Computers, Materials & Continua
DOI: 10.32604/cmc.2023.032290
Article

Tech Science Press

Translation of English Language into Urdu Language Using LSTM Model

Sajadul Hassan Kumhar[1], Syed Immamul Ansarullah[2], Akber Abid Gardezi[3], Shafiq Ahmad[4], Abdelaty Edrees Sayed[4] and Muhammad Shafiq[5,*]

[1]SSSUTMS, Sehore, 466001, India
[2]Department of Computer Science, GDC Sumbal, J&K, 193502, India
[3]Department of Computer Science, COMSATS University Islamabad, Islamabad, 45550, Pakistan
[4]Industrial Engineering Department, College of Engineering, King Saud University,
P.O. Box 800, Riyadh, 11421, Saudi Arabia
[5]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan, 38541, Korea
*Corresponding Author: Muhammad Shafiq. Email: shafiq@ynu.ac.kr

Received: 13 May 2022; Accepted: 04 August 2022

Abstract: English to Urdu machine translation is still in its beginning and lacks simple translation methods to provide motivating and adequate English to Urdu translation. In order to make knowledge available to the masses, there should be mechanisms and tools in place to make things understandable by translating from source language to target language in an automated fashion. Machine translation has achieved this goal with encouraging results. When decoding the source text into the target language, the translator checks all the characteristics of the text. To achieve machine translation, rule-based, computational, hybrid and neural machine translation approaches have been proposed to automate the work. In this research work, a neural machine translation approach is employed to translate English text into Urdu. Long Short Term Short Model (LSTM) Encoder Decoder is used to translate English to Urdu. The various steps required to perform translation tasks include preprocessing, tokenization, grammar and sentence structure analysis, word embeddings, training data preparation, encoder-decoder models, and output text generation. The results show that the model used in the research work shows better performance in translation. The results were evaluated using bilingual research metrics and showed that the test and training data yielded the highest score sequences with an effective length of ten (10).

Keywords: Machine translation; Urdu language; word embedding

1 Introduction

English is a West Germanic language that was first adopted in early medieval England and became the lingua franca long thereafter. English is mainly written in Latin script, but there are other scripts such as B. Anglo-Saxon Rues, English Braille, and Unified English Braille. The language has 26 basic letters, each with upper and lower case letters. English is written from left to right. Urdu, commonly known as Lashkar, is the standard Persian record of Hindustani. It is the official language of Pakistan and Indian-controlled Jammu and Kashmir, and is one of the 22 official languages recognized by the Indian constitution. Urdu is written in Persian-Arabic script known as Naskh and Taliq script. Urdu has 39 basic letters called "Tahajji" which are written from right to left.

The art of translating works into written texts is as old as written literature. Machine translation is a branch of artificial intelligence in which a computer or any other machine contains a dictionary as well as programs and instructions to make choices, make logical decisions for this purpose, translate sentences, phrases and words of textual utterances in natural language. Machine translation uses ideas from interdisciplinary fields such as linguistics, computer science, artificial intelligence, statistics, and mathematics. The source language is the morphological enigma of the translated text, and the target language is the translated language translation of the source text. The taste and perception of text input is maintained throughout the translation process, and the purpose and melody of the information is preserved throughout the sentence. Machine translation can be bilingual or multilingual. Bilingual translation only selects two languages for machine translation, while multilingual translation is more than one language for machine translation. Machine translation can be decoded and rearranged the acquired meaning in the target language. During the process of decoding the source text into the target language, the translator examines all the characteristics of the text. This approach requires a detailed analysis and understanding of the source and target languages in terms of syntax, semantics, and syntax. Two fundamental needs have arisen in the history of machine translation: one is to disseminate and empower a person to understand the information provided by a foreign language. Another is to achieve communication between people of different backgrounds. Therefore, in order to make knowledge accessible to the general public, there should be a mechanism and tools in place to make things understandable by translating from the source language to the target language in an automated manner. Machine translation has achieved this goal with encouraging results. Recent developments in computational linguistics and natural language processing can efficiently process large volumes of textual utterances, making machine translation a reality.

There are four types of translation models for translation from source language to target language, including rule-based, computational, hybrid, and neural machine translation models. Rule-based translation models deal with linguistic properties of source and target languages. After processing these attributes, a rule-based translation model converts words, phrases, and sentences into rules. The whole process requires a thorough analysis of the input source language and output target language. Compared to rule-based translation, statistical translation is more robust and does not require linguistic information. Statistical machine translation relies on models such as pattern recognition, statistical decision theory, and machine learning models. This type of machine translation consists of three main parts. The translation module is the first component to contain probabilistic mappings between source and target languages. The language engine of the second component is responsible for generating fluent target sentences, and the last component is the decoder, which uses the first two modules to translate sentences at runtime. Hybrid translation is a mix of statistical and rule-based machine translation. Hybrid machine translation uses translation memory, making it more accurate in terms of accuracy. This method of translation has many drawbacks, the most important of which is the need for thorough human editing. To create mathematical models of computer translation, neural machine translation relies on neural network models built on top of the human brain. The advantage of the neural machine translation paradigm is the use of a single model to train and decode both source and target languages.

