Reinforcement of low-resource language translation with neural machine translation and backtranslation synergies

Padma Prasada[1,2], Malode Vishwanatha Panduranga Rao[3]
[1]Department of Electronics Engineering, Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bengaluru, India
[2]Department of Artificial Intelligence & Data Science, Sri Dharmasthala Manjunatheshwara Institute of Technology, Karnataka, India
[3]Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bengaluru, India

Article Info

Article history:

Received Dec 29, 2023
Revised Jan 24, 2024
Accepted Feb 10, 2024

Keywords:

Backtranslation
Low resource language
Neural machine translation
NLP data augmentaation
Pretrained language models

ABSTRACT

This research investigates challenges and advancements in neural machine translation (NMT), specifically targeting English-to-Kannada translation. Emphasizing the scarcity of data and linguistic complexity in low-resource languages (LRL), particularly Kannada, the study underscores the need for specialized techniques. Starting with exploration of Kannada's historical and cultural significance, the paper highlights critical importance of linguistic comprehension. The primary objective is to develop robust NMT models for precise and contextually relevant translations in low-resource scenarios. The novelty of this research lies in its innovative approach to Kannada NMT challenges, incorporating comprehensive examination of historical and cultural context to establish strong linguistic foundation. Motivated by the urgency to address translation needs in LRL, the paper proposes novel strategies, advocating notably for backtranslation to generate synthetic parallel corpora. Rigorous testing, including bilingual evaluation understudy (BLEU) score assessments, evaluates effectiveness of these proposed approaches. Beyond assessing backtranslation, the study explores challenges faced by Kannada NMT in handling dialectical and spelling variations. The research reports substantial 83-percentage-point average increase in BLEU scores, contingent on aligning unique Kannada terms with the same domain as existing occurrences. This study contributes significantly to Kannada natural language processing by offering novel insights into NMT intricacies and providing practical solutions for enhancing translation accuracy in low-resource settings.

Corresponding Author:

Padma Prasada
Department of Electronics Engineering, Faculty of Engineering & Technology, JAIN (Deemed-to-be University)
Kanakapura Main Road, Bengaluru 562112, Karnataka, India
Email: ppjain15@gmail.com

1. INTRODUCTION

Recent years have seen notable advances in machine translation (MT), owing mostly to the development of neural machine translation (NMT) models that leverage deep learning techniques [1]–[3]. While these models excel at translating between languages, their effectiveness is dependent on the availability of large amounts of parallel training data. Notably, NMT outperforms in high-resource language pairs such as English-Spanish and English French, owing to its large bilingual corpora [4]. Unfortunately, this abundance of translation resources is not ubiquitous, providing significant problems for low-resource languages (LRLs) to achieve accurate and fluent translations. Kannada, a Dravidian language spoken mostly in Karnataka, India, has limited resources in its linguistic setting. Despite its huge native speaker population, Kannada has gotten less attention in MT research than languages with larger speaker numbers [5]. The biggest impediment to developing successful English-Kannada translation systems is the scarcity of parallel corpora and linguistic resources. This deficiency becomes especially difficult in environments requiring exact translation, such as the government, educational, and commercial sectors [6].

Two major variables are driving this investigation. To begin, there is an urgent need to bridge the language gap between English and Kannada speakers to improve communication efficiency, allow knowledge transfer, and promote cross-cultural contacts. Because English is a widely recognized international language, many Kannada speakers want to interact with English speaking groups for educational, vocational, and personal reasons [7]. English speakers who want to learn about or interact with Kannada culture, on the other hand, have difficulties due to a lack of translation resources. Second, the research is motivated by the larger goal of enhancing low-resource MT. The MT community faces a unique difficulty when it comes to dealing with LRL pairs, demanding novel ways to handle data shortages [8]. This study aims to fulfil an immediate need while also making significant contributions to the broader field of low-resource MT research by evaluating the suitability of NMT models and employing backtranslation specifically in the English-Kannada translation domain. This study's techniques and conclusions may provide a useful framework for future initiatives focusing on languages that are currently underrepresented in research.

