Improving Multilingual Neural Machine Translation System for Indic Languages

SUDHANSU BALA DAS, National Institute of Technology (NIT), Rourkela, India
ATHARV BIRADAR, Pune Institute of Computer Technology (PICT), India
TAPAS KUMAR MISHRA, National Institute of Technology (NIT), Rourkela, India
BIDYUT KR. PATRA, Indian Institute of Technology (IIT), Varanasi, India

The Machine Translation System (MTS) serves as an effective tool for communication by translating text or speech from one language to another language. Recently, neural machine translation (NMT) has become popular for its performance and cost-effectiveness. However, NMT systems are restricted in translating low-resource languages as a huge quantity of data is required to learn useful mappings across languages. The need for an efficient translation system becomes obvious in a large multilingual environment like India. Indic languages (ILs) are still treated as low-resource languages due to unavailability of corpora. In order to address such an asymmetric nature, the multilingual neural machine translation (MNMT) system evolves as an ideal approach in this direction. The MNMT converts many languages using a single model, which is extremely useful in terms of training process and lowering online maintenance costs. It is also helpful for improving low-resource translation. In this article, we propose an MNMT system to address the issues related to low-resource language translation. Our model comprises two MNMT systems, i.e., for English-Indic (one-to-many) and for Indic-English (many-to-one) with a shared encoder-decoder containing 15 language pairs (30 translation directions). Since most of IL pairs have a scanty amount of parallel corpora, not sufficient for training any machine translation model, we explore various augmentation strategies to improve overall translation quality through the proposed model. A state-of-the-art transformer architecture is used to realize the proposed model. In addition, the article addresses the use of language relationships (in terms of dialect, script, etc.), particularly about the role of high-resource languages of the same family in boosting the performance of low-resource languages. Moreover, the experimental results also show the advantage of back-translation and domain adaptation for ILs to enhance the translation quality of both source and target languages. Using all these key approaches, our proposed model emerges to be more efficient than the baseline model in terms of evaluation metrics, i.e., BLEU (BiLingual Evaluation Understudy) score for a set of ILs.

CCS Concepts: • Computing Methodologies → Machine Translation System;

Additional Key Words and Phrases: Multilingual neural machine translation system (MNMT), Indic languages (ILs), low resource language, corpus, BLEU score

This work is partially funded by Meity (Ministry of Electronics and Information Technology, Government Of India) for project sanction no. 13 (12)/2020-CC & BT dated 24.04.2020.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2375-4699/2023/06-ART169 $15.00
https://doi.org/10.1145/3587932

ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 22, No. 6, Article 169. Publication date: June 2023.
1 INTRODUCTION

Language is the base of communication among humans, and numerous languages persist around
the globe that differ from region to region. Interestingly, every language has its own lexical inter-
pretation with a vast vocabulary (words and phrases) and rules (grammar) that differ from others.
Except in some special cases, it is not possible for an individual to be familiar with more than
one language. For this reason, it becomes difficult to understand an unfamiliar language without
proper translation/interpretation, which is generally done by a third party, i.e., human. Due to fac-
tors like affordability, dependency, human error, time, and others in case of a human interpreter,
machine translation has been the preferred choice[22]. India is home to numerous ancient and
morphologically rich languages used distinctly over various regions. Apart from having a set of di-
ialects and accents, every regional language has its own identity and use. On the contrary, English
is the most common language across the globe as a medium of sharing information among citi-
zens not only for administrative works but also for exchanging sentiments, emotions, ideas, and
actions over global media (social platform). In the vast sector of Information Technology, English
is preferred over other natural languages since the information provided in the English language
by the standard ASCII symbols is relatively simple for computers to process[59]. However, those
who are less conversant with English find it difficult to cope with and often need proper transla-
tion/interpretation for clarity at every step. In order to bridge the gap, a machine-based translation
is a perfect approach that can achieve all these tasks with little human involvement.

Machine Translation (MT) generally refers to an independent process of translating, primar-
ily through a computer application, from one language as source/input to another language(s)
as target(s)/output. With the advent of natural language processing (NLP), a component of
artificial intelligence (AI), computers detect and sense the intent of input language and translate
accurately as per the format of output language. The MT process is very effective in terms of time
(speed), volume, and cost[3]. In consideration with the Indian environment, as discussed earlier,
the development of a quality machine translation system (MTS) for the Indian languages
(ILs) is in huge demand. Still, it remains a challenging task, since many ILs are individually low
in terms of resources, resulting in an adverse impact on their translation quality. Therefore, trans-
lation patterns, both from English to IL and from IL to English, face problems in morphological as
well as structural analysis. Even the variation in word orders and the dissimilarity in the sentence
size (of the source and target) create an issue in the word alignment and reduce the translation
quality. However, recent research proposes that proper usage of the parallel and monolingual
corpus increases the translation quality for low-resource languages. The Neural Machine
Translation (NMT) system[5], [68] has gained popularity and shown better results than the
Statistical Machine Translation (SMT) [32] system. Although a traditional NMT model can successfully translate between a single language pair (e.g., Hindi and English), training a discrete
model for every language pair is impractical due to a large number of languages spoken worldwide.
This problem is resolved with the recent development of the multilingual neural machine
translation (MNMT) model through various approaches. One such approach is transfer-learning,
where a model is first trained on a high-resource language pair and then parameter values are trans-
scribed from that model and further fine-tuned on low-resource data [47]. This technique helps
to increase the amount of corpus as well as the translation quality for low-resource languages. Hence, the main objective of the MNMT system is to utilize all the corpora of different language pairs and make one model that can give a qualitative translation [9]. Recent research works on MNMT evoke promising methods for improving quality with low-resource languages trained in a multilingual setting. It also works well with translation between languages that do not have any parallel corpus at the time of training, i.e., zero-shot translation [9]. Both the source and target languages are included in the training and the MNMT system can translate between unknown pairs automatically without any manual intervention in this mode of translation. In case of low-resource languages, lexical and orthographic relationships among languages may be used to make the quality of translation better [36]. The links made between words are known as lexical relationships, whereas orthographic relationships address punctuation rules and the link between written and spoken language.

In the domain of ILs being restricted with low resources, the use of MNMT, with some associated strategies, is preferred to generate an ideal model for obtaining a qualitative translation system. Applying a string of procedures over a transformer-based [68] MNMT model makes this possible, which is thoroughly explained in this article. We propose to build an MNMT system for 15 languages by utilizing an entire parallel as well as monolingual corpus [54]. The translation outputs are also reviewed using automatic evaluation metrics, i.e., Bilingual Evaluation Understudy (BLEU) score [51].

1.1 Motivation

MT systems have significantly improved in recent years, but only on a few language pairs with abundant parallel data [22]. Out of nearly 19,500 languages spoken in India, MT systems have been built for a few ILs to date [26]. Many of the ILs are low-resource languages, with insufficient data available. It has been discovered that knowledge transfer from high-resource languages benefits low-resource languages [53]. Multilingual MT gains from the knowledge transfer caused by the fusion of various languages. Eventually, it motivates us to explore an ideal MTS for ILs, with optimum quality, employing state-of-the-art neural architecture through a novel approach. For this, MNMT with the transformer model [68] is utilized with extensive training for 15 ILs with English (EN) in both directions. Various processes like noise reduction, language similarity, back-translation, and domain adaptation are also incorporated toward a qualitative output.

1.2 Our Contribution

Our novel approach toward realization of a qualitative MT system for ILs to English (and back) comprises various evaluation processes with satisfactory test results. Moreover, in the course of this research work, the following major contributions have been achieved:

1. As a first attempt ever, to the best of our knowledge, this novel work explores the MT of 15 EN-IL and IL-EN pairs (both directions), including both the Dravidian language and Indo-Aryan groups, in a multilingual environment using the cutting-edge transformer architecture, and assesses their effectiveness.
2. Apart from trials over the linguistic features commonly occurring in the ILs, different data filtration techniques are explored to clean the data for betterment of translation quality.
3. We check the language relationship approach in the MT system, which plays a major role in Indic languages. It is noticed that language mix-up techniques (based on IL similarity, i.e., script, dialects, etc.) help the low-resource languages to achieve better quality translation.
4. We also examine the effectiveness of back-translation and the domain adaptation technique in achieving better results in translating both low-resource and high-resource ILs.
This article establishes the performance superiority of our model through the results in comparison with the fine-tuned OPUS-MT (pretrained) model for the majority of IL pairs using the PMI corpus.