The types of machine translation discussed above have both positive and negative aspects. Rule-based machine translation requires the study of source and target languages. This analysis is required for every language pair considered in machine translation. Machine translation engines are created based on linguistic methods, so the resulting translations are better than normal machine translations. Statistical machine translation methods cannot take into account all aspects of the syntactic and morphological diversity of the source and target languages. Neural machine models eliminate the need for advanced systems typical of the other three forms of machine translation. Urdu is a language that is spoken and written not only in South Asia but also in the West. Urdu is the official language of India and Pakistan. Urdu is used as a language of instruction in schools, lower levels of government, Indian sub-content and newspapers. Urdu is also spoken in Bangladesh, Nepal, and Afghanistan, and has become the lingua franca of South Asians living outside the subcontinent, especially in the Middle East, Europe, the United States, and Canada. When it comes to foreign languages, English is very important. Almost 45% of the world's knowledge is written or spoken in English, while the remaining 55% is written or spoken in Russian, French, German, Arabic, Farsi and Urdu. Urdu and English are very different languages. Urdu is written from left to right whereas English is written opposite to Urdu, from right to left. When writing or speaking a sentence, the subject is written first in grammatical order, followed by the verb and object, whereas in Urdu, the subject is written in grammatical order, then the object, and then the verb. Urdu is a resource-poor and morphologically rich language whereas English is a resource-rich and morphologically poor language.

2 Literature Review

Machine translation has been extensively studied by computational linguistics, natural language programmers, social scientists, and many other scholars of various bilingual and multilingual language pairs. Various machine translation models, methods, theories, and tools have been developed based on the language under consideration. Most machine models are translations of generative languages. Such a model requires a lot of language skills. Nagao [1] proposed a machine translation model that can efficiently translate samples fed into the system. The model is fed a bilingual corpus of source and target languages. DeNeefe et al. [2] DrivTool is recommended for translation. DrivTool is an interactive translation visualization tool that provides users with rich options to select or examine the decoding process in grammar-based translations. Koehn [3] proposed a statistical machine translation toolkit. The toolkit is open source and performs statistical translation. Rule-based translation is performed when the toolkit comes with obfuscation and factor decoding. Kumhar et al. In 2021 [4], they proposed a word generation model for embedding words in Urdu using a vector representation they called word2vec. However, the model distributes the words directly as vectors, without translation.

Mathur et al. [5] proposed the matching ontology evaluation tool used by Joshi et al. [6] proposed a machine translation engine. Tahir et al. [7] proposed a knowledge base machine translation model. The model performs data mining and text mining techniques to translate English to Urdu. Ata et al. [8] proposed a rule-based translation model. The translation model translates English to Urdu. The proposed translation model is suitable for a transfer method that handles phrasal and verb postpositions through Panama grammatical concepts. Perform rule-based machine translation for the best results with optimization complexity. Gupta et al. [9] proposed a rule-based translation model in 2016. The model worked with Stemmer to develop Stemmer rules for Urdu. In addition, Stemmer is used to rate English-Urdu translations. See King et al. [10] extended the stemmer for derivation rules to an influential stemmer. Gupta et al. [11] proposed a scoring-based machine translation model from English to Urdu. The research model focuses on the consistency of Urdu machine translation as interpreted by various technologies such as Ijunoon, Babylon, and Google. Machine translation scoring is done through both human scoring and automatic scoring. Kumhar et al. [12] collected corpora of multilingual Roman Urdu and English texts in 2021 using various Python corpus collection techniques. However, the model did not translate English to Urdu.

Dubey et al. [13] proposed direct machine translation from Hindi to Dori through word-to-word machine translation in 2017. To enable translation, a dictionary model is employed. Irvine [14] proposed a statistical translation model for bilingual small datasets with few resources other than parallel data. Shahnawaz et al. [15] proposed a neural machine translation model from English to Urdu. The proposed model is based on case-based reasoning (CBR). CBR is used to select Urdu as a training method for introductory English sentences. Sharma et al. [16] proposed a statistical machine translation model based on a Hindi-English bilingual corpus, which uses multiple algorithms such as independent location rate, translation error rate, word error rate, and n-gram to achieve translation from English to Hindi translate. Jawaid et al. [17] proposed a language-based statistical translation model from English to Urdu. Experiments were conducted on two Indo-European languages using Moses' statistical machine translation model. They also propose a new paradigm for reorganizing phrases in the syntactic tree of English sentences. The method shows significant improvement on the bilingual evaluation understudy (BLEU) metrics and human judgment. They further propose a baseline for hierarchical and phrase-based machine translation models and report results on three official datasets. In the reported study, the hierarchical output performed better than other models. Ali et al. [18] proposed Moses decoder-based statistical machine translation and supporting tools for English to Urdu translation. Singh [19] proposed a statistical translation model in 2013. The proposed model performs part-of-speech (POS) tagging for Marathi to learn the model, followed by Singh et al. [20] added supervised learning to the same POS tagger.

Khan et al. [21] proposed a hierarchical translation of morphologically rich Urdu records. Experiment using K-fold cross-validation method by selecting prepare, tune, and test from parallel corpora. Narayan et al. [22] proposed a neural translation model from Hindi to English. Neural pattern recognition has been used to recognize and learn patterns in corpora. The model uses Sanskrit Hindi and English word and sentence pairs to learn and recognize patterns in the corpus. Chand [23] conducted a comparative study of different tools and techniques that have been developed for machine translation. In the proposed model, the tested tools include rule-based Angla-Bharat and Anubaad, IM-based Babylon, and statistics-based Bing and Google and so on. Alabau [24] proposed a translation model for Graphical User Interface (GUI). The model's GUI functionality automates MT server and CAT server translation. Kumhar et al. [25] proposed a Urdu sentiment analysis technique using word2vec and Long Short-Term Memory (LSTM) techniques in 2020. However, the model does not inform translations of texts collected from social media platforms.

