Machine Translation and Transliteration for Indo-Aryan Languages: A Systematic Review

Sameera Perera
Informatics Institute of Technology
Colombo 006
Sri Lanka
sameeraperera827@gmail.com

T.G.D.K. Sumanathilaka
Swansea University
Wales
United Kingdom
deshankoshala@gmail.com

Abstract

With the advent of Web 2.0, digital platforms have become increasingly multilingual. Non-English speakers are rapidly adopting their native languages on social media, highlighting the need for robust translation and transliteration models to facilitate effective communication. This systematic review paper provides an overview of recent machine translation and transliteration developments for Indo-Aryan languages spoken by a large South Asian population. The paper examines advancements in translation and transliteration systems for a few language pairs that have appeared in recently published papers in the last half a decade. The review summarizes the current state of these technologies, providing a worthwhile resource for anyone who is doing research in these fields to understand and find existing systems and techniques for translation and transliteration. The current challenges and limitations in the current systems are identified, and possible directions are suggested.

1 Introduction

The Indo-Aryan languages constitute a main branch of the Indo-European language family, predominantly spoken in Central and North India as well as in neighbouring countries such as Sri Lanka, Pakistan, Nepal, Maldives, Bangladesh and Bhutan (Pal and Zampieri, 2020). The large linguistic varieties within the Indo-Aryan language family make it challenging to communicate both outside and within the region. Machine translation and transliteration systems help to bridge language barrier,s enabling effective communication between different linguistic societies.

The goal of this review paper is to provide an overview of the current state of machine translation and transliteration techniques for Indo-Aryan languages. The review discusses diverse techniques used in the recently published translation and transliteration systems which handle the various scripts and linguistic features of Indo-Aryan languages.

The contribution of this study can be summarized as performing a systematic review of existing translation and transliteration techniques related to Indo-Aryan languages, highlighting the significant contributions and developments made by researchers in this constantly developing field. Going forward, the review is structured to clearly look into the recent developments in machine translation and transliteration for Indo-Aryan languages. Starting with the methodology explains how studies were selected based on their relevance. The following sections dive into various translation and transliteration approaches and outline the challenges faced in the field.

2 Methodology

A systematic approach was adopted in this review to choose the relevant studies on machine translation and transliteration for Indo-Aryan languages. A comprehensive search was conducted across several major academic databases, including IEEE Xplore and Google Scholar. In addition to the academic database searches, several key papers were identified from references cited in already published research, ensuring a wide-ranging collection of studies relevant to the focus of the review. Keywords such as "machine translation", "transliteration" and "Romanized languages" were used to identify relevant literature. To avoid redundancy, duplicate publications across different databases were identified and removed.

This review focused on papers published from 2018 to the available 2024 publications to ensure that the recent advancements were included. Studies were chosen based on the relevance to machine translation and transliteration within the context of Indo-Aryan languages. This review also includes

Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (IndoNLP2025), pages 23–33
January 20, 2025. ©2025 Association for Computational Linguistics

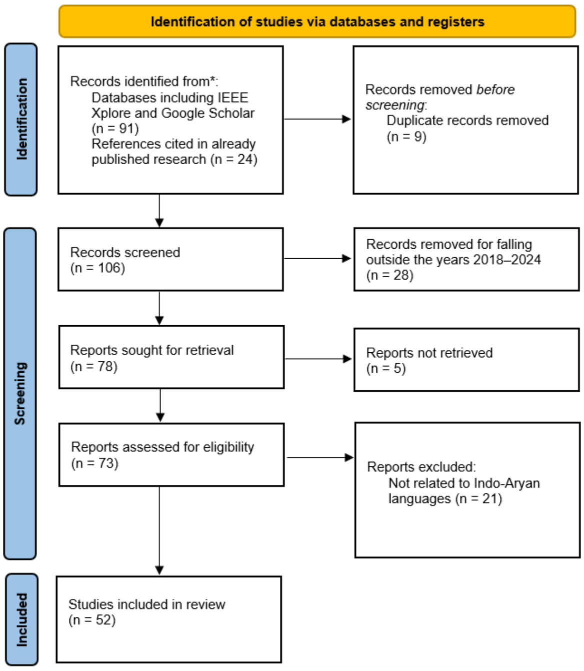
Identification of studies via databases and registers
Identification
Records identified from*:
Databases including IEEE Xplore and Google Scholar (n = 91)
References cited in already published research (n = 24)
Records removed *before screening*:
Duplicate records removed (n = 9)
Screening
Records screened (n = 106)
Records removed for falling outside the years 2018–2024 (n = 28)
Reports sought for retrieval (n = 78)
Reports not retrieved (n = 5)
Reports assessed for eligibility (n = 73)
Reports excluded:
Not related to Indo-Aryan languages (n = 21)
Included
Studies included in review (n = 52)

Figure 1: PRISMA Flow of the Paper Selection Process

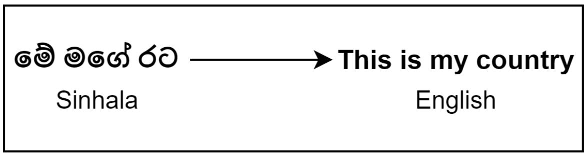

Figure 2: Sinhala to English Translation

papers that have proposed and utilized relevant techniques as part of their work while not directly focused on translation or transliteration. Specifically, the papers which proposed novel methodologies or made outstanding contributions to the field were prioritized. Figure 1 illustrates the systematic flow of the paper selection process.