This study influences language, culture, technology, and practice. The study improves translation quality in this language pair with limited resources to meet the need for accurate English-Kannada cross-lingual communication. Knowledge flow, educational possibilities, and meaningful cross-cultural encounters may increase. Better translations enable reciprocal access to English and Kannada language materials, fostering cultural and intellectual exchange. This study also contributes to low resource MT by ensuring fair access to information and opportunity for speakers of languages with little global representation. MT technology is democratising, bridging the digital language divide and promoting globalisation. The study's broad ramifications are evident, notably in education, where better translation quality increases Kannada-speaking students' access to high-quality instructional resources. More accurate translation technologies will improve communication with English-speaking peers, boost economic growth, and facilitate international collaboration for Karnataka businesses and government agencies. Beyond cultural preservation and promotion, the project could help translate legacy literature and cultural artefacts. This contribution considerably promotes linguistic and cultural diversity. To achieve our research goals, this study examined backtranslation using monolingual data to extend parallel corpora and enhance translation quality [9]. This study uses NMT models in low-resource contexts and intentional back-translation for data augmentation. This research aims to improve English-Kannada translations. Additionally, our research aims to advance low-resource MT methods.

This work addresses NMT challenges, particularly translating from English to Kannada, LRL [10]. Kannada language resources are a major concern. Therefore, NMT models must be constructed to make correct, context-appropriate translations. To overcome these challenges, the programme will investigate novel ways, focusing on leveraging backtranslation to supplement training data and improve Kannada NMT systems. Kannada, LRL, embodies its rich language tradition and data scarcity. Our research aims to overcome difficulties and improve NMT systems for English-to-Kannada translation. By exploring this language pair's intricacies, NMT's potential can be reconciled with Kannada's particular limits. The main goal is to construct robust NMT models that can translate accurately and contextually with limited linguistic resources. These goals are a systematic approach to Kannada's data shortage and dialectal diversity challenges. Investigating novel tactics like backtranslation is expected to improve NMT systems and improve context-aware translation procedures. The initiative aims to develop NMT for LRLs by focusing on Kannada's complexity.

- Develop advanced NMT models specifically optimized for translating from English to Kannada, considering the limitations provided by Kannada's status as LRL.
- To study and apply effective approaches for addressing the restrictions of data scarcity in training NMT models for Kannada translation, with a primary focus on using backtranslation.
- Perform systematic experiments and detailed studies to quantify and validate the efficiency of backtranslation in improving translation accuracy, especially in the context of Kannada's linguistic complexities.

2. BACKGROUND

Advancements in MT have been propelled by the introduction of NMT models. This section critically examines pertinent academic literature, specifically delving into NMT, the translation challenges inherent in languages with limited resources, and the application of backtranslation methods. The review particularly addresses the intricacies of English to Kannada translation—a language pair characterized by resource constraints. By surveying this body of literature, aiming to distill key insights into the transformative impact of NMT, the nuances of translation challenges in LRL, and the efficacy of backtranslation strategies in mitigating complexities specific to English to Kannada translation.

2.1. Why backtranslation?

In this study backtranslation is chosen as the preferred NLP data augmentation methodology over other methods because of its unique advantages and versatility in overcoming issues associated with limited linguistic resources. Backtranslation is different from other technologies that add to data by using paraphrased datasets that already exist because it can create artificial parallel corpora by translating monolingual data [11], [12]. This distinguishing feature is especially important in the case of LRLs because such paraphrased datasets may be rare or non-existent. Backtranslation is the process of translating statements from the target language back into the source language, resulting in extra training instances. This method not only increases the amount of training data accessible, but it also assures that the synthetic sentences generated are contextually appropriate and diverse [13]. This technique exposes the NMT model to a broader range of language variances and expressions, thereby improving its adaptability and overall performance. Furthermore, backtranslation is adaptable because it does not rely on the availability of specific linguistic resources for the destination language. This makes it particularly well-suited to addressing the issues faced by LRLs like Kannada, where a lack of parallel corpora and linguistic variation can make training effective NMT models difficult [14].