This article is organized as follows. In Section 2, we concisely illustrate a few prominent works on MNMT systems. Section 3 talks about the approach used. Section 4 explains the model’s architecture, details about the corpus, pre-processing steps, linguistic features of ILs, and the procedure of training the model. Section 5 explains the model overview and all subsequent training approaches. In Section 6, the performance of fine-tuning the OPUS model with our model is compared. Results are shown in Section 7, followed by the conclusion in Section 8.

2 RELATED WORK

The first MT study was undertaken in the 1950s, and since then, a significant amount of work has been reported on MT [25]. The MT system initially determines the translation of a text in the source language by matching words and their meanings in the source language to those in the target language with a set of rules. Researchers utilize a variety of methodologies, including rule-based [17], corpus-based [11, 60], and hybrid-based techniques [55]. There are advantages and disadvantages to each strategy. Due to their inability to capture the diverse sentence structures found in the language, these approaches could not produce satisfactory results. This lengthy translation process also calls for individuals who are fluent in both languages. In order to address the shortcomings of rule-based systems, corpus-based translation approaches such as SMT [32] and NMT [5, 7, 63] have been developed. In SMT [62], the preprocessed data are examined with statistical metrics to produce the desired outcome in language processing tasks. This method searches for statistical relationships in pre-processed data (such as probability, distance metric, etc.). When translating a document, the probability distribution function $P(m/n)$ is used as the basis for translation. The probability of converting a sentence $n$ from the source language $N$ (e.g., Hindi) to the sentence $m$ target language $M$ (e.g., Telugu) is represented by the $P(m/n)$. Due to the scarcity of high-quality parallel corpora, a lot of research has been done over machine translation from English to IL, primarily relying on rule-based techniques. Significant efforts have been done on using statistical and hybrid approaches to translate text from English to IL(s) despite the lack of appropriate parallel corpora. In SMT tasks, the outcomes are not noticeable. NMT, on the other hand, is a new technique that has significantly improved translation outcomes but can be used for one-to-one translation languages. Most NMT systems are supervised deep learning systems, which are extremely data-hungry. Despite years of research, only a small number of languages spoken worldwide have access to high-quality annotated MT resources [50]. So, one of the major issues that arises from MT’s multilingualism and linguistic diversity is data scarcity. As per Arivazhagan et al. [4] and Lample et al. [37], the types of languages utilized to train the model have a significant impact on the effectiveness of the supervised MNMT models. MNMT systems have the capability to train a single model for multiple language pairs at a time. The first MNMT [14] is based on a Multiple-task learning framework used for translating in a one-to-many language direction where for target languages, different language-dependent decoders and attention mechanisms were applied. Their approach to one-to-many language showed better results than individual language translation. Then, a many-to-many language model for the multilingual machine translation system was proposed by Firat et al. [16]. Their model is based on a shared attention mechanism with many encoders and decoders. In case of enhancing the quality of low-resource languages, Aharoni et al. [2] added a transfer interference tradeoff and found that it is more efficient in a many-to-one (English) direction. To boost the performance of the MNMT models, there are different ways that work significantly with the training method as well as with the model architecture of Wang.
et al. [70], Aharoni et al. [2], and Lin et al. [40]. Approaches on augmentation of the corpus (both parallel and monolingual) like back-translation, transliteration, etc., for low-resource languages enhance the quality of translation in the MNMT model [49]. A brief summary of all the techniques, architecture selection, and other parameters relevant to MNMT as well as issues related to it was described by Dabre et al. [9]. In recent scenarios [15], pre-trained language models and multiple languages have proved favorable for the MNMT system. In case of low-resource languages like IL, many researchers [1], [61] have trained and tested MNMT models on IL corpora from online websites in different domains. An extensive MNMT model that can work with 102 languages was proposed by Aharoni et al. [2], which focuses on training models on the corpora of many language pairs, with English as a source or target language. Even with incredible progress of MNMT, it is sensitive to the noise in the corpus [6]. Hence, various filtering techniques have gained popularity and have become crucial to clean the corpus and to remove the unwanted contents, symbols, etc. Liu et al. [41], Pinnis et al. [52], and Li et al. [39] have experimented with different filtering techniques with parallel corpora and then trained multilingual models to achieve progressive results. mBART [42] is the first pre-trained multilingual model based on sequence-to-sequence architecture. In the mBart model, filtered corpora of different languages (from noises) are used to train the model for multiple languages, achieving outstanding results in terms of BLEU score. This shows the importance of noise removal for a substantially enhanced performance of both supervised and unsupervised MT, applicable to the both sentence level and document level. So, lots of techniques, approaches, and methods have been used by researchers to achieve translation quality.

2.1 Background

This section gives background information with particular emphasis on traditional bilingual NMT and MNMT systems.

2.1.1 Neural Machine Translation (NMT) System. A radical improvement over earlier MT techniques is NMT. In addition to embracing the probabilistic framework, NMT offers a data-driven approach to MT [9]. Provided a parallel dataset C, the NMT reduces the translation task into finding the probability distribution \( p \) of a target language \( b \) given source language \( a \), given by Equation (1):

\[
p(b \mid a; \omega) = \prod_{i=1}^{y} p\left(b_i \mid b_{(i-1),...1}, a; \omega\right),
\]

where \( a = a_1, ..., a_x \) is an input source language sentence of \( x \) words, whereas the translated sentence that is the target language of \( y \) words is \( b \), where \( b = b_1, ..., b_y \), \( \omega \) is the parameter to be learned, \( b_j \) is the presently produced word, and \( b_{(i-1),...1} \) are the previously created words. The log-likelihood \( S \) with respect to the parameter set \( \omega \) is maximized during training of model of Equation (1) by Equation (2):

\[
S(\omega) = \sum_{(a,b) \in C} \log p(b \mid a; \omega).
\]

In contrast to the early studies on NMT that focused on developing translation systems between bilinguals, nowadays researchers find that the NMT framework can naturally include numerous languages. This approach has demonstrated cutting-edge performance for different language pairs. As a result, research on Multilingual MT systems has significantly increased [9].

2.1.2 Multilingual Neural Machine Translation (MNMT) System. The MNMT system is capable of translating between multiple language pairs [9]. It significantly increases the translation quality of low-resource languages. The low-resource languages learn additional information by training on high-resource languages. MNMT training can be done in a variety of ways:
Transfer Based [72]: Translation knowledge transfer learning uses learned attributes from well-resourced language pairs to train related language pairs with fewer resources.

Pivot Based [8, 44, 67]: When direct parallel data between the source and target language are unavailable, the pivot-based system uses a pivot language \( p \) to link the translation between the two.

Multiway Based [9]: Using parallel corpora for several language pairs, the aim of multiway translation is to build a single NMT system for many-to-one, one-to-many, or many-to-many translation.

MNMT’s main objective is to create a model that facilitates translation between multiple language pairs, i.e., any source- and target-language pairs, i.e., \( N_{(\text{source})m}, N_{(\text{target})m} \), respectively. The goal of multiway NMT training is to maximize the log-likelihood \( S(\omega) \) (where \( \omega \) is the parameter to be learned) of training data taken collectively for all language pairs (different weights may be assigned to the likelihoods of different language pairs) as shown in Equation (3):

\[
S(\omega) = \frac{1}{M} \sum_{m=1}^{M} S^{N_{(\text{source})m},N_{(\text{target})m}}(\omega),
\]

where \( S^{N_{(\text{source})m},N_{(\text{target})m}}(\omega) \) is the loss of individual language pairings calculated by Equation (2), \( m \) is the currently used language pair, and \( M \) is the total number of acceptable language pairs.