3 Machine Translation (MT) and Transliteration

Machine Translation (MT) is the study of how to use machines to translate from a source language into another target language. This concept was first put forward by Warren Weaver in 1947 (Wang et al., 2022). From then on, MT has been one of the most challenging tasks in the natural language processing (NLP) field. Figure 2 is an example of machine translation between Sinhala and English.

Machine transliteration is the process of words transformation from one language into their phonetic equivalent of another. There are two types of machine transliteration: forward and backward transliteration. forward transliteration is the process of transliterating a word to a foreign language from the language from which it originated. On the other hand, when a word is converted back to the language of its origin from a foreign language, it is known as backward transliteration (Kaur and Garg, 2022). Figure 3 illustrates the difference between forward and backward transliteration using Romanized Sinhala.

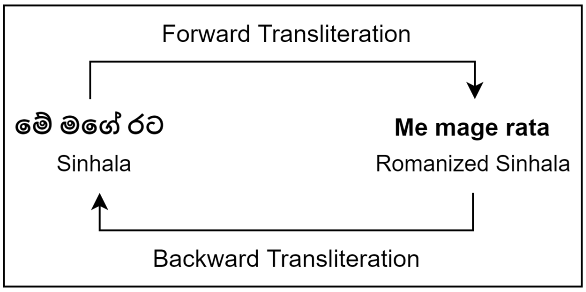

Figure 3: Forward and Backward Transliteration

4 Approaches in Machine translation and Transliteration

Many machine translation and transliteration systems have been implemented for Indo-Aryan languages. Since transliteration is considered a form of translation, both translation and transliteration systems have used similar approaches. The following section will discuss the various machine translation and transliteration approaches found in the literature. Here, the ISO 15919 standard[1] for the transliteration of Devanagari and related Indic scripts into Latin characters has not been the focus, but it may be relevant in rule-based approaches.

4.1 Rules-based Machine Translation (RBMT)

RBMT (Rules-Based Machine Translation) is a type of MT system which translates languages based on the rules which represent linguistic knowledge. Large number of linguistic terms can be applied to using the Rules-Based Machine Translation methodology in 3 stages: analyzing, transferring, and generating. Programmers and linguists who have already spent a significant amount of time to understand the principles and patterns between 2 languages have established rules. RBMT methods only produce good results only if the translation rules are applied correctly. Transfer-based machine translation and Interlingual machine translation are

[1] https://www.unige.ch/biblio_info/files/5116/3775/9122/ISO_15919_en.pdf

two main types of RBMT (Khepra et al., 2023).

Transfer-based machine translation: This MT type breaks down the process of translation into several subtasks, such as morphological analysis, syntactic parsing, and semantic analysis, and then translates the meaning of source input into the target languages. This approach is useful to handle complex grammatical structures and idiomatic expressions (Khepra et al., 2023).

Interlingual machine translation: This approach involves using an intermediary language to translate between the source and target languages and then translate it into the target language. One of the major advantages of this approach is that it can handle multiple languages at once, and it may bring down errors in the output (Khepra et al., 2023).

4.2 Corpus-based Machine Translation (CBMT)

Corpus-based machine translation (CBMT) relies on large amounts of parallel corpus (bilingual text) to train statistical models for translation. The models are trained to learn patterns in the data, then use those patterns to make translations. The study by Khepra et al. (2023) describes two types of CBMTs: Example-based machine translation and Statistical Machine Translation (SMT).

Example-based machine translation (EBMT): This type of MT uses a bilingual sentence pairs database to translate text. The system gets the most similar sentence pair from the database and use it to generate the target sentence. This approach is useful to handle less common language pairs or rare languages (Khepra et al., 2023).

Statistical Machine Translation (SMT): In Statistical Machine Translation, the model is developed completely from the information in corpora without user intervention. It was the dominant paradigm up until the beginning of 2010. A computer requires examples which provide information about the translation of the phrases (the bilingual word mappings) and the appropriate placements of the converted words in the targeted phrase (alignment) to learn how to translate (Khepra et al., 2023).

4.3 Knowledge-based Machine Translation (KBMT)

This kind of MT uses a predetermined set of grammatical and lexical rules to translate text. A different name for it is a rule-based machine translation. KBMT is especially advantageous in its capability to handle specific domains such as legal or technical texts where the text structure is well-defined (Khepra et al., 2023).

4.4 Neural Machine Translation (NMT)

Neural Machine Translation (NMT) uses deep learning techniques to train an MT model on large amounts of parallel data. Typically, NMT models are more accurate than rule-based or statistical models but also need more computational resources for the training process (Khepra et al., 2023).

4.5 Hybrid Machine Translation

This approach is one of the latest approaches in machine translation systems. This will be developed with the combination of more than one existing MT-based approach. Two or more approaches discussed in the above sections can be used in the Hybrid approach to produce accurate results (Sumanathilaka et al., 2023).

4.6 Discussion of MT Approaches

Each machine translation (MT) approach has different strengths and limitations according to their underlying mechanisms. To address some of the issues with these single approaches, researchers have used a combination of these approaches to overcome those issues. RBMT and KBMT approaches depend on predefined rules, making them effective for structured texts. However, those approaches struggle with unseen text which does not follow predefined rules. CBMT approaches, including SMT and EBMT, utilize large parallel corpora, offering more adaptability, but these approaches require substantial data. NMT can be identified as the most commonly used approach recently. Both NMT and CBMT face the challenge of data scarcity for low-resource languages. When corpus size is small, SMT performs better than the NMT according to results obtained by Tennage et al. (2017). Recently, there has been an outstanding trend to use transformers (Vaswani et al., 2017), which is one of the latest NMT approaches.