In summary, the choice of backtranslation is based on its ability to overcome the constraints associated with data scarcity in LRL contexts. Back-translation is a recommended alternative for augmenting NLP data in the context of NMT due to its capacity to generate synthetic training data without relying on pre-existing paraphrased datasets and its effectiveness in increasing the contextual relevance of the training set. The following section presents a comprehensive summary of earlier efforts, with a particular emphasis on NMT for LRLs. It focuses on the difficulties connected with low resource translation and the complexities of data augmentation approaches customized for NMT applications.

2.2. Overview of neural machine translation for low resource language

MT, especially NMT, has grown in popularity due to intercultural communication needs. NMT relies on large parallel corpora, which is difficult for languages with low resources [15], [16]. The study examined LRL problems using “Kazakh English” MT models. To increase model performance, the authors artificially augmented corpora using OpenNMT [17]. A monolingual dataset from the source language improves NMT system performance because LRLs are uncommon. Using artificially constructed and actual parallel datasets and self-learning and fine-tuning, authors translated “Wolaytta-English” bidirectionally [18], [19]. Recurrent neural networks (RNNs) with sequence-to-sequence (Seq2Seq) models and the encoder-decoder mechanism with long short-term memory (LSTM) as the RNN unit have been shown to be effective for MT in LRL scenarios [20], [21]. Convolutional and sequence-to-sequence models trained on conditional distribution translate well but suffer as input phrase length grows [22]. A multi-source neural model with two independent encoders is designed to translate agglutinative languages with complex morphology and limited resources. These encoders put a language layer in the input embedding layer and consider lemma, POS tag, and morphological tag [23]. Kannada to Telugu MT should use a dictionary-based approach where semantic changes are acceptable but not professional translation. A position aware transformer (P-transformer) may increase absolute and relative location information in self- and cross-attention, according to researchers. To develop a Doc2Doc NMT model that uses sequence-to-sequence transformation to generate target documents from input documents, utilise the P-transformer. Doc2Doc NMT models enhance bilingual evaluation understudy (BLEU) scores and discourse coherence, especially when addressing discourse issues [24].

2.3. Low-resource translation challenges

In languages like Kannada, low-resource MT makes it difficult to build accurate and fluent translation models. Low resource translation is complicated, and this section explores its main challenges for researchers and practitioners. Data scarcity is the biggest obstacle to low-resource translation. Durable MT models, especially neural network-based ones, require a lot of parallel data [25]. The data are pairs of sentences in the source (English) and target (Kannada) languages. Many languages, including Kannada, lack parallel corpora [26]. Limited training data makes it hard to develop precise models, resulting in poor translation. Languages with few resources may have a lot of unusual and specialised vocabulary not in training databases. OOV words are difficult for NMT models to translate [27]. Translators with limited resources must use specialised methods to manage OOV terminology. Kannada has complex inflectional and agglutinative morphology. Due to its complexity, translating words effectively, managing word order, and preserving meaning subtleties is difficult [28]. To ensure contextually proper translations, NMT models must master these difficulties. Pre-trained models and large language models provide a solid foundation for NMT in languages with abundant resources [29]. LRL have fewer pre-existing models, which requires more attention and creativity in model construction.

2.4. Data augmentation for neural machine translation applications

Backtranslation successfully resolves data constraints in language pairs with limited resources in MT [30]. Monolingual source and destination data generates parallel model training data. Li and Specia [31] devised unique data augmentation methods to expand constrained and noisy data, improving NMT model resilience. Strategies aim to compact models. Using causal interpretation of language models and phrasal alignments, Liu *et al.* [32] created larger parallel translation datasets to add data to NMT. Graça *et al.* [33] validated backtranslation in cross-entropy optimisation of an NMT model and explained its mathematical assumptions and approximations. This study examines synthetic data production methods for three LRL pairings: Spanish Portuguese, Czech-Polish, and Hindi-Nepali. The study examines how backtranslated data affects new MT systems, using real-world scenarios and data selection techniques to optimise synthetic corpora [34]. It focuses on monolingual data integration. Depending on the MT methods and the number of words in the corpora, backtranslated data from different sources may function better [35]. Self-training reverse NMT models using forward translation of the reverse model's output improves model performance, especially when parallel data is scarce [36].

