Comparative Analysis of Machine Translation for Hindi-Dogri Text Using Rule-Based, Statistical, and Neural Approaches

Joginder Kumar[1*], Manik Rakhra[2], Preeti Dubey[3]

[1,]School of Computer Applications, Lovely Professional University, Punjab, India.

[2,]Department of Computer Science and Engineering, Lovely Professional University, Punjab, India.

[3]Department of Compute Science GCW, Parade, Jammu., India

Emails: Manik Rakhra [2] (rakhramanik786@gmail.com) and Preeti Dubey [3](preetidubey2000@yahoo.com)

Corresponding author: Joginder Kumar[1] (e-mail:.joginder1981@gmail.com)

ABSTRACT Machine translation has made significant progress in several Indian languages, but some, known as computationally low-resourced languages, have seen very little work in this field. Dogri language which is listed in the 8th Schedule of the Indian Constitution is one such language. The authors has developed a Machine Translation System for the Hindi-Dogri pair in the fixed news domain using three approaches: Rule-Based Machine Translation (developed using linguistic rules), Statistical Machine Translation (built using the Moses toolkit), and Neural Machine Translation (developed using neural networks). A comparison of all three approaches is presented in this paper. The paper also discusses various research challenges identified in each approach used for machine translation. A corpus of around 0.1 million sentences in the news domain was used for the development of corpus-based techniques, i.e., SMT and NMT models. The authors also addressed the question of whether NMT produces equivalent or better results compared to the SMT and RBMT approaches. Evaluation of the test results was performed by language experts along with the Bilingual Evaluation Understudy (BLEU) metric. In expert evaluation, it was observed that the NMT and SMT models' results are less ambiguous compared to RBMT. The BLEU score of RBMTS is (79.65), SMT is (52.39) and Bidirectional Embedding LSTM model of NMT is (52.46). The performances of the SMT and NMT models can improve further with the increase in dataset (bilingual parallel corpus).

KEYWORDS Machine Translation, Hindi-Dogri Language Pair, Low-Resourced Languages, Neural Machine Translation (NMT), Statistical Machine Translation (SMT), Rule-Based Machine Translation (RBMT).

I. INTRODUCTION

The technological advancements have enabled digitization in every sphere of life, yet there remains a digital divide due to the language barrier. Every person without the difference of gender, age, geographical domain needs access to various kinds of information and applications available for use to make daily tasks easier and time saving. The Government of India has taken many digital initiatives to provide access to each person in their regional language, but computationally low resourced languages like Dogri are still not visible on these applications, which lead to dependency of major non English knowing population on other people or with the only resort the use of manual sources of information. It has been observed that no government website in the state currently displays content in the regional language Dogri, despite it being declared an official language. One of the hindrances in making content available in local languages is the manual effort required to convert the content into Dogri. This highlights the need for developing state-of-

the-art (SOTA) automated machine translation systems. Such systems not only speed up the process but are also cost-effective. It can aid in the translation of various documents such as manuals, newspapers, academic content, literature, and other necessary content in less time and in a cost-effective manner. With intent to develop a state-of-the art (SOTA) machine translation system (MTS) for Hindi-Dogri language pair, the authors have worked on the three major approaches of machine translations: a system based on rule based approach, statistical MTS and a system based on deep learning models. Machine translation (MT) is a method that uses computer software to translate source language text (such as Hindi) to a target language (such as Dogri) while preserving the original meaning of the source language. Translation is a challenging task for both humans and machines, as it requires proper syntax and semantic knowledge of both languages, but it has emerged over the past ten years as a useful tool [1] for overcoming communication barriers in natural language processing. Machine translation (MT) methods are generally divided into two categories: rule-based and corpus-based approaches. Rule-based methods dominated the field from the inception of MT until the 1990s [2][3]. Rule-based machine translation (RBMT) systems rely on bilingual dictionaries and manually created rules to translate text between languages. In this study, the authors employed the direct approach of rule-based machine translation, one of three approaches alongside indirect and Interlingua methods. The direct approach, also known as the first generation of machine translation, relies on large dictionaries and word-by-word translation with simple grammatical adjustments. It is designed for specific language pairs, particularly closely related ones, making development easier due to shared grammar and vocabulary. However, this approach is limited to bilingual, unidirectional translation and struggles with ambiguous source texts.

With the emergence of bilingual corpora, corpus-based approaches became the dominant approach to translate text from one language to another after the 2000s. Mainly three approaches of corpus-based MT techniques are very popular that are example-based machine translation (EBMT), statistical machine translation (SMT), and neural machine translation (NMT). EBMT, established in the mid-1980s, operates by retrieving similar sentence pairs from a bilingual corpus to translate source texts [4]. If similar sentence pairs can be retrieved, EBMT algorithms produce high-quality translations. However, EBMT approaches have low translation coverage because bilingual corpora cannot include all the linguistic phenomena of the language pairings.

In 1990, Brown [5] introduced the concept of Statistical Machine Translation , where machines learn translation patterns from large datasets, removing the need for human experts to manually define rules. By 1993, this concept was formalized into five distinct SMT models [6]. Initially, a rule-based approach was developed [2][3] followed by corpus-based approaches using statistical machine translation (SMT) [5] and neural machine translation (NMT) [7][8]. In this study, the authors employed all three approaches for translating Hindi text into Dogri text and analyzed the performance based on adequacy, fluency, BLEU score, and ambiguity in the translated text. The paper is structured into several sections, including: methodology adopted for Machine Translation Systems (MTS) for the Hindi to Dogri language pair using RBMTS, SMT, and NMT; the necessary datasets; the experimental setup for RBMTS, SMT, and NMT models; research findings and challenges; a comparative analysis of the results; and a conclusion.

A brief about the languages under study:

Hindi: Hindi is one of the two official languages of India. Apart from India, the majority of people in Nepal speak Hindi. It is also a protected language in South Africa and the third official court language in UAE. It is the fourth most spoken language in the world (Wikipedia, “Hindi”] [9]).

