International Journal of Management, IT & Engineering
Vol. 14 Issue 06, June 2024,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Hindi To Dogri Machine Learning Translation System using Embedding Networks

Joginder Kumar
Department of Computer Applications
Govt. Degree College Paloura
joginder1981@gmail.com

Abstract— Machine Translation (MT) has become a critical tool in the digital world, yet many low-resource languages remain underrepresented. Dogri, an Indo-Aryan language spoken primarily in India's Jammu region, is one such language with minimal prior work in MT. This paper presents a Hindi-to-Dogri neural machine translation system designed to address this gap. Leveraging deep sequence learning techniques, the system learns translation patterns from a parallel corpus of annotated Hindi-Dogri sentence pairs. The study evaluates multiple architectures—Embedding LSTM, Bidirectional LSTM (BiLSTM), Embedding BiLSTM, Encoder-Decoder GRU, and BiLSTM with Repeat Vector—focusing on translation accuracy and fluency. The Embedding BiLSTM model achieves the best performance (95% accuracy, BLEU score of 52.46), demonstrating the potential for effective Hindi-Dogri translation. The results indicate that expanding the parallel corpus could further enhance model performance.

Keywords— Machine Translation System, Natural Language Processing, Hindi-Dogri Language Pair, Machine Learning, Neural Machine Translation (NMT).

I. INTRODUCTION

A longstanding goal in computer science has been to enable real-time, accurate machine translation between languages. Advances in artificial intelligence (AI), natural language processing (NLP), and computational power have brought us closer to achieving this objective. Over the past decade, machine translation (MT) has evolved from a theoretical challenge into a practical tool, transforming cross-linguistic communication in academia, industry, and everyday life. [1] , [2].

Figure 1. Advantages of Machine Translation system with NLP

As a branch of artificial intelligence, natural language processing enables effective cross-linguistic communication through computational methods, and has emerged as a powerful tool for overcoming language barriers. This interdisciplinary domain draws upon computer science, linguistics, statistics, and related fields. By the late 20th century, what was once a 17th-century dream - machine translation of natural languages - had become a reality [1]. While debates persist regarding the reliability and quality of machine translation (Figure 1 illustrates its key benefits), its advantages are undeniable.

The most notable benefit is speed: modern systems can translate large volumes of text rapidly [3], [4]. Although machine translation cannot yet match human translators in accuracy, its efficiency makes it indispensable for many applications. While the initial investment in translation software might appear substantial, it proves cost-effective compared to ongoing human translation services. Moreover, users gain permanent access after purchase, and numerous free alternatives exist for basic needs. Modern systems also demonstrate learning capabilities, memorizing terminology and applying it consistently [5], [6].

Despite numerous digital initiatives by the Indian government to enhance regional language accessibility, computationally low-resourced languages like Dogri remain inadequately represented on these platforms. Consequently, non-English speakers must depend on intermediaries or manual information sources. Notably, even with Dogri's official language status, no government websites in Jammu and Kashmir currently offer content in this language. The primary challenge lies in the labour-intensive process of manual content conversion. This gap underscores the critical need for advanced automated machine translation systems capable of efficiently and cost-effectively translating diverse materials – including government documents, educational resources, literary works, and media content – while significantly reducing the time and expense associated with human translation.

The development of a more sophisticated translation system for the Dogri language has been difficult for the researchers. This is mostly due to the language being relatively unexplored and the lack of digital linguistics tools. Dogri is still a growing language, and computing linguistics and natural language processing have not yet been applied to it. Devanagari scripts are used with certain extra symbols for the Dogri language [7]. Preeti D. [8] developed a rule-based

International Journal of Management, IT & Engineering
Vol. 14 Issue 06, June 2024,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

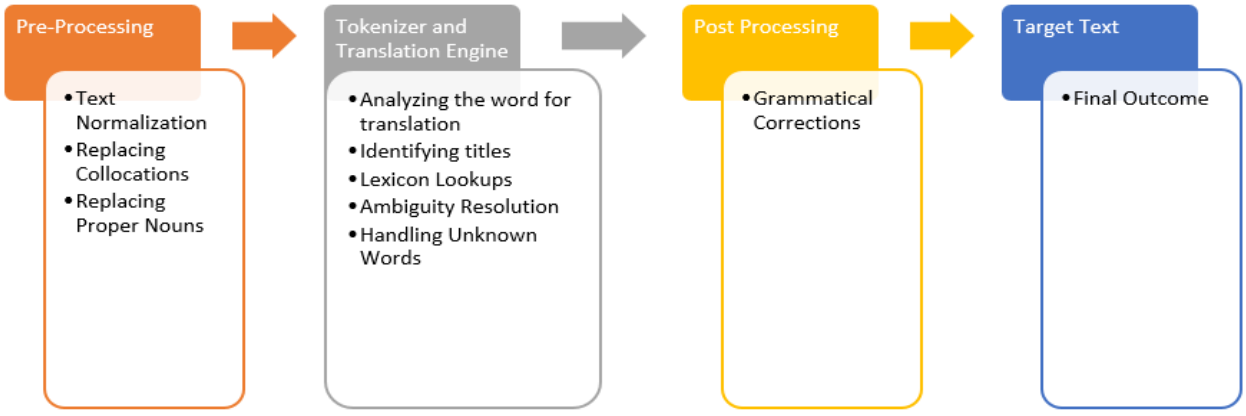

Figure 2. Basic Working of Hindi to Dogri Translation

machine translation system (RBMTS) that utilizes bilingual dictionaries and handcrafted linguistic rules to convert Hindi text into Dogri text is shown in Figure 2. The text is first pre-processed in the original language. The preprocessing stage comprises normalising the Hindi text, identifying proper nouns, and locating collocations within the same text. In Hindi, there are numerous words that have the same meaning but are spelt differently. This is known as text normalisation, whereas collocation refers to sentences that cannot be translated word for word because the words' meanings would alter. After the text has been further provided for by the tokenizer and translation engine, grammatical fixes are made in post-processing, and we get our desired text in the Dogri language. The proposed system generates ambiguous Dogri text for polysemes Hindi words.

Machine translation systems may be developed using a variety of methods [9], [10]. A statistical-based translation system that creates translations based on statistical models whose variables are obtained from bilingual text analysis is recommended in [11] (English-Hindi, Hindi-Marathi, Hindi-Konkani, and Bengali-Hindi language pairs), [12] (Sanskrit to Hindi), and [13] (Hindi, Tamil, English, and Marathi). Finding links between words, sentences, and paragraphs in the source and target texts is its main objective. Statistical machine translation (MT) should only be utilized for basic translation since it ignores context, which frequently results in inaccurate translations.

Rule-based machine translation systems provide a different approach that gets over these drawbacks. A Dogri and Gujarati Hindi language NLP strategy that is based on a rule-based approach is proposed in [14], [15] study to create output utilizing linguistic rules. It necessitates a thorough comprehension of the languages used. It makes use of computer-readable dictionaries. Grammar rules are used in rule-based machine translation (MT). It does a grammatical examination of the source and the target languages to produce the translated sentence. Rule-based MT has a number of significant flaws. First, because of its dependency on lexicons, effectiveness is only obtained after a considerable amount of time. Second, languages have to be individually added. Third, because of its low quality, it requires a large amount of post-editing by humans. As a result, it has certain uses in simple circumstances when a rapid understanding of meaning is needed.