Moreover, MNMT may be classified into three types depending on the alignment of the source and target languages, namely (1) Many to Many [16]: translation among multiple source and target languages is feasible in this category, such as the translation of European languages into one another in a single model; (2) One to Many [14]: a single source language is translated using the MNMT approach into numerous target languages, for instance, the translation of English into other European languages in a single model; and (3) Many to One [38]: the model is trained to translate from various source languages into a single target language in this context, like the translation of other European languages to English in a single model. The next section describes the approaches we have used to build our MT system for 15 EN-IL and IL-EN pairs i.e., 30 directions.

3 APPROACH USED

Previous studies on NMT focus on building several encoders and decoders for every source/target language. As per cutting-edge research by Johnson et al. [27], employ a single MNMT model for translation between multiple language pairs. It also affirms that mixing low-resource languages with high-resource languages significantly improves translation effectiveness on low-resource bi-texts. In case of Indic-to-English language pairs, we use a single shared encoder and one decoder on the target language (English), which we need to translate for all source languages. For the English-to-Indic translation task, we use a single encoder for the source (English) and a shared decoder for target languages. For example: In case of multilingual NMT, the translation of the English-to-Hindi sentence pair: \(<\text{hi}>\text{Where are you?}\>\) can be written as \(<\text{hi}>\text{आप कहाँ हैं?}\>\). Figure 1 describes the MNMT model for the one-to-many and many-to-one MNMT system.

4 PROCESSING OF INDIC LANGUAGES FOR TRANSLATION

This section illustrates the architecture details of the system, corpus/dataset, data preprocessing and filtering steps, and model training details.

4.1 Architecture

The proposed MT system has been realized with transformer architecture as described by Vaswani et al. [68]. In the architecture at each step it assigns a self-attention mechanism, which passes the
Fig. 1. One-to-many and many-to-one MNMT system.

Fig. 2. Transformer architecture [68].

information among the encoders and decoders efficiently, as shown in Figure 2. Indeed, our system is designed in consideration of the following basic advantages of transformer architecture over other NMT architectures in a low-resource MNMT scenario. Transformer models are highly parallelizable, which makes them incredibly compute-optimal and enables to train extremely big models. In order to establish dependency between input and output, transformer architecture relies on self-attention instead of recurrence and convolution. Such mechanism provides transformer architecture direct access to data from every step, unlike sequential models [71]. Apart from having a traditional six-encoder-decoder architecture, the transformer primarily relies on attention layers to translate source sentences into target sentences. It employs stacked encoders, each of which has two sublayers: a feed-forward neural network (FFNN) and a self-attention layer to learn the representations of source sentences. The importance of each token in the self-attention layer is
Table 1. Parallel and Monolingual Corpus Statistics

| Language (to English) | Parallel Corpus (Sentences) | Monolingual Corpus (Sentences) |
|----------------------|----------------------------|--------------------------------|
| Bengali (bn)         | 8.52M                      | 39.9M                          |
| Oriya (or)           | 1.00M                      | 6.94M                          |
| Gujarati (gu)        | 3.05M                      | 41.1M                          |
| Kannada (kn)         | 4.07M                      | 53.3M                          |
| Marathi (mr)         | 3.32M                      | 34.0M                          |
| Hindi (hn)           | 8.56M                      | 63.1M                          |
| Malayalam (ml)       | 5.85M                      | 50.2M                          |
| Telugu (te)          | 4.82M                      | 47.9M                          |
| Punjabi (pa)         | 2.42M                      | 29.2M                          |
| Assamese (as)        | 0.14M                      | 1.39M                          |
| Tamil (ta)           | 5.16M                      | 31.5M                          |
| Urdu (ur)            | 6.1M                       | NA                             |
| Nepali (ne)          | 0.7M                       | NA                             |
| Sinhala (si)         | 6.3M                       | NA                             |
| Sindhi (sd)          | 1.7M                       | NA                             |

learned by analyzing the other appropriate tokens in the sentence. The decoder focuses on particular features in the source sentence during translation. Since a transformer model can handle all the data concurrently, both preceding and succeeding elements are handled at the same time. This leads to less processing time and more effective training (prime requirement for a low-resource language). For our novel method of realizing a qualitative MTs toward ILs-English (in both directions) using transformer architecture, the complete assignment is done employing six encoder-decoder layers with eight attention heads, 2,048-layer FFNN, and 512 embedding dimensions.

4.2 Dataset

The Samanantar Corpus [54] is used to train our MNMT system. This corpus includes more than 40 million sentence pairs between English to ILs from where Assamese (as), Malayalam (ml), Bengali (bn), Marathi (mr), Gujarati (gu), Kannada (kn), Hindi (hn), Oriya (or), Punjabi (pa), Telugu (te), and Tamil (ta) are used. The parallel data included in the corpus are collected from various sources such as PMIndia, UFAL EnTam, IITB 3.0, Uka Tarsadia, JW, NLPC, OpenSubtitles, Bibleuedin, Wiki Titles, MTEnglish2Odia, WikiMatrix, OdiEnCorp 2.0, CVIT-PIB, and TED [46]. For Nepali (ne), Sindhi (sd), Sinhala (si), and Urdu (ur) languages, we use the OPUS corpus [64]. For testing purposes, Flores101 [18] and FLoRes [21] datasets (only for Nepali and Sinhala) are used. No other corpus or any other resources related to linguistics were used for our experiments. Table 1 gives full detail about both the monolingual [29] and parallel [54] corpus, where NA represents Not Available.

4.3 Linguistic Features of Indian Languages [11, 12]

The ILs pose difficulties for MT and NLP tasks as they exhibit rich morphology in nature [69]. The word order is a key structural distinction between English and ILs [33]. While most ILs primarily utilize subject-object-verb (SOV), English uses the subject-verb-object (SVO) sequence. Some of the linguistic features of ILs are:

- Duplication: This refers to a phenomenon where a word is used in repetition to convey a variety of speech acts, including intensity, plurality, emphasis, and so forth. For example, the word “at village village” in a Hindi sentence गाँव गाँव में is repeated to imply plurality, i.e., across villages in English. Therefore, effort is required for the translation process in order
Improving Multilingual Neural Machine Translation System for Indic Languages

4.4 Transliteration for Similar Languages

The purpose of transliteration is to retain as much of the original pronunciation of the source word as possible while adhering to the phonological structures of the target language. For example, the sentence in Hindi (Devanagari) हम किताबें पढ़ते हैं is transliterated as “hum kitaaben padhte hain” in English. A transliteration system T accepts a source word S and returns a ranked list R with \((M_i, K_i)\) tuples as its elements. \(M_i\) is the \(i\)th transliteration of the source word S obtained with the \(i\)th highest likelihood \(K_i\) in each tuple. There are numerous scripts used for Indian languages. Transliteration between scripts of related languages can improve the quality of multilingual models, as described by Haddow et al. [23] and Goyal and Sharma [20]. This approach is used in similar languages determined based on the similarity in the languages [39] such as their dialects, script, word order, and so forth. We have used the related-language transliteration.
technique for the languages that fall into a similar group. For transliteration, we utilized the Indic NLP Library [35]. For instance, Odia, Bengali, and Assamese are the languages that have similarities. As compared to Oriya and Assamese, the Bengali language is high in the corpus. So, Bengali training data is transliterated to Oriya and Assamese and then gets enhanced using training data of the Oriya and Assamese languages. The training data for low-resource languages has been supplemented with relatively high-resource-related language training data that have been transliterated into the low-resource language. Similarly, Gujarati and Punjabi, which are low-resource languages, are similar to Hindi. Therefore, Hindi training data is transliterated to Gujarati and Punjabi and then added to training data of Gujarati and Punjabi.