Salunkhe et al. [26] proposed a hybrid translation model for translating English websites into Marathi. Websites translated by Hybrid Translation include agricultural websites, medical reporting websites, and tourism-related information websites. A hybrid model is a combination of statistical and rule-based translation models. In this proposed model, rule-based translation is implemented using a mapper algorithm. Additionally, Salunkhe et al. [27] proposed another hybrid translation model. In the proposed model, statistical translation methods work together with rule-based methods for better results. Marathi WordNet is used in the model to extend the dictionary for better translations. Ayesha et al. [28] proposed a model to compare the performance of online translation models such as Google Translation, Being, and Babylon. The input is the Urdu to be modeled and the output is the Arabic translation. Check and compare Arabic sentences in the proposed model. The results of the comparative study show that Google Translate outperforms the other two translation methods. Zafar et al. [29] proposed a rule-based translation model. The proposed model gives users the flexibility to customize the user's perception of translation. The machine translation model supports features such as idioms and homographs on bilingual corpora. The model shows remarkable performance in machine translation.

Dubey et al. [30] proposed a Dogri-to-Hindi machine translation model. Godase et al. [31] studied different translation models suitable for Indian language translation in detail. Among them, Language has a detailed examination of Urdu-English and English-Urdu. Kumhar et al. [32] studied word embedding generation methods and tools in detail, but the model did not translate social media texts. Sinha et al. [33] proposed a pseudo-Interlingua machine translation model. The model uses AnglaBharti modules and abstract examples. The model achieved 90% accuracy. 5. Goya et al. [34] proposed a statistical translation model from English to Hindi, which they called a statistical engine. In the model, a preprocessing step re-syntaxes the source language to remove long-range motion and morphologically to process the corpus. The model achieves efficiency by effectively separating word order and suffixes. Choudhury et al. [35] proposed a neural machine translation model from English to Tamil. The model uses techniques such as word embeddings and byte-pair encoding to remove obstacles to smooth translation, such as lack of vocabulary. 5. Goya et al. [36] proposed a hybrid translation model from Hindi to Punjabi. The model uses a morphology analyzer developed by IIIT-Hyderabad combined with a hybrid translational model. The model achieved an accuracy of 87.60%. Kaur et al. [37] proposed a hybrid English-to-Punjabi translation model.

In the proposed model, a rule-based translation method is used to analyze the source text to generate intermittent representations. Khan et al. [38] proposed a scalable bivariate feature extraction and extended the Harris method image hashing method research to identify malware attacks for obfuscation. However, the proposed model exclusively uses bivariate feature extraction and Haris algorithm for malware detection and image-optimized hashing. Binti et al. [39] proposed cluster-purpose-based access control in big data environments to achieve long-term data protection. However, this approach fails to account for different types of privacy rules in shared environments. Gumaei et al. [40] proposed an edge computing model by creating deep learning-based human activity recognition (DL-HAR) framework. However, the framework cannot detect actions based on time-series data from online linked sensors. Al-Wesabi [41] proposed a Zero Watermark and Natural Language Processing (CAZWNLP) technique to ease the recognition of English text shared on the Internet. However, the model does not account for all attack types and increases in attack rates.

3 Research Methodology

English to Urdu machine translation is still in its infancy and lacks basic translation methods to provide motivating and acceptable English to Urdu translation. For translating English snippets into Urdu script, it is agreed to use character-level machine translation as suggested by Lee et al. suggest [42] because it outperforms the statistical machine translation model proposed by S.K. Mahata. In the research, a Long Short Term Memory Encoder decoder for a recurrent neural network model for English to Urdu translation is proposed. Advantages of using an encoder-decoder model according to Chung et al. [43], is the ability to handle morphological changes, resolve out-of-lexical issues, and translate without segmentation. The various steps required to perform translation tasks include preprocessing, tokenization, grammar and sentence structure analysis, word embeddings, training data preparation, encoder-decoder models, and output text generation. The workflow of the translation process is shown in Fig. 1.

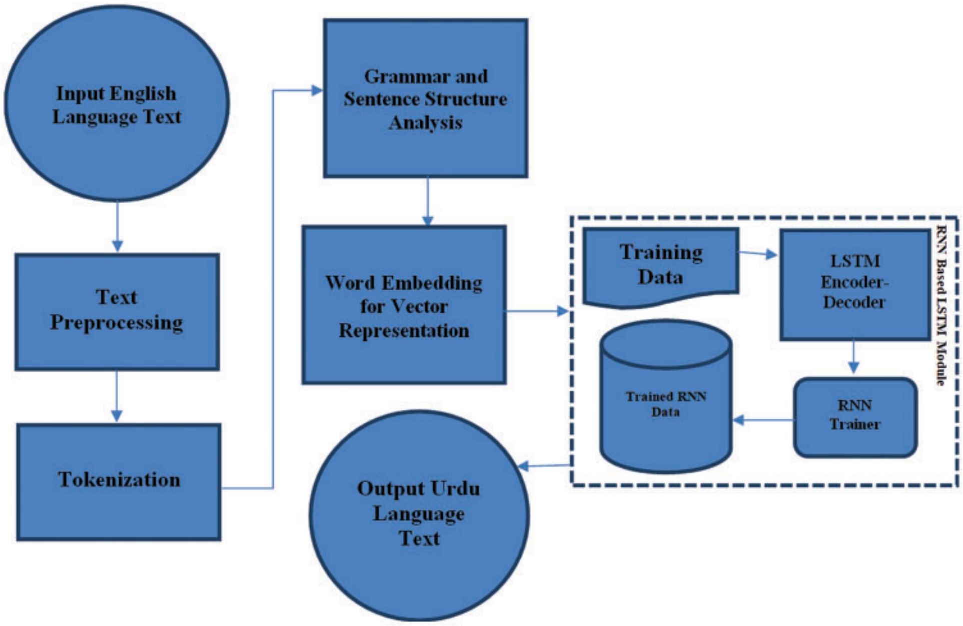
Input English Language Text
Text Preprocessing
Tokenization
Grammar and Sentence Structure Analysis
Word Embedding for Vector Representation
Training Data
LSTM Encoder-Decoder
Trained RNN Data
RNN Trainer
RNN Based LSTM Module
Output Urdu Language Text