Dogri: In India, the Dogri language is spoken by over 5 million people in northern India (particularly in Jammu & Kashmir, Himachal Pradesh, and some parts of Punjab) and parts of Pakistan as a Pahari language (WIKIPEDIA, “DOGRI.” [10]). Dogri got the status of an official language of the Union Territory of Jammu and Kashmir by the Jammu and Kashmir re-organization act 2019. It got added to the 8th Schedule of the Indian constitution in 2003. Both Hindi and Dogri Language are written in the Devanāgarī script [11] from left to right; however, Dogri has certain unique characteristics that distinguish it from Hindi and few are mentioned below:

i) Phonetically, some consonants produce different sounds in Dogri compared to Hindi as shown in the below table

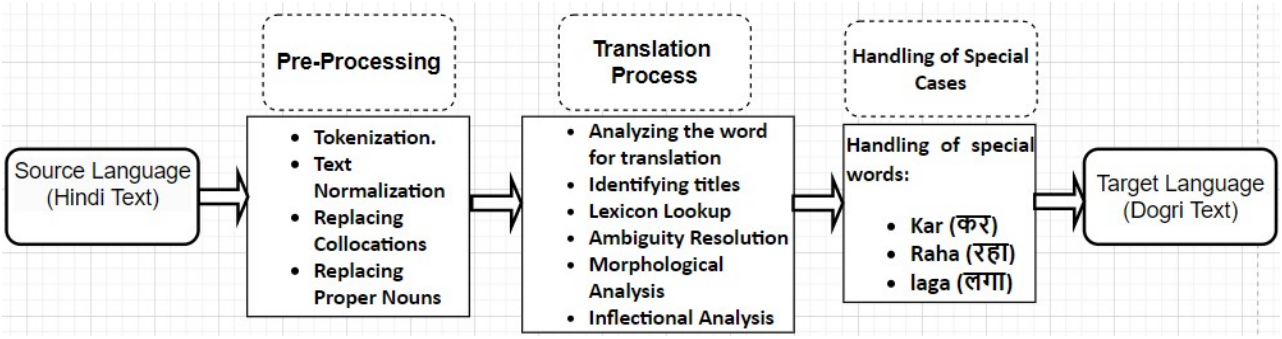

Figure 1 Architecture of Rule-Based Machine Translation System

TABLE I
PHONETICALL DIFFERENCE BETWEEN HINDI AND DOGRI

Hindi (Phoneticall y)	Dogri (Phoneticall y)	Examples of some Hindi words that are pronounce d differently in Dogri.	Phonetic Pronunciatio n in Dogri
घ (gha)	क (ka)	घर (ghara)	कर (kara)
झ (jha)	च (ca)	झंडा (jhaṃḍā)	चंडा (caḍā)
ढ (ḍha)	ट (ṭa)	ढाबा (ṭābā),	टाबा (ṭābā)
ध (dha)	त (ta)	धन(dhana)	तन (tana)
भ (bha)	प (pa)	भाग (pāga)	पाग (pāga)
ढ (ḍha)	ट (ṭa)	ढाबा (ṭābā)	टाबा (ṭābā)

ii) With the change in the tone of the word, the entire meaning of the word changes. The apostrophe comma '" is used to represent the tone, and it changes the meaning of the words when it is placed on top of the word, as mentioned in the table II

TABLE II
MEANING OF THE WORD CHANGES WITH TONE

Dogri Word	Hindi Meaning	Same Dogri word with tone change	Hindi Meaning
कुन (kuna)	घुन (ghuna)	कु'न (ku'na)	कौन (kauna)
खल्ल (khalla)	खाल (khāla)	ख'ल्ल (kha'lla)	नीचे (nīce)
फड़ (phaḍa)	पकड़ (pakaḍa)	फ'ड़ (pha'ḍa)	शेखी (śekhī)

iii) In Dogri, the Hindi symbols chandrabindu (ँ) and visarga (ः) are not used.

iv) The letters क्ष, ष, ऋ, and ज्ञ are solely used to transliterate Sanskrit words.

v) Extra-long vowels are indicated by the sign (ऽ) eg. चनाऽ (canā') - election, ब्हाऽ (bhā') - marriage, ग्रांऽ (grāṃ') - village.

vi) In Dogri, triple use of consonants is also seen in some words, such as ननान (nanāna) - sister in law, लग्गग (laggaga) - in use, मन्नना (mannannā) - agree, सस्स (sassa) - mother in law, बब्ब (babba) - father.

vii) In Dogri, nasalization (ँ) is also a phonemic. The following examples dem-onstrate how nasalization affects the meaning of words.

viii) तां (tāṃ) – then, ता (tā) - heat , बांग (bāṃga) - the crowing of a cock , बाग (bāga) - garden

II. RULE-BASED MACHINE TRANSLATION SYSTEM (RBMTS)

A direct approach of rule-based machine translation system for Hindi language text to Dogri language text was developed by creating bilingual dictionaries and a large collection of linguistic rules for Hindi and Dogri languages where both the languages follow the same grammar structure[12]. The grammatical structure of the Hindi language text is transferred into the Dogri language text using these intricate rules. The system's primary components include pre-processing (tokenization, normalization, replacing the collocation and proper nouns), lexicon lookup, ambiguity resolution, Inflectional and Morphological Analysis followed by the handling of the special cases like Kar, Raha and Laga. The figure 1 depicts the architecture of the system.

A. DATASET USED FOR RBMTS

A dataset of 22,801 words and phrases was collected under different categories, such as the creation of a Hindi-to-Dogri dictionary, collection of collocation phrases, collection of named entities, and standard words. Details are provided in the below table III for the development of the rule-based Hindi to Dogri machine translation system.

TABLE III
DATASET COLLECTED FOR THE DEVELOPMENT OF RBMTS

S. No	Category	Total Word / Phrases	Meaning
1	Dictionary (Hindi to Dogri)	18524	In this dictionary each Hindi word is represented by an equivalent Dogri word.
2	Collocation Phrases	1834	Collocation phrases that need to be translated as a single unit not individually are identified.
3	Named Entities	2130	For proper noun identification
4	Standard Words	412	Standard words are the single common variant for multiple words of similar meaning.