4.5 Data Preprocessing

All of our experiments are done using Byte Pair Encoding (BPE), which is an efficient technique for data division, i.e., splitting up the words into sub-words [58]. BPE works very well for Indian languages having morphological richness features. After normalization and pre-tokenization are complete, BPE training begins by computing the distinct set of words used in the corpus. The vocabulary is then built by utilizing all of the symbols used to represent those words. BPE has the advantage of making UNKs obsolete. In MT tasks, the UNK symbol denotes the words that are not present in the vocabulary. Before learning the BPE codes, the data from all 15 Indic languages is merged for training the one-to-many and many-to-one models. The BPE codes of the one-to-many and many-to-one models are learned using 48,000 and 6,400 merge operations, respectively.

4.6 Tokenization

The initial step in any machine translation application involves tokenizing the raw text, which involves converting the given text into lexical units, which are the most fundamental components [13]. Each lexical unit is designated as a token after tokenization. Depending upon the requirement, tokenization may occur at the phrase or word level. Three different types of tokenization are (1) sentence-level tokenization, (2) word-level tokenization, and (3) subword-level tokenization, as shown in Figure 3. Tokenization at sentence level addresses issues like ambiguity, word sequences, and the detection of sentence endings, whereas in word-level tokenization, words serve as lexical units. The entire document is tokenized to a set of words as applications such as language processing and text processing frequently use the word-level tokenization. The n-gram tokenization is an n-words token, where “n” is the lexical units of the total number of words. For instance, “n” is a unigram when “n” equals 1, it is a diagram when it equals 2, it becomes a trigram when it equals 3, and so on. We use a sentence piece tokenizer to take advantage of the morphological richness attribute of ILs. SentencePiece [34] enhances direct training using raw sentences to include sub-word units and a unigram language model. We create a full end-to-end system using SentencePiece that is independent of language-specific processing (both pre and post). The
Table 2. Filtered Parallel Corpora Statistics

| EN to Indic | Filtered | After Filtration (Sentences) |
|------------|----------|-----------------------------|
| en-bn      | 11.53%   | 7,537,644                   |
| en-or      | 3.12%    | 968,800                     |
| en-gu      | 6.87%    | 2,840,465                   |
| en-kn      | 6.92%    | 37,788,356                  |
| en-mr      | 6.44%    | 3,106,192                   |
| en-hi      | 8.21%    | 7,857,224                   |
| en-ml      | 8.71%    | 5,340,465                   |
| en-te      | 7.45%    | 4,460,910                   |
| en-pa      | 3.86%    | 2,326,588                   |
| en-as      | 4.24%    | 132,487                     |
| en-ta      | 7.62%    | 4,766,808                   |
| en-ur      | 9.46%    | 5,522,940                   |
| en-ne      | 2.56%    | 682,080                     |
| en-si      | 9.72%    | 5,687,640                   |
| en-sd      | 5.03%    | 1,614,490                   |

SentencePiece tokenizer combined with a 48K-word vocabulary of 15 target ILs along with English has been used for En-XX Multilingual Translation, wherein a character coverage of 1.0 has been utilized. For XX-En, the Sentencepiece tokenizer has been employed using a combined vocabulary of 64K words from the 15 source Indic languages with 1.0 character coverage.

4.7 Data Filtering

Currently there is an increasing number of parallel corpora accessible for MT training. However, we can never be certain about data quality when obtaining corpora from different sources, which is crucial for an MT system’s effectiveness. A short scan of the Samantar corpus, generally used for translation task (for sentence pairings that are not noise-free), like the other major corpus used for MT tasks, provides an overview of the noise and further helps in the application of a set of heuristics to remove a good amount of noise from those sentence pairings. Many methods/experiments have been done to enhance and filter the Samantar corpus. Some of the methods used are:

- Removing sentence pairings with an empty source or destination language
- Removing low-likelihood sentence pairs (based on factors like sentence lengths, identified languages, etc.)
- Removing sentences containing characters from a certain language pairs’ Unicode range

Table 2 describes the filtered parallel corpora statistics. After all Indian scripts get standardized, tokenized, and transliterated using the Indic NLP library, the experiment is carried out using both methods: (1) by using a noise-free corpus and (2) without removing the noisy corpus. It becomes clear that none of these methods used on the parallel corpus lead to a significant quality improvement. Hence, the filtered corpus is preferred for the translation task using different approaches.

5 MODEL OVERVIEW

An efficient strategy to increase the number of training instances for MT is data augmentation, which is becoming a standard procedure for MTs with limited resources. In the next subsection, different techniques used in our experiments to check the quality of translation are described.
Table 3. Primal Model Training Results with Evaluation Metrics as BLEU Score

| EN->Indic | Baseline | Proposed Primal Model | Indic->EN | Baseline | Proposed Primal Model |
|-----------|----------|-----------------------|----------|----------|-----------------------|
| en-bn     | 2.00     | 27.4                  | bn-en    | 4.50     | 30.2                  |
| en-or     | 21.2     |                       | or-en    |          | 33.8                  |
| en-gu     | 29.9     |                       | gu-en    |          | 35.3                  |
| en-kn     | 26.1     |                       | kn-en    |          | 37.0                  |
| en-mr     | 21.9     |                       | mr-en    |          | 36.7                  |
| en-hi     | 39.9     |                       | hi-en    |          | 44.2                  |
| en-ml     | 20.9     |                       | ml-en    |          | 37.4                  |
| en-te     | 28.0     |                       | te-en    |          | 43.3                  |
| en-pa     | 35.9     |                       | pa-en    |          | 40.6                  |
| en-as     | 10.20    | 16.6                  | as-en    | 15.50    | 27.3                  |
| en-ta     | 18.9     |                       | ta-en    |          | 41.9                  |
| en-ur     | 19.60    | 28.7                  | ur-en    | 20.50    | 26.9                  |
| en-ne     | 6.6      |                       | ne-en    | 8.00     | 14.0                  |
| en-si     | 9.50     | 15.3                  | si-en    | 8.20     | 15.7                  |
| en-sd     | 6.30     | 15.5                  | sd-en    | 15.40    | 18.5                  |

5.1 Model Primal Training

In order to train the system, two distinct MNMT models are constructed: (1) one (English)-to-many (15 IL) model and (2) many (15 IL)-to-one (English) model. The transformer model [68] is employed in our one-to-many approach, with a single shared encoder and decoder. The decoder employs a shared vocabulary of all the Indic languages, whereas the encoder uses the English language’s vocabulary. For the many-to-one model, the transformer architecture is used with a single decoder and a single shared encoder. Here, the shared vocabulary of all the ILs has been employed for the encoder, and English vocabulary is used for the decoder. A specific token is added to the input language for both of these MNMT models before the phrase. For the implementation of the multilingual system, the fairseq [48] library is preferred. Adam [31] optimizer is utilized for training with betas of (0.9,0.98). With 8,000 warm-up updates and an initial learning rate of 5e-0.4, the inverse square root learning rate scheduler has been employed. The criterion used is label smoothed cross-entropy with a label smoothing of 0.1, and the dropout probability value has been set to 0.6. For the multilingual models, we prefer to adopt an update frequency of 15 and apply the beam search algorithm during the decoding process, with a beam length of 20. The one-to-many model has been trained for 12 epochs, whereas the many-to-one model has received 13 epochs. Based on the correctness of the validation set, all of our models have been trained with early stopping criteria. During testing, all the translated BPE segments are reassembled, which are changed back to the original language scripts. Finally, BLEU is used to assess the precision of our translation models [51]. Results of the primal training model are compared with the baseline in Table 3 [10]. As shown in Table 3, the majority of languages, those that are used in our experiment being newly added to the dataset, have no baseline in the Evaluation Result Server of WAT 2022 [45].