Figure 1: Flow of process in the machine translation from english to Urdu language

3.1 Preprocessing

The English corpus was preprocessed and cleaned by removing hashtags and retweets, extra spaces before and after lines of text and numbers, and replacing multiple spaces with a single space. Noise is additional information about texts collected from social media sites and the internet. We use Python API web scraping techniques and Selenium for web scraping. Information in text form contains noise, and while maintaining the quality of the English corpus, unnecessary and illegal text data such as cluttered code, hypertext markup language tags, redirects, web format information, etc. are removed from the data. Additionally, we removed the repeated dates and diacritics, as they are optional and only contain the altered pronunciation. In the next step of preprocessing and cleaning, the original text is tokenized, segmented, and POS-tagged.

3.2 Tokenization

Tokenization is a vital and essential process in machine translation. Tokenization is used to break sentences down into parts of speech called words. These smaller units are also fragments or fragments. Python programming is used to tokenize English and Urdu languages.

3.3 Grammar and Sentence Structure Analysis

The grammar and sentence structure analysis process assigns the correct grammatical graphemes and correct structures to the word segments. In grammar and sentence structure analysis, the above segment obtained by word segmentation is marked as the correct grapheme, such as B. Word noun word plus N, adjective plus ADJ, etc.

3.4 Embedding Generation

Word embeddings aim to represent words in the form of vectors. Word embeddings represent words in the vector space that have similar meanings on one side and less similar meanings on the other. There are various word embedding techniques for word vector representation, including Bert, fastText, Word2Vec, and Glove. In the study, Word2Vec's word distribution representation was used.

The word embedding method shown in Fig. 2 focuses on maximizing the probability of surrounding words on a given center word, which is given by,

$$p(w_{i-c},\ w_{i-c+1}, \ldots,\ w_{i-1},\ w_{i+1}, \ldots,\ w_{i+c-1},\ w_{i+c}|w_i) \tag{1}$$

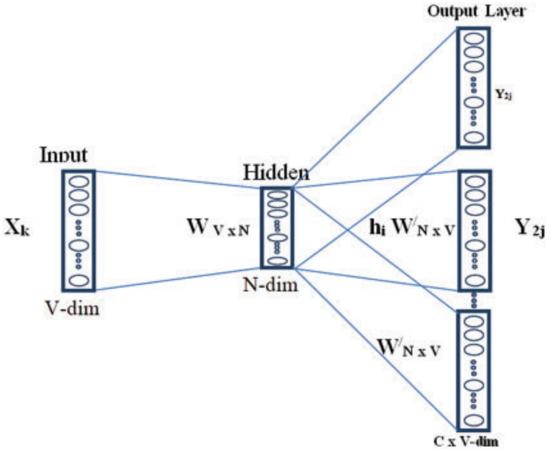

Figure 2: Word embedding by skip gram model

The optimization mechanism process for the Skip-Gram model is similar to the Continuous Bag of Words (CBOW) model with reverse context words. The softmax function that generates the vector word distribution can be written as,

$$p\left(w_c|w_i\right) = \frac{\exp\left(v_{mc}^{T} v_{w_i}\right)}{\sum_{w=1}^{|w|} \exp\left(v_{w}^{T} v_{w_i}\right)} \tag{2}$$

This function is not efficient because it needs to sum all W words for normalization. Global co-occurrence statistics improve the performance of the function.

3.5 Word2Vec Method

The Word2Vec method is a flat two-layer neural network that reconstructs the linguistic context of words. The Word2vec method represents each individual word as a real number stored in a vector. This method takes a large amount of input text and creates a vector space with hundreds of dimensions. Each individual word is assigned a corresponding vector in the dimension space. The vectors used to store words are chosen such that more meaningful numbers represent the semantic similarity between the words represented. The above vectors are also known as word embedding generation.

Word2Vec methods can use CBOW or Skip-Gram architectures to generate word vectors for distribution in vector space. In a nonstop bag of words architecture, the next word is forecasted from the context of nearby surrounding words, and the order of words does not affect the prediction of new vector words. The Skip-Gram model uses the current word to predict the context of nearby surrounding words. In Skip-Gram, neighboring words influence the prediction of the next neighboring word. Compared with the Skip-Gram architecture, the CBOW architecture is the most efficient.

3.6 Encoder

Encoder is an LSTM cell with memory. It accepts an input as an element in sequence at a single time stamp processes the input element for collection of information and propagates it forward. The LSTM encoder process a single word at single time stamp. Therefore, if a sentence has n words or padding of sentence is of length L so it will require n number of time stamps to process it. Encoder generates the thought vector (Context vector) to represent the meaning of distributed words of source language. Some of the notation that are used with encoder are: x_t which is the input to the encoder at time stamp t; h_t and c_t are the internal states of LSTM at time stamp t which are initially set to zero; y_t is the output produced from encoder at time stamp t.

Let us take an example of a sentence: "what is your name?" The sequence of input sentence will be taken as five stamp words as x_1 = what, x_2 = is, x_3 = your, x_4 = name and x_5 = ?. The encoder will the above sequence in five (5) time stamp steps as shown in Fig. 3.