B. EVALUATION / FINDING OF RBMT SYSTEM

The current rule-based machine translation system relies on a lexicon lookup dictionary containing approximately 22,900 words and phrases. Because of this limited size, many Hindi words and phrases remain un-translated and are directly carried over into the output in Dogri text without any change. This reduces the accuracy of the final output when translating proper nouns, collocations and named entities. Additionally, in the case of polysemous words in Hindi, such as 'से' (se), 'और' (aur), दिया (diyā), की (kee) etc., where the exact translation depends on the context of the conversation, the system generates output with ambiguity. Table IV shows the output of RBMTS, where the system does not recognize named entities, resulting in incorrect Dogri translations. Table V displays a collection of polysemous words that can take multiple forms depending on the context of the conversation, resulting in ambiguous output.

TABLE IV
SHOWS SYSTEM'S WRONG TRANSLATION RESULTS

Source Text (Hindi)	Target Text (Dogri) Translated by RBMTS	Accurate Dogri Text
उत्तर अमेरिका (uttara amerikā)	**जवाब** अमेरिका (**Javāb** amerikā)	उत्तर अमेरिका (uttar amerikā)
आम आदमी (āma ādamī)	**अंब** आदमी (**Aanba** ādamī)	आम आदमी (ām ādamī)
विजय कुमार (vijaya kumāra)	**जित्त** कुमार (**Jitta** kumāra)	विजय कुमार (vijaya kumāra)

TABLE V
COLLECTION OF AMBIGUOUS WORDS THAT CAN TAKE MULTIPLE FORMS DEPENDING ON THE CONTEXT OF THE CONVERSATION

Source Text (Hindi)	Target Text (Dogri) Translated by RBMTS	The translation of Dogri words varies depending on the context
क्या (Kyā)	केह् (keh)	केह् (keh), केह्डा (kehdaa), कुस (kusa)
से (Se)	कोला (kolā)	कोला (kolā), थमां (thamāan), उप्पर (uppara), जेहे (jehe), कन्नै (kannai), चा (chā), दा (dā), शा (shā), चाल्ली (chāllī), गै (gai), उप्पर (uppara)
और (Aura)	ते (te)	ते (te), ओर (ora), बक्खी (bakkhī)
दिया (Diyā)	ओड़ेआ (odeaa)	ओड़ेआ (odeaa), दित्ता (dittā) , कीता (kītā)
की (Kī)	कीती (kītī)	कीती (kītī), आसेआं (āseāāan), दी(dī)

The following paragraph presents the translation of Hindi text into Dogri text using RBMT. It contains several incorrect translations of named entities, collocations, and polysemous words. The text marked with strikethrough indicates the incorrect translations produced by the system, while the bold text represents the expected translations:

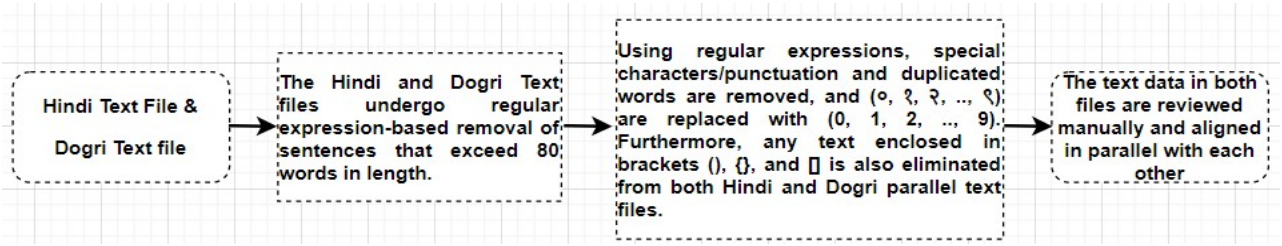

Figure 2 Preprocessing of the Hindi - Dogri Text Corpus

Hindi Text (Input)

पर्यटन विभाग के निदेशक डॉ .विवेकानंद राय ने सोमवार को बसोहली क्षेत्र के पर्यटन स्थलों का दौरा कर कुछ जरूरी दिशा निर्देश जारी किए हैं। उनहोने साथ में आम लोगों के साथ मुलाक़ात की। इस दौरान उनके साथ डीडीसी अध्यक्ष प्रशांत किशोर, सीओ रोहित सरदाना, सहायक निदेशक विजय शर्मा, बीडीसी अध्यक्ष सुषमा जमवाल और पर्यटन विभाग के अन्य अधिकारी मौजूद रहे। नई योजनाओं पर विचार विमर्श किया। डीडीसी अध्यक्ष ने निदेशक से इलाकों को पर्यटन की दृष्टि से विकसित करने के लिए प्रोजेक्ट बनाने को कहा। इस मंदिर की चारदीवारी करीब एक साल से क्षतिग्रस्त है। इसके अलावा, उन्होंने टूरिज्म रिसेप्शन सेंटर की इमारत का भी जायजा लिया। अंत में पृथ्वी शॉ ने सभी को स्वतंत्रता दिवस की अग्रिम शुभकामनाएं दीं। जहां तीन नदियां गंगा, यमुना और भूमिगत सरस्वती का विलय होता है।

(Paryaṭan vibhāg ke nideshak ḍa॰ Vivekānanda rāya ne somavār ko basohalī kṣhetra ke paryaṭan sthaloan kā daurā kar kuchh jarūrī dishā nirdesh jārī kie haian। unahone sāth mean ām logoan ke sāth mulāक़āt kī। is daurān unake sāth ḍīḍīsī adhyakṣha prashāanta kishora, sīo rohit saradānā, sahāyak nideshak vijaya sharmā, bīḍīsī adhyakṣha suṣhamā jamavāl aur paryaṭan vibhāg ke anya adhikārī maujūd rahe। naī yojanāoan par vichār vimarsha kiyā। ḍīḍīsī adhyakṣha ne nideshak se ilākoan ko paryaṭan kī dṛuṣhṭi se vikasit karane ke lie projekṭa banāne ko kahā। is mandir kī chāradīvārī karīb ek sāl se kṣhatigrasta hai। isake alāvā, unhoanne ṭūrijma risepshan seanṭar kī imārat kā bhī jāyajā liyā। aanta mean pṛuthvī sha॰ ne sabhī ko svatantratā divas kī agrim shubhakāmanāean dīan। jahāan tīn nadiyāan gangā, yamunā aur bhūmigat sarasvatī kā vilaya hotā hai।)