5.2 Different Approaches

As discussed earlier, ILs have similarities, being inherited from Aryan and Dravidian languages. Based on this, ILs are categorized into two groups as shown in Table 4. Hindi, Urdu, Punjabi, Gujarati, Marathi, Oriya, Bengali, and Sindhi belong to Group A, whereas Telugu, Tamil, Kannada, and Malayalam are in Group B [56]. Combined use of languages with lexical and structural
Table 4. Similar Languages

| Group | Languages Belong |
|-------|------------------|
| A     | Hindi, Urdu, Punjabi, Gujarati, Marathi, Oriya, Bengali, Sindhi |
| B     | Telugu, Tamil, Kannada, Malayalam |

similarities help the MNMT system for effective translation. In this regard, various enhancement approaches are used viz. back-translation and domain adaptation, which are discussed further in subsequent subsections. The training of these approaches is done by restoring the best checkpoint of our primal model.

5.2.1 Exploiting Language Relationships. ILs have a rich history of scripts. These scripts are descended from the historic Brahmi script. Hindi, Bengali, Tamil, and Telugu are among the major languages that use Brahmi scripts. However, several languages use Arabic script. Both Arabic and Brahmi-derived scripts are used in Punjabi and Sindhi. Additionally, assuming that the populations of surrounding regions frequently mix, languages of such areas also exhibit some degree of resemblance. Hence, the effect of the language mix-up technique has been considered in increasing precision by training only those languages that have some similarities among one another [39]. To apply this, methods suggested by Goyal et al. [19] are followed; for example, all Indo Aryan languages (like Hindi, Bengali, Gujarati, Marathi, etc.) have been trained at a time to translate them to English (i.e., many-to-one) as well as vice versa (i.e., English to Indo Aryan languages). Similarly, for languages of Dravidian origin (Tamil, Kannada, Telugu, and Malayalam), we train all languages in both the directions to check the effect of the language relationship. One more approach is also implemented by reducing the number of languages being compensated with similar languages for training the model to check the effect. For example, Hindi and Marathi are quite similar in their writing style, and even Gujarati grammar also shares several similarities with Hindi. Hence, Group A languages have been trained together to check the translation quality, as shown in Table 4. The same method has been applied for Kannada and Malayalam from Group B, which are similar. The model for these languages has been trained to generate translations for English and vice versa. With completion of training of both Group A and B languages, an average increment of 1.6 in BLEU score has been noticed using both the techniques, classifying the languages according to their families in terms of origin and similarities.

5.2.2 Domain Adaptation. This is the process of changing a model, previously trained on any common domain, to a new domain. It is used in multi-domain problems, where systems are modified with addition of tags at the sentence/word level to provide the system with more meta-information, allowing it to produce translations using vocabulary and style that are acceptable for the domain. In order to adapt an MNMT model to a different domain, the model is typically trained on the entire parallel corpus before fine-tuning its parameters on a smaller in-domain corpus [43, 72]. The same model configuration as described in Section 5.1 is used for implementing the domain adaptation technique, emphasizing domains such as PMI, CVIT, DD-National [54], and Anuvaad. A minor improvement of 0.86 in average BLEU score is achieved using this approach for language translation in both the directions. Algorithm 1 describes our steps for domain adaptation.

5.2.3 Back-translation (BT). The statistics of massive parallel corpora, or collections of related phrases in both the source and target languages, are what machine translation relies on. Bitext has several limitations, whereas there is a lot more data available in Monolingual. Here, Bitexts are a simple depiction of the source text and the translated version, whereas Monolingual data has texts for only one language. Language models have historically been trained using monolingual data, which increased the accuracy of NMT systems. A typical method for enhancing machine
Algorithm 1: Domain Adaptation Algorithm

**Input:** Domain 1-Parallel Corpus $C_{p1}$, Domain 2-Parallel Corpus $C_{p2}$, Target corpus $C_t$

- **Let:** No. of epoch $= x$
- Train model $\Delta$ on $C_{p1}$ for $x$ epochs
- Generate last checkpoint, $L_c$, for the model $\Delta$
- Train model $\pi$ on $C_{p2}$ for $x$ epochs restoring the checkpoint $L_c$
- Generate best checkpoint $L_M$ for model $\pi$
- Use best checkpoint of $\pi$ to get translation results for $C_t$

**Output:** Domain adapted model $\pi$

Algorithm 2: Back-translation Algorithm

**Input:** Parallel Corpus $C_p$, Monolingual Corpus $C_m$, target corpus $C_t$

1. **Let:** $T_{r\rightarrow} = C_p$
2. **loop**
3. Train $\text{Lang}_1$ to $\text{Lang}_2$ model $\Delta_{\rightarrow}$ on $T_{r\rightarrow}$
4. With $\Delta_{\rightarrow}$, create new corpus $N_{\rightarrow}$ using $C_m$
5. Let $T_{r\rightarrow} = C_p \cup N_{\rightarrow}$
6. Train $\text{Lang}_2$ to $\text{Lang}_1$ model $\Delta_{\leftarrow}$ on $T_{r\rightarrow}$
7. With $\Delta_{\leftarrow}$, create new corpus $N_{\leftarrow}$ using $C_t$
8. Let $T_{r\rightarrow} = C_p \cup N_{\leftarrow}$
9. **until** convergence condition is achieved.

**Output:** optimized & updated models $\Delta_{\rightarrow}$ and $\Delta_{\leftarrow}$

Translation is BT of the monolingual corpus. BT plays a major role in more resource languages. More training data were produced using the BT technique of Sennrich et al. [57] to enhance translation model performance. To produce extra synthetic parallel data from the monolingual target data, it is necessary to train a target-to-source system. This data supplements human bitext to train the intended source-to-target system. BT does not require any change of the MT training algorithms, which makes it simple and straightforward to use. Algorithm 2 describes our BT approach. In order to train NMT systems, we provide an approach of iterative BT, a technique for creating synthetic parallel data that is progressively improved from monolingual data. Our experiments are conducted on both high-resource languages and low-resource languages (according to parallel corpora statistics as shown in Table 1). For these languages, with the available corpora from PMI Domain [24], the monolingual corpora have been subsampled for about twice the size of the parallel training corpus. Bahdanau et al.’s [5] neural machine translation systems are deployed to translate the monolingual data. Our configuration is similar to Sennrich et al. [57]; however, the fast Marian toolkit [28] is used for training. Two different BT model approaches have been applied:

- Training NMT model on parallel corpus (15K iterations stopping after 0.20 epochs (the model stops after training over 20% of batches with batch size of 128))
- 172K iterations where the minimal loss has been achieved, i.e., convergence is achieved

Performance of the 15K iteration model is not satisfactory, and its synthetic parallel corpus does not perform better for back-translated data. Longer trained systems have significantly higher translation quality, and their synthetic parallel corpora are effective. The back-translation system
Table 5. Comparison with Fine-tuned OPUS-MT Model with Evaluation Metrics as BLEU Score

| Language Pair | Our Results | Fine-tuned OPUS Model | Our Results | Fine-tuned OPUS Model |
|---------------|-------------|-----------------------|-------------|-----------------------|
| en-bn         | 27.4        | 4.7                   | 30.2        | 19.1                  |
| en-or         | 21.2        | NR*                   | 33.8        | 13.8                  |
| en-gu         | 29.9        | 4.7                   | 35.3        | 14.3                  |
| en-kn         | 26.1        | 3.4                   | 37.0        | 13.8                  |
| en-mr         | 21.9        | 2.9                   | 36.7        | 14                    |
| en-hi         | 39.9        | 10.5                  | 44.2        | 19.6                  |
| en-ml         | 20.9        | 3                     | 37.4        | 10                    |
| en-te         | 28          | 4.7                   | 43.3        | 15.3                  |
| en-pa         | 35.9        | 7.4                   | 40.6        | 18.2                  |
| en-ta         | 18.9        | 3.3                   | 41.9        | 13.7                  |
| en-as         | 16.6        | NR*                   | 27.3        | NR*                   |
| en-si         | 15.3        | 7.4                   | 15.7        | NR*                   |
| en-ne         | 6.6         | NR*                   | 14          | NR*                   |
| en-ur         | 28.7        | NR*                   | 26.9        | NR*                   |
| en-sd         | 15.5        | NR*                   | 18.5        | NR*                   |

NR* - NOT recorded. The OPUS-MT executable model was not found for these languages.

that has undergone 172K iterations of training delivers noticeable advantages (±3.2 BLEU in both the language pair direction). Hence, it is clearly visible that using all the above techniques, the BLEU score for all language pairs (available in the PMI corpus) gets better in both directions.