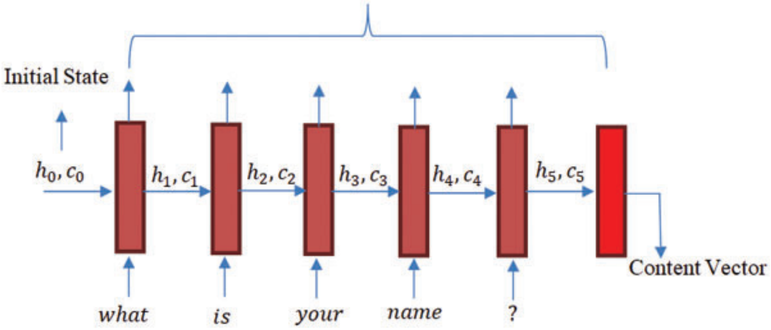

Figure 3: Time stamp taken for reading a sentence by encoder

At time stamp t_1, the LSTM cell remembers that the content "what" has been read, at time stamp t_2, the LSTM recalls that it had read "what is" and similarly at time stamp t_5, the LSTM cell recalls that it has read all the sentence.

The initial states h_0 and c_0 are set to zero vectors. The LSTM encoder takes a set of sequence of words $x_s = x_s^1 + x_s^2 + \ldots + x_s^n$ and calculates the thought vector $v = [h_c, v_c]$ where as the h_c represents the final hidden state of LSTM cell. The h_c is obtained after processing cell v_c. Which is represented mathematically a $v_c = C_l$ and $v_h = h_l$.

3.7 Decoder

Decoders are also an important part of neural machine translation. The decoder translates the intermediate context vector information into the desired language. The decoder is also an LSTM neural network. The encoder and decoder both have the same weights. However, in the research work, two different networks are used to encode and decode textual information. This combination of using both encoder and decoder LSTMs yields a more efficient translation mechanism. The decoder LSTM architecture is shown in Fig. 4. The decoder takes as input a one-hot vector (context vector) and generates an output sequence based on the information encoded by the encoder. The states in LSTM cells of Decoder are initialized with one hot vector $v = \{v_h, v_c\}$ as $h_0 = v_h$ and $c_0 = v_h$. Where h_0 and c_0 are the internal states of LSTM which are initially set to zero. Since the only the link from end-end is on hot vector which is the only available information to the decoder about the source text. The mth predication is calculated by the decoder as follows,

$$C_m,\ h_m = Decoder(yT^{m-1}|v,\ y_T^1,\ y_T^2 \ldots\ldots,\ y_T^{m-1}) \tag{3}$$

where as y_T^m is SoftMax function which is calculated as

$$y_T^m = softmax\left(w_{softmax} *\ h_m + b_{softmax}\right) \tag{4}$$

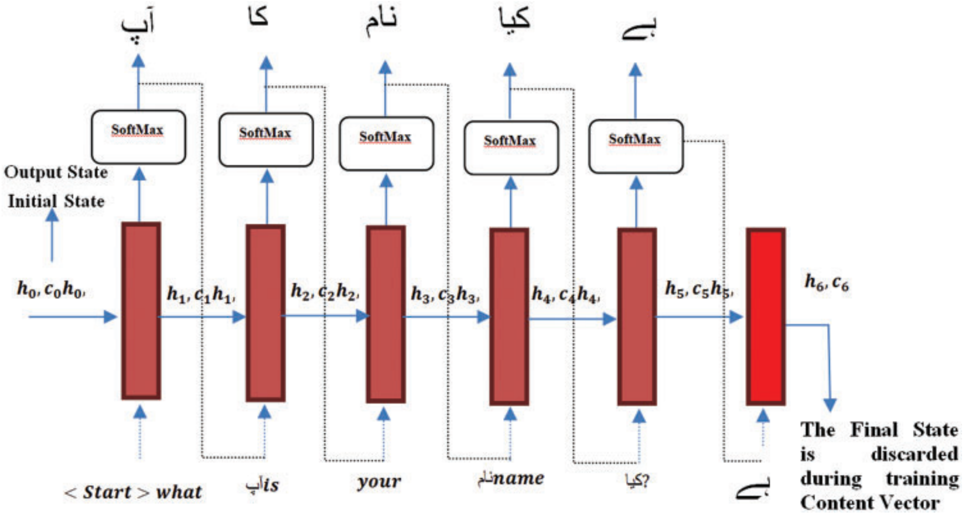

Figure 4: Time stamp taken for reading a sentence by encoder

The research work model includes two parts: encoder and decoder. The input is an English sequence, and the corresponding result in our case is also an Urdu sequence. The research model consists of LSTM cells and is resistant to vanishing gradients. These LSTM units show better results on longer sequences, the internal state is retained on the model's encoder, and the output is discarded. The main motivation for preserving internal states at coding sites and discarding them at output is to keep the cell's internal contextual knowledge intact. These states are transmitted to the decoder as initial states in the encoder-decoder transition process. The initial states are used with the LSTM cells as secret states from the encoder to build the decoder, and they are programmed to return sequences and states. Fig. 5 shows the encoder-decoder translation from English to Urdu.

The compact form of equation of LSTM with forged gated having forward pass are given as follows,

$$f_t = \sigma_g(W_i x_t + U_f h_{t-1} + b_f \tag{5}$$

$$i_t = \sigma_g\ (W_i x_t + U_i h_{t-1} + b_i) \tag{6}$$

$$O_t = \sigma_g\ (W_0 x_t + U_0 h_{t-1} + b_0) \tag{7}$$

$$\widetilde{c}_t = \sigma_h\ \ (W_c x_t + U_c h_{t-1} + b_c) \tag{8}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \widetilde{c}_t \tag{9}$$

$$h_t = O_t\ \circ\ \sigma_h\ (c_t) \tag{10}$$

where c0 = 0 and h0 = 0 are the initial values and ∘ is Hadamard Product (element-wise product). The subscript t indicates time steps. We need to mention that $x_t \in R^d$ is input vector to LSTM, $f_t \in R^h$ is activation vector to forged gate, $i_t \in R^h$ is input/update gate to activation vector, $O_t \in R^h$ is output gat's activation vector, $h_t \in R^h$ is the output state or hidden state of LSTM, $\widetilde{c_t} \in R^h$ is cell input activation vector, and $c_t \in R^h$ is cell state vector. W $\in R^{h\,X\,d}$, U $\in R^{h\,X\,h}$ and b $\in R^h$. The weight matrices and bias vector parameters which need to be learned during training. The activation functions that have been used are σ_g which is known as sigmoid function, σ_c a hyperbolic tangent function and σ_h is the prehole function or hyperbolic tangent function.

