Bi-LSTM-Based Machine Translation from Dogri to English: An In-Depth Study of Neural Approach for Low-Resource Language

[1]Vijay Singh Sen, [2]Shubhnandan S. Jamwal
[1]Research Scholar,[2]Professor
[1,2]Department of Computer Science and IT
[1,2]University of Jammu, Jammu & Kashmir, India
[1]senvjju@gmail.com, [2]jamwalsnj@gmail.com

***Abstract*— Translating low-resource languages like Dogri presents significant challenges, primarily due to the limited availability of parallel corpora and the complex linguistic structure of the language. Dogri, an Indo-Aryan language spoken mainly in northern India, is one of the twenty-two scheduled languages in Indian Constitution but remains underrepresented in computational linguistics. This study presents the first experimental investigation into the application of Bidirectional Long Short-Term Memory (Bi-LSTM) networks for translating text from Dogri to English. A manually generated parallel corpus consisting of 85,000 Dogri-English sentence pairs was developed and pre-processed to train the Bi-LSTM-based neural machine translation model. The model leverages the bidirectional context of input sequences to improve translation accuracy, particularly for a low-resource language like Dogri. Experimental results demonstrate the effectiveness of the proposed approach, achieving a BLEU score of 46.03, and highlighting its potential for integrating underrepresented languages like Dogri into modern computational and technological frameworks.**

***Index Terms*—Machine Translation, Dogri-to-English, Bi-LSTM, Low-Resource Languages, Neural Machine Translation, BLEU Score.**

I. Introduction

Generative Large Language Models (LLMs) have brought transformative advancements to natural language processing (NLP), excelling in a wide range of applications [1]. These models demonstrate remarkable capabilities in open-domain question answering by generating accurate and coherent responses and performing instruction-based tasks such as code completion, error detection, and correction [2]. Additionally, LLMs are highly effective in tasks like essay writing, grammar correction, and text summarization, producing outputs of exceptional quality [3]. However, much of this progress has been concentrated on English, leaving many other languages underexplored or underdeveloped [4]. For machine translation (MT) and related subtasks, encoder decoder-based models remain state-of-the-art, while decoder-only models require further advancement to effectively handle low-resource languages, particularly those spoken in India. India's linguistic diversity is immense, comprising 22 scheduled languages from linguistic families such as Indo-Aryan, Dravidian, Tibeto-Burman, and Austroasiatic. There are over 559 classified mother tongues spoken across the country, reflecting the cultural and linguistic richness of the nation. This diversity, however, poses significant challenges in communication, education, business, healthcare, tourism and governance. Advances in encoder-decoder and LLM-based models hold the potential to address these linguistic challenges by enabling cross-lingual communication and fostering India's multilingual ecosystem [5]. Translation plays a vital role in addressing linguistic diversity and enabling effective cross-lingual communication in multilingual societies like India. However, the distinct characteristics of Indian languages present unique challenges. Many Indian languages exhibit complex morphological structures, often involving agglutinative or inflectional morphology, which makes them syntactically diverse and distinct from each other and English. The use of diverse scripts, such as Devanagari, Tamil, Bengali, and Gurmukhi, adds further