Hybrid techniques are suggested in [16], [17], [18] studies to address these drawbacks. Multiple techniques are used in the development of hybrid machine translation systems. A system created via this technique may make use of linguistic rules from corpora as well. MT hybrid HMT, as its name suggests, combines rule-based and statistical MT. Because translating memory is used, the quality has greatly increased. Hybrid MT still has certain disadvantages despite combining the benefits of both empirical and rule-based MT. Among the most crucial is the requirement for thorough proofreading by human translators. These days, academics are focusing on more sophisticated techniques, such as deep learning machine translation systems (MTS). The chosen strategy is determined by the languages utilized and the accessibility of computer capabilities. The well-known deep neural-based MT system for translating Sanskrit to Hindi was suggested and shown in [12] For this language pair, we also compare how neural MT performs better than the statistical baseline system. Another deep learning approach is also addressed in [19] for English to Punjabi translation.

The improvement of a novel translator that translates between Hindi and Dogri language is the paper's contribution. The following is a description of the main characteristics of this interpretation methodology:

1. In this paper, a Hindi-Dogri corpus is collected and developed.

2. The paper suggested a Hindi-to-Dogri translation system based on machine learning and natural language processing that identifies the appropriate relevance of an objective term.

3. The paper utilized the sentence structure of each language to determine the actual meaning of the phrase.

4. Then the paper has implemented the five deep learning models with an embedding encoder and decoder module for Hindi-Dogri translation.

5. The suggested model successfully translates words from Hindi to Dogri with high accuracy and produces superior word predictions.

The remainder of the paper is organized into four sections: Section 2 discusses related work in the field of machine language translation (MLT). Section 3 outlines the proposed methodology along with the corresponding flowcharts. Section 4 presents a comparative analysis of the results obtained from various state-of-the-art models. Finally, Section 5 provides the conclusion and outlines potential directions for future research.

II. LITERATURE REVIEW

Thukroo and Bashir [20] proposed a CNN-based approach to identify six languages spoken in Jammu and Kashmir—Kashmiri, Urdu, Dogri, Hindi, English, and Ladakhi—using a hybrid dataset of TV recordings (for regional languages) and standardized corpora (IIIT-H for Hindi, VoxForge for English). The model pre-processed speech segments (5-second duration, noise-filtered), converted them into mel-spectrograms via FFT, and achieved 100% training/testing accuracy at 100 epochs, demonstrating resilience to variations in speech duration, speaker demographics, and acoustic conditions

Gupta and Jamwal [21] developed the Dogri language's stemmer using an unsupervised machine learning methodology. In addition to being widely utilised in morphology analyses, stemming and lemmatisation are important NLP methods that are also applied in information extraction for query optimisation, text summarisation, word meaning identification, document segmentation, and computational linguistics. The majority of NLP activities are rule-based, need linguistic monitoring, and have an average accuracy of 69%. Additionally, it is believed that by including language-specific rules into the domain expertise, the degree of accuracy may be significantly raised.

Dubey [8] created a machine translation system (MTS) to translate Hindi text into Dogri, which is included in this publication. The preprocessing stage, part of speech tagging, transcription, and other crucial steps are all part of the machine process of translating; however, they are not all required in every MTS. This research has aided in the choice of the translation strategy, the preparation tasks to be performed on the source document, the creation of guidelines for inflectional assessment, as well as other tasks for handling certain unique Dogri circumstances. This research analyses Hindi and Dogri's grammatical and lexical structures. 98.54% of the phrases are understandable, according to the users' evaluation. The average of all sentences with a score of two or above is shown below. These are sentences that scored a 2 or higher on the accuracy test. Consequently, 98.71% of the test sheet's statements were correct.

Jamwal et al. [22] recognised the finite adjectives and all of their inflectional variants, as well as determining if the finite verb in the literature is appropriate in its situation. The hybrid approach for Dogri verb production is presented and used to produce all of the verb's inflected forms. Researchers discovered throughout the process of the research that there is no one widely accepted model which can account for the derivational events for the formation of the verbs. The way that the contextual qualities are expressed syntactically varies amongst Indian languages. For the automated production and identification of verbs, the linkage between syntax and semantics is crucial. The findings of the discovery of the suggested model for the creation of the verbs in Dogri are tested using the corpus-based technique. They evaluated the model on five different sets of data, and after computing the average of all the results for accuracy, precision, recall, and F-score, we found that they were, respectively, 69.213, 93.861, 75.278, and 83.500.

Bawa [23] used a hybridised type of direct and rule-based translation software to deliver an MTS between Sanskrit and English. The disparity between Sanskrit and English is also discussed in this paper along with a suggestion on how to address it. Sanskrit-English and Sanskrit-UNL multilingual lexicons, a labelled Sanskrit corpus, a Sanskrit analytic rule basis, and an ELGR base have all been employed in the developed framework. Additionally, a unique technique that builds a parse tree from the parse tables is presented in this paper. The target language phrase is produced using the ELGR basis and bilingual lexicon. The suggested system received a BLEU rating of 0.7606, a fluency score of 3.63, and an adequacy score of 3.72 when assessed using Python's Natural Language Toolkit API. The proposed method works better than current systems when compared to cutting-edge systems, according to the assessment.

Singh et al. [24] suggested a brand-new corpus-based translation method for Sanskrit to Hindi conversion using the Bhagavad Gita—the Lord's Song—as data input. In this study, deep neural networks are utilised for training. After data processing and analysing, the input data are provided to the neural networks, which subsequently carry out auto-tuning to improve the model. Using this proposed framework, the target text is created with an improved BLEU score and word error rate. The suggested MTS performs better than the rule-based MTS by 39.6% because it has a lower WER.

Goyal et al. [25] created an MTS based on CNN and bidirectional gated recurrent units (Bi-GRU). Additionally, they conduct a number of studies using multilingual Punjabi and Hindi data. The experimental tests were carried out in this work's results show that the bidirectional GRU and CNN-based design, alongside enhanced word embeddings (EWE), has excelled, with precision at 0.926, recall at 0.907, and F-score at 0.9164. Despite the absence of a huge feature set or any kind of spatial information, improved word embeddings enhance the performance of a named entity identification system built on the foundation of the proposed method.

Bisht and Solanki [26] suggested effort seeks to illustrate several deep learning algorithms and methods that may be used for the most accurate and effective caption translation from English to Hindi. According to the findings, transformer-

based methods perform better than sequence-to-sequence methods on all measures (with around 5-20% higher accuracy ratings). Additionally, pre-trained transformer-based techniques are also particularly good at handling ambiguity. Findings further demonstrate that text-only techniques are adequate in these low-resource circumstances, but multimodal techniques are not able to enhance translation quality. Therefore, for the majority of English-to-Hindi picture caption translation systems, text-only pre-trained converters are advised.

Ahmad and Singla [27] utilised multinomial Naive Bayes, decision trees, and SVM to identify languages at the word level in social media material that was code-mixed with English, Hindi, and Urdu. The challenge of NLP is made more difficult and complicated by the fact that people, particularly in multilingual countries, prefer to write in numerous languages and strategies to communicate their ideas effectively. This diversity in language usage poses a significant obstacle for machine translation systems, which must be able to accurately interpret and translate text in multiple languages. As a result, researchers continue to explore new methods and technologies to improve the accuracy and efficiency of multilingual translation systems. An MTS is therefore very necessary for developing sophisticated NLP systems employing code-mixed data. Machine learning techniques have recently attracted a lot of interest in the area of categorisation issues. SVM outperforms the other two methods with an accuracy of 83.58% for Hindi-English and 75.79% for Urdu-English, correspondingly.