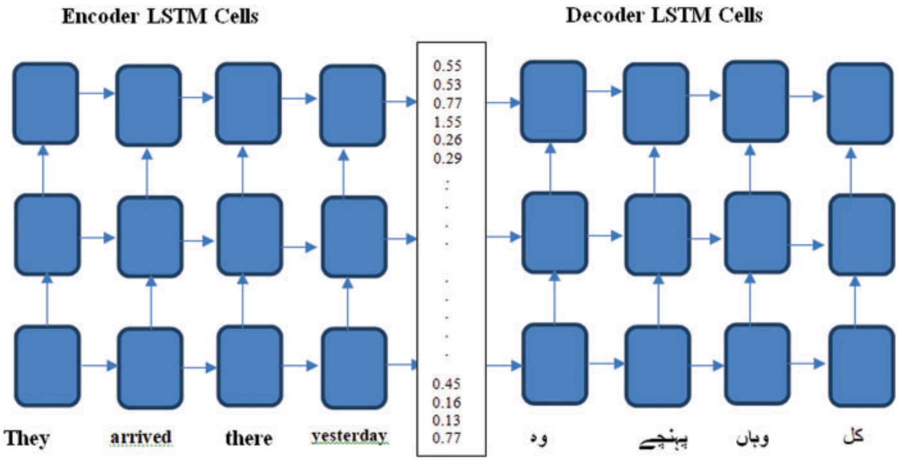

Figure 5: Translation using encoder-decoder method of translation

4 Results and Discussion

The results show that the model used in the research work shows better performance in terms of transliteration and translation. To improve the results using a sequence-to-sequence model, a three (3) layer LSTM architecture was implemented to fix the total sequence length to fifteen (15). To learn the most important parameters in training and testing, SoftMax loss function and Adam optimizer are combined with adaptive learning rate. Analyzing the results using the BLEU metric, the sequences that yield the highest scores for both test and training data have an effective length of ten (10). The validation step was performed on the model chosen in the research work, and the numerical estimates obtained using the BLEU metric were 50.86 and 47.06 for both training and test data. The small differences in BLEU metric values compared to other models used in the study demonstrate the effective performance of the model chosen in the study. Fig. 6 shows BLEU metric scores along with sequence length.

It can be observed that as the sequence length increases, the BLEU score also increases, but until the sequence length is 10 and from that point on the sequence length, the BLEU score drops abruptly. This clearly shows that the model is working well. However, as can be seen from Fig. 6, the BLEU score drops abruptly after burst lengths over 10, which clearly shows that the encoder-decoder model is estimating reasonable burst lengths. It is important to note that the system did not generate any new vocabulary; all words were drawn from the corpus already found in the encoder-decoder model.

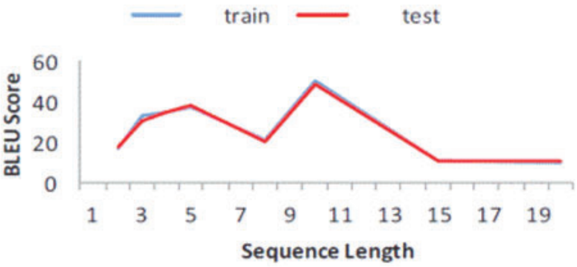

Figure 6: BLEU score for varying input sequence

There are no unreasonable words for this quality assurance of our models. Despite the low BLEU score, our system performs translation and transliteration based on quantitative analysis, which inspires interesting patterns. Tab. 1 shows the three correct and three incorrect datasets from the training and testing datasets.

Table 1: Sentences with correct results and inaccurate results from our dataset

	Sentence from Train dataset	Sentence from Test dataset
Accurate results	Tum donoo achy ladkyhai Transliteration تم دونوں اچھے لڑکے ہے-	Is he a good boy? Transliteration کیا وہ ایک اچھا لڑکا ہے؟
	Is it possiblefor me to catch train? Translation کیا یہ میرے لیے ممکن ہے کہ میں ریل گاڑی کو پکڑسکوں؟	He brought new bat for playing- Translation **اس نے کھیلنے کیلئے نیا بلا لایا ہے-**
Inaccurate Results	Human should make themselves accounts. Translation انسان کو اپنے آپ کا محاسبہ کرنا چاہئے-	Crordand oppressor personsunmoved Transliteration **ذہین** اور جابر **مجرموں** کو ہٹایا-
	AdilHassanarrivedfromJammutoKashmirbyeroad. Translation عادل حسن زمینی راستے جموں ریل گاڑی پر ذریعے کشمیر آئے-	AskingforGod'smercyin-frontofscholarsisworship. Translation اہل علم کے سامنے خداکےرحمت مانگنا عبادت ہے-

Fig. 7 shows the percentage of responses received from English and Urdu experts regarding verification of translations from English to Urdu, with 53% rated as excellent, 38% good, 8% average, 1% Bad and 0% are very bad.