5 Current State of Machine Translation for Indo-Aryan Languages

This section provides an overview of the current state of MT approaches developed for diverse Indo-Aryan language pairs.

5.1 Hindi-English Translation

Recently, NMT has been broadly explored for this language pair. Singh et al. (2019) proposed LSTM

(Long Short-Term Memory) based NMT system for English-Hindi translation showing promising results, especially for shorter sentences. Further advancements include the study by Tiwari et al. (2020), who suggested 2 other NMT approaches, which are ConvS2S and LSTM Seq2Seq, with the ConvS2S model outperforming the proposed LSTM model. Similarly, Gogineni et al. (2020) proposed an NMT model based on Bidirectional LSTM (BiLSTM), outperforming the traditional SMT approaches in terms of BLEU scores. Attention mechanisms have also been a major focus in enhancing the performance of NMT systems. Laskar et al. (2019) studied the comparison between two NMT approaches, one based on the modern transformer model, which is based on a recently introduced self-attention mechanism and the other on the LSTM. The results demonstrated that the transformer-based model outperformed the LSTM-based model. Rose et al. (2023) showed that incorporating an attention mechanism into an Encoder-Decoder-based LSTM model significantly improved the translation. The use of a guided transformer model proposed by Bisht et al. (2023) further increased the translation performance by integrating dependency parsing into the encoder. For addressing challenges in long sentence translation, Sarode et al. (2023) explored the Recurrent Neural Networks (RNN) and Gated Recurrent Units (GRU) usage in a Seq2Seq architecture with an attention mechanism. Lastly, Watve and Bhalekar (2023) implemented a transformer-based English-to-Hindi translator, contributing to the improvement of work in this area.

5.2 Sinhala-English Translation

To improve the accuracy of Sinhala to English translation, Nugaliyadde et al. (2019) proposed a novel approach using an Evolutionary Algorithm (EA). This method iteratively refines the translation ensuring that the final output is meaningful and grammatically correct. According to their paper, this is one of the early efforts to apply EA in MT for Sinhala-English language pairs. Fonseka et al. (2020) introduced a transformer-based translation system particularly developed for translating official government documents between English and Sinhala. To address one of the common issues in MT, which is the out-of-vocabulary (OOV) issue, they implemented Byte Pair Encoding (BPE). Further advancements were made by researchers who explored the document alignment in Sinhala and English. For example, research extended the Si-Ta (Ranathunga et al., 2018) system (Will be discussed in the next section) to include SMT techniques improving the alignment process between Sinhala and English texts (C et al., 2020). To enable Sinhala speakers to search English web content effectively, Hisan et al. (2020) focused on a cross-language information retrieval system using word embeddings to enhance the translation of Sinhala queries into English. Additionally, Sandaruwan et al. (2021) addressed the challenge of translating Romanized Sinhala into English. They built a Seq2Seq NMT model with an attention mechanism that effectively handled the various spelling variations in Singlish. In this system, a deep multi-layer RNN, which consists of bidirectional LSTMs, is considered recurrent units.

5.3 Sinhala-Tamil Translation

The first dedicated MT system for Sinhala and Tamil official documents was Si-Ta which is proposed by Ranathunga et al. (2018). Nissanka et al. (2020) further explored Neural Machine Translation for this pair of languages using Byte Pair Encoding (BPE) to address the OOV problem as described above in the study by Fonseka et al. (2020). In their approach, they combined monolingual and parallel corpus data utilizing transformer architecture to improve translation accuracy. In a study comparing different translation models done by Pramodya et al. (2020), they found that the introduction of the Incrementally Filtered Back-Translation technique, which was proposed by Arukgoda et al. (2019), enabled NMT models to surpass SMT models, especially in low-resource conditions. They compared different translation models, including RNNs, SMT and Transformer models for Tamil to Sinhala translation. Thillainathan et al. (2021) extended this line of research by fine-tuning modern pre-trained large language models such as mBART for extremely low-resource translation tasks. They showed that fine-tuning these models significantly enhanced the quality of translation for Sinhala-Tamil, especially in domain-specific contexts (such as official government documents) compared to traditional Transformer-based NMT models.

5.4 Punjabi-English Translation

SMT-based system for Punjabi-English language pair using the Moses toolkit has been studied by Jindal et al. (2018). That involved creating a 20,000-

sentence parallel corpus encompassing diverse domains and utilizing GIZA++ for word alignment.

5.5 Bengali-English Translation

Research on Bengali-English translation has been focused on both NMT and SMT approaches. Rahman et al. (2018) proposed an MT system which uses a corpus-based method with an N-gram language model. The results of this system have been shown to outperform Google Translate in terms of computational efficiency and accuracy. More recently, Paul et al. (2023) evaluated four different Seq2Seq models, which are LSTM, GRU, BiLSTM and Bidirectional GRU (BiGRU), concluding that the BiLSTM model performed well achieving high BLEU scores.