Jamwal and Sen [28] performed the transformation of the Dogri language query to its corresponding SQL query format. The basic and frequently utilised selection of queries that the developed scheme receives as input and afterwards transforms have been highlighted. After preprocessing the data by removing stopwords and tokenising the data, the Dogri language question was divided into clauses. Due to the considerable ambiguity of natural languages, 150 enquiries were tried, with each inquiry being tested in around five different ways. The suggested model successfully transformed Dogri language queries with an accuracy of 88%.

Shen [28] examined an intelligent data-driven MTS system. It is advised to refer to English linguistic expressions as functioning noun phrases based on their brief syntactic properties. In other words, systemic functional grammar bases its definition of noun phrase borders on the structural role of linguistic expressions in phrases. The F value is 98.88%, and the classification result is good. The investigation's findings may make sentences easier to understand, make analysis less hard and complicated, serve as a foundation for further research, and improve MTS's capacity to handle structural ambiguity.

III. METHODOLOGY

A. *Overview of Machine Translation System Using Deep Learning*

Machine translation (MT) refers to the automated process of converting text from a source language to a target language. Our work focuses specifically on Hindi-to-Dogri translation, which presents several unique challenges: significant differences in alphabets and grammatical structures between the languages, the inherent complexity of sequence-to-sequence conversion (particularly for phrases rather than numerical data), and the absence of single "correct" translations (e.g., Hindi's gender-neutral pronouns complicate conversion to Dogri).

When implementing machine translation, several key considerations emerge: First, while MT quality has improved significantly, human translators remain the gold standard for accuracy; we recommend incorporating human review into the translation workflow. Second, despite common perception, machine translation is not a new field - systematic efforts to develop automated translation systems date back to the 1970s. Three main methods have evolved throughout time:

- Rule-based Machine Translation (RBMT): 1970s-1990s
- Statistical Machine Translation (SMT): 1990s-2010s
- Neural Machine Translation (NMT): 2014-till date

Rule-based machine translation systems (RBMTS) suffer from several limitations, including: heavy reliance on comprehensive dictionaries, the labour-intensive process of manual rule creation (requiring linguistic expertise), and increasing system complexity as additional rules are incorporated. Statistical approaches present their own challenges, particularly: dependence on large bilingual corpora, difficulty in correcting specific translation errors, and poor handling of language pairs with substantial syntactic divergence in word order..

Consequently, researchers are increasingly focusing on deep learning-based machine translation systems to overcome these limitations. In recent years, deep learning has transformed numerous fields, ranging from computer vision to artificial intelligence in gaming. Following this trend, machine translation (MT) has shifted from rule-based systems and statistical phrase-based approaches to neural-based techniques using deep learning. Modern neural machine translation (NMT) models leverage vast training datasets and unprecedented computational power to comprehensively analyze source sentences, automatically identifying relevant features at each stage of target text generation. This advancement eliminates previous independence assumptions, leading to substantial improvements in translation quality. In some cases, neural translation has even approached human-level performance for isolated phrases.

To create a better machine translation system, many deep learning methods and libraries are needed. The system that will translate the sentence from the source language to the target language is trained using RNNs, LSTMs, etc. It is a wise decision to adapt the appropriate networks and deep learning algorithms since it tailored the system to maximise the translation system's accuracy in comparison to others. We are proposing Hindi-to-Dogri translation using deep learning.

International Journal of Management, IT & Engineering
Vol. 14 Issue 06, June 2024,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

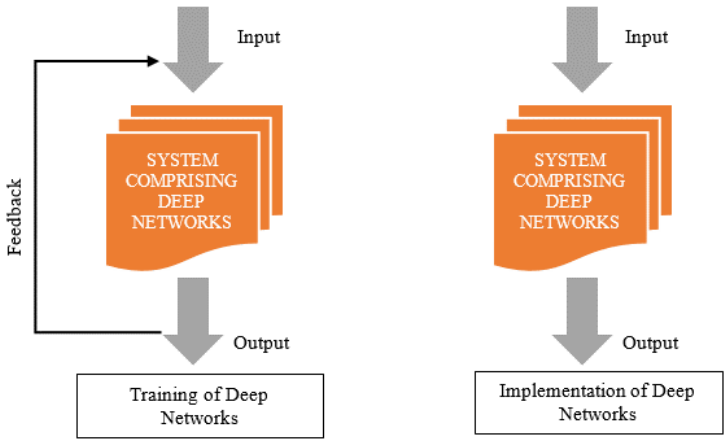

Figure 3. Training and implementing of MT using Deep leaning

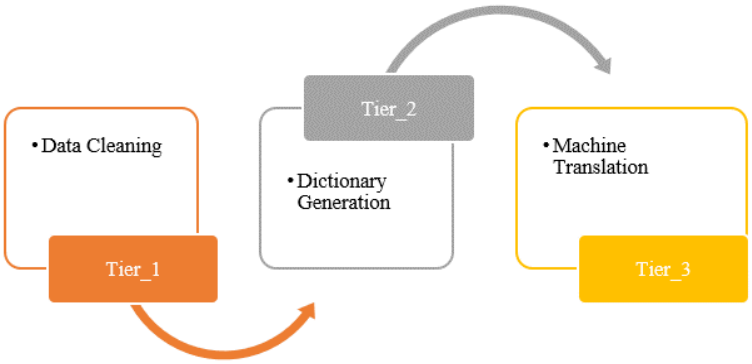

Figure 4. Workflow

Deep learning is a cutting-edge method that is used in many machine learning applications. With training, it allows the system to learn just like a human and to become more effective. Even deeper and more abstract layers may be used in deep learning techniques to represent features utilising supervised and unsupervised learning. Deep learning is being employed in voice recognition, machine translation, data analysis, image applications, etc. Deep neural networks are neural networks with several hidden layers (DNNs). The above connections first enter into the training stage, then are executed to deal with the issue, as shown in Figure 3. The provided task affects the structure and training process of DNNs.

The major goal here is to create a system that serves as a translator. Unlike traditional approaches that rely on extensive rule libraries, our method employs a trained deep neural network that leverages contextual patterns and learned linguistic representations. Alternative machine translation techniques include word alignment, rule-based reordering, and language modeling, among others.

B. Proposed MTS

The Hindi-to-Dogri MTS proposed in our work is divided into three tiers (Figure 4), i.e., data cleaning, dictionary generation and finally machine translation using deep learning translation.

Figure 5. Data Cleaning

1) Dataset / Corpus Used in the Study

The study utilizes a Hindi-Dogri parallel corpus containing approximately 100000 sentence pairs (approximately 0.1 million), with each sentence limited to 80 words or fewer. The dataset was compiled from multiple sources like: conversational sentence pairs from the Dogri-Hindi Conversation Book published by the Central Hindi Directorate (digitized using OCR techniques), sentences extracted from major Hindi newspapers (Amar Ujala, Dainik

International Journal of Management, IT & Engineering
Vol. 14 Issue 06, June 2024,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Jagran, BBC Hindi, Dainik Bhaskar, and Hindustan), and vocabulary items including collocations, dictionary entries, and proper nouns. This comprehensive corpus totals 510,086 Hindi tokens and 522,747 Dogri tokens, providing balanced coverage of conversational, journalistic, and lexical content. The same dataset was used to five neural machine translation (NMT) models, enabling direct performance comparison between approaches. The dataset creation methodology was originally published in [29].

2) Data Cleaning

Data cleansing is the process of identifying and removing invalid data from a dataset. Invalid data may include incorrect, duplicate, incomplete, corrupt, or improperly formatted entries. This process is critical for maintaining data integrity and ensuring the reliability of subsequent analyses.