complexity to text processing tasks. Additionally, code-mixing, a common phenomenon in India where speakers frequently blend multiple languages in communication, complicates translation workflows. The scarcity of high-quality linguistic resources, annotated corpora, and evaluation benchmarks for many Indian languages exacerbates these challenges, hindering the development of robust MT systems. Addressing these challenges requires a systematic approach. Developing translation systems for Indian languages is crucial to facilitate communication and preserve linguistic diversity. With the translation system, robust evaluation frameworks are needed to assess their quality, with or without reference translations. Identifying and categorizing translation errors is essential for system improvement, enabling targeted refinements based on predefined error groups. Automatic post editing systems can then help correct translation errors efficiently, either independently or in collaboration with human translators. Incorporating these advancements across general and domain-specific applications will greatly benefit the language research community and end users. Translation plays a vital role in addressing linguistic diversity and enabling effective cross-lingual communication in multilingual societies like India [6]. However, the distinct characteristics of Indian languages present unique challenges. Like Many Indian languages Dogri exhibit complex morphological structures, often involving agglutinative or inflectional morphology, which makes it syntactically diverse and distinct from English [7]. The use of diverse scripts such as Devanagari adds further complexity to text processing tasks. Additionally, code-mixing, a common phenomenon in India where speakers frequently blend multiple languages in communication, complicates translation workflows. The scarcity of high-quality linguistic resources, annotated corpora, and evaluation benchmarks for Indian languages exacerbates these challenges, hindering the development of robust MT systems. The remainder of this paper is organized as follows. Section 2 presents the related work in the field of machine translation. Section 3 describes the development of the Dogri–English parallel corpus used in this study. Section 4 explains the proposed methodology and model architecture. Section 5 presents the experimental results and their analysis. Finally, Section 6 concludes the paper and discusses future research directions.

II. Related Work

The paper by Das et al. presents a Statistical Machine Translation (SMT) system for Assamese–English and English–Assamese translation using a parallel corpus of around 2,500 tourism-related sentence pairs. Implemented using the Moses toolkit, IRSTLM, and GIZA++, the system achieved better BLEU scores for Assamese-to-English due to the complexity of Assamese morphology. To handle proper nouns and out-of-vocabulary words, a statistical transliteration module was incorporated. The study highlights the viability of SMT for low-resource languages and emphasizes the importance of transliteration in improving translation quality for Indian language pairs with rich morphological structures [11].

The paper proposes a corpus-based machine translation system for Sanskrit-to-Hindi translation utilizing deep neural networks. Leveraging a parallel corpus, the authors develop a deep learning model designed to capture complex linguistic patterns inherent in Sanskrit and Hindi. Their approach employs neural architectures that improve translation accuracy by learning contextual relationships within sentences. Experimental results demonstrate the model's effectiveness in handling the morphological richness and syntactic differences between the two languages. This study highlights the potential of deep neural networks in enhancing translation quality for classical and morphologically complex language pairs like Sanskrit and Hindi [12].

Koehn and Knowles identify six key challenges facing neural machine translation (NMT) systems: domain adaptation, handling rare words, interpretability, beam search issues, robustness to noise, and efficient use of limited training data. The paper discusses how these challenges impact NMT performance and suggests potential directions for research to address them. By highlighting these obstacles, the authors aim to guide future improvements in NMT architectures, making them more practical and effective for diverse real-world translation tasks. The work remains influential in understanding and advancing the state of neural machine translation [13].

Liu presents a case study on low-resource neural machine translation (NMT) focusing on Cantonese, a language with limited parallel data. The study explores various strategies to improve translation quality despite data scarcity, including transfer learning, data augmentation, and model architecture adaptations.

Experimental results show that leveraging related language resources and innovative training techniques can significantly enhance NMT performance for Cantonese. This work contributes valuable insights into addressing challenges inherent to low-resource languages, demonstrating practical approaches to develop effective NMT systems where data availability is limited, and supporting broader efforts to include underrepresented languages in machine translation research [14].

Paul et al. propose a Bengali-English neural machine translation (NMT) system utilizing deep learning techniques to improve translation accuracy. The study employs advanced neural network architectures to capture the linguistic nuances of both languages, addressing challenges such as syntax differences and vocabulary richness. Experiments demonstrate that the model achieves promising BLEU scores, indicating effective translation quality. The paper highlights the importance of deep learning in overcoming difficulties inherent in translating low-resource language pairs like Bengali-English. This work contributes to the growing research on neural approaches for Indian language machine translation, paving the way for more accurate and efficient NMT systems [15].