Dogri Text Output (Translated using RBMTS)

~~सैर-सपाटा मैह्कमा~~ **पर्यटन मैह्कमें** दे निदेशक डा विवेकानंद ~~राऽ~~ **राय** नै ~~सोमबार~~ **सोमवार** गी बसोहली खेतर दे ~~सैर-सपाटा-स्थलों~~ **पर्यटन स्थलें** दा ~~दौराइयै~~ **दौरा करियै** किश ~~ज़रूरी बक्खी~~ **जरूरी दिशा** निर्देश जारी कीते न। ~~उनहोने~~ **उ'नें** कन्नै ~~च~~ **गै** ~~अंब~~ **आम** लोकें दे कन्नै मुलाकात कीती। इस ~~दुसन~~ **दौरान** उं'दे कन्नै डीडीसी प्रधान प्रशांत ~~नबालग~~ **किशोर**, सीओ रोहित ~~ठंडोना~~ **सरदाना**, ~~मददगार~~ **सहायक** निदेशक ~~जिस~~ **विजय** शर्मा, बीडीसी प्रधान ~~सुषमा~~ **सुशमा** जमवाल ते ~~सैर सपाटा मैह्कमा~~ **पर्यटन मैह्कमें** दे होर अधिकारी मजूद रेह्। नमीं योजनाओं पर बिचार ~~सलाह्~~ **विमर्श** कीता। डीडीसी प्रधान नै निदेशक ~~कोला~~ **गी** ~~लाकैं~~ **लाके** गी ~~सैर सपाटा~~ **पर्यटन** दी नज़र शा **कन्नै** विकसत करने दे आस्तै प्रोजेक्ट बनाने गी आखेआ। इस मंदर ~~कीती~~ **दी** चार-दवारी करीब इक ब'रें शा क्षतिग्रस्त ऐ। एह्दे अलावा, ~~उ'नैं~~ **उ'नें** टूरिज्म रिसेप्शन सेंटर दी अमारत ~~दा~~ **दी** बी ~~परख पड़ताल~~ **जांच-पड़ताल** ~~लैता~~ **कीती**। ~~अंत~~ **खीर** च ~~मृत्युलोक~~ **पृथ्वी** शा नै सारें गी अजादी दिन ~~दी~~ **दियां** शुभकामनाओ ~~दीं~~ **दितिया**। जित्थै त्रै नदियां गंगा, जमना ते भूमिगत सरस्वती दा ~~मौत~~ **मेल** होंदा ऐ।

III. DATASET FOR CORPUS-BASED MT APPROACHES:

The dataset comprises Hindi to Dogri parallel text and contains a total of 0.01 million sentence pairs, each with fewer than 80 words per sentence [13]. The same corpus is used to train both SMT and four different NMT models, providing a true comparison of the two approaches. To perform a comparative analysis of all machine translation approaches, 100 random Hindi sentences are picked from various sources and used to test the RBMTS, SMT, and four NMT models. Table VI provides statistics on the corpus.

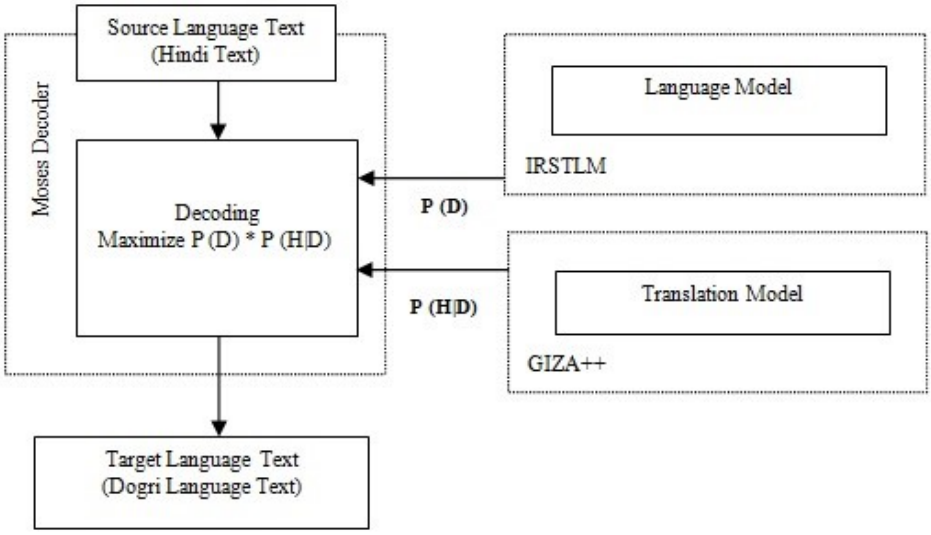

Figure 3 Architecture of Hindi to Dogri SMT System

TABLE VI
HINDI-DOGRI PARALLEL CORPUS STATISTICS USED FOR THE DEVELOPMENT OF SMT AND NMT MODELS

Dataset Division	**Hindi to Dogri Text (Sources)**	**Total no. of Hindi - Dogri parallel sentences**	**Hindi words**	**Unique Hindi words**	**Dogri words**	**Unique Dogri words**
Total Corpus used for Training, Validation and Testing	The corpus collected from various sources like news papers, books, Standard words, Hindi to Dogri dictionary, Dogri Names etc.	**100,000**	**7,71,930**	**67,332**	**7,77,401**	**66,184**
Testing Corpus	**10 percent of the Total Corpus**					
Validation Corpus	**10 percent of the Total Corpus**					
Corpus for Comparative analysis of RBMTS, SMT and NMT	Random text from News portals	**100**	**1741**	**156**	**1742**	**158**

A. PREPROCESSING OF CORPUS

To create a high-quality Hindi to Dogri parallel corpus, data is collected from various sources [13] and then proofread by a professional linguist. The data undergoes preprocessing, which includes removing repeated words or sentences, noisy data, incomplete words, special characters/punctuation. Additionally, Devanagari numerals (०, १, २, . . , ९) are replaced with numeric digits (0, 1, 2, .., 9). Sentences with more than 80 words in both

Hindi and Dogri are excluded from the corpus during the preprocessing phase. Figure 2 depicts the various stages of preprocessing required to prepare the raw text data for training the corpus-based models. A sample of the final Hindi to Dogri parallel corpus is presented in Figure 6.