5.6 Sanskrit-Hindi Translation

For the Sanskrit-Hindi language pair, a Corpus-Based Machine Translation (CBMT) system using deep neural networks to translate Vedic texts and other sacred writings was proposed by Singh et al. (2020). This system was able to handle phrasal and idiomatic expressions, achieving a BLEU score of 41.17. Lastly, Bhadwal et al. (2020) explored an RBMT model which utilizes a direct (dictionary-based) approach for translating text from Hindi to Sanskrit.

5.7 Sanskrit-Gujarati Translation

Raulji et al. (2022) introduced a novel framework to translate Sanskrit to Gujarati using a symbolic approach. They focused on keeping grammatical structures through a sequential process involving morphological and syntactic analysis, lexical transfer and grammatical transfer. This system achieved a BLEU score of 58.04 despite the challenge of scarcity of resources, which demonstrated the effectiveness of this system for low-resource languages.

5.8 Urdu-English Translation

A study proposed by Naeem et al. (2023) evaluated the performance of different neural network models (RNN, GRU, and LSTM) for translation between English and Urdu languages and the results showed that the GRU model outperformed the others.

5.9 Marathi-English Translation

Recent research on the Marathi-English translation has been relatively limited. For the Marathi-English translation, Gunjal et al. (2023) proposed a Seq2Seq transformer model, which was trained on a large dataset of parallel English-Marathi sentences and achieved a BLEU score of 41. 99.

5.10 Kashmiri-English Translation

Research on the Kashmiri-English language translation has also been relatively limited. A study (Giri et al., 2024) proposed an RNN-based MT system focusing on the tourism domain. This system is structured on an Encoder-Decoder model, indicating initial efforts for this pair of languages, especially in domain-specific contexts.

5.11 Other Multilingual Translation

A study proposed by Sen et al. (2018) introduced two multilingual Transformer architecture-based NMT models: many-to one (7 Indic languages to English) and one-to-many (English to 7 Indic languages). The results showed that multilingual NMT performs better than separate bilingual NMT models if the target side has only one language (English). When the target has many languages, multilingual NMT performance degrades compared to bilingual models for relatively high-resource languages. Further advancements in multilingual language translation involve the inclusion of Hindi, Telugu, Kannada and English within a single system (Chimalamarri et al., 2020). This study improved transformer-based NMT models by incorporating source-side morpho-linguistic features, which are word-based, BPE-based, and morpho-lexical features with POS tags. The results showed significant enhancements in the translation process for all language pairs by incorporating source-side morpho-linguistic features, especially morpho-lexical features with POS tags. Another important translation system based on pre-trained mT5 transformer was fine-tuned to translate between Hindi, Bengali, and English (Jha et al., 2023). That system leveraged the extensive multilingual capabilities in the mT5 model, achieving high BLEU scores for Bengali-English and English-Bengali translations.

6 Current State of Machine Transliteration for Indo-Aryan Languages

The transliteration of Indo-Aryan languages has been a challenge of research for several decades. There are various models proposed to address the complexities of converting text from one script to another. Over the years, the transliteration approaches have improved from traditional rule-based methods to modern neural and hybrid models, reflecting the increasing computational capabilities. From 2018 to 2024, there were more studies on transliteration systems for Sinhala compared to other Indo-Aryan languages.

In 2018, significant contributions were made to transliteration with the development of rule-based and modern machine-learning approaches. A rule-based transliteration system for Romanized Sinhala was proposed, using phonetic and transliteration rule bases to transliterate Romanized text into native Sinhala script. While effective, the system faced limitations in handling ambiguities, particularly with proper nouns(Vidanaralage et al., 2018). Another study experimented with Seq2Seq and LSTM models to develop a scalable transliteration pipeline for Indian languages and evaluated different language transliterations. The results showed that the Seq2Seq models outperformed traditional LSTM models, although they need large datasets for effective training (Joshi et al., 2018). Additionally, a character-level transliteration tool was created to improve Tamil to Sinhala NM,T demonstrating the utility of rule-based methods in translation tasks (Tennage et al., 2018). According to their literature, that was the first Tamil to English and Sinhala to English transliteration tool that used a rule-based approach.

In 2019, Priyadarshani et al. (2019) introduced a hybrid approach using SMT and machine learning to transliterate personal names in the Sri Lankan context using Moses SMT toolkit for Sinhala, Tamil and English languages. This system showed the importance of incorporating ethnic origin classification for personal name transliteration to improve accuracy. Another significant transliteration approach was the Gurmukhi to Roman transliteration, which used character mapping and handcrafted rules for the transliteration of Punjabi to English with a good accuracy of 99.27% (Singh and Sachan, 2019).

There were further advancements in transliteration techniques during 2020 and 2021, particularly with the use of neural networks. A rule-based method which is proposed by UCSC is combined with a trigram model trained on social media text to improve the Sinhala transliteration accuracy in the study by Liwera and Ranathunga (2020). Another study in 2021 introduced a rule-based approach for Singlish to Sinhala transliteration with an error correction module to improve accuracy (Silva and Ahangama, 2021). Singh and Bansal (2021) experimented with various neural architectures for the transliteration of Hindi and Punjabi languages. Out of those, a model with a character/grapheme level bidirectional encoder and auto-regressive decoder proved to be the best-performing architecture. In the same year, a systematic approach employing phrase-based statistical machine translation (PB-SMT) to create an English-Hindi parallel database for transliteration was introduced (Mogla et al., 2021). Another work in 2021 was the development of a Python-based algorithm to transliterate between Devanagari or Roman scripts and Brahmic scripts, and vice versa (Nair and Ahammed, 2021). Additionally with the introduction of a method for normalizing and back-transliterating Hindi-English code-switched text, this field saw further innovation (Parikh and Solorio, 2021). This system first normalized Romanized Hindi with the use of the Seq2Seq model based on an LSTM encoder-decoder architecture and then syllabified the tokens to map them to the Devanagari script. This approach could handle informal typing variations and phonetic discrepancies, improving the transliteration.