Ephrem develops a bidirectional machine translation system for Amharic and Tigrinya using recurrent neural networks (RNNs). Addressing the low-resource nature of these closely related Ethiopian languages, the study constructs parallel corpora and implements RNN-based models to capture linguistic patterns for effective translation. Experimental results demonstrate the system's ability to handle morphological complexities and improve translation accuracy in both directions. The research contributes to the advancement of machine translation technology for underrepresented African languages and highlights the potential of deep learning methods in overcoming challenges posed by limited data availability in low-resource language pairs [16].

Tennage et al. present a neural machine translation (NMT) system for Sinhala and Tamil, two widely spoken South Asian languages with limited computational resources. The study employs sequence-to-sequence neural architectures to model the linguistic features and syntactic differences of these languages. Using available parallel corpora, the authors train and evaluate the NMT models, demonstrating improved translation accuracy compared to traditional methods. The work addresses challenges related to low-resource language pairs and contributes valuable insights for developing effective NMT systems for underrepresented languages, supporting broader efforts to enhance language technology in the South Asian region [17].

Luong, Pham, and Manning (2015) explore effective methods to improve attention mechanisms in neural machine translation (NMT). They propose novel global and local attention models that enhance the alignment between source and target language sequences during translation. Their approaches address limitations of previous attention techniques by focusing on relevant parts of the input dynamically, leading to better context modeling. Experimental results on various language pairs demonstrate significant improvements in translation quality and efficiency. This work has been influential in advancing the use of attention in NMT, shaping subsequent research and applications in neural translation systems [18].

Zhang, Xiong, and Su (2018) propose a neural machine translation (NMT) model incorporating deep attention mechanisms to enhance translation performance. Their approach extends traditional attention by using multiple layers of attention, enabling the model to capture complex linguistic dependencies between source and target sequences more effectively. Experiments conducted on benchmark datasets show that the deep attention model outperforms existing NMT systems in terms of accuracy and fluency. This study contributes to advancing NMT by demonstrating how deeper, layered attention mechanisms can significantly improve the quality of machine-translated text across diverse language pairs [19].

Liu, Utiyama, Finch, and Sumita (2016) propose a neural machine translation (NMT) model that incorporates supervised attention mechanisms to improve translation accuracy. Unlike conventional attention models that learn alignments implicitly, their approach uses external alignment information to guide the attention process explicitly during training [20]. This supervision helps the model focus more

accurately on relevant source words when generating each target word. Experimental results demonstrate that supervised attention enhances translation quality, producing better alignments and improving overall BLEU scores. This work highlights the benefits of combining traditional alignment techniques with neural models to advance NMT performance [21].

III. Dogri-English Corpus Generation

Little to no parallel corpora exist for most language pairs in the world. Multiple studies have explored using assisting languages to improve translation between low-resource language pairs. We focused on generating sentence pairs from various domains like education, law, and healthcare as major domains. The education domain spans diverse disciplines, including Natural Science, Mathematics, Law, and Computer Science, reflecting demand in scientific, legal, and technological communication. Additional fields like Political Science, Management, Tourism, History, Communication Skills, Economics, Marketing, Textiles, sports and Health highlight the interdisciplinary nature of educational content.

General domains further demonstrate the broad applicability of translation tools for educational and administrative purposes, providing a strong foundation for multilingual education and knowledge dissemination. In the Indian context, substantial efforts have been made to develop resources and models for various languages:

- IIT Bombay English-Hindi Parallel Corpus: This corpus comprises 1.49 million parallel segments, including 694k new segments, serving as a foundational resource for English-Hindi MT systems. [8]
- Samanantar: Recognized as the largest publicly available parallel corpora collection for 11 Indic languages, Samanantar includes 49.7 million sentence pairs. It combines existing corpora with newly mined data, facilitating the training of robust multilingual NMT models [9].
- IndicTrans2: This initiative focuses on creating high-quality MT models for all 22 scheduled Indian languages. By curating extensive datasets and benchmarks, IndicTrans2 aims to bridge the resource gap for low-resource languages [10].