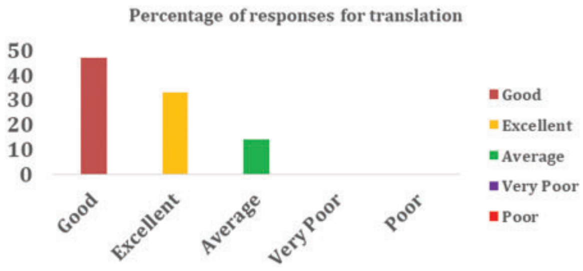

Figure 7: Performance of translation English into Urdu

The general opinion of experts indicates that English to Urdu language translation is a reliable and feasible model for performing the desired language translation. After analyzing all the reviews from different experts, it was concluded that most experts generally agree that the proposed method should be used effectively to translate English datasets into Urdu monolingual texts. Last but not least, evaluations by various experts help to validate the proposed methods and tools through experimental validation on real projects.

5 Conclusions

In this work, an encoder-decoder model is used to translate the English text into plain Urdu. This process is implemented by an encoder-decoder using a sequence-2-sequence model, which receives a sequence of English sentences as input and produces a sequence of Urdu sentences as output. We evaluate the results using the BLEU metric, which shows that the sequences that yield the highest scores for both test and training data have an effective length of 10. The validation step was performed on the model chosen in the research work, and the numerical estimates obtained using the BLEU metric were 50.86 and 47.06 for both training and test data. The small differences in BLEU metric values compared to other models used in the study demonstrate the effective performance of the model selected in the study. Unlike Python text, social media text has been used to analyze items such as political beliefs, products and services. The model proposed in the research paper can be applied to other Romance languages, as well as to different types of linguistic texts to translate languages contained in sentences.

Acknowledgement: The authors extend their appreciation to King Saud University for funding this work through Researchers Supporting Project number (RSP-2021/387), King Saud University, Riyadh, Saudi Arabia.

Funding Statement: This work was supported by King Saud University through Researchers Supporting Project number (RSP-2021/387), King Saud University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle," in *Artificial and Human Intelligence*, 1st ed., vol. 1, Kyoto, Japan: Elsevier Science Publishers, pp. 351–354, 1984.

[2] S. DeNeefe, K. Knight and H. H. Chan, "Interactively exploring a machine translation model," in *Proc. ACL Interactive Poster and Demonstration Sessions*, Ann Arbor, Michigan, USA, pp. 97–100, 2005.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch and M. Federico *et al.*,"Moses: Opensource toolkit for statistical machine translation," in *Proc. Association for Computational Linguistics*, Prague, Czech Republic, pp. 177–180, 2007.

[4] S. H. Kumhar, M. M. Kirmani, J. Sheetlani and M. Hassan, "Word embedding generation for Urdu language using word2vec model," *Materials Today: Proceedings*, vol. 1, 2021.

[5] I. Mathur, N. Joshi, H. Darbari and A. Kumar, "Automatic evaluation of ontology matchers," in *Proc. Int. Conf. on Transport Science*, Udaipur, India, pp. 1–6, 2016.

[6] N. Joshi, I. Mathur and S. Mathur, "Translation memory for Indian languages: An aid for human translators," in *Proc. Int. Conf. & Workshop on Emerging Trends in Technology*, Mumbai, Maharashtra, India, pp. 711–714, 2011.

[7] G. R. Tahir, S. Asghar and N. Masood, "Knowledge based machine translation," in *Proc. Int. Conf. on Information and Emerging Technologies*, Karachi, Pakistan, pp. 1–5, 2010.

[8] N. Ata, B. Jawaid and A. Kamran, "Rule based English to Urdu machine translation," in *Proc. Conf. on Language and Technology*, Barigali, Karachi, Pakistan, pp. 1–7, 2007.

[9] V. Gupta, N. Joshi and I. Mathur, "Design and development of a rule-based Urdu lemmatizer," in *Proc. Int. Conf. on ICT for Sustainable Development*, Singapore, pp. 161–169, 2016.

[10] B. King and S. Abney, "Labelling the languages of words in mixed-language documents using weakly supervised methods," in *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, pp. 1110–1119, 2013.

[11] V. Gupta, N. Joshi and I. Mathur, "Subjective and objective evaluation of English to Urdu machine translation," in *Proc. Int. Conf. on Advances in Computing, Communications and Informatics: Human Language Technologies*, Mysore, India, pp. 1520–1525, 2013.

[12] S. H. Kumhar, W. Qadir, M. M. Kirmani, H. Bashir and M. Hassan, 2021, "Effectiveness of methods and tools used for collection of roman Urdu and English multilingual corpus," *Design Engineering*, vol. 2021, no. 7, pp. 13428–13438, pp. 2021.

[13] P. Dubey, "Natural language processing," *JK Knowledge Initiative*, vol. 1, no. 2, pp. 89–91, 2017.

[14] A. Irvine, "Statistical machine translation in low resource settings," in *Proc. Int. Conf. on Advances in Computing, Communications and Informatics: Human Language Technologies*, Atlanta, Georgia, AG, USA, pp. 54–61, 2013.

[15] N. A. Shahnawaz and R. B. Mishra, "An English to Urdu translation model based on CBR, ANN and translation rules," *International Journal of Advanced Intelligence Paradigms*, vol. 7, no. 1, pp. 1–23, 2015.

[16] N. Sharma, S. Seth, M. Jain, M. Syed and N. Bharti, "Development of English-Hindi interactive machine translation," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, pp. 4185–4189, 2019.

[17] B. Jawaid, A. Kamran and O. Bojar, "English to Urdu statistical machine translation: Establishing a baseline," in *Proc. Workshop on South and Southeast Asian Natural Language Processing*, Dublin, Ireland, pp. 37–42, 2014.

[18] A. Ali, A. Hussain and M. K. Malik, "Model for English-Urdu statistical machine translation," *World Applied Sciences Journal*, vol. 24, no. 10, pp. 1362–1367, 2013.