Moving into 2022, Swa-Bhasha (Athukorala and Sumanathilaka, 2022) proposed a novel approach using a combination of rule-based methods and fuzzy logic to transliterate Singlish to Sinhala even when vowels are omitted. This system has introduced a new numeric coding system to use with the Romanized Sinhala letters by matching with the recognized typing patterns. Fuzzy logic-based implementation has been used for the mapping process. Another back-transliteration system for Romanized Sinhala to Sinhala was proposed by Nanayakkara et al. (2022) utilizing a Transliteration Unit (TU) based model and a BiLSTM encoder combined with an LSTM decoder. Moreover, in 2022, a bilingual RBMT system was developed for Sanskrit-English. This system allowed users to type Sanskrit using English orthography and transliterate Sanskrit text into the English script (Sethi et al., 2022).

In 2023, Sharma et al. (2023) introduced a Generative Adversarial Networks (GANs) based system using Pix2Pix GAN architecture to transliterate ancient Indian scripts (images) like Nandinagari and Sharda into modern Devanagari script (images). Yadav and Kumar (2023) proposed a hybrid approach to transliterate Hindi to English which includes image processing and a model trained with attention. The final phase of the proposed system, which is the transliteration phrase, used the Python Indicate Transliteration library to transliterate Hindi characters into the Roman script. In the same year, Swa-Bhasha hybrid approach combining statistical methods with a Trigram and rule-based model was proposed for Singlish back transliteration (Sumanathilaka et al., 2023). Additionally, it incorporated a Trie data structure to generate word suggestions. The work by Athukorala and Sumanathilaka (2022) has achieved 0.64-word level accuracy while Liwera and Ranathunga (2020) achieved 0.52-word level accuracy. This Swa-Bhasha system has performed much more accurately with 0.84-word level accuracy compared to the existing transliteration works for Sinhala. By applying a similar hybrid approach, another back-transliteration system for Romanized Tamil, TAMZHI, was proposed by Mudiyanselage and Sumanathilaka (2024). This system achieved 93% accuracy at the character level and 70% at the word level, further demonstrating the effectiveness of this method.

In 2024, further advancement was made with the introduction of Swa Bhasha 2.0 (Dharmasiri and Sumanathilaka, 2024), which is developed to address the ambiguities of Romanized Sinhala back transliteration using GRU-based NMT. Also, the study of Swa-Bhasha Dataset (Sumanathilaka et al., 2024) introduced a rule-based transliteration tool which can annotate Sinhala words into Romanized Sinhala. This system can accommodate the various ad hoc typing patterns used by the community. Finally, in 2024, another model was proposed for accurate cross-script conversion, focusing on the hybrid model development for transliteration. This study compared two models: a hybrid of Seq2Seq with LSTM and a hybrid of rule-based and NMT approaches. Seq2Seq with an LSTM-based model demonstrated superior performance, especially in back-transliterating English text into different Indic languages (Shukla et al., 2024).

7 Gaps and Challenges in Machine Translation and Transliteration for Indo-Aryan Languages

Despite significant advancements in the field of machine translation (MT) and transliteration for Indo-Aryan languages, there are still several challenges and gaps that can be identified. Addressing these will be important to develop reliable systems for any language. This section describes some of the identified gaps and challenges in this field.

7.1 Data Scarcity

Data scarcity in low-resource languages presents significant challenges to machine translation and transliteration, especially when using neural machine translation and corpus-based translation approaches like statistical machine translation (SMT). This problem gets worse in NMT approaches because these models are even more data-hungry than SMT. Some studies have shown that when corpus size is small, SMT performs better than the NMT (Tennage et al., 2017). Even though the transformer architecture, one of the latest NMT approaches, has shown outstanding results with high-resource language pair translation, recent studies have still conducted only a small number of works on Indo-Aryan languages because of data scarcity problems.

7.2 Complex Morphological and Syntactic Structures

The complex grammatical structures and rich morphology of Indo-Aryan languages, where a single word can have multiple forms depending on tense, gender, and case, pose challenges to translation systems. Syntactic differences between Indo-Aryan languages and other language families like English also complicate the translation process, especially with idiomatic expressions.

7.3 Out-of-Vocabulary (OOV) Words

The "out of vocabulary" (OOV) issue in this field refers to the problem which occurs when a source language word is not present in the vocabulary ofthe translation/transliteration system, meaning it has not been seen or learned during training. OOV words might include rare terms, names or new slang. Techniques such as Byte Pair Encoding (BPE) have been used to address this issue in recent systems, but this issue still persists in some developments.

7.4 Code-Mixing

A significant number of people use social media in various native languages other than English. However, most of these people do not use Unicode characters to represent their languages. Instead, they use phonetic typing with the English alphabet. Therefore, people express their native languages using the English alphabet, and they even insert English words mixed up with the native language words. This phenomenon is known as code-mixing (Smith and Thayasivam, 2019). Also, sometimes, people write in their native script and insert English words using the English alphabet. Some of the current MT and transliteration systems struggle to handle mixed language inputs.