IV. STATISTICAL MACHINE TRANSLATION SYSTEM (SMT)

Developing and maintaining rules in rule-based approach is time-consuming, and transferring them across different domains or languages is a complex task. As a result, scaling rule-based systems for open-domain or multilingual translation is challenging. In 1990, Brown et al. introduced the SMT [5] model for machine translation, and later, an SMT model specifically for Hindi to Dogri translation was developed by the authors using the Moses toolkit. The Moses toolkit by Koehn et al [14] trains the translation model using aligned-text of both Hindi and Dogri languages. Once the training is complete, the

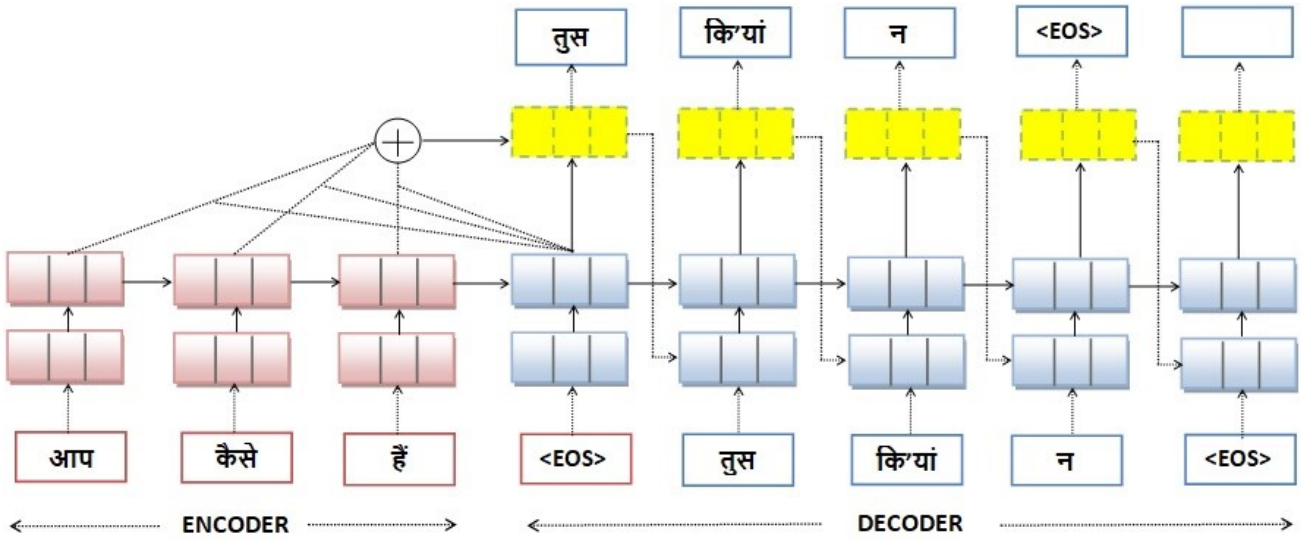

Figure 4 Encoder- Decoder Architecture of NMT

decoder uses Beam Search to translate the source text to target language text. The Beam search algorithm selects the translation with the highest probability. Figure 3 illustrates the architecture of the Hindi to Dogri SMT system. SMT analyzes bilingual text corpora to create translation rules, with translation accuracy depending on the quality and size of the bilingual corpora.

A. EVALUATION / FINDING OF SMT SYSTEM

The SMT model is trained with a parallel corpus of approximately 0.1 million sentences as shown in the table IV. The translation results are generally quite accurate and fluent, barring the translation of rare or unknown. The system is producing UNK for words which are not part of the training corpus. The System handled ambiguity in a better way as compared to RBMTS. The detailed comparison of results is shown in the table VI

V. NEURAL MACHINE TRANSLATION SYSTEM (NMT

Statistical Machine Translation (SMT) methods can enhance translation quality significantly; however they rely on log-linear models that include several manually built components such as translation model, language model, and reordering model. This frequently results in substantial reordering issues, particularly with distant language combinations. However, as deep learning techniques advanced in disciplines such as voice and vision, researchers began to incorporate them into machine translation.

Nowadays, academics and researchers are focusing on more sophisticated techniques, such as Deep Learning machine translation system (MTS) and it is producing better results [15]. In this paper authors have implemented four deep learning models with embedding encoder and decoder module for Hindi-Dogri translation. The figure 4 shows the general architecture of the neural machine translation with encoder decoder. Deep learning based Neural MT (NMT) models can now access all of the information included in the source phrase and automatically learn which parts are helpful at which stages of synthesizing the output text due to the large amounts of training dataset and unparalleled computer power. 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, GRUs, etc.

A. TRAINING SETUP FOR NMT MODELS

The machine language translation system for Hindi-Dogri has been designed and implemented using four models: embedding LSTM, Bidirectional LSTM (BiLSTM), embedding BiLSTM, and Encoder-Decoder GRU. The Tensor Flow framework has been used for the implementation task to achieve better performance. The dataset is split into training and validation sets in a ratio of 80:20, with 80% of the dataset used for training, 10% for validation and 10 % for testing. The framework has been implemented in Python using Google Colab, with a minimum batch size of 32 for the training module. All

four models were trained for 50 epochs. The implementation was done in Keras, using TensorFlow as the backend, and all networks were trained on Tesla P100-PCIE GPU. The table below shows the training and validation Accuracy and Loss for each of the four models. The figures numbered 5 to 8 below show the training and validation accuracy, as well as the losses, for all four models. %. All networks achieved near about same accuracy and loss but bidirectional LSTM achieved the best as shown in the table VII.