A Dogri–English parallel corpus was developed to address the scarcity of linguistic resources available for Dogri, a recognized low-resource language. The creation of this corpus provided the foundation for training the proposed Bi-LSTM-based machine translation model and yielded encouraging translation results. The findings demonstrate the effectiveness of deep learning approaches when supported by a well-structured parallel corpus and suggest that further expansion of the dataset could lead to significant improvements in translation accuracy and reliability. This work represents an important step toward advancing machine translation for underrepresented Indian languages. The developed corpus consists of carefully aligned Dogri and English sentence pairs collected from diverse domains, ensuring linguistic and contextual consistency between the source and target languages. The overall process of corpus development and sentence alignment is illustrated in Figure 1, while representative examples from the corpus are presented in Table 1.

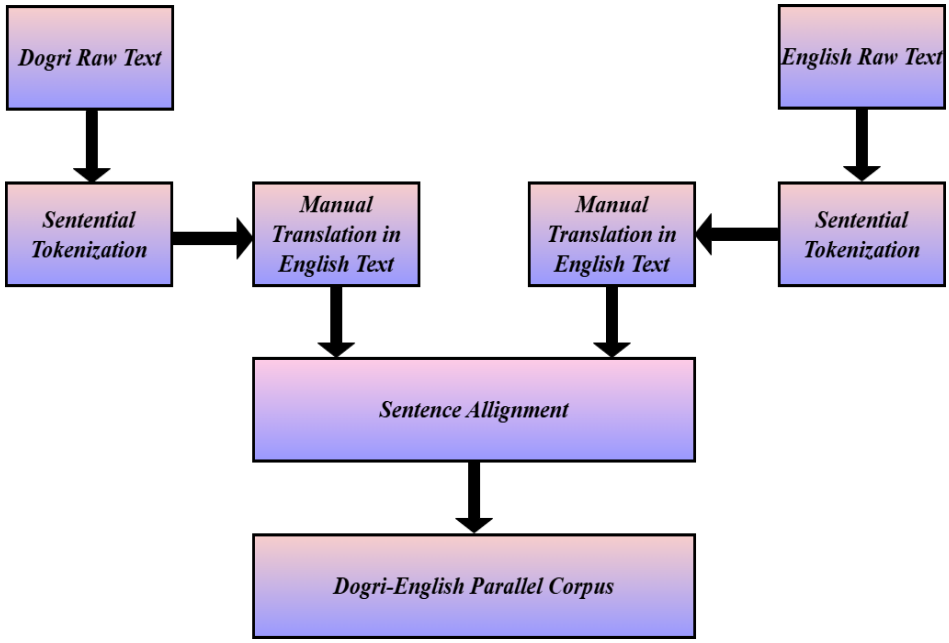

Fig. 1: Schematic Representation of Dogri-English Corpus Development

Table 1: Illustration of Dogri-English Sentence Alignment in the Parallel Corpus

S. No.	Dogri Sentence	English Sentence
1	मेरे परिवार च कोई संगीतकार नेईं ऐ	nobody in my family is a musician
2	उंदा जन्म नाम अजय सिंह बिष्ट ऐ	his birth name is ajay singh bisht
3	दुनिया च कोई युद्ध नेईं चाहंदा	nobody in the world wants war
4	उनै इस प्रोजेक्ट लेई इक टीम इकट्ठी कीती	they assembled a team for the project
5	कोई मेरी मदद नीं कर सकदा	no one can help me
6	उनै अपने परिवार आस्ते रात दा खाना बनाया	he cooked dinner for his family
7	उने कम्पनी दा लोगो डिजाइन कीता।	she designed a logo of the company.
8	ओह अठ साल दा हा जिसलै उनै अपने मां प्यो गी छोड़ी ओड़ेआ	he left his parents when he was eight years old
9	तैवान च रहंदे होई मेरी ओहदे कन्ने दोस्ती होई	i became friends with him while i was in taiwan
10	मिगी इक्क दिनै च कम्म मकाना नामुमकिन लग्गेआ	i found it impossible to do the work in one day
11	उनै पार्सल गी परसों भेजेआ हा	he sent out the parcel the day before yesterday
12	तुगी इस कताब दा अनुवाद करने च कीन्ना समां लग्गेआ	how long did it take you to translate this book