This study addresses English-Kannada MT problems in LRL instances as shown by this literature review. We use NMT models with backtranslation to improve translation quality and progress low-resource MT systems. This study adds to the knowledge of NMT, low-resource MT, and English-Kannada translations. Table 1 also compares data augmentation methods, assessing their efficacy in our research.

A dominant data augmentation method, backtranslation, generates synthetic parallel corpora using monolingual data. This enhancement greatly expands seq2seq NMT model training datasets, especially in low-resource environments with minimal training data. Since the synthetic examples are generated by the same NMT model, they preserve the context and linguistic nuances of the monolingual data. Accurate and meaningful seq2seq translations require context preservation. LRLs benefit from this method since it solves simultaneous corpora training problems. Backtranslation makes the seq2seq NMT model more robust, enabling it to handle multilingual translation. Backtranslation is more independent than data augmentation methods that use paraphrased datasets or external resources. It works well in sparse linguistic situations due of its independence. Backtranslation exposes seq2seq NMT models to more language variants, improving their generalisation. This exposure helps the model adapt to different language patterns by producing more accurate translations for a variety of inputs.

Table 1. Comparison of merits and demerits of different NLP data augmentation techniques

Augmentation Technique	Merits	Demerits
Backtranslation [13]	Effective in generating synthetic parallel corpora by translating monolingual data	May introduce noise or incorrect translations, particularly if the initial model is not robust
Paraphrasing [37]	Diverse augmentation method by rephrasing sentences	Difficulty in obtaining high-quality paraphrased datasets; risk of losing original context
Word embedding-based Synonym replacement [38]	Preserves sentence structure while introducing variability	Limited to synonymous substitutions; may not capture context nuances
Data mixup [39]	Blends multiple sentences to create novel examples	Potential loss of original sentence coherence; risk of generating unrealistic examples
Synthetic data generation [40]	Creates entirely new examples using generative models	Quality heavily depends on the generative model; may introduce unrealistic examples

3. INTEGRATING BACK-TRANSLATION INTO NEURAL MACHINE TRANSLATION

For the first time back-translation in NMT has been presented by Sennrich *et al.* [41]. The proposed approach involves utilizing monolingual data in the target language to create synthetic parallel data, thereby enhancing the training of the NMT models, and improving the quality of translations. This strategy has since emerged as a fundamental approach in tackling the issue of limited data availability in the field of MT. In the field of NMT, researchers focus on the task of translating a given source phrase $S_1^J = S_1, \dots, S_j, \dots S_J$ into a desired target sentence $T_1^I = T_1, \dots, T_i, \dots T_I$. To achieve the desired outcome, the process of translation is represented using a neural model $p\theta\left(T_i \middle| S_1^J, T_1^{i-1}\right)$ that is characterized by parameters θ. To obtain the most effective optimization criterion for an NMT model, it is necessary to access the accurate joint distribution of the origin and destination parallel corpora, denoted as $P_r(S_1^J, T_1^I)$. The approximation is obtained by utilizing the empirical distribution $\hat{p}\left(S_1^J, T_1^I\right)$, which is produced from a dataset consisting of bilingual data $\left(S_{1,s}^{Js}, T_{1,s}^{Is}\right)_{s=1}^{S}$. During the training process, the model's parameters are adjusted to limit the cross-entropy. The cross-entropy was subsequently normalized based on the number of target tokens. This procedure was regularly implemented.

$$L(\theta) = -\frac{1}{S}\sum_{s=1}^{S} \frac{1}{I_s}\log\, p_\theta\left(eT_{1,s}^{I_s} \mid S_{1,s}^{J_s}\right) \qquad (1)$$

One approach to including monolingual data is to create a pseudo-parallel source corpus using methods such as the backtranslation technique. This study examined the significance of generators as a component of the optimization criteria in NMT models. In Addition, it delves into practical approximations commonly used in this context. The concept under consideration can be characterized as a sampling distribution denoted as $q(S_1^J, T_1^I; p)$, which is parameterized by the target-to-source model p.

$$L(\theta) = -\sum_{e_1^I} \hat{p}(T_1^I) \cdot \frac{1}{I}\sum_{f_1^J} q\left(S_1^J \mid T_1^I; p\right) \cdot \log\, p_\theta\left(T_1^I \mid S_1^J\right) \qquad (2)$$