In machine translation systems, which learn linguistic patterns from training data, certain elements can hinder algorithm performance. Non-translatable components like emojis, usernames, and quoted text may confuse the model, as can excessive capitalization and irregular punctuation. The cleansing process typically begins with quote removal, followed by data normalization - which involves eliminating language-agnostic elements (e.g., numbers, special characters) and standardizing formatting (e.g., removing extra spaces). These steps help optimize the training data quality, enabling more effective model learning and improved translation accuracy.

Data cleansing typically accounts for over 30% of the time required to achieve data integrity. This substantial investment is justified by its crucial role in ensuring dataset quality: without proper cleansing, data-driven systems risk producing erroneous outputs that undermine decision-making processes.

3) Dictionary Generation

An effective machine translation system typically requires three key dictionaries: a source language dictionary, a bilingual dictionary, and a target language dictionary. Constructing comprehensive dictionaries is significantly more time- and data-intensive than developing grammatical rules. Translation accuracy improves substantially with more extensive bilingual dictionary coverage, yet most Indian language systems lack electronic bilingual dictionaries, necessitating manual or automated entry addition.

For automated dictionary construction, a substantial parallel corpus is essential. When unavailable, manual entry becomes the only option. Domain-specific systems allow focused entry addition, while domain-independent systems require exhaustive vocabulary coverage

4) Machine Translation using Deep Learning

Machine learning translation is the final stage of machine translation. Figure 6 depicts the model's overall design. The encoding and decoding module, in addition to the attention module, are the primary components of the neuro-machine translation system for Hindi to Dogri that was designed and implemented in this section. Figure 6 presents the overall architecture of the system. The end-to-end neuro-machine translation process is only possible thanks to the collaborative efforts of all of the modules. The following is a description of the functions that each component possesses:

Encoding module: such a module is included as a component of the "encoder-decoder" framework that is used for end-to-end neuro-machine translation. For instance, a bidirectional LSTM/GRU network is typically employed for the purpose of mapping the word ID sequence corresponding to a sentence onto a continuous, dense, high-dimensional vector representation.

Attention module: When generating words in the target language, the attention module pays dynamic and selective attention to different parts of the source language sentence. This is done to obtain more accurate context vectors, which serve as the basis for translation.

Another component of the "encoder-decoder" framework, the decoding module is analogous to the encoding module and serves the same purpose. The semantic representation in vector form will have to be remapped to sentences in natural language as one of this module's functions.

All these modules are divided into the following layers:

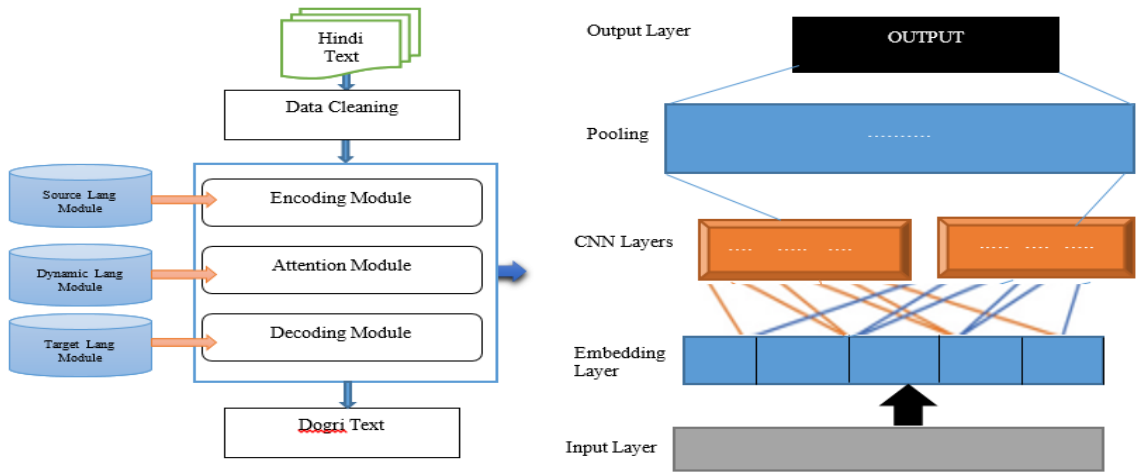

Figure 6. Proposed Machine Language Translation

International Journal of Management, IT & Engineering
Vol. 14 Issue 06, June 2024,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

- Input Layer: User-defined as h for Hindi text features and dictionary-defined items as I. The first layer, known as the input layer, used this vector, I.
- Embedding Layer: The interaction vector, x^{cnn}, was created.
- Convolution Layer: The convolution layer was used to extract context data. Using the interaction vector x^{cnn}, the context of the input vectors was evaluated using the convolution structure.

$$z_n^m = f(w_z^m * x_{(:n:(n+w-1))}^{cnn} + p_z^m) \quad (1)$$

Where, f is an activation function, * is a convolution operator. The following might be used to create the feature latent vectors:

$$z^m = [z_1^m, z_2^m, z_3^m \dots \dots \dots z_{1n}^m, z_{1k-w+1}^m] \quad (2)$$

Pooling Layer: This layer extracts relevant features from the convolutional model, where a contextual FV describes an interaction.

$$x_f^{cnn} = [\max(z^1), \max(z^2) \dots \dots \max(z^j)] \quad (3)$$

The m_{th} shared weight obtains the contextual feature vector. The vector is constructed as follows:

$$x_f^{cnn} = f(x_f^{cnn}) \quad (4)$$

Output Layer: After analyzing the data from the previous layer, this layer calculates the probability score for the final prediction. It is shown mathematically as:

$$\hat{P}_{hi} = \sigma[h^T x_f^{cnn}] \quad (5)$$

In which σ is the sigmoid function given by $\sigma(a) = \frac{1}{1+e^{-a}}$ is the weights of the output layer.

Encoding Mechanism: The context information of the source language translation should be carefully considered throughout the translation process. By sequentially scanning the input sequence, the cyclic NN may memorize context data and keep temporal information. Furthermore, the word vector produced using this approach only considers the context on the left side of the word, ignoring the context on the right. Translation by neuro-machines is carried through using bidirectional recurrent NN. Two circular NN can the input data in various ways, and the hidden layer state sequence is collected in two of such directions. By combining the two directions, a whole background vector sequences are finally produced. Depending on this concept, this research builds the encoder component of the neuro-machine translation system utilizing the bidirectional LSTM/RNN/GRU network structure. The structure is a forward- and backward-transferring. The structure is a forward- and backward-transferring. The forward network creates the forward hidden layer state sequences $\vec{h} = (\overrightarrow{h_1}, \overrightarrow{h_2}, \overrightarrow{h_3}$ --) by scanning the input sequence X=(x_1, x_2----x_n) from front to backward. The backward network simultaneously creates the reverse hidden layer state sequence $\vec{h} = (\overrightarrow{h_1}, \overrightarrow{h_2}, \overrightarrow{h_3}$ --) and examines the input sequence from backward towards the front, or from x_n to x_1. The source language end words are expressed as h by connecting different hidden states. Because of this form, the background vectors used this to describe the input sequence data may simultaneously contain all the context's data.

Attention Mechanism: Although neuro-machine translation has significantly increased effectiveness, accurate coding remains a problem. Because the encoder transfers the sentence to a fixed-dimension vector regardless of how long it is, in order to address this issue, attention-based neuro-machine translation is presented, which uses an attention model to concurrently train the alignment and translation processes. Instead of using the entire source language sentence each time, the decoder dynamically considers the context of the source language sentence while generating the target language word y_i. An end-to-end neural machine translation model with an attention mechanism.

The design of an attention-based context vector C is essential for the integration of an attention mechanism in neuro-machine translation. The source language hidden state sequence e and the attention weight an are added to produce the context vector C_i that corresponds to the word Y_i at the source, whereas an is produced by the interactions of the target hidden state d_{t-1} and the source hidden state "e" at time t-1. The mathematical model explains how the three mentioned intermediary variables were computed:

$$c_t = \sum_{j=0}^{n} at, jej \quad (6)$$

$$a_{i,j} = \frac{\exp(b_{i,j})}{\sum_{j=0}^{n} \exp(b(t,k))}$$

$$b_{i,j} = m\,(d_{t-1}, ej)$$

Where h and f are non-linear activation functions and exp is an exponential function depending on e and m. The conditional probability distributions of the hidden state t d of the current time decoding and the target word t y to be created at the current time may be solved by the following expression, accordingly, after getting the context vector t c:

$$d_t = hdecoder(d_{t-1}, y_{t-1}, c_t) \quad (7)$$

$$p(yt\ I\ t < t, X) = softmax(f(d_{t-1}, y_{t-1}, c_t)$$

Decoding Mechanism: It will consist of two reasonably natural languages, or two relatively independent language models, because it is a machine translation technology. The upgraded bidirectional network in the encoding module is used to generate the hidden layer sequence. The identical decoding module s' requires extra output gates so this function has already been represented in the output module. However, if this were the only step, there would be no communication between the encoder and decoder and no translation:

$$z_i = \sigma(W_z\, e_{i-1}^y + U_s\, S_{i-1} + C_z\, C_i) \quad (7)$$

$$r_i = \sigma(W_r\, e_{i-1}^y + U_s\, S_{i-1} + C_r\, C_i)$$

$$s'_i = tanh(W_r\, e_{i-1}^y + U_r\, [r_i\, o\, S_{i-1}]\ C_s\, C_i)$$

$$s'_i = (1 - z_i) o\, S_{i-1} + z_i\, o\, s'_i)$$

where $e_i^y \in R^m$ is the m-dimensional word embedding vector of target language; o is point computation; $W_r, W_z, W_s \in R^{mxn}$, $U_r, U_z, U_s \in R^{mxn}$ and $C_r, C_z, C_s \in R^{mxn}$ are weight matrix; Because the output module will output the target language

sequence in translation one at a time, the decoder's computing for the bi-directional language modeling is not necessary.

IV. Results and Discussion

The designed machine language translation for Hindi-Dogri language is designed and implemented with five models, such as embedding LSTM, bidirectional LSTM (BiLSTM), embedding BiLSTM, encoder-decoder GRU and BiLSTM with repeat vectors. The designed framework is implemented in Python using Google Colab. The minimum batch size of the training module was 32. Adam optimiser for training the models. The training is done for the 20 epochs. The implementation is done in Keras, using TensorFlow as the backend. All networks are trained on the Tesla P100-PCIE GPU.

A. Dataset Description

The dataset consists of approximately 100000 Hindi Dogri sentences in the corpus. The TensorFlow framework has been used for the implementation task for the performances. Fig. 7 shows the Hindi-Dogri sample of corpus datasets. The dataset will be split for training and testing purposes in a ratio of 70:30. 70% of the dataset was used for training, and 30% of the dataset is used for testing.

B. Result Analysis

In this section, five deep learning models were implemented and evaluated using the Hindi–Dogri bilingual corpus. The models include: Embedding LSTM, Bidirectional LSTM (BiLSTM), Embedding BiLSTM, Encoder–Decoder GRU, and Bidirectional LSTM with Repeat Vector.

Figure 8 presents the training and validation accuracy of the Embedding LSTM model. Over 20 epochs, the accuracy gradually increases and stabilizes around 0.95. Figure 9 shows the corresponding loss, which decreases from 5.0 to 0.6, with a final validation loss of 0.64.

Figure 10 illustrates the training and validation accuracy for the Bidirectional LSTM model, which also stabilizes around 0.95 after 20 epochs. As shown in Figure 11, the loss decreases from 2.0 to 0.4, with a final validation loss of 0.42.

Figure 12 displays the accuracy for the Embedding BiLSTM model, which reaches approximately 0.953. The loss curve in Figure 13 decreases from 10 to 2, with an average validation loss of 2.27.

The Encoder–Decoder GRU model's accuracy remains steady at around 0.95 across 20 epochs, as shown in Figure 14. The corresponding loss, shown in Figure 15, decreases from 0.5 to 0.3, with an average validation loss of 0.408.

Figure 16 shows the accuracy curve for the Bidirectional LSTM with Repeat Vector model, which also converges around 0.95. Figure 17 indicates a loss reduction from 0.7 to 0.4, with an average validation loss of 0.47. Table 1. summarizes the performance of all models.

Figure 18 compares model accuracies. Bidirectional LSTM achieves the highest accuracy (95.5%), followed closely by the other models, all of which exceed 95%. Although all models perform comparably, Bidirectional LSTM demonstrates slightly superior accuracy.

Figure 19 presents the loss comparison. Bidirectional LSTM records the lowest loss (0.426), indicating the best generalization among the models. Figure 20 provides an overall comparison of the proposed models with existing state-of-the-art approaches

International Journal of Management, IT & Engineering
Vol. 14 Issue 06, June 2024,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

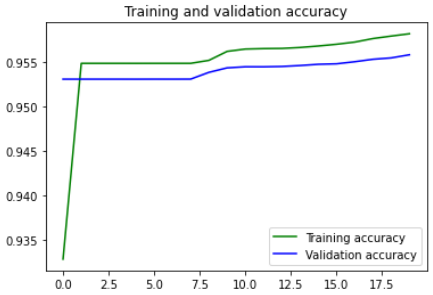

Figure 8. Training and Validation Accuracy of Embedding LSTM

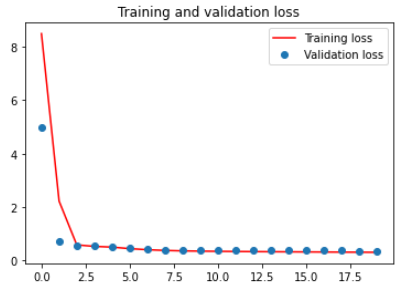

Figure 9. Training and Validation Loss of Embedding LSTM

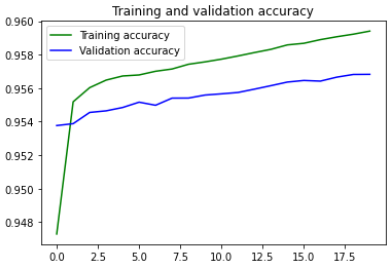

Figure 10. Training and Validation Accuracy of Bidirectional LSTM

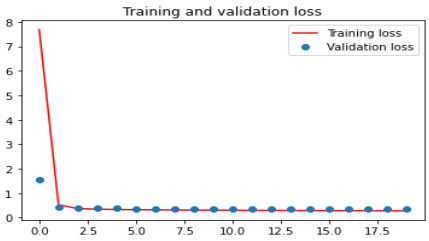

Figure 7. Training and Validation Loss of Bidirectional LSTM

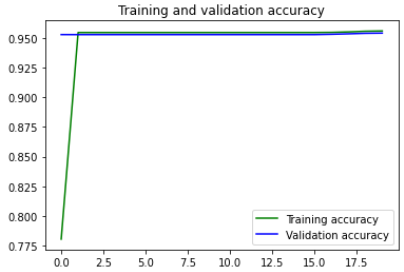

Figure 12. Training and Validation Accuracy of Embedding Bidirectional LST

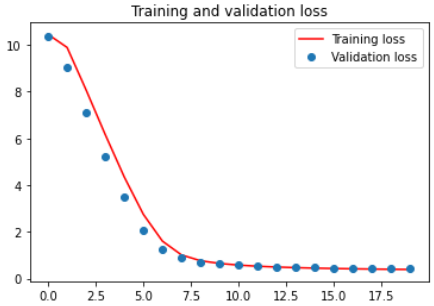

Figure 13. Training and Validation Loss of Embedding Bidirectional LSTM

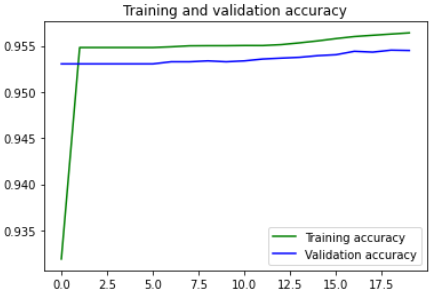

Figure 14. Training and Validation Accuracy of Encoder-Decoder GRU

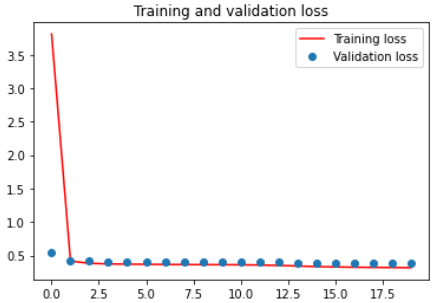

Figure 15. Training and Validation Loss of Bidirectional LSTM

International Journal of Management, IT & Engineering
Vol. 14 Issue 06, June 2024,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

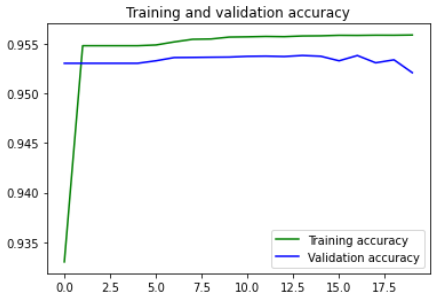

Figure 16. Training and validation accuracy of bidirectional LSTM wit repeat vector

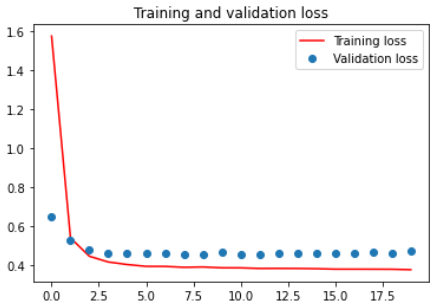

Figure 17. Training and Validation Loss of Bidirectional LSTM

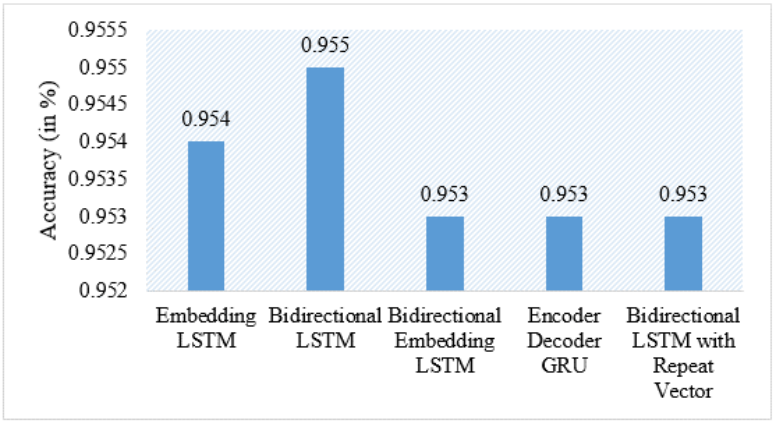

Figure 18. Accuracy Comparison of Different Techniques

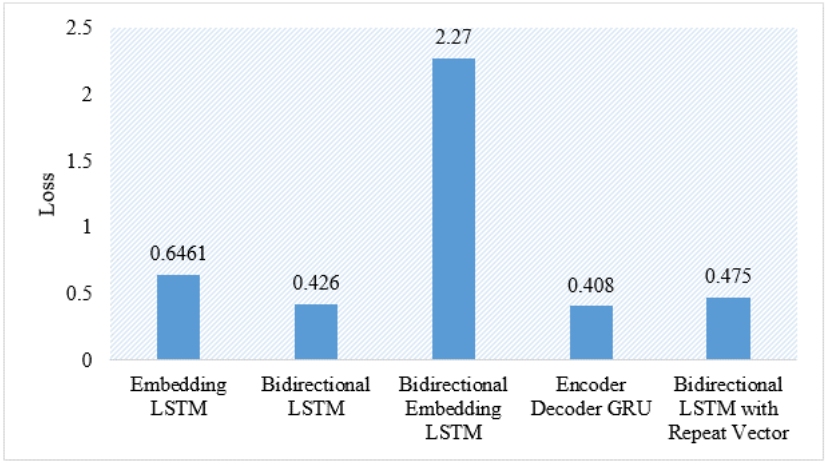

Figure 19. Loss Comparison of Different Techniques

International Journal of Management, IT & Engineering
Vol. 14 Issue 06, June 2024,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

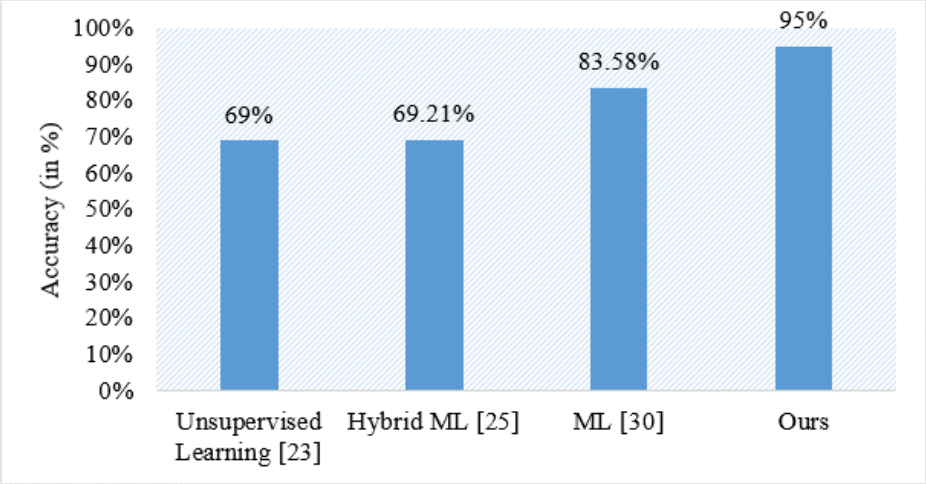

Figure 20. Comparative State-of-Art

Hindi Language Sentences	Dogri Language Sentences
इस वर्ष की शुरुआत से अभी तक कश्मीर घाटी में सुरक्षाबलों को 66 आतंकियों को मार गिराने में सफलता मिली है	इस बरे दी शुरुआत थमां हून तगर कश्मीर घाटी च सुरक्षाबलें गी 66 आतंकियें गी मारने च सफलता मिली ऐ
पिछले महीने जून में 11 दहशतगर्द ढेर किए गए थे	पिछले म्हीने जून मच 11 दहशतगर्द ढेर कीते गे हे
सबसे ज्यादा 17 आतंकी अप्रैल के महीने में मारे गए थे	सारें शा मते 17 आतंकी अप्रैल म्हीने च मारे गे हे
पुलवामा जिले के राजपोरा इलाके में 15 घंटे से अधिक चली मुठभेड़ में आतंकी संगठन लश्कर ए तैयबा के पांच दहशतगर्दों को सुरक्षाबलों ने मार गिराया	पुलवामा जिले दे राजपोरा इलाके म 15 घंटे थमां मते चिर चली मुठभेड़ च आतंकी संगठन लश्कर ए तैयबा दे पंज दहशतगर्दें गी सुरक्षाबलें मारेआ
इनमें जिला कमांडर व एक पाकिस्तानी आतंकी शामिल है	इंदे च जिला कमांडर ते इक पाकिस्तानी आतंकी शामल न
ऑपरेशन में सेना के एक जवान शहीद जबकि दो अन्य घायल हुए हैं	ऑपरेशन च सेना दे इक जवान शहीद जद के दो होर घायल होए न
मुठभेड़ स्थल से हथियार भी बरामद किए गए हैं	मुठभेड़ आह्ली जगहा थमां हथियार बी बरामद कीते गे न
मुठभेड़ शुरू होते ही जिले में मोबाइल इंटरनेट सेवाएं स्थगित कर दी गई थीं	मुठभेड़ शुरू होंदे गै जिले च मोबाइल इंटरनेट सेवाएं स्थगित करी दित्ती गेई ही
सुरक्षाबलों को गुरुवार देर रात राजपोरा के हान्जिन गांव में आतंकियों के एक बड़े दल के छुपे होने की सूचना मिली थी	सुरक्षाबलें गी बीरवार रातीं राजपोरा दे हान्जिन ग्रांऽ च आतंकियें दे इक बड्डे जत्थे दे छप्पे होने दी सूचना मिली ही
इस इनपुट के आधार पर एसओजी ने सेना की 44 राष्ट्रीय राइफल्स और सीआरपीएफ की 182 और 183 बटालियन के जवानों के साथ मिलकर इलाके की घेराबंदी की	इस इनपुट दे आधार उप्पर एसओजी ने सेना दी 44 राष्ट्रीय राइफल्स ते सीआरपीएफ दी 182 ते 183 बटालियन दे जवानें कन्नै मिलियै लाके दी घेराबंदी कीती
इस दौरान एक मकान में छुपे आतंकियों ने जवानों पर ताबड़तोड़ गोलियां बरसाईं	इस दौरान इक मकाने च छप्पे दे आतंकियें जवानें उप्पर ताबड़तोड़ गोलियां बरसाइयां
जवानों ने मोर्चा संभाला और कई बार आतंकियों को आत्मसमर्पण करने का मौका दिया	जवानें मोर्चा संभालेआते केई बार आतंकियें गी आत्मसमर्पण करने दा मौका दित्ता
उन्होंने हर बार उसे ठुकराया और जवानों पर फायरिंग जारी रखी	उ ने हर बारी उस्सी ठुकराया ते जवानें उप्पर फायरिंग जारी रक्खी
इस पर जवानों की जवाबी कार्रवाई से मुठभेड़ शुरू हो गई	इस उप्पर जवानें दी जवाबी कार्रवाई च मुठभेड़ शुरू होई गेई
गुरुवार रात के वक्त तो कोई आतंकी नहीं मारा गया, लेकिन शुक्रवार को पांच आतंकियों को मार गिराने में सफलता मिली	बीरवार रातीं ते कोई बी आतंकी नेईं मारेआ गेआ, पर शुक्रवार गी पंज आतंकियें गी मारने च कामयावी मिली
कश्मीर रेंज के आईजी विजय कुमार ने पांच आतंकियों के मारे जाने की पुष्टि करते हुए बताया कि मारे गए आतंकियों में एक लश्कर का जिला कमांडर निशाज लोन उर्फ खिताब, पाकिस्तानी आतंकी अबु रेहान उर्फ तौहीद, दानिश मंजूर शेख , आमिर वागे व मेहरान मंजूर शामिल है	कश्मीर रेंज के आईजी विजय कुमार नै पंजे आतंकियें दे मारे जाने की पुश्टि करदे होए सनाया जे मारे गे आतंकियें च इक लश्कर का जिला कमांडर निशाज लोन उर्फ खिताब, पाकिस्तानी आतंकी अबु रेहान उर्फ तौहीद, दानिश मंजूर शेख , आमिर वागे ते मेहरान मंजूर शामल न
राजपोरा इलाके में पहले मुठभेड़ में दौरान पत्थरबाजी की घटनाएं देखने को मिलती थीं जो इस बार नहीं देखने को मिली	राजपोरा लाके च पैह्ले मुठभेड़ दे दौरान पत्थरबाजी दी घटना दिखने गी मिलदियां हियां ओह् इस बार नेईं दिखने गी मिलियां
मारे गए आतंकियों से एक एसएलआर भी बरामद की गई है जिसे आतंकियों ने लोअर मुंडा में टीवी टावर के गार्ड से 2016 में लूटी थी	मारे गए आतंकियें शा इक एसएलआर बी बरामद कीती गेई ऐ जिस्सी आतंकियें लोअर मुंडा थमां टीवी टावर दे गार्ड कोलां 2016 च लुट्टेआ हा

Figure 8. Sample of Hindi-Dogri Corpus

TABLE 1. PERFORMANCE EVALUATION

Models	Accuracy	Losses
Embedding LSTM	0.954	0.6461
Bidirectional LSTM	0.955	0.426
Bidirectional Embedding LSTM	0.953	2.27
Encoder Decoder GRU	0.953	0.408
Bidirectional LSTM with Repeat Vector	0.953	0.475

V. Challenges in Machine Translation

Effective machine translation requires more than word-for-word substitution; it must achieve human-level accuracy in conveying meaning. This presents significant computational and linguistic hurdles, including:

Resource Scarcity: Developing a machine translation system for Hindi to Dogri is particularly challenging due to the scarcity of linguistic resources. Essential components such as bilingual dictionaries, parallel corpora, and morphological analyzers are either limited or nonexistent for Dogri, a recognized yet low-resource language. As a result, creating these foundational resources from scratch requires significant time, effort, and financial investment. This scarcity not only impacts the quality and scalability of machine translation models but also highlights the urgent need for focused resource development to support Dogri and similar under-resourced languages.

Lexical Ambiguity: Natural languages like Hindi are rich in polysemous words—terms that carry multiple meanings depending on context. Accurately translating such words into Dogri poses a significant challenge, as the correct interpretation often depends on subtle contextual cues. This ambiguity becomes especially problematic in low-resource language pairs, where limited training data may restrict the model's ability to learn nuanced distinctions. Effectively addressing this issue remains a core difficulty in developing reliable and context-aware Hindi-to-Dogri machine translation systems.

VI. Conclusion

This research presents a neural machine translation (NMT) system for Hindi-to-Dogri, a language pair with shared syntax but limited digital resources. Among the five NMT models evaluated, the Bidirectional Embedding LSTM demonstrated superior performance. While deep learning offers promising results, the scarcity of high-quality Dogri corpora remains a significant challenge. Our work not only introduces a functional Hindi-Dogri translation system but also highlights unique linguistic aspects of Dogri that require further exploration. Future efforts should focus on expanding the training dataset, improving ambiguity resolution, and enhancing translation accuracy to bridge the resource gap for low-resource languages like Dogri. With larger and more refined datasets, NMT systems can achieve even greater precision, solidifying their role as the future of machine translation.

References

[1] M. Singh, R. Kumar, and I. Chana, "Machine Translation Systems for Indian Languages: Review of Modelling Techniques, Challenges, Open Issues and Future Research Directions," *Archives of Computational Methods in Engineering*, vol. 28, no. 4, pp. 2165–2193, Jun. 2021, doi: 10.1007/s11831-020-09449-7.