6 FINE-TUNING PRETRAINED OPUS-MT

The MT model can be fine-tuned to fit a specific domain or style. A set of bilingual sentences representing the domain or style that the MT model should adapt to may be needed for fine-tuning. For MT fixed-domain problems, fine-tuning on robust pre-trained models (on limited, verified sets) has grown to be a preferred strategy. There are currently more than 1,000 pretrained neural MT models available in the repository of OPUS MT [66] ranging over different language pairs. Both bilingual and multilingual models are supported by OPUS-MT and its main objective is to create open-source materials and applications for MT. So, we employ opus-mt-en-mul model and opus-mt-mul-en model for fine-tuning all the IL pairs in both directions, i.e., EN-XX and XX-EN. In both the terms, mul represents multiple languages, XX indicates any Indic language, and English is represented by en or EN. We make use of the PMI dataset [24], which is an MT Indic dataset composed of a variety of political data sources, such as news commentary and parliamentary hearings. After fine-tuning, it is clearly visible in Table 5 that our NMT system performs better then the OPUS-MT model.

7 RESULTS AND OBSERVATION

The effectiveness of the MT system has been evaluated using automated evaluation measures. BLEU refers to "Bilingual Evaluation Understudy" and is used as an automatic evaluation metric for the MT system [51]. The number of words in the MT output that match the reference translation is used to compute the BLEU score. The BLEU score ranges between 0 and 1 (or 0 and 100), where 0 denotes no similarities and 1 denotes identity, which is not always attainable for a model. To determine the BLEU score for all the translation files generated from our model, we prefer to use SacreBLEU [51]. Using the WAT 2022 server [45], we have tested and also evaluated all our translation files and checked the BLEU score [51]. Table 3 lists the BLEU score results and shows...
that our system is better than the baseline model [45] as well as from fine-tuning the OPUS-MT model for both many-to-one and one-to-many directions as shown in Table 5. For English-to-IL translation, our MNMT model gives a BLEU score between 6.6 and 39.90, with Nepali and Hindi receiving the lowest and highest scores, respectively (refer to Table 3). For the IL-to-English direction, the BLEU score ranges between 14.0 and 44.2, with Nepali having the lowest and Hindi having the highest BLEU score (refer to Table 3).

7.1 Translation Files

This subsection gives an overview of the translation of all the languages generated by our model.

(1) English to Assamese

**English:** Until 1960, Brzezinski worked as an advisor to John F. Kennedy, then to Lyndon B. Kennedy, on behalf of the Johnson administration.

**Generated:** ১৯৬০ চল প্রজটিনকে জন এফ কেনেডির উপদেষ্টা আক তার পিছু হিসাবে জ্ঞানসন প্রশাসনে কাজ করেছিল।

**Reference:** ১৯৬০ সাল হইতে, প্রজটিনকে জন এফ, কেনেডির জন্য উপদেষ্টা হিসেবে কাজ করেছিলেন এবং জ্ঞানসন প্রশাসনে কাজ করেছিল।

(2) English to Bengali

**English:** Until 1960, Brzezinski worked as an advisor to John F. Kennedy, then to Lyndon B. Kennedy, on behalf of the Johnson administration.

**Generated:** ১৯৬০-এর দশকে, প্রজটিনকে জন এফ কেনেডির উপদেষ্টা এবং তারপরে হিসাবে জ্ঞানসন প্রশাসনে কাজ করেছিল।

**Reference:** ১৯৬০ সাল হইতে, প্রজটিনকে জন এফ, কেনেডির জন্য উপদেষ্টা হিসেবে কাজ করেছিলেন এবং জ্ঞানসন প্রশাসন এর সাথে।

(3) English to Gujarati

**English:** The discovery also provides a deeper understanding of the evolution of bird feathers.

**Generated:** આ શોધ પક્ષીઓના પાયંઓ ઉત્કૃષ્ટિની પાણી સમજા આપે છે.

**Reference:** આ શોધ પક્ષીઓના પાયંઓ ઉત્કૃષ્ટિની જુદૂ સમજા પાણી આપે છે.

(4) English to Hindi

**English:** In late 2017, Siminoff appeared on the shopping television channel QVC.

**Generated:** 2017 के अंत में, सिमिनोफ शॉपिंग टेलीविजन चैनल के प्रसारण पर दिखाई दिया।

**Reference:** 2017 के अंत में, सिमिनोफ शॉपिंग टेलीविजन चैनल QVC में दिखाई दिया।

(5) English to Kannada

**English:** The discovery also offers insight into the evolution of bird feathers.

**Generated:** ಆ ಶೋಧ ಪಕ್ಷೀಯನಾ ಪಾಯಂಗಳ ಉತ್ಕೃಷ್ಟತೆಯ ಪಡಿತಪಡಿಯುವಣೆಗಳಲ್ಲಿ ಮಹಾಶಾಯ ಮಾತ್ರವೇ ಮಾತ್ರ ಮಾತ್ರವೇ ಮಾತ್ರವೇ.

**Reference:** ಆ ಶೋಧ ಪಕ್ಷಿಯನಾ ಪಾಯಂಗಳ ಉತ್ಕೃಷ್ಟತೆಯ ಪಡಿತಪಡಿಯುವಣೆಗಳಲ್ಲಿ ಮಹಾಶಾಯ ಮಾತ್ರವೇ ಮಾತ್ರವೇ ಮಾತ್ರವೇ.
(6) English to Malayalam