In (2) emphasizes the need for the generation technique $q(S_1^J, T_1^I; p)$, which should yield a distribution of origin sentences that closely resembles the original distribution $P_r(S_1^J, T_1^I)$. The utilization of back-translation and its many forms predominantly adheres to the original methodology. Each target sentence that is considered authentic is paired with a single synthetic source sentence. The newly acquired dataset was subsequently utilized in a manner that assumes multilingual capabilities. Approximated the estimation of the summation across the complete collection of potential source phrases (in (2)) within a search area of N phrases. However, the expenses of data creation and training increase with N, which discourages the selection of larger volumes. The pseudo corpora remained static during training, indicating that synthetic phrases did not change over epochs. This static nature negates the benefits of sampling-based techniques, forcing real-time phrase production, which complicates implementation and significantly slows training. Regardless, the approximation has no effect on backtranslation because the model continuously provides identical translations. In (3) encapsulates this approximation.

$$L(\theta) \approx -\sum_{s=1}^{S} \frac{1}{N \cdot I_s}\sum_{n=1}^{N} \log p_\theta\left(T_{1,s}^{I_s} \mid S_{1,s,n}^{J_s,n}\right) \qquad (3)$$

It is hypothesized that these criteria become less troublesome when a substantial quantity of monolingual data is available. This phenomenon can be linked to the principle known as the rule of large numbers. According to this principle, when a specific phrase is repeated multiple times, the resulting distribution of source sentences tends to align with the underlying probability distributions $q(S_1^J, T_1^I; p)$.

4. NEURAL MACHINE TRANSLATION TRAINING FROM KANNADA BITEXT

4.1. Challenges in Kannada natural language processing and neural machine translation

As a Dravidian language spoken in Karnataka, southern India, Kannada is highly variable. NLP and NMT systems struggle with Kannada's dialectal and orthographic variances, despite Tamil's standing as a classical language in India and its linguistic prominence in the region. To demonstrate Kannada language processing's difficulties, this argument will provide insights and examples. Indian language grammar books list eight vibhakti cases, but Kannada's polysyllabic and agglutinative structure shows that the six cases represent several inflections. The language's inflectional complexity makes word forms difficult to determine, distinguishing it from English [42]. Sentence type does not affect word order in Kannada due to cases and inflections. By carefully applying cases to each category, sentences in both active and passive forms are generated while maintaining the same subject-object-Verb (S-O-V) sequence, a common structure in Kannada (Table 2).

Kannada, an agglutinative language, has a grammatical phenomenon known as 'Sandhi,' which represents the fusion or union of two or more words or morphemes to form a single composite word. Kannada text is written from left to right, and it uses the same punctuation marks as English. It is critical to emphasize the language's lack of distinct letter styles, such as italics, uppercase, or lowercase.

arasana + mane = aramane
(king's + residence = palace)

– Dialectal differences: Kannada dialects vary across Karnataka. Phonetic, lexical, and syntactic differences characterise these differences. The Kannada language has several regional variations, including Mysuru, North Karnataka, and Coastal. Dialectal distinctions present issues. At diverse locales, words and sounds are pronounced differently. In some dialects, the term "Giḍa" (meaning "tree") may be pronounced as "Giḍā". Regional dialects may use distinct vocabulary words not found in standard Kannada. Standard Kannada uses "Dēvālaya" for "temple" whereas certain dialects use "Dēvasthāna".
– Orthographic variations: orthographic variants in Kannada give intricacy to its letters and characters. Kannada orthographic issues include the following. Conjunct Characters: Kannada script uses two or more consonants to construct a character. Conjunct characters vary by dialect. In some dialects, the word "Bālaka" (meaning "boy") may be transcribed as "Bālka". Vowel Length: Kannada script marks vowel length, but dialects use them differently. In some dialects, the word "Kannaḍa" (meaning "Kannada") may be spelled as "känada".