TABLE VII
TRAINING AND VALIDATION ACCURACY AND LOSS FOR FOUR NMT MODELS

Model Name	Training and Validation Accuracy	Training and Validation Loss
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

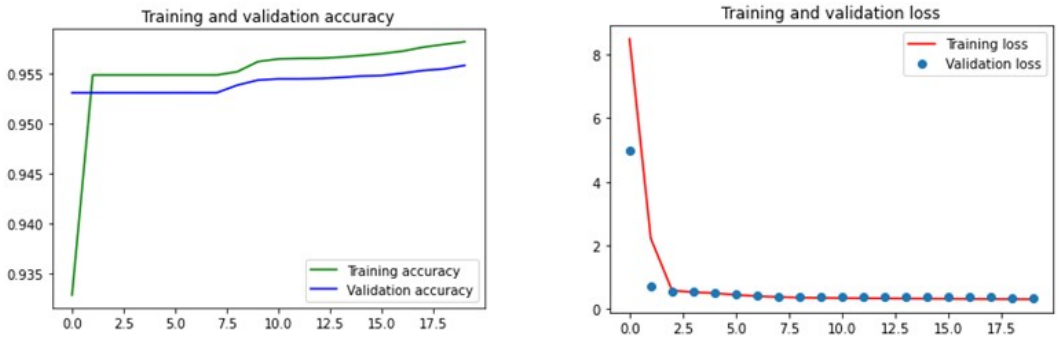

Figure 5 Showing Training and Validation Accuracy and Loss for the Embedding LSTM

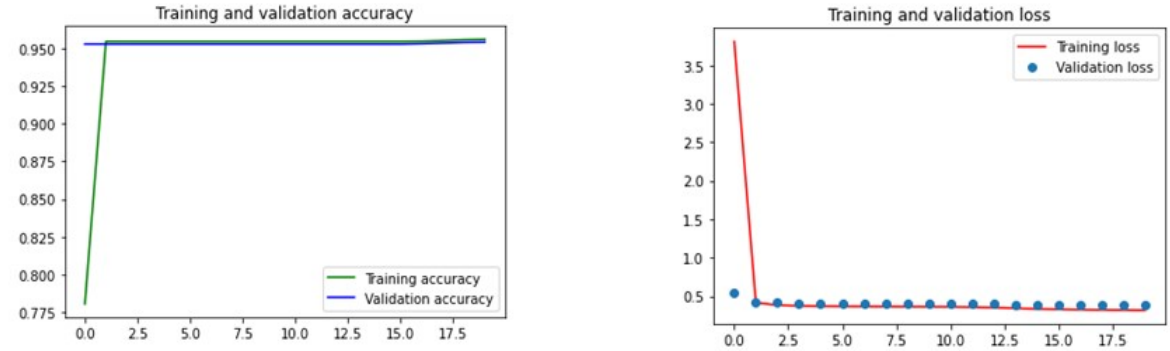

Figure 6 Showing Training and Validation Accuracy and Loss for the Bidirectional LSTM

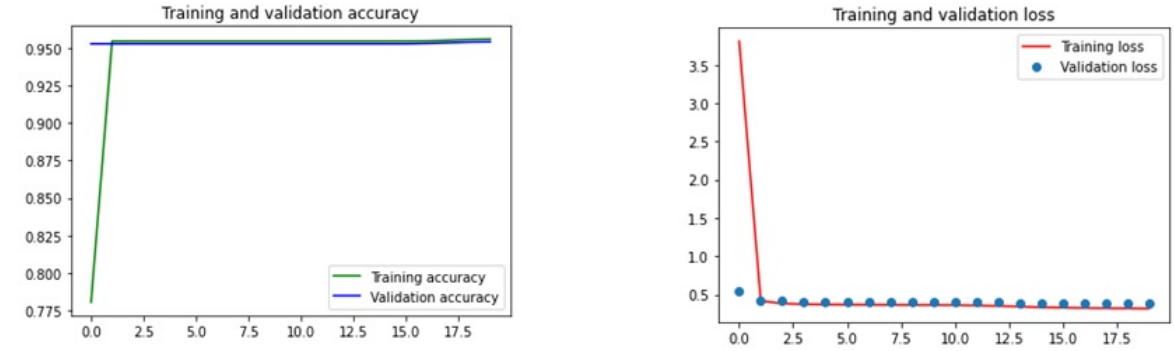

Figure 7 Showing Training and Validation Accuracy and Loss for the Bidirectional Embedding LSTM

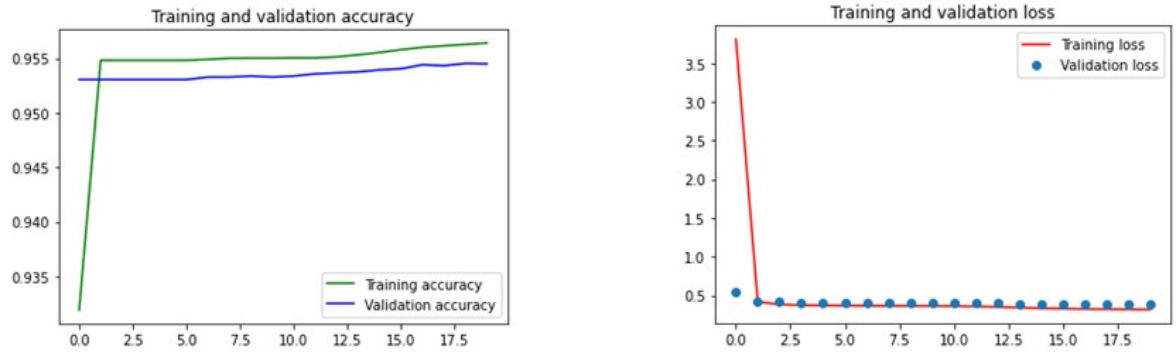

Figure 8 Showing Training and Validation Accuracy and Loss for the Encoder-Decoder GRU

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 ब'रे 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 मकानै 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'नें 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 कामयाबी 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 5 Sample of Bilingual Hindi to Dogri Parallel Coprus

.

A. EVALUATION / FINDING OF NMT SYSTEM

NMT uses neural networks, with each neuron processing data mathematically. Initially, the network is trained by feeding bilingual Hindi to Dogri parallel text corpra and adjusting the weights of the neurons based on the error in translation. NMT systems continuously fine-tune themselves, leading to improved results. NMT is more reliable than SMT, particularly for low-resourced languages, as it better accounts for context and produces more human-like translations.

VI. COMPARATIVE ANALYSIS OF TRANSLATION RESULTS

Recent research has documented the differences between various MT systems with respect to the output quality and error types. Some researchers [7] [16] have used automatic evaluation metrics such as TER [17] and BLEU [18] and Others have assessed MT systems based on adequacy and fluency through human evaluations of the translation output [19]. Some studies have also combined human evaluation methods with automatic evaluation metrics (AEMs) for a more comprehensive analysis [20] [21]. In this study the results of all the three approaches of MT are evaluated by linguist experts (human evaluation) and by using automatic evaluation system (BLEU) technique.

TABLE VIII
COMPARATIVE ANALYSIS OF RESULTS OF ALL THE APPROACHES DEVELOPED IN THIS STUDY

MT-System	Adequacy	Fluency	BLEU Score
RBMTS	3.8	4.0	79.65
SMT	2.8	2.9	52.39
Embedding LSTM	2.7	2.9	45.05
Bidirectional LSTM	2.5	2.5	43.06
Bidirectional Embedding LSTM	2.8	3.0	52.46
Encoder Decoder GRU	2.1	2.3	41.47

A. HUMAN EVALUATION OF HINDI TO DOGRI TEXT

The text results are evaluated on the basis of **adequacy** of the output text to the extent it conveys the same meaning as of the input text and **fluency** that evaluates the output text's grammatical accuracy and naturalness on a scale of 1 to 5. Higher the value means better results as mentioned in the table no.

B. AUTOMATIC EVALUATION USING BLUE SCORE

Out of various tools available for automatic translated text evaluation, we used BUEL score in our study for the comparison of text results of three systems as shown in the table no VIII

C. SAMPLE OUTPUT

The section below demonstrates the output generated by all three machine translation approaches on 100 sample texts gathered from various news portals. The accuracy of the output is evaluated by comparing it with the reference text in Dogri translated by professional linguists to gain a better understanding of the distinct variations among the three suggested approaches

Hindi Text 1: ताकि लोगों को सुविधा मिल सके (tāki logoan ko suvidhā mil sake)

Dogri Reference Text 1: *तांजे लोकें गी सुविधा थ्होई सकै (tāanje lokean gī suvidhā thhoī sakai)*

Dogri Text (RBMTS): तांजे लोकें गी सुबधा मिल्ल सकै (tāanje lokean gī subadhā milla sakai)

Dogri Text (SMT): तांजे लोकां गी सुबधा थ्होई सकै (tāanje lokāan gī subadhā thhoī sakai)

Dogri Text (NMT): तांजे लोकें गी सुबधा मिल्ल सकै (tāanje lokean gī subadhā milla sakai)

Hindi Text 2: एस एस पी राजेश्वर सिंह का कहनाहै कि सांबा जम्मू कठुआ राजोरी और पुंछ जिलों से 10 लोगों ने हमारे पास शिकायत दर्ज कराई है (es es pī rājeshvar sianha kā kahanāhai ki sāanbā jammū kaṭhuā rājorī aur puanchha jiloan se 10 logoan ne hamāre pās shikāyat darja karāī hai)

Dogri Reference Text 2: *एस एस पी राजेश्वर सिंह दा आखना ऐ जे सांबा जम्मू कठुआ राजोरी ते पुंछ जिलें थमां 10 लोकें साढ़े कोल शकैत दर्ज कराई ऐ (es es pī rājeshvar sianha dā ākhanā ai je sāanbā jammū kaṭhuā rājorī te puanchha jilean thamāan 10 lokean sāḍha़e kol shakait darja karāī ai)*

Dogri Text 2(RBMTS): एस एस पी सम्राट सिंह दा आखना ऐ जे सांबा जम्मू कठुआ राजोरी ते पुंछ ज़िलें कोला 10 लोकें नै साढ़े कोल शकैत दर्ज कराई ऐ (ah es es pee samraaṭ sinh daa aakhanaa ai je saanbaa jammoo kaṭhuaa raajoree te punchh zailen kolaa 10 loken nai saaḍhxe kol shakait darj karaa_ii ai)

Dogri Text 2(SMT): एस एस पी राजेश्वर सिंह दा छानना ऐ जे साम्बा जम्मू कठुआ रजोरी ते पुंछ जिले दे 10 लोकां साढे कोल शकैत दर्ज कराई (ah es es pee raajeshvar sinh daa chhaananaa ai je saambaa jammoo kaṭhuaa rajoree te

punchh jile de 10 lokaan saaḍhe kol shakait darj karaa_ii)

Dogri Text 2(NMT): एस एस पी सिंह सिंह दा आखना ऐ जे सांबा जम्मू कठुआ राजोरी ते पुंछ ज़िलें शा 10 लोकें नै साढ़े कोल शकैत दर्ज ऐ (ah es es pee sinh sinh daa aakhanaa ai je saanbaa jammoo kaṭhuaa raajoree te punchh zailen shaa 10 loken nai saaḍhxe kol shakait darj ai)

Hindi Text 3: जिसके आधार पर एक संयुक्त एफ आई आर दर्ज की गई है (jisake ādhār par ek sanyukta ef āī ār darja kī gaī hai)

Dogri Reference Text 3: *जेह्दे बुनियाद पर इक सांझी एफ आई आर दर्ज कीती गेई ऐ (jehde buniyād par ik sāanzī ef āī ār darja kītī geī ai)*

Dogri Text 3(RBMTS): जेह्दे बुनियाद पर इककट्ठे एफ आई आर दर्ज दी गेई ऐ (jehde buniyād par ikakaṭṭhe ef āī ār darja dī geī ai)

Dogri Text 3(SMT): जिस दे बुनियाद उप्पर इक सांझी एफ आई आर दर्ज कीती गेई ऐ (jis de buniyād uppar ik sāanzī ef āī ār darja kītī geī ai)

Dogri Text 3(NMT): जेह्दे बुनियाद पर इक कट्ठे एफ आई आर दर्ज कीती गेई ऐ (jehde buniyād par ik kaṭṭhe ef āī ār darja kītī geī ai)