7.5 Variations in Transliteration

When people use transliterated text, especially Romanized forms of Indo-Aryan languages, the writing patterns they use to express their native language vary from person to person. Also, these typing patterns change depending on the time and the mood of uthe ser (Sumanathilaka et al., 2024). Common variations in transliterated text include ambiguous consonant transliteration, vowel dropping, long vowel transliteration, double consonant transliteration, slang and abbreviations (Parikh and Solorio, 2021). These inconsistencies make it challenging to convert the transliterated text back into the native script. Few recent developments have focused on addressing these typing variations.

7.6 Word Ambiguity

Word ambiguity, where a single word can represent multiple meanings based on the context of the sentence, remains a key challenge. Addressing this problem is known as word sense disambiguation. While SMT and NMT approaches, such as LSTM and GRU models, can retain contextual information to some extent, they have not provided an optimal solution. The transformer architecture can offer a better approach. However, only a few translation/transliteration systems have been developed with this architecture, and it seems they have not given much direct attention to this problem.

8 Conclusion

The review highlights significant advancements in machine translation and transliteration for Indo-Aryan languages. Translation systems have seen notable improvements in accuracy with the advancement of natural language processing. In transliteration, there has been progress in converting text between different scripts by managing the phonetic variations. Notably, both translation and transliteration have seen significant enhancements with the advent of transformer architecture variations, which is marking a promising direction for future research in this field. These developments are important in improving effective communication and access to information across different Indo-Aryan language communities.

Limitations

This systematic review has several limitations that need to be considered. Considering only papers published between 2018 and 2024 might have left out earlier important studies which could provide more details on how machine translation and transliteration related to Indo-Aryan languages have evolved. This review only included papers which are freely available. As a result, it might have missed important studies published in less accessible journals or conference proceedings. Additionally, using specific keywords to find relevant studies might have caused important studies which do not use these exact keywords to be missed.

References

Anupama Arukgoda, A. Weerasinghe, and Randil Pushpananda. 2019. Improving Sinhala-Tamil Translation through Deep Learning Techniques.

Maneesha Athukorala and Deshan Sumanathilaka. 2022. Swa Bhasha: Message-Based Singlish to Sinhala Transliteration.

Neha Bhadwal, Prateek Agrawal, and Vishu Madaan. 2020. A Machine Translation System from Hindi to Sanskrit Language using Rule based Approach. *Scalable Computing: Practice and Experience*, 21(3):543–554.

Akhilesh Bisht, Deepa Gupta, and Shantipriya Parida. 2023. Guided Transformer for Machine Translation: English to Hindi. In *2023 IEEE 20th India Council International Conference (INDICON)*, pages 636–641, Hyderabad, India. IEEE.

Rajitha M. D. C, Piyarathna L.L. C, Nayanajith M. M.D. S, and Surangika S. 2020. Sinhala and English Document Alignment using Statistical Machine Translation. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 29–34, Colombo, Sri Lanka. IEEE.

Santwana Chimalamarri, Dinkar Sitaram, Rithik Mali, Alex Johnson, and K A Adeab. 2020. Improving Transformer based Neural Machine Translation with Source-side Morpho-linguistic Features. In *2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pages 1–5, Hyderabad, India. IEEE.

Sachithya Dharmasiri and T.G.D.K. Sumanathilaka. 2024. Swa Bhasha 2.0: Addressing Ambiguities in Romanized Sinhala to Native Sinhala Transliteration Using Neural Machine Translation. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 241–246, Belihuloya, Sri Lanka. IEEE.

Thilakshi Fonseka, Rashmini Naranpanawa, Ravinga Perera, and Uthayasanker Thayasivam. 2020. English to Sinhala Neural Machine Translation. In *2020 International Conference on Asian Language Processing (IALP)*, pages 305–309, Kuala Lumpur, Malaysia. IEEE.

Kaiser J. Giri, Nawaz Ali Lone, Rumaan Bashir, and Javaid Iqbal Bhat. 2024. English Kashmiri Machine Translation System related to Tourism Domain. In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1713–1717, New Delhi, India. IEEE.

Saikiran Gogineni, G. Suryanarayana, and Sravan Kumar Surendran. 2020. An Effective Neural Machine Translation for English to Hindi Language. In *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, pages 209–214, Trichy, India. IEEE.

Om Gunjal, Saurav Garje, Samir Aghav, Lavesh Jaykar, Sunil Sangve, and Saurabh Gunge. 2023. An Enhanced English to Marathi Translator using sequence-to-sequence Transformer. In *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, pages 1–5, Bangalore, India. IEEE.

M. H. M. Hisan, A. R. Weerasinghe, and B. H. R. Pushpananda. 2020. Cross Language Information Retrieval for Accessing the English Web in Sinhala. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 244–249, Colombo, Sri Lanka. IEEE.

Abhinav Jha, Hemprasad Yashwant Patil, Sumit Kumar Jindal, and Sardar M N Islam. 2023. Multilingual Indian Language Neural Machine Translation System Using mT5 Transformer. In *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, pages 1–5, Nagpur, India. IEEE.

Shishpal Jindal, Vishal Goyal, and Jaskarn Singh Bhullar. 2018. English to Punjabi statistical machine translation using moses (Corpus Based). *Journal of Statistics and Management Systems*, 21(4):553–560.