Table 2. Active and passive sentence with S-O-V sequence

Sentence (Kannada and its Corresponding English Version)	Sentence sequence
avanu māvina haṇṇu tindanu (He mango ate)	Active (S-O-V)
avaninda māvina haṇṇu tinnalapaṭṭitu (Him by mango eaten)	Passive (S-O-V)

4.2. Training the model

Using Samanantar, 4094 English-Kannada sentence pairs are generated. Diacritics and writing systems were standardised using rule-based deterministic methods, reducing author-specific variance. This method makes the text compatible with several recent orthographies. Standardising functional words was also a priority. Each language used a separate byte pair encoding (BPE) model for phrase pairings. These encoding models were created with a 2000-token vocabulary. Each cycle partitioned the dataset into training 80%, validation 10%, and testing 10% subsets. Ten rounds of random shuffling, re-splitting, and retraining increase dataset validation. Trained verified transformer encoder/decoder models for translation tasks using these sets. To avoid overfitting on the little dataset, few training steps are used. PyTorch and GPU-based translation models are tested at a Google Collaboratory.

4.3. Evaluation metrics

The BLEU metric is a popular NLP metric for assessing the authenticity of machine-generated translations. It counts shared n-grams (word sequences) to determine the similarity between a computer-generated translation and one or more reference translations. To provide a normalized number, the count is adjusted based on the length of the machine-generated text. A higher BLEU score implies better translation quality, with '1' indicating a perfect match. This study uses the BLEU metric because it is the leading automated evaluation measure in the area. The BLEU ratings are calculated in this study using a specific sample as following examples.

NMT Translation: "Bēku majjige ide"
Reference Translation: "Bēku majjige ide "

Higher BLEU scores indicate greater translation quality, and the score itself is often given as a number between zero and one. Because the provided translation was a perfect match to the source, the BLEU score was '1'. The BLEU score for the following example was 0.5, as there was a disparity in vocabulary selection.

NMT translation: "Nānu ōduvudannu prītisuttēne"
Reference translation: "Nānu pustakagaḷannu ōduvudu iṣṭavide"

4.4. Experimental architecture

Figure 1 shows a backtranslation experimental architecture that created a synthetic dataset by translating monolingual English data to Kannada and then creating a model using the EN-KN parallel corpus. A model from the EN-KN parallel dataset translated the monolingual corpus into Kannada. The translated text was then backtranslated using a KN-EN model. Used duplication removal and smoothing on the corpus. A complete model was generated by training the concatenated parallel corpus.

Model vocabularies are generated using BPE to extract subwords from the dataset. The early English-to-Kannada translation models created erroneous data for later backtranslation models. These models translated phrases into monolingual data for the parallel corpora. BLEU metric to evaluate translation results since greater BLEU scores correlate with better translation quality. Backtranslation is used to enhance NMT data using EN-KN parallel corpora and generated datasets. After translating monolingual English data into Kannada, a synthetic dataset is backtranslated into English. Backtranslated corpus post-processing comprises duplicate removal and smoothing. The larger EN-KN parallel corpus is connected to the original dataset. The concatenated parallel corpus is used to train a complete model with BPE vocabulary. English-to-Kannada translation methods generate false data by transforming phrases into monolingual data. The system evaluates translation quality using the BLEU metric, and higher scores indicate better translation. This comprehensive method enhances the NMT paradigm, which benefits LRLs like Kannada. Figure 2 details this approach's algorithm.

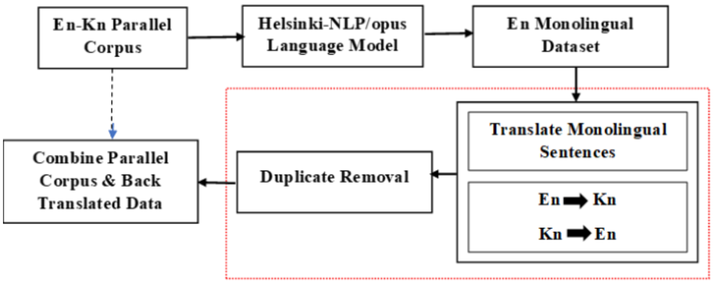
En-Kn Parallel Corpus
Helsinki-NLP/opus Language Model
En Monolingual Dataset
Translate Monolingual Sentences
En ➡ Kn
Kn ➡ En
Duplicate Removal
Combine Parallel Corpus & Back Translated Data