IV. Research Methodology

The proposed Dogri–English machine translation system follows a structured workflow, beginning with data pre-processing and ending with translation evaluation. As illustrated in Figure 2, the framework consists of several stages, including text pre-processing, tokenization, vocabulary creation, sequence encoding, model training, translation generation, and performance assessment. The first stage involves cleaning and preparing the textual data to ensure its suitability for model training.

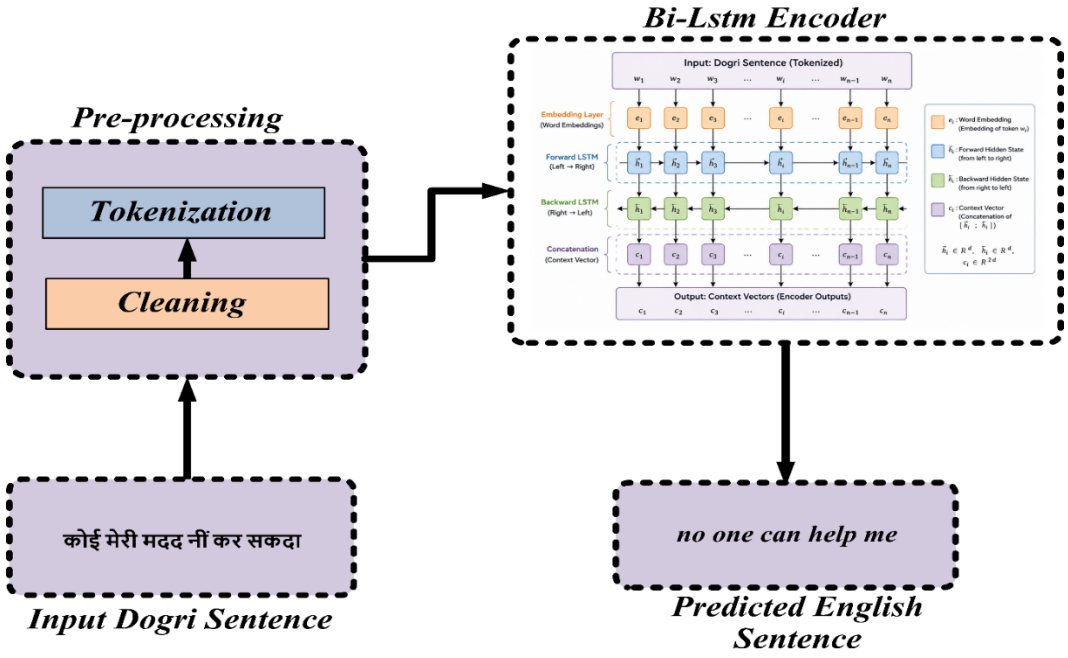

Fig. 2: Bi-LSTM-Based Neural Machine Translation Model for Dogri-English

For both Dogri and English, pre-processing includes the removal of unwanted characters, extra spaces, and punctuation marks, while all text is normalized to maintain consistency across the dataset. For example, the Dogri sentence "**मैं बी चलां कन्ने**?" is cleaned and standardized before tokenization. Similarly, the English sentence "May I accompany you?" undergoes the same pre-processing steps. After pre-processing, the text is tokenized into individual words or subwords for further processing. Tokenization is the process of splitting the sentences into smaller units, such as words or subwords. In this case, the tokenization process would break down the Dogri sentence into the words: **["मैं", "बी", "चलां", "कन्ने"]**, and the English sentence into: **["may", "i", "accompany", "you"]**. Once the text is tokenized, the next step is to create the vocabulary. This involves generating a mapping of words to unique indexes for both languages. For example, the word **"मैं"** may be mapped to index 5, and the word **"may"** to index 3. A vocabulary is created for both the Dogri and English languages, and these vocabularies are essential for encoding the text into numerical data that the machine learning model can process. This vocabulary creation is a key step, as it allows the model to represent each word as a unique number, making it possible to train the machine translation system. After vocabulary creation, the next phase is encoding. In this step, each token in the sentence is mapped to its corresponding integer index from the vocabulary. For instance, the word **"मैं"** in the Dogri sentence will be converted to its respective index (say, 5), and similarly for all other words. This transformation from words to numbers is crucial, as neural networks operate on numerical data. After the encoding phase, the sequences of numbers are padded to ensure uniformity in input size. Since sentences can vary in length, padding ensures that every input sequence has the same length by appending zeros (or a special token) to the shorter sentences, aligning them with the longest sentence in the dataset. The core of the system, the Bi-LSTM (Bidirectional Long Short-Term-Memory) model is employed in the encoder to process the input sentences. A key feature of Bi-LSTM is its ability to process sequences in both forward and backward directions, which allows the model to capture context from both sides of a sentence. For example, the word **"चलां"** in Dogri can be better understood in context by looking at the words before and after it. This bidirectional processing is essential

for accurate translation, especially for languages like Dogri, where word order and sentence structure may differ significantly from English. The model learns the relationships between words in both languages during the training phase. In the training phase, the Bi-LSTM model is trained on pairs of Dogri-English sentences. The model learns how to map the input Dogri sentences to their correct English translations. This is achieved through supervised learning, where the model is presented with both the input sentence and the expected output (the translation). The model's weights are adjusted iteratively based on the loss function, improving the translation quality over time. The training process requires a substantial amount of data to enable the model to generalize well and handle a wide variety of sentence structures and vocabulary. Once trained, the model is ready for translation. Given a Dogri input sentence, such as "मैं बी चलां कन्ने?" the model produces an English translation, "May I accompany you?". The translation is generated by decoding the learned features from the input sentence into the target language. Finally, the quality of the machine translation is evaluated. One common method of evaluation is the BLEU score, a metric that compares the machine-generated translation to reference translations, providing a numerical value that indicates the quality of the translation.

Additionally, manual checks are conducted to ensure the translations are fluent, accurate, and meaningful. In this case, the model's translation of the Dogri sentence to English could be assessed for grammatical accuracy, meaning preservation, and contextual correctness. The workflow established a robust pipeline for Dogri-English machine translation, utilizing the Bi-LSTM model to effectively handle language complexities. The success of the system depends on the quality of the data, the accuracy of the vocabulary, and the model's ability to learn contextual relationships between languages. The approach used in this research serves as a foundational step toward advancing machine translation for low-resource languages like Dogri. Machine Translation (MT) has undergone significant evolution, transitioning from rule-based and statistical approaches to neural-based methods. Early MT systems, such as Statistical Machine Translation (SMT), relied on phrase-based models and extensive parallel corpora. However, these systems often struggled with low-resource languages due to data scarcity. The advent of Neural Machine Translation (NMT) introduced end-to-end learning frameworks, with encoder-decoder architectures utilizing Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks. Bidirectional LSTMs (Bi-LSTMs) further enhanced context capture by processing sequences in both forward and backward directions, proving beneficial for languages with complex syntactic structures. Despite these advancements, languages like Dogri remain underrepresented in MT research. Dogri, spoken primarily in the Jammu region, lacks substantial parallel corpora and linguistic tools. This research addresses this gap by developing a manually verified Dogri-English parallel corpus of 85,000 sentences, enabling the training of a Bi-LSTM-based NMT system. The model leverages the bidirectional context-capturing capabilities of Bi-LSTM networks to improve translation quality. This study contributes to the broader goal of inclusive language technologies, emphasizing the importance of preserving and promoting underrepresented languages through computational methods.