[19] J. Singh, N. Joshi and I. Mathur, "Development of Marathi part of speech tagger using statistical approach," in *Proc. Int. Conf. on Advances in Computing, Communications and Informatics*, Mysore, India, pp. 1554–1559, 2013.

[20] J. Singh, N. Joshi and I. Mathur, "Marathi parts-of-speech tagger using supervised learning," *Intelligent Computing, Networking & Informatics*, vol. 243, pp. 251–257, 2014.

[21] N. Khan, W. Anwar, U. I. Bajwa and N. Durrani, "English to Urdu hierarchical phrase based statistical machine translation," in *Proc. Workshop on South and Southeast Asian Natural Language Processing*, Nagoya, Japan, pp. 72–76, 2013.

[22] R. Narayan, V. P. Singh and S. Chakraverty, "Quantum neural network based machine translator for Hindi to English," *The Scientific World Journal*, vol. 2014, no. 485737, pp. 1–8, 2014.

[23] S. Chand, "Empirical survey of machine translation tools," in *Proc. Int. Conf. on Research in Computational Intelligence & Communication Networks*, Kolkata, India, pp. 181–85, 2016.

[24] V. Alabau, C. Buck, M. Carl, F. Casacuberta and M. García-Martínez *et al.,* "CASMACAT: A computer-assisted translation workbench," in *Proc. Workshop on Computational Approaches to Causality in Language*, Gothenburg, Sweden, pp. 25–28, 2014.

[25] S. H. Kumhar, M. M. Kirmani, J. Sheetlani and M. Hassan, "Sentiment analysis of Urdu language on different social media platforms using word2vec and LSTM," *Turkish Journal of Computer & Mathematics Education*, vol. 11, no. 3, pp. 1439–1447, 2020.

[26] P. Salunkhe, A. D. Kadam, S. Joshi, S. patil and D. Thakore *et al.,* "Hybrid machine translation for English to Marathi: A research evaluation in machine translation," in *Proc. Int. Conf. on Electrical, Electronics, and Optimization Techniques*, Chennai, India, pp. 924–931, 2016.

[27] P. Salunkhe, M. Bewoor and S. Patil, "A research work on English to Marathi hybrid translation system," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 3, pp. 2557–2560, 2015.

[28] M. A. Ayesha, S. Noor, M. Ramzan, H. U. Khan and M. Shoaib, "Evaluating Urdu to arabic machine translation tools," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 90–96, 2017.

[29] M. Zafar and A. Masood, "Interactive English to Urdu machine translation using example-based approach," *International Journal on Computer Science and Engineering*, vol. 1, no. 3, pp. 275–282, 2009.

[30] P. Dubey, "Machine translation system for Hindi-Dogri language pair," in *Proc. Int. Conf. on Machine Intelligence and Research Advancement, SMVDU*, Jammu and Kashmir, India, pp. 422–425, 2013.

[31] A. Godase and S. Govilkar, "Machine translation development for Indian languages and its approaches," *International Journal on Natural Language Computing*, vol. 4, no. 2, pp. 55–74, 2015.

[32] S. H. Kumhar, M. M. Kirmani, J. Sheetlani and M. Hassan, "Word embedding generation methods and tools: A critical review," *International Journal of Innovative Research in Computer & Communication Engineering*, vol. 8, no. 10, pp. 4015–4026, 2020.

[33] R. M. K. Sinha and A. Jain, "AnglaHindi: An English to Hindi machine-aided translation system," in *Proc. Association for Machine Translation in the Americas*, New Orleans, USA, pp. 1–5, 2002.

[34] V. Goyal and G. S. Lehal, "Advances in machine translation systems," *Languages in India*, vol. 9, no. 11, pp. 139–150, 2009.

[35] H. Choudhary, A. K. Pathak, R. R. Shah and P. Kumaraguru, "Neural machine translation for English-Tamil," in *Proc. Conf. on Machine Translation*, Belgium, Brussels, pp. 770–775, 2018.

[36] V. Goyal and G. S. Lehal, 2011, "Hindi to Punjabi machine translation system," in *Proc. Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, pp. 1–6, 2011.

[37] H. Kaur and V. Laxmi, "A web based English to Punjabi mt system for news headlines," *International Journal of Advanced Research in Computer Science & Software Engineering*, vol. 3, no. 6, pp. 1092–1094, 2013.

[38] M. A. R. Khan and M. K. Jain, "Feature point detection for repacked android apps," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1359–1373, 2020.

[39] N. Binti, M. Ahmad, Z. Mahmoud and R. M. Mehmood, "A pursuit of sustainable privacy protection in big data environment by an optimized clustered-purpose based algorithm," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1217–1231, 2020.

[40] A. Gumaei, M. Al-Rakhami, H. AlSalman, S. Md and A. Alamri, "Dl-har: Deep learning-based human activity recognition framework for edge computing," *Computers, Materials & Continua*, vol. 65, no. 2, pp. 1033–1057, 2020.

[41] F. N. Al-Wesabi, S. Alzahrani, F. Alyarimi, M. Abdul and N. Nemri *et al.,* "A reliable NLP scheme for English text watermarking based on contents interrelationship," *Computer Systems Science & Engineering*, vol. 37, no. 3, pp. 297–311, 2021.

[42] J. Lee, K. Cho and T. Hofmann, "Fully character-level neural machine translation without explicit segmentation," *Transactions of the Association for Computational Linguistics*, vol. 5, no. 2, pp. 365–378, 2017.

[43] J. Chung, K. Cho and Y. Bengio, "A Character-level decoder without explicit segmentation for neural machine translation," in *Proc. Association for Computational Linguistics*, Berlin, Germany, pp. 1693–1703, 2016.