Figure 1. Proposed backtranslation model

```
        1.  Initialization:
             SyntheticDatasetEN←{}SyntheticDataset$_{EN}$←{}
             MonolingualDatasetKN←{}MonolingualDataset$_{KN}$←{}
        2.  Back Translation Process:
            a.  Translation from English to Kannada:
            SyntheticDataset$_{EN}$←NMT$_{EN \rightarrow KN}$(MonolingualDataset$_{EN}$) \$_{quad}$ ▹
            b.  Backtranslation from Kannada to English:
            MonolingualDataset$_{KN}$←NMT$_{KN \rightarrow EN}$(SyntheticDataset$_{EN}$) \$_{quad}$ ▹
        3.  Post-Processing:
    a.  Remove duplicate sentences from MonolingualDatasetKNMonolingualDataset$_{KN}$
    b.  Apply smoothing methods on MonolingualDatasetKNMonolingualDataset$_{KN}$
        4.  Concatenation:
          AugmentedEN-KN Parallel Corpus←EN-KN Parallel Corpus ∪
                    MonolingualDataset$_{KN}$ \quad ▹
        5.  Vocabulary Construction
        6.  Evaluation (BLEU)
        7.  AugmentedEN-KN Parallel Corpus (OUTPUT)
```

Figure 2. Proposed algorithm

5. RESULTS AND DISCUSSIONS

Back translation augmentation involves translating text into another language and back again. This technique can produce textual data with different wording while keeping context and meaning. Researchers compare MT and reference human translation using the BLEU statistic. Synthetic data added to sentences expands datasets. Table 3 compares the original sentences and their suggested backtranslations. Figure 3 shows how training perplexity (PPL) evolves over epochs, demonstrating the model's learning dynamics. The PPL is high at 7 at Epoch 1-10, indicating model uncertainty. PPL decreases with training. By Epoch 20, the PPL drops rapidly to 3, and from Epoch 30 to Epoch 60, it peaks at 2.2. PPL has decreased because the model can anticipate and interpret training data better, showing greater learning and convergence. The constancy in PPL from Epoch 50 to Epoch 60 suggests that the model has achieved a saturation barrier, meaning extra training may not enhance it. This study highlights the model's excellent learning trajectory and provides guidance for training stopping places.

Table 3. Sample backtranslations generated by proposed model

Source sentence	Back translation
The plant has an annual production capacity of 3.153 MT of saleable steel	The yearly capacity of the plant to produce marketable steel is 3,153 metric tonnes
A case has been registered in Byndoor police station	At the Byndoor police station, a case has been filed
The court was hearing a plea filed by activist Harsh Mander highlighting the condition of those living in detention centres in Assam	Activist Harsh Mander brought attention to the condition of those residing in Assam detention camps during the court hearing

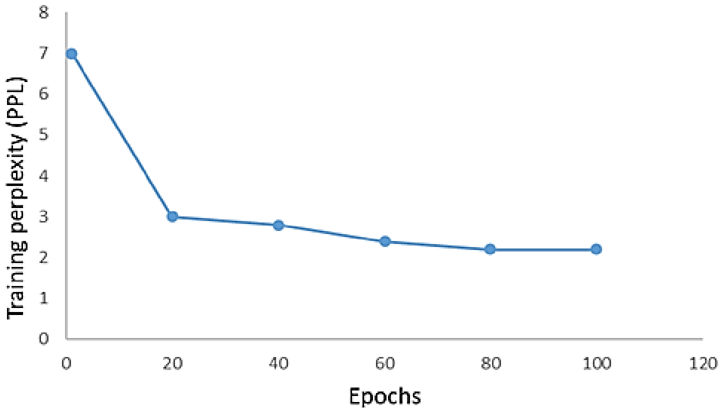

Figure 3. Training PPL vs epochs curve

The "sentence length vs. BLEU scores" graph (Figure 4) shows how backtranslation impacts English-Kannada NMT model translation quality for sentences of varying lengths. The backtranslated model consistently outperforms the baseline model in BLEU scores for sentences of all lengths, improving translation accuracy. BLEU values of 42 and 46 for shorter sentences (10 and 20) are significantly higher than baseline scores of 30 and 35 for the backtranslated model. Backtranslation effectively addresses shorter phrase issues, boosting the model's competency. Both models had closer BLEU scores for lengthier sentences (40 and 50). The improvement is moderate. This detailed research shows how backtranslation affects sentence length translation quality, providing ideas for model refinement and optimisation.