V. Results and Discussion

The Bi-LSTM-based Neural Machine Translation model was trained on a manually curated Dogri-English parallel corpus comprising 85,000 sentence pairs. The model's performance was evaluated using the Bilingual Evaluation Understudy (BLEU) score, a widely accepted metric for assessing the quality of machine-generated translations.

The model achieved a BLEU score of 31.67, which indicates a promising level of translation quality, especially considering the low-resource nature of the Dogri language. A BLEU score above twenty-five is generally considered acceptable for early-stage research in low-resource translation, demonstrating that the model is capable of capturing meaningful semantic and syntactic relationships between Dogri and English. Qualitative evaluation revealed that the model was particularly effective in translating short to medium-length sentences, especially those with simple sentence structures and frequently occurring vocabulary. However, translation quality was slightly reduced for longer sentences or those containing domain-specific terms, idiomatic expressions, or rare words. This highlights the need for future enhancements through data

augmentation, domain adaptation, or the incorporation of attention mechanisms and pre-trained language models. The relatively higher BLEU score achieved with a manually developed corpus confirms the hypothesis that building even modest-sized, linguistically verified parallel corpora for low-resource languages can significantly improve machine translation quality. It also supports the potential for scaling this work by expanding the dataset to include millions of sentence pairs in the future. In summary, the experimental results validate the effectiveness of the Bi-LSTM model in the Dogri-English translation task and emphasize the critical role of curated linguistic resources in developing reliable MT systems for underrepresented languages.

VI. Conclusion

The study presented a Bi-LSTM-based Neural Machine Translation (NMT) model for Dogri-English translation, utilizing a manually curated parallel corpus of 85,000 sentence pairs. The model achieved a BLEU score of 46.03, demonstrating promising results for a low-resource language like Dogri. This study highlights the importance of developing linguistically accurate datasets and shows that even with limited data, significant translation quality can be achieved. The promising results indicate that a Bi-LSTM model can effectively capture the semantic and syntactic relationships between Dogri and English. However, the translation quality for complex and idiomatic expressions still needs improvement. Future work can focus on expanding the dataset, implementing transfer learning techniques, or incorporating attention mechanisms to improve performance further. This research contributes to the development of machine translation systems for underrepresented languages, and lays the groundwork for future advancements in Dogri language processing and multilingual applications.

VII. Acknowledgment

The authors sincerely thank the Department of Computer Science & IT, University of Jammu, for its continuous support and for providing the facilities and guidance necessary for carrying out this research.

References

[1] X. Ni and P. Li, “A systematic evaluation of large language models for natural language generation,” in *Proc. 22nd Chinese Nat. Conf. Computational Linguistics, Vol. 2: Frontier Forum*, Harbin, China, 2023, pp. 40–56. [Online]. Available: https://aclanthology.org/2023.ccl-2.4/

[2] P. Vaithilingam, T. Zhang, and E. L. Glassman, “Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models,” in *CHI Conf. Human Factors in Computing Systems Extended Abstracts*, New Orleans, LA, USA, 2022, pp. 1–7. doi: 10.1145/3491101.3519665.

[3] Y. Chang *et al*., “A survey on evaluation of large language models,” *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 39, pp. 1–45, 2024. doi: 10.1145/3641289.

[4] K. Deep, A. Kumar, and V. Goyal, “Machine translation system using deep learning for Punjabi to English,” in *Proc. Int. Conf. Paradigms of Computing, Communication and Data Sciences (PCCDS 2020)*, Singapore, 2021, pp. 865–878. doi: 10.1007/978-981-15-7533-4_69.