Hindi Text 4: रक्षाबंधन का पर्व भाई और बहन के बीच अट्टू प्रेम का प्रतीक है (rakṣhābandhan kā parva bhāī aur bahan ke bīch aṭṭū prem kā pratīk hai)

Dogri Reference Text 4*: रक्खड़ी दा पर्व भ्राऽ ते भैन च अट्टू प्यार दा प्रतीक ऐ (rakkhaड़ī dā parva bhrā′ te bhain ch aṭṭū pyār dā pratīk ai)*

Dogri Text 4(RBMTS): रक्खड़ी दा पर्व भ्राऽ ते भैन दे बिच्च अट्टू प्यार दा नशानी ऐ (rakkhaड़ī dā parva bhrā′ te bhain de bichcha aṭṭū pyār dā nashānī ai)

Dogri Text 4(SMT): रक्खड़ी दा पर्व भ्राऽ ते भैन च अट्टू प्यार दा प्रतीक ऐ (rakkhaड़ī dā parva bhrā′ te bhain ch aṭṭū pyār dā pratīk ai)

Dogri Text 4(NMT): राजकी दा पर्व भ्राऽ ते भैन दे बिच्च ऐ प्यार दा ऐ (rājakī dā parva bhrā′ te bhain de bichcha ai pyār dā ai)

VII. CONCLUSION

Each machine translation method has its own set of advantages and disadvantages, making it imperative to choose the one that aligns with the unique requirements and constraints of each translation task, including the language pairs, the subject matter domain, and available resources. In our research on translating Hindi to Dogri text, where both languages follow the same syntax structure and training data is limited, Rule-Based systems have proven to be more effective than SMT and NMT systems. Rule-Based machine translation (RBMT) is now overshadowed by statistical and neural methods, it was pivotal in the early development of the field. Its focus on linguistic precision and rule-based modeling remains important for certain use cases, particularly in domain-specific translations and for languages with limited resources. Among the four models of NMT, the Bidirectional Embedding LSTM model outperforms the other three neural models and the statistical model. Deep learning models are the future of machine translation, and with larger high-quality datasets, better results are expected. In the future, the author plans to experiment with hybrid MT systems that combine the domain-specific expertise of RBMT systems with the natural language processing capabilities of NMT systems to produce more robust translation. Throughout the study, any unknown words were left un-translated and remained as they appeared in the target text in all three machine translation approaches. This approach highlights the need for a more effective technique to handle unknown words in the future. It is a wise decision to adapt the appropriate networks and deep learning algorithms since it tailored the system to maximize the translation system's accuracy in comparison to others.

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," *Arch. Comput. Methods Eng.*, vol. 28, no. 4, pp. 2165–2193, 2021, doi: 10.1007/s11831-020-09449-7.

[2] G. V. Garje, A. Bansode, S. Gandhi, and A. Kulkarni, "Marathi to English Sentence Translator for Simple Assertive and Interrogative Sentences," *Int. J. Comput. Appl.*, vol. 138, no. 5, pp. 42–45, 2016, doi: 10.5120/ijca2016908837.

[3] P. Dubey, "The Hindi to Dogri machine translation system: grammatical perspective," *Int. J. Inf. Technol.*, vol. 11, no. 1, pp. 171–182, 2019, doi: 10.1007/s41870-018-0085-4.

[4] M. Nagao, *A framework of a mechanical translation between Japanese and English by analogy principle*.

[5] P. Brown *et al.*, "A Statistical Approach to French/English Translation," *Biol. Artif. Intell. Syst.*, vol. 16, no. 2, pp. 547–561, 1988, doi: 10.1007/978-94-009-3117-6_27.

[6] N. C. Nas, "3 0 1963," *Comput. Linguist.*, vol. 10598, no. August, pp. 263–311, 1963.

[7] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[8] S. K. Mahata, S. Mandal, D. Das, and S. Bandyopadhyay, "SMT vs NMT: A Comparison over Hindi & Bengali Simple Sentences," Dec. 2018, [Online]. Available: http://arxiv.org/abs/1812.04898

[9] Wikipedia, "Wikipedia, 'Hindi,' 2014. https://en.wikipedia.org/wiki/Hindi (accessed July. 10, 2024)." [Online]. Available: https://en.wikipedia.org/wiki/Hindi

[10] wikipedia, "wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Dogri_language

[11] V. Gupta, *Dogri Vyakaran.* Jammu and Kashmir Academy of Art, Culture and Language, 2004, 2004.

[12] P. Devanand and D. Preeti, "Study and Development of Machine Translation System from Hindi Language to Dogri Language: An Important Tool to Bridge the Digital Divide," University of Jammu, 2013.

[13] 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.

[14] P. Koehn *et al.*, "Moses : Open Source Toolkit for Statistical Machine Translation," no. June, pp. 177–180, 2007.

[15] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [Review Article]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, 2018, doi: 10.1109/MCI.2018.2840738.

[16] M. Dowling, T. Lynn, A. Poncelas, and A. Way, "SMT versus NMT: Preliminary comparisons for Irish," *Proc. AMTA 2018 Work. Technol. MT Low Resour. Lang. (LoResMT 2018)*, pp. 12–20, 2018.

[17] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," *AMTA 2006 - Proc. 7th Conf. Assoc. Mach. Transl. Am. Visions Futur. Mach. Transl.*, no. August, pp. 223–231, 2006.

[18] G. Wentzel, "Funkenlinien im Röntgenspektrum," *Ann. Phys.*, vol. 371, no. 23, pp. 437–461, 1922, doi: 10.1002/andp.19223712302.

[19] L. Bentivogli, A. Bisazza, M. Cettolo, and

M. Federico, “Neural versus phrase-based machine translation quality: A case study,” *EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 257–267, 2016, doi: 10.18653/v1/d16-1025.

[20] Y. Jia, M. Carl, and X. Wang, “Post-editing neural machine translation versus phrase-based machine translation for English–Chinese,” *Mach. Transl.*, vol. 33, no. 1–2, pp. 9–29, 2019, doi: 10.1007/s10590-019-09229-6.

[21] M. S. Vilelmini Sosoni, “Un rre cte d Au tho r P roo f cte d Au tho r P roo f Un,” no. March, 2020, [Online]. Available: https://www.researchgate.net/publication/339302764