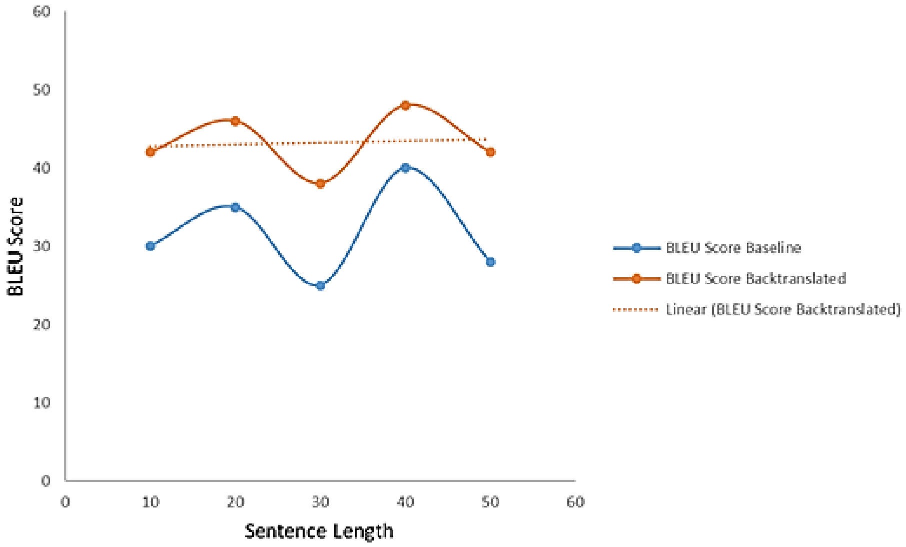

Figure 4. Sentence length vs BLEU score

Table 4 summarises the findings of our automated experimental evaluation using BLEU scores. Backtranslation improves strong baseline models like the English-to-Kannada model by 83%. The efficient

KN-EN translation model derived from the Samanantar corpus and enhanced monolingual data translations are responsible for this improvement. Backtranslation improves MT models, according to the research. English's simpler morphology makes models better at translating sentences into English than Kannada. Kannada to English translations is simpler due to this fundamental linguistic characteristic. These complex findings emphasise the significance of adapted procedures for distinct language pairs and the different effects of language factors on translation model performance.

Table 4. BLEU scores of translations

Model	BLEU Score
EN-KN (without Backtranslation)	34.5
EN-KN with Backtranslation	42.8
KN-EN	48.57

Metrics like BLEU enable user-friendly and convenient automated evaluation, but meaning may limit their ability to capture the nuanced semantic connection between a translated language and a reference text. Understanding these measurements' limits is crucial to judging translation quality. Manual evaluation is necessary for more nuanced and accurate BLEU score validation due to language's complexity. Human assessment adds a qualitative layer to automated measurements, making translation quality evaluation more comprehensive and aligned with human language intuition.

6. CONCLUSION

This study introduces a new NMT approach that emphasises English-to-Kannada translation and language pairs, including LRLs. The study examines NMT's intricacies and advancements in Kannada, a historically and culturally significant language. Kannada both a challenge and an opportunity for low-resource NMT languages. This research aims to improve NMT models that can translate accurately and contextually even with limited data. The work innovates by introducing backtranslation, a versatile technique that builds synthetic parallel corpora to improve NMT system performance in resource-limited situations. This study investigated numerous approaches to improve translation quality, make models more sensitive to domain-specific data, and address Kannada's dialectal and orthographic variances to achieve these goals. The study conducted detailed research and rigorous trials to see if backtranslation may increase translation accuracy, overcome data restrictions, and make cross-linguistic interactions more inclusive. This trip emphasises the need for context-aware NMT for Kannada and adapting translation methods to its unique linguistic features. It ensures the translation retains cultural subtleties and is accurate. NMT aims to elevate all languages, regardless of resource status, in global discourse. This work enhