantic knowledge about meanings of words and combinations.
8) Example based translation—This approach reuses the examples of already existing translations. This involves translation by analogy and is trained on a prior knowledge base of bilingual corpus.

4. Proposed methodology

The raw input text of the source language may be in structured/unstructured form. There are a various language processing technique including pre-processing of texts and machine translation of text using neural machine translation that is applied to translate the source language into target Indian languages (**Figure 1**).

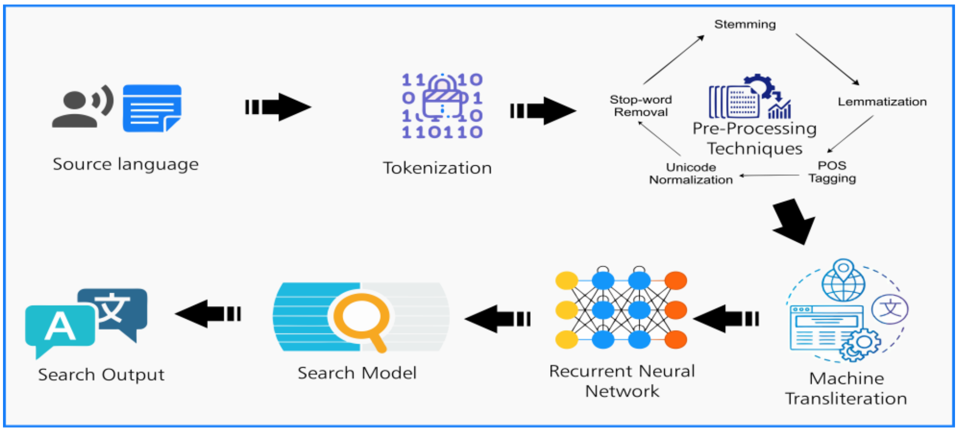

Figure 1. Language processing.

The methodology used for multi-lingual machine translation and language processing is described as below:

Step 1: Tokenization: The input source language serves as a raw text which is tokenized into lexical/basic units at word or sentence level in text processing applications. The lexical/basic units are termed as tokens. The sentence level tokenization is used for detection of the sentences based on sentence ending and boundary ambiguity and the word level tokenization, is used for tokenization of all the set of words in the whole document which serves as a lexical unit.

Step 2: Pre-processing techniques: The source language input text is further processed using various pre-processing techniques such as stemming, stop-word removal, lemmatization, POS tagging, Unicode normalization, etc.

Step 3: Neural network: Recurrent Neural Network (RNN) based Neural Machine Translation (NMT) is used for transliteration of language from source language to various Indian languages. This is a sequence-to-sequence based model which helps in machine translation and text summarization considering the context of the source language.

RNN neural network consists of hidden states h and optional output y which operates on a variable length sequence $x = (x_1, x_2, \ldots, x_T)$. At each time step t, hidden state h_t of the RNN is updated by

$$H_t = f(h_{t-1}, x_t)$$

RNN is a natural generalization of feedforward neural networks to sequences. Given a sequence of inputs ($x1$, ..., xT) a standard RNN computes a sequence of outputs ($y1$, ..., yT) by iterating the following equation:

$$h_t = sigm(W^{hx} x_t + W^{hh} h_{t-1})$$

$$y_t = W^{yh} h_t$$

RNN can easily map sequences to sequences when the alignment between inputs is known ahead of time. RNN Encoder-Decoder framework is best technique used for the training of the model. The encoder converts the source sentence into vectors which holds the meaning of the sentence and further the vectors are processed by the decoders to generate translation output. The RNN model consists of multi-layers along with Long Short-Term Memory (LSTM) which captures long dependency such as syntax structure and Gated Recurrent Unit (GRU). The LSTM predicts the next words of the target sequence given the previously translated words from the sequence. The bi-directional LSTM and BLEU metric is used for evaluation.

Step 4: Search engine model: Cross-lingual information retrieval (CLIR) model is used that performs multilingual text mining for cross-lingual text retrieval. CLIR model helps in retrieval of the relevant information from the document collection written in different languages. CLIR searches for the relevant term/word in other document datasets. Query translation is performed based on corpus-based method using parallel corpus having a set of identical text written in multiple languages. CLIR and multi-lingual text mining approach analyze the multi-lingual textual data employing techniques from information retrieval, NLP, and machine learning.

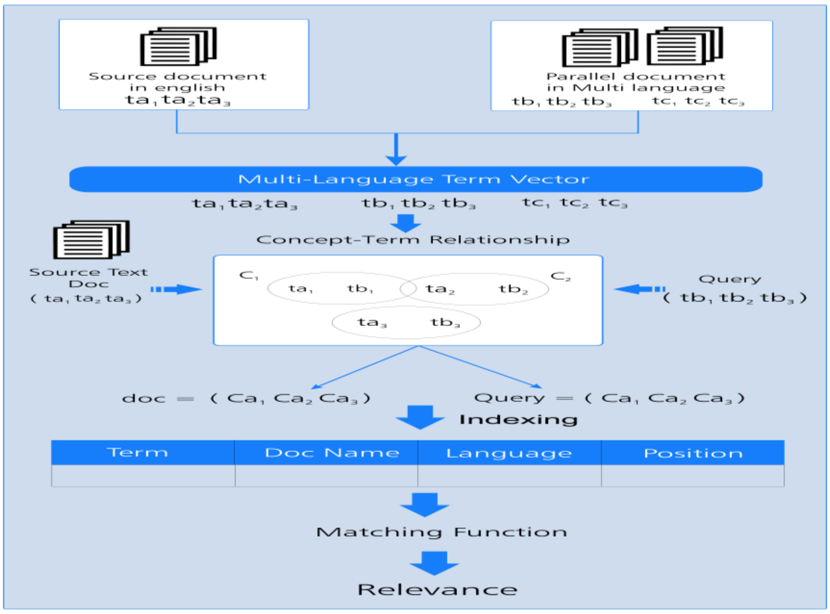

Figure 2. Search engine model.

As described in **Figure 2** above, the terms present in the source document in English and terms present in parallel multi-language documents such as = Hindi and Tamil languages are identified, and a multi-language term vector is created. Further, the multi-lingual concept-term relationship is identified among the multilingual terms of the source document with the parallel multi-lingual documents from the parallel corpus. The modified CLIR model proposed in this paper searches the input term/word in English language with the multi-lingual document datasets. Based on the concept-term relationship, the terms are indexed and details about the term/word, document name, language type and position of term/word in the document is captured for further matching the correct word using search model based on the term relevance.

Step 5: Multi-lingual search output: The source language is converted to multi-lingual languages after processing through RNN model and incorporating various machine transliteration techniques.

5. Experimental setup

The developed multi-lingual translation system is based on deep learning-based algorithms such as Neural Machine Translation (NMT) based Recurrent Neural Network (RNN) for translation of the source language into other multi-lingual Indian languages. The source language text is tokenized, pre-processed, matched with the trained Machine Transliteration based RNN model. Thereafter multi-lingual search is performed to identify the multi-lingual word position in the target document and sharing the results.

Some of the python libraries that are used for development of the model includes langdetect, textblob, englishtohindi, indicnlp, indicscripts, phonetic_sim, Indic_normalize, English_script Indic_tokenize, indic_detokenize, sentence_tokenize, sinhala_transliterator, Unicode_transliterate, acronym_transliterator, script_unifier.

6. Experimental results and analysis

The proposed multi-lingual translation model is evaluated based on following performance metrics such as accuracy, precision, recall and F1-score. The developed system is tested on a dataset size of more than 49 million sentence pair consisting of multi-lingual documents in various Indian languages. The below table name **Table 1** summarized the results obtained with the developed multi-lingual translation model for conversion of different language pairs from source language to target Indian language: The largest parallel corpus collection of Indic languages, including Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu, is available to the public via Samanantar.

Dataset analysis in terms of words/stop-words/unique words.

Table 1. Details the statistics (https://ai4bharat.iitm.ac.in/samanantar/).

Sr. No	Parallel corpora	Words	Lines
1	English-Bengali	9,80,12,782-8,61,96,471	84,35,356-84,35,356
2	English-Punjabi	17,17,791-13,79,148	1,38,354-1,38,354
3	English-Tamil	3,07,76,956-2,84,91,118	30,19,564-30,19,564

Table 2. Performance accuracy calculations.

Sr. No.	Language pairs	Accuracy	Precision	Recall	F1-score
1	English-Bengali	85%	74%	78%	74%
2	English-Punjabi	75%	70%	88%	70%
2	English-Tamil	70%	66%	62%	66%

The above table named **Table 2** presents performance metrics for a machine translation system across different language pairs. Each row corresponds to a specific language pair, and the columns provide various evaluation metrics, including accuracy, precision, recall, and F1 score. As per the results obtained, it is observed that the accuracy, precision, recall and F1-scores are better for multi-lingual translation of English language to Hindi language compared to translation to other Indian language such as Tamil, Bengali and Punjabi. The efficiency can further be increased by increase in the size of the dataset. English-Hindi language pair size is 3.21 GB. So, it achieves top performance for English-Hindi language pair and comparable results for other cases.

7. Conclusion and future work

This paper describes the methodology used for language transliteration from source English language to other regional languages (Tamil, Bengali, Punjabi, etc.) using machine transliteration techniques, RNN model and multi-lingual search processing model.

This system currently supports the information retrieval related to three languages namely English, Hindi and Tamil. In the future, the developed system shall support language translation for other major Indian regional languages such as Bengali, Telugu, Gujarati, Urdu, Kannada, Punjabi and Marathi.

Along with this, further efficiency shall be enhanced of the developed system.

For Indic languages primarily, the shift of MT from the SMT platform to the NMT platform represents a paradigm shift in NLP research. NMT models continue to have an insatiable appetite for data despite their progress, and recent findings on low- and zero-resource languages indicate that there is still more work to be done.

Author contributions

Conceptualization, RKD and OP; methodology, RKD; software, RKD; validation, PN, OP and RKD; formal analysis, RKD; investigation, RKD; resources, RKD; data curation, RKD; writing—original draft preparation, OP; writing—review and editing, OP; visualization, OP; supervision, PN; project administration, PN; funding acquisition, RKD. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Soumyadeep K, Sayantan P, Santanu P. A deep learning based approach to transliteration. In: Proceedings of the Seventh Named Entities Workshop, Melbourne, Australia. Association for Computational Linguistics. 2018. pp. 79–83.
2. Harish BS, Rangan RK. A comprehensive survey on Indian regional language processing. SN Applied Sciences. 2020; 2(7). doi: 10.1007/s42452-020-2983-x
3. Kunchukuttan A. An Introduction to Machine Translation & Transliteration. Available online: www.cse.iitb.ac.in/~anoopk (accessed on 19 June 2023).
4. Vidya PV, Raj PCR, Jayan V. Web Page Ranking Using Multilingual Information Search Algorithm - A Novel Approach. Procedia Technology. 2016; 24: 1240-1247. doi: 10.1016/j.protcy.2016.05.102
5. Narayan R, Singh VP, Chakraverty S. Quantum Neural Network Based Machine Translator for Hindi to English. The Scientific World Journal. 2014; 2014: 1-8. doi: 10.1155/2014/485737
6. Sheshadri SK, Gupta D, Costa-Jussà MR. A Voyage on Neural Machine Translation for Indic Languages. Procedia Computer Science. 2023; 218: 2694-2712. doi: 10.1016/j.procs.2023.01.242
7. Islam SI, Indika Devi MI. A Study on Various Applications of NLP Developed for North-East Languages.
8. Bhattacharyya P, Murthy H, Ranathunga S, et al. Indic language computing. Communications of the ACM. 2019; 62(11): 70-75. doi: 10.1145/3343456
9. Godase A, Govilkar S. Machine Translation Development for Indian Languages and its Approaches. International Journal on Natural Language Computing. 2015; 4(2): 55-74. doi: 10.5121/ijnlc.2015.4205
10. Khan A, Sarfaraz A. RNN-LSTM-GRU based language transformation. Soft Computing. 2019; 23(24): 13007-13024. doi: 10.1007/s00500-019-04281-z
11. Mallick R, Susan S, Agrawal V, et al. Context- and sequence-aware convolutional recurrent encoder for neural machine translation. Proceedings of the 36th Annual ACM Symposium on Applied Computing. Published online March 22, 2021. doi: 10.1145/3412841.3442099
12. Manogaran DrG, Qudrat-Ullah DrH, Xin DrQ, et al. Special Issue on Deep Structured Learning for Natural Language Processing. ACM Transactions on Asian and Low-Resource Language Information Processing. 2021; 20(1): 1-2. doi: 10.1145/3436206.
13. Philip J, Siripragada S, Namboodiri VP, et al. Revisiting Low Resource Status of Indian Languages in Machine Translation. Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD). Published online January 2, 2021. doi: 10.1145/3430984.3431026
14. Ramesh A, Parthasarathy VB, Haque R, et al. Comparing Statistical and Neural Machine Translation Performance on Hindi-To-Tamil and English-To-Tamil. Digital. 2021; 1(2): 86-102. doi: 10.3390/digital1020007
15. Singh V pal, Kumar P. Word sense disambiguation for Punjabi language using deep learning techniques. Neural Computing and Applications. 2019; 32(8): 2963-2973. doi: 10.1007/s00521-019-04581-3
16. Srivastava S, Govilkar S. A Survey on Paraphrase Detection Techniques for Indian Regional Languages.

International Journal of Computer Applications. 2017; 163(9): 42-47. doi: 10.5120/ijca2017913757
17. Vathsala MK, Holi G. RNN based machine translation and transliteration for Twitter data. International Journal of Speech Technology. 2020; 23(3): 499-504. doi: 10.1007/s10772-020-09724-9
18. Yu Z, Yu Z, Guo J, et al. Efficient Low-Resource Neural Machine Translation with Reread and Feedback Mechanism. ACM Transactions on Asian and Low-Resource Language Information Processing. 2020; 19(3): 1-13. doi: 10.1145/3365244
19. Zhou L, Zhang J, Kang X, et al. Deep Neural Network--based Machine Translation System Combination. ACM Transactions on Asian and Low-Resource Language Information Processing. 2020; 19(5): 1-19. doi: 10.1145/3389791