**English:** In late 2017, Siminoff appeared on shopping television channel QVC.

**Generated:** 2017 ഓക്ടോബർ, സിമിൻഫോ ചീസ് ടെലിവിഷൻ ചാനൽ എക്സോസ് നിയന്ത്രണം ചെയ്യുന്നു.

**Reference:** 2017-ചെതെ ഓക്ടോബർ‌മാസിലാണ്, സിമിൻഫോ സിപ്പ് ടെലിവിഷൻ ചാനൽ എക്സോസ് നിയന്ത്രണം ചെയ്തു.  

(7) English to Marathi

**English:** Ring settled a lawsuit with a rival security company, ADT Corporation.

**Generated:** रिंगने स्पष्टक मुद्दा संबंधी, एडीटी कंपनीसंबंधी खबर देखील सुलभ केला.

**Reference:** रिंगने प्रतिस्पर्धी मुद्दा संबंधी, एडीटी कंपनीसंबंधी खबर देखील सुलभ केला.

(8) English to Oriya

**English:** Gosling and Stone received nominations for Best Actor and Best Actress, respectively.

**Generated:** ଗୋସଲିଙ ଓ ସ୍ନ୍ତନ ପାଇଁ ଭେଟିକୀ ବିଶେଷତା ଅନ୍ତରେ ଅନୁମାନ କରାଲା ଏବଂ ସାମ୍ପଲାଣ କରାଲା।

**Reference:** ଗୋସଲିଙ ଓ ସ୍ନ୍ତନ ପାଇଁ ଭେଟିକୀ ବିଶେଷତା ଅନ୍ତରେ ଅନୁମାନ କରାଲା ଏବଂ ସାମ୍ପଲାଣ କରାଲା।

(9) English to Punjabi

**English:** Now we have 4-month-old mice that are non-diabetic that are used for diabetics, he added.

**Generated:** ਹੁਣ ਮਾਹੀ ਵੈਸਾ 4 ਮਹਿਨਾ ਦੇ ਹੁਣ ਵਰ ਨੇ ਤੀਜਾ-ਤਫ਼ਿਲਟੀ ਵਰ ਨੇ ਪਾਬੀ ਤੀਜਾ ਸਟੀਕ ਵਰ, ਹੌਂ ਸੀ ਕਈ।

**Reference:** ਹੁਣ ਮਾਹੀ ਵੈਸਾ 4 ਮਹੀਨਾ ਹੁਣ ਡੈਬੀਟਸ ਵਰ ਨੇ ਪੂਜਾ ਵਾਲੀ ਵਰ ਨੇ ਪ੍ਰਤਿਲਿੰਧ ਵਰਤਿਆ ਗਿਆ ਕਾਂ ਇਸੇ ਵਰ ਹੌਂ ਸੀ ਕਈ।

(10) English to Tamil

**English:** The Iraq Study Group delivered its report at 12.00 GMT today.

**Generated:** இராஜ் பானின் ஊடார் குழு அறிமுகத்துறை விளக்கம் 12.00 GMT-கு நியமாக வெளியிட்டன.

**Reference:** இராஜ் பானின் ஊடார் குழு அறிமுகத்துறை விளக்கம் 12.00 GMT-கு நியமாக வெளியிட்டன.

(11) English to Telugu

**English:** He built a WiFi doorbell. He said.

**Generated:** ఇతను విడిఫైబర్ డోర్బెల్ ని నిర్మించాడు. తప్పాడు.

**Reference:** ఇతను WiFi డోర్బెల్ ని నిర్మించాడు. తప్పాడు.
(12) English to Sinhala

**English:** The crew of the ship was divided into three main sections.
**Generated:** දේවි නාමක මාලාවල මාල කළුම් වන ලදී හැකියාවක් කර ලැබේ.
**Reference:** දේවි නාමක මාලාවල මාල කළුම් වන ලදී හැකියාවක් කර ලැබේ.

(13) English to Sindhi

**English:** In late 2017, Summons of Shopping appeared on the television channel QVC.
**Generated:** تے ظاهر تی ویو QVC 2017
**Reference:** تے ظاهر تی ویو QVC 2017

(14) English to Urdu

**English:** He said he had invented a Wi-Fi doorbell.
**Generated:** سے ہو گیا ہے کہ وی ای فی ڈری بال اختراع کیا ہے
**Reference:** سے ہو گیا ہے کہ وی ای فی ڈری بال اختراع کیا ہے

(15) English to Nepali

**English:** He said he had invented a Wi-Fi doorbell.
**Generated:** उनले व्याइफाई डर्बेल आक्षेप गरेको बताए
**Reference:** उनले व्याइफाई डर्बेल आक्षेप गरेको बताए

**Indic to English**

(1) Assamese to English

**Assamese:** এই যাত্রাতে বিভিন্ন সময়ে হিংসাজহ্রাস্কর বিপদ পরিচিত।
**Generated:** On this trip, Iwasaki was in danger at various times.
**Reference:** During his trip, Iwasaki ran into trouble on many occasions.

(2) Bengali to English

**Bengali:** আনুসারে পাঠ্য পাঠকের বিবর্তনের বিষয়েও পরিজ্ঞান দেয়।
**Generated:** the discovery also gives insight into the evolution of bird feathers.
**Reference:** The find also grants insight into the evolution of feathers in birds.

(3) Gujarati to English

**Gujarati:** તેમાં વીજી ડોર બનાવ્યો હતો, તેને તેમનો હાલ મળ્યો હતું.
**Generated:** he made a wifi bell, he said.
**Reference:** He built a WiFi door bell, he said.
(4) Hindi to English

Hindi: रिंग ने अपनी प्रतिस्पर्धी कंपनी ADT कोष्ठी के साथ एक मुकदमा निपटाया है।
Generated: ring has settled a lawsuit with its rival adt corporation.
Reference: Ring also settled a lawsuit with competing security company, the ADT Corporation.

(5) Kannada to English

Kannada: ರಿಂ ಹೆಸರು ಇತರ ಸಂತತಿಯನ್ನು ಅಪ್ರತ್ಯೇಕವಾಗಿ ಅಲ್ಲದೇ ಸಾಮಾನ್ಯವಾಗಿದ್ದು ನಿವೃತ್ತಿಗೆ ತನ್ನ ಸಂತತಿಯನ್ನು ತಲುಪಿಸುತ್ತಾ." ರಿಂಗ್ ಅತ್ಯಂತ ನಿವೃತ್ತಿಯಾಯಿತು.
Generated: Earlier it was diabetic, but now we have 4-month mice without diabetes, he said.
Reference: We now have 4-month-old mice that are non-diabetic that used to be diabetic,” he added.

(6) Malayalam to English

Malayalam: രിംഗ് ന്റെ പ്രതിരീക്ഷിത ആയ 4-സിമ്മസ് മെയിൻ ഒക്കെ അപൂർവമാണ്‌.
Generated: he said he made a wifi door bell.
Reference: He built a WiFi door bell, he said.

(7) Marathi to English

Marathi: रिंग ने अपनी प्रतिस्पर्धी कंपनी ADT कोष्ठी के साथ एक मुकदमा निपटाया है।
Generated: US president donald trump announced that his troops were leaving syria.
Reference: Late on Sunday, the United States President Donald Trump, in a statement delivered via the press secretary, announced US troops would be leaving Syria.

(8) Oriya to English

Oriya: ରିଂଗ୍କୁ ଅଧିକାରିତ ଉତ୍ତର ଅଧିକାରିତ ସମ୍ପର୍କୀତ ବିଷୟର କରି ହେକୁ ବିକ୍ରେଖାନୀ କରିବା ପଦ୍ଧତି ଥାକନ୍ତି ରଜାଙ୍କୁ ଛିଟିକାଯାଂ।
Generated: Gosling and stone were nominated for best actor and best actress respectively.
Reference: Gosling and Stone received nominations for Best Actor and Actress respectively.

(9) Punjabi to English

Punjabi: ਆਪਣੀ ਪਰਤੀਕਾਲ ਦੇ ਵੇਖਾਂ, ਸੀਮਾਨਾਵਾਂ ਵਾਲੀ ਬਣ ਨਾਲਾਂਡ ਹੀ ਦਿੱਖਣਾ।
Generated: during his journey, iwasaki was sometimes in trouble.
Reference: During his trip, Iwasaki ran into trouble on many occasions.

(10) Tamil to English

Punjabi: புகழ் தனிக்கு கை, குறுக்கு மிக்கவும் விளக்காமல் 12.00 GMT'இல் பொய்மிக்கும்.
Generated:
8 CONCLUSION

We presented an MNMT for 15 ILs i.e., English to Indic and vice versa. Different techniques have been explored in this article such as fine-tuning, exploiting language relationships, back-translation, and domain adaptation. With experiments on Samantar datasets, our methods significantly outperform the best individual systems and cutting-edge traditional system combination techniques. In comparison with the pretrained OPUS-MT NMT model (for the majority of
language pairs) using the PMI corpus, results confirm better performance of our MNMT model than the pretrained one. The effectiveness of language relatedness was explored, analyzed, and validated to be useful for the low-resource language(s), with support of high-resource language(s), for ascertaining a qualitative translation by increasing the corpus. Further, implementation of BT enhanced the translation quality of both sources and targets over original sentences (by +3.2 BLEU score in both directions). A minor improvement of 0.86 average BLEU score was achieved using the Domain Adaptation technique for language translation in both directions. Moreover, through experiments, we noticed that substantial progress would be achieved for low-resource languages like ILs, in terms of translation quality. For our future work, we would like to do more in-depth analysis to explore the abilities of our MNMT system. Our focus will remain to use our MNMT system for the ILs without sufficient datasets or zero-resource scenarios. In order to take our work to the next level, our future experiments will point toward the techniques for a qualitative translation of mixed language sentences (i.e., code-mixing).