Akshat Joshi, Kinal Mehta, Neha Gupta, and Varun Kannadi Valloli. 2018. Indian Language Transliteration Using Deep Learning. In *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 103–107, Thiruvananthapuram, India. IEEE.

Palakpreet Kaur and Kamal Deep Garg. 2022. Machine Transliteration for Indian languages: Survey. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 462–467, Solan, Himachal Pradesh, India. IEEE.

Shaveta Khepra, Priya Kumari, Raj Gupta, Abhishek, and Vijendra Singh Bramhe. 2023. A Survey of Punjabi Language Translation using OCR and ML. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 136–144.

Sahinur Rahman Laskar, Abinash Dutta, Partha Pakray, and Sivaji Bandyopadhyay. 2019. Neural Machine Translation: English to Hindi. In *2019 IEEE Conference on Information and Communication Technology*, pages 1–6, Allahabad, India. IEEE.

W.M.P. Liwera and L. Ranathunga. 2020. Combination of Trigram and Rule-based Model for Singlish to Sinhala Transliteration by Focusing Social Media Text. In *2020 From Innovation to Impact (FITI)*, pages 1–5, Colombo, Sri Lanka. IEEE.

Radha Mogla, C. Vasantha Lakshmi, and Niladri Chatterjee. 2021. A Systematic Approach for English-Hindi Parallel Database Creation for Transliteration of General Domain English Words. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 1–5, Kuala Lumpur, Malaysia. IEEE.

Anuja Dilrukshi Herath Herath Mudiyanselage and T. G. Deshan K. Sumanathilaka. 2024. TAM: Shorthand Romanized Tamil to Tamil Reverse Transliteration Using Novel Hybrid Approach. *The International Journal on Advances in ICT for Emerging Regions*, 17(1).

Muhammad Naeem, Abu Bakar Siddique, Raja Hashim Ali, Usama Arshad, Zain ul Abideen, Talha Ali Khan, Muhammad Huzaifa Shah, Ali Zeeshan Ijaz, and Nisar Ali. 2023. Performance Evaluation of Popular Deep Neural Networks for Neural Machine Translation. In *2023 International Conference on Frontiers of Information Technology (FIT)*, pages 220–225, Islamabad, Pakistan. IEEE.

Jayashree Nair and Riyaz Ahammed. 2021. English to Indian Language and Back Transliteration with Phonetic Transcription for Computational Linguistics Tools based on Conventional Transliteration Schemes. In *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–6, Erode, India. IEEE.

Rushan Nanayakkara, Thilini Nadungodage, and Randil Pushpananda. 2022. Context Aware Back-Transliteration from English to Sinhala. In *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 051–056, Colombo, Sri Lanka. IEEE.

L. N. A. S. H. Nissanka, B. H. R. Pushpananda, and A. R. Weerasinghe. 2020. Exploring Neural Machine Translation for Sinhala-Tamil Languages Pair. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 202–207, Colombo, Sri Lanka. IEEE.

A. Nugaliyadde, J.K. Joseph, W.M.T. Chathurika, and Y. Mallawarachchi. 2019. Evolutionary Algorithm for Sinhala to English Translation. In *2019 National Information Technology Conference (NITC)*, pages 26–30, Colombo, Sri Lanka. IEEE.

Santanu Pal and Marcos Zampieri. 2020. Neural Machine Translation for Similar Languages: The Case of Indo-Aryan Languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 424–429, Online. Association for Computational Linguistics.

Dwija Parikh and Thamar Solorio. 2021. Normalization and Back-Transliteration for Code-Switched Data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 119–124, Online. Association for Computational Linguistics.

Nipun Paul, Ishmam Faruki, Mutakabbirul Islam Pranto, Md. Tanvir Rouf Shawon, and Nibir Chandra Mandal. 2023. Bengali-English Neural Machine Translation Using Deep Learning Techniques. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6, Chittagong, Bangladesh. IEEE.

Ashmari Pramodya, Randil Pushpananda, and Ruvan Weerasinghe. 2020. A Comparison of Transformer, Recurrent Neural Networks and SMT in Tamil to Sinhala MT. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 155–160, Colombo, Sri Lanka. IEEE.

H.S. Priyadarshani, M.D.W. Rajapaksha, M.M.S.P. Ranasinghe, K. Sarveswaran, and G.V. Dias. 2019. Statistical Machine Learning for Transliteration: Transliterating names between Sinhala, Tamil and English. In *2019 International Conference on Asian Language Processing (IALP)*, pages 244–249.

Mohammad Masudur Rahman, Md. Faisal Kabir, and Mohammad Nurul Huda. 2018. A Corpus Based N-gram Hybrid Approach of Bengali to English Machine Translation. In *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pages 1–6, Dhaka, Bangladesh. IEEE.

Surangika Ranathunga, Fathima Farhath, Uthayasanker Thayasivam, Sanath Jayasena, and Gihan Dias. 2018. Si-Ta: Machine Translation of Sinhala and Tamil Official Documents. In *2018 National Information Technology Conference (NITC)*, pages 1–6, Colombo. IEEE.

Jaideepsinh K. Raulji, Jatinderkumar R. Saini, Kaushika Pal, and Ketan Kotecha. 2022. A Novel Framework for Sanskrit-Gujarati Symbolic Machine Translation System. *International Journal of Advanced Computer Science and Applications*, 13(4).