[2] J. Bansal, P. Bansal, and R. K. Chakrawarti, "Analysis of Hindi-English Poetry Translation through Machine Translation Systems," in *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, IEEE, Oct. 2021, pp. 1–8. doi: 10.1109/ISCON52037.2021.9702364.

[3] Sitender, S. Bawa, M. Kumar, and Sangeeta, "A comprehensive survey on machine translation for English, Hindi and Sanskrit languages," *J Ambient Intell Humaniz Comput*, vol. 14, no. 4, pp. 3441–3474, Apr. 2023, doi: 10.1007/s12652-021-03479-0.

[4] M. Chouhan and D. K. Srivastava, "Neural approach-based quality estimation in improving translation of English to Hindi using machine translation under data science," in *2021 International Conference on Computational Performance Evaluation (ComPE)*, IEEE, Dec. 2021, pp. 035–039. doi: 10.1109/ComPE53109.2021.9751729.

[5] M. Akter, M. Shahidur Rahman, M. Zafar Iqbal, and M. Reza Selim, "A Review of Statistical and Neural Network Based Hybrid Machine Translators," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, IEEE, Sep. 2018, pp. 1–6. doi: 10.1109/ICBSLP.2018.8554792.

[6] M. K. Rohil, S. Saini, and R. K. Rohil, "An Interactive System leveraging Automatic Speech Recognition and Machine Translation for learning Hindi as a Second Language," in *2022 3rd International Conference for Emerging Technology (INCET)*, IEEE, May 2022, pp. 1–4. doi: 10.1109/INCET54531.2022.9824940.

[7] P. Dubey, "The Hindi to Dogri machine translation system: grammatical perspective," *International Journal of Information Technology (Singapore)*, vol. 11, no. 1, pp. 171–182, Mar. 2019, doi: 10.1007/s41870-018-0085-4.

[8] Preeti Dubey, "The Hindi to Dogri Machine Translation System," pp. 19–20, 2020.

[9] A. H. Patil, S. S. Patil, S. M. Patil, and T. P. Nagarhalli, "Real Time Machine Translation System between Indian Languages," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, Apr. 2022, pp. 1778–1783. doi: 10.1109/ICOEI53556.2022.9777103.

[10] K. Mrinalini, V. P, and N. Thangavelu, "SBSim: A Sentence-BERT Similarity-Based Evaluation Metric for Indian Language Neural Machine Translation Systems," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 30, pp. 1396–1406, 2022, doi: 10.1109/TASLP.2022.3161160.

[11] D. Banik, A. Ekbal, and P. Bhattacharyya, "Machine Learning Based Optimized Pruning

Approach for Decoding in Statistical Machine Translation," *IEEE Access*, vol. 7, pp. 1736–1751, 2019, doi: 10.1109/ACCESS.2018.2883738.

[12] M. Singh, R. Kumar, and I. Chana, "Neural-Based Machine Translation System Outperforming Statistical Phrase-Based Machine Translation for Low-Resource Languages," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, IEEE, Aug. 2019, pp. 1–7. doi: 10.1109/IC3.2019.8844915.

[13] S. Saxena, U. Chaurasia, N. Bansal, and P. Daniel, "Improved Unsupervised Statistical Machine Translation *via* Unsupervised Word Sense Disambiguation for a Low-Resource and Indic Languages," *IETE J Res*, vol. 69, no. 12, pp. 8848–8858, Dec. 2023, doi: 10.1080/03772063.2022.2098189.

[14] S. Dutta and B. Arora, "Parts of Speech (POS) Tagging for Dogri Language," 2021, pp. 529–540. doi: 10.1007/978-981-16-0733-2_37.

[15] D. P. Nitesh Patel, "Implementation Approach of Indian Language Gujarati Grammar's Concept 'Sandhi' Using the Concepts of Rule-Based NLP," in *8th Int. Conf. on Computing for Sustainable Global Development (INDIACom)*, 2021, pp. 481–485.

[16] S. Sreelekha, P. Bhattacharyya, and D. Malathi, "Statistical vs. Rule-Based Machine Translation: A Comparative Study on Indian Languages," 2018, pp. 663–675. doi: 10.1007/978-981-10-5520-1_59.

[17] Md. A. Islam, Md. S. H. Anik, and A. B. M. A. Al Islam, "Towards achieving a delicate blending between rule-based translator and neural machine translator," *Neural Comput Appl*, vol. 33, no. 18, pp. 12141–12167, Sep. 2021, doi: 10.1007/s00521-021-05895-x.

[18] M. Singh, R. Kumar, and I. Chana, "Improving Neural Machine Translation Using Rule-Based Machine Translation," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, IEEE, Jun. 2019, pp. 1–5. doi: 10.1109/ICSCC.2019.8843685.

[19] K. Deep, A. Kumar, and V. Goyal, "Machine Translation System Using Deep Learning for Punjabi to English," 2021, pp. 865–878. doi: 10.1007/978-981-15-7533-4_69.

[20] I. A. Thukroo and R. Bashir, "Spoken Language Identification System for Kashmiri and Related Languages Using Mel-Spectrograms and Deep Learning Approach," in *2021 7th International Conference on Signal Processing and Communication (ICSC)*, IEEE, Nov. 2021, pp. 250–255. doi: 10.1109/ICSC53193.2021.9673212.

[21] P. Gupta and S. S. Jamwal, "Designing and Development of Stemmer of Dogri Using Unsupervised Learning," 2021, pp. 147–156. doi: 10.1007/978-981-16-1048-6_11.

[22] S. S. Jamwal, P. Gupta, and V. S. Sen, "Hybrid Model for Generation of Verbs of Dogri Language," 2021, pp. 497–508. doi: 10.1007/978-981-15-9873-9_39.

[23] Sitender and S. Bawa, "A Sanskrit-to-English machine translation using hybridization of direct and rule-based approach," *Neural Comput Appl*, vol. 33, no. 7, pp. 2819–2838, Apr. 2021, doi: 10.1007/s00521-020-05156-3.

[24] M. Singh, R. Kumar, and I. Chana, "Corpus based Machine Translation System with Deep Neural Network for Sanskrit to Hindi Translation," *Procedia Comput Sci*, vol. 167, pp. 2534–2544, 2020, doi: 10.1016/j.procs.2020.03.306.

[25] A. Goyal, V. Gupta, and M. Kumar, "A deep learning-based bilingual Hindi and Punjabi named entity recognition system using enhanced word embeddings," *Knowl Based Syst*, vol. 234, p. 107601, Dec. 2021, doi: 10.1016/j.knosys.2021.107601.

[26] P. Bisht and A. Solanki, "Exploring Practical Deep Learning Approaches for English-to-Hindi Image Caption Translation Using Transformers and Object Detectors," 2022, pp. 47–60. doi: 10.1007/978-981-19-4831-2_5.

[27] G. I. Ahmad and J. Singla, "Machine learning approach towards language identification of Code-Mixed Hindi-English and Urdu-English Social Media Text," in *2022 International Mobile and Embedded Technology Conference (MECON)*, IEEE, Mar. 2022, pp. 215–220. doi: 10.1109/MECON53876.2022.9751958.

[28] H. Shen, "Intelligent Recognition English Translation System Based on Data Analysis Algorithm," 2022, pp. 900–906. doi: 10.1007/978-3-031-05484-6_119.

[29] J. Kumar, M. Rakhra, and P. Dubey, "Bilingual Parallel Corpora: A Major Resource for Developing Computational Tools for Automatic Processing of Hindi-Dogri Language Pair," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, IEEE, Oct. 2022, pp. 1–6. doi: 10.1109/ICRITO56286.2022.9964875.