REFERENCES

[1] Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. arXiv preprint arXiv:2204.08776 (2022).
[2] Melvin Aharoni, Roeo Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1. ACL, 3874–3884.
[3] Hessam Amini, Farhood Farahnak, and Leila Kosseim. 2019. Natural language processing: An overview. Frontiers in Pattern Recognition and Artificial Intelligence 1 (2019), 35–55.
[4] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roeo Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. arXiv preprint arXiv:1903.07091 (2019).
[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).
[6] Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. arXiv preprint arXiv:1711.02173 (2017).
[7] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014).
[8] Raj Dabre, Chenhui Chu, Fabien Cromieres, Toshiaki Nakazawa, and Sadao Kurohashi. 2015. Large-scale dictionary construction via pivot-based statistical machine translation with significance pruning and neural network features. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. 289–297.
[9] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. ACM Computing Surveys (CSUR) 53, 5 (2020), 1–38.
[10] Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kumar Patra. 2022. NIT Rourkela machine translation (MT) system submission to WAT 2022 for MultiIndicMT: An Indic language multilingual shared task. In Proceedings of the 9th Workshop on Asian Translation. 73–77.
[11] Niladri Sekhar Dash. 2009. Translation Corpora and Machine Aided Translation. Translation Today 6, 1 (2009), 134–153.
[12] Niladri Sekh Dar Dash. 2008. Corpus linguistics: An empirical approach for studying a natural language. In Language Forum, Vol. 34. Bahri Publications, 5–26.
[13] Miguel Domingo, Mercedes Garcia-Martinez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2018. How much does tokenization affect neural machine translation? arXiv preprint arXiv:1812.08621 (2018).
[14] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 1723–1732.
[15] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptichinsky, Sergery Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-centri multilingual machine translation. The Journal of Machine Learning Research 22, 1 (2021), 4839–4886.
[16] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. arXiv preprint arXiv:1601.01073 (2016).
[17] Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics. 376–384.
[18] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics 10 (2022), 522–538.

[19] Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 162–168.

[20] Vikrant Goyal and Dipti Misra Sharma. 2019. The iit-h Gujarati-English machine translation system for wmt19. In Proceedings of the 4th Conference on Machine Translation. 191–195.

[21] Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Fino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English. arXiv preprint arXiv:1902.01382.

[22] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindrich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. Computational Linguistics 48, 3 (2022), 673–732.

[23] Barry Haddow, Nikolay Bogoychev, Denis Emelin, Ulrich Germann, Roman Grundkiewicz, Kenneth Heafield, Antonio Valerio Miceli-Barone, and Rico Sennrich. 2018. The University of Edinburgh’s submissions to the WMT18 news translation Task. In Proceedings of the 3rd Conference on Machine Translation. 399–409.

[24] Barry Haddow and Faheem Kirefu. 2020. PMIndia - A collection of parallel corpora of Languages of India. CoRR (2020).

[25] W. Hutchins. 2000. The first decades of machine translation. Early Years in Machine Translation 1 (2000), 1–15.

[26] Shubhnandan S. Jamwal and Vijay Singh Sen. 2022. Natural language interface in Dogri to database. In Machine Intelligence and Data Science Applications. Springer, 843–852.

[27] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics 5 (2017), 339–351.

[28] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in C++. arXiv preprint arXiv:1804.00344 (2018).

[29] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, N. C. Gokul, Akiv Bhattacharyya, Mitesh M. Khatra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Findings of the Association for Computational Linguistics: EMNLP 2020. 4948–4961.

[30] Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. arXiv preprint arXiv:1805.12282 (2018).

[31] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[32] Philipp Koehn. 2009. Statistical Machine Translation. Cambridge University Press.

[33] Prudhvi Kosaraju, Sruthilaya Reddy Kesidi, Vinay Bhargav Reddy Ainavolu, and Puneeth Kukkadapu. 2010. Experiments on Indian language dependency parsing. In Proceedings of the NLP Tools Contest: Indian Language Dependency Parsing. 40–45.

[34] Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018).

[35] Anoop Kunchukuttan. 2020. The indicnlp library. Indian Language NLP Library.

[36] Anoop Kunchukuttan and Pushpak Bhattacharyya. 2020. Utilizing language relatedness to improve machine translation: A case study on languages of the Indian subcontinent. arXiv preprint arXiv:2003.08925 (2020).

[37] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora. arXiv preprint arXiv:1711.00043 (2017).

[38] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. Transactions of the Association for Computational Linguistics 5 (2017), 365–378.

[39] Bei Li, Yinqiao Li, et al. 2019. The NiuTrans machine translation systems for wmt19. In Proceedings of the 4th Conference on Machine Translation. 257–266.

[40] Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 293–305.

[41] Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 3044–3049.
[42] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics 8 (2020), 726–742.

[43] Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign. 76–79.

[44] Rohit More, Anoop Kunchukuttan, Pushpak Bhattacharyya, and Raj Dahre. 2015. Augmenting Pivot based SMT with word segmentation. In Proceedings of the 6th International Conference on Natural Language Processing. 303–307.

[45] Toshiaki Nakazawa, Hideta Mino, Isao Goto, Raj Dahre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In Proceedings of the 9th Workshop on Asian Translation (WAT’22). Association for Computational Linguistics.

[46] Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dahre, Shohei Higashiyama, Hideta Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, et al. 2021. Overview of the 8th workshop on Asian translation. In Proceedings of the 8th Workshop on Asian Translation (WAT’21). 1–45.

[47] Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Asian Federation of Natural Language Processing, 296–301.

[48] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. arXiv preprint arXiv:1904.01038 (2019).

[49] Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 244–258.

[50] John C. Paoliilo and Anupam Das. 2006. Evaluating language statistics: The ethnologue and beyond. Contract Report for UNESCO Institute for Statistics.

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 311–318.

[52] Márcis Pinnis. 2018. Tilde’s parallel corpus filtering methods for WMT 2018. In Proceedings of the 3rd Conference on Machine Translation. 939–945.

[53] Annie Rajan, Ambuja Salgaonkar, and Ramprasad Joshi. 2020. A survey of Konkani NLP resources. Computer Science Review 38 (2020), 1–5.

[54] Gowtham Ramesh, Sumanth Doddapaneni, Aravind Bheemaraj, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. Transactions of the Association for Computational Linguistics 10 (2022), 145–162.

[55] J. Sangeetha, S. Jothilakshmi, and Devendra Kumar. 2014. An efficient machine translation system for English to Indian languages using hybrid mechanism. International Journal of Engineering and Technology 6, 4 (2014), 1909–1919.

[56] Debapriya Sengupta and Goutam Saha. 2015. Study on similarity among Indian languages using language verification framework. Advances in Artificial Intelligence 2015 (2015), 2 pages.

[57] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709 (2015).

[58] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015).

[59] Max Silberstein. 2016. Formalizing Natural Languages: The NooJ Approach. John Wiley & Sons.

[60] Muskaan Singh, Ravinder Kumar, and Indever Chana. 2020. Corpus based machine translation system with deep neural network for Sanskrit to Hindi translation. Procedia Computer Science 167 (2020), 2534–2544.

[61] Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C. V. Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, 3743–3751.

[62] S. Sreelekh, Pushpak Bhattacharyya, and D. Malathi. 2018. Statistical vs. rule-based machine translation: A comparative study on Indian languages. In International Conference on Intelligent Computing and Applications. Springer, 663–675.

[63] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems 27 (2014), 1–9.

[64] J. Tiedemann. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In Recent Advances in Natural Language Processing, Vol. 5. 237–248.
[65] J. Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In Lrec, Vol. 2012. 2214–2218.
[66] Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the world. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. European Association for Machine Translation. 479–480.
[67] Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In The Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 484–491.
[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in Neural Information Processing Systems 30 (2017), 1–11.
[69] Shweta Vikram. 2013. Morphology: Indian languages and European languages. International Journal of Scientific and Research Publications 3, 6 (2013), 1–5.
[70] Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP’20). Association for Computational Linguistics, 1022–1034.
[71] Hua-Peng Wei, Ying-Ying Deng, Fan Tang, Xing-Jia Pan, and Wei-Ming Dong. 2022. A comparative study of CNN-and transformer-based visual style transfer. Journal of Computer Science and Technology 37, 3 (2022), 601–614.
[72] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 1568–1575.

Received 14 November 2022; accepted 2 March 2023