Dafni Rose, K. Vijayakumar, D. Kirubakaran, R. Pugalenthi, and Gotti Balayaswantasaichowdary. 2023. Neural Machine Translation Using Attention. In *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, pages 1–7, Chennai, India. IEEE.

Dinidu Sandaruwan, Sagara Sumathipala, and Subha Fernando. 2021. Neural Machine Translation Approach for Singlish to English Translation. *International Journal on Advances in ICT for Emerging Regions (ICTer)*, 14(3):36–42.

Sonia Sarode, Raghav Thatte, Kajal Toshniwal, Jatin Warade, Ranjeet Vasant Bidwe, and Bhushan Zope. 2023. A System for Language Translation using Sequence-to-sequence Learning based Encoder. In *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 1–5, Pune, India. IEEE.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2018. IITP-MT at WAT2018: Transformer-based Multilingual Indic-English Neural Machine Translation System. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation*, pages 1003–1007. Association for Computational Linguistics.

Nandini Sethi, Amita Dev, and Poonam Bansal. 2022. A Bilingual Machine Transliteration System for Sanskrit-English Using Rule-Based Approach. In *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)*, pages 1–5, Delhi, India. IEEE.

Anshumani Sharma, Ayushi Verma, Chetan Shahra, and S. Indu. 2023. Ancient Indian Script Transliteration Using GANs. In *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 621–626, Noida, India. IEEE.

Aditya Shukla, Pragati Agrawal, and Sweta Jain. 2024. Delineating Indic Transliteration: Developing A Robust Model for Accurate Cross-Script Conversion. In *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–6, Bhopal, India. IEEE.

Lahiru de Silva and Supunmali Ahangama. 2021. Singlish to Sinhala Transliteration using Rule-based Approach. In *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*, pages 162–167, Kandy, Sri Lanka. IEEE.

Aryan Singh and Jhalak Bansal. 2021. Neural Machine Transliteration Of Indian Languages. In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, pages 91–96, Chennai, India. IEEE.

Muskaan Singh, Ravinder Kumar, and Inderveer Chana. 2020. Corpus based Machine Translation System with Deep Neural Network for Sanskrit to Hindi Translation. *Procedia Computer Science*, 167:2534–2544.

Shailendra Kumar Singh and Manoj Kumar Sachan. 2019. GRT: Gurmukhi to Roman Transliteration System using Character Mapping and Handcrafted Rules. *International Journal of Innovative Technology and Exploring Engineering*, 8(9):2758–2763.

Shashi Pal Singh, Hemant Darbari, Ajai Kumar, Shikha Jain, and Anu Lohan. 2019. Overview of Neural

Machine Translation for English-Hindi. In *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, pages 1–4, GHAZIABAD, India. IEEE.

Ian Smith and Uthayasanker Thayasivam. 2019. Language Detection in Sinhala-English Code-mixed Data. In *2019 International Conference on Asian Language Processing (IALP)*, pages 228–233, Shanghai, Singapore. IEEE.

Deshan Sumanathilaka, Nicholas Micallef, and Ruvan Weerasinghe. 2024. Swa-Bhasha Dataset: Romanized Sinhala to Sinhala Adhoc Transliteration Corpus. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 189–194, Belihuloya, Sri Lanka. IEEE.

T.G.D.K. Sumanathilaka, Ruvan Weerasinghe, and Y.H.P.P. Priyadarshana. 2023. Swa-Bhasha: Romanized Sinhala to Sinhala Reverse Transliteration using a Hybrid Approach. In *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, pages 136–141, Belihuloya, Sri Lanka. IEEE.

Pasindu Tennage, Achini Herath, Malith Thilakarathne, Prabath Sandaruwan, and Surangika Ranathunga. 2018. Transliteration and Byte Pair Encoding to Improve Tamil to Sinhala Neural Machine Translation. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 390–395, Moratuwa. IEEE.

Pasindu Tennage, Prabath Sandaruwan, Malith Thilakarathne, Achini Herath, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. 2017. Neural machine translation for sinhala and tamil languages. In *2017 International Conference on Asian Language Processing (IALP)*, pages 189–192, Singapore. IEEE.

Sarubi Thillainathan, Surangika Ranathunga, and Sanath Jayasena. 2021. Fine-Tuning Self-Supervised Multilingual Sequence-To-Sequence Models for Extremely Low-Resource NMT. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 432–437, Moratuwa, Sri Lanka. IEEE.

Gaurav Tiwari, Arushi Sharma, Aman Sahotra, and Rajiv Kapoor. 2020. English-Hindi Neural Machine Translation-LSTM Seq2Seq and ConvS2S. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 871–875, Chennai, India. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint*. ArXiv:1706.03762 [cs].

A.J. Vidanaralage, A.U. Illangakoon, S.Y. Sumanaweera, C. Pavithra, and S. Thelijjagoda. 2018. Sinhala Language Decoder. In *2018 National Information Technology Conference (NITC)*, pages 1–5, Colombo. IEEE.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in Machine Translation. *Engineering*, 18:143–153.

Abhinav Y. Watve and Madhuri A. Bhalekar. 2023. English to Hindi Translation using Transformer. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pages 993–1000, Uttarakhand, India. IEEE.

Mahima Yadav and Ishan Kumar. 2023. Transliteration from Hindi to English Using Image Processing. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–6, Hubli, India. IEEE.