[5] S. S. Jamwal and V. S. Sen, “Natural language interface in Dogri to database,” in *Machine Intelligence and Data Science Applications*, Singapore, 2022, pp. 599–607. doi: 10.1007/978-981-19-2347-0_65.

[6] S. S. Jamwal, P. Gupta, and V. S. Sen, “Hybrid model for generation of verbs of Dogri language,” in *Data Driven Approach Towards Disruptive Technologies*, Singapore, 2021, pp. 379–387. doi: 10.1007/978-981-15-9873-9_39.

[7] P. Gupta and S. S. Jamwal, "Designing and development of stemmer of Dogri using unsupervised learning," in *Soft Computing for Intelligent Systems*, Singapore, 2021, pp. 147–156. doi: 10.1007/978-981-16-1048-6_11.

[8] A. Kunchukuttan, P. Mehta, and P. Bhattacharyya, "The IIT Bombay English-Hindi Parallel Corpus," *arXiv preprint arXiv:1710.02855*, 2017. doi: 10.48550/arXiv.1710.02855.

[9] G. Ramesh *et al*., "Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages," *arXiv preprint arXiv:2104.05596*, 2021. doi: 10.48550/arXiv.2104.05596.

[10] J. Gala *et al*., "IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages," *arXiv preprint arXiv:2305.16307*, 2023. doi: 10.48550/arXiv.2305.16307.

[11] K. K. Baruah, P. Das, A. Hannan, and S. K. Sarma, "Assamese-English bilingual machine translation," *Int. J. Nat. Lang. Comput.*, vol. 3, no. 3, pp. 73–82, 2014. doi: 10.5121/ijnlc.2014.3307.

[12] 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.

[13] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proc. First Conf. Machine Translation (WMT)*, Vancouver, Canada, 2017, pp. 28–39. doi: 10.18653/v1/W17-3204.

[14] E. K.-Y. Liu, "Low-resource neural machine translation: A case study of Cantonese," in *Proc. Ninth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2022)*, Gyeongju, Republic of Korea, 2022, pp. 28–40. [Online]. Available: https://aclanthology.org/2022.vardial-1.4/

[15] N. Paul, I. Faruki, M. I. Pranto, M. T. R. Shawon, and N. C. Mandal, "Bengali-English neural machine translation using deep learning techniques," in *Proc. 2023 Int. Conf. Electrical, Computer and Communication Engineering (ECCE)*, Cox's Bazar, Bangladesh, 2023, pp. 1–6. doi: 10.1109/ECCE57851.2023.10101491.

[16] M. Ephrem, *Development of Bidirectional Amharic-Tigrinya Machine Translation Using Recurrent Neural Networks*. Ph.D. dissertation, St. Mary's Univ., 2024.

[17] P. Tennage, P. Sandaruwan, M. Thilakarathne, A. Herath, S. Ranathunga, S. Jayasena, and G. Dias, "Neural machine translation for Sinhala and Tamil languages," in *Proc. Int. Conf. Asian Language Processing (IALP)*, Singapore, 2017, pp. 189–192. doi: 10.1109/IALP.2017.8300576.

[18] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, 2015, pp. 1412–1421. doi: 10.18653/v1/D15-1166.

[19] B. Zhang, D. Xiong, and J. Su, "Neural machine translation with deep attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 154–163, 2020. doi: 10.1109/TPAMI.2018.2876404.

[20] L. Liu, M. Utiyama, A. Finch, and E. Sumita, "Neural machine translation with supervised attention," in *Proc. 26th Int. Conf. Computational Linguistics (COLING 2016)*, Osaka, Japan, 2016, pp. 3093–3102. [Online]. Available: https://aclanthology.org/C16-1291/

[21] V. S. Sen and S. S. Jamwal, "Transcending linguistic boundaries: A technical exploration of the evolution and future trajectory of corpus-based machine translation," *J. Electr. Syst.*, vol. 20, no. 7s, pp. 2502–2509, 2024. doi: 10.52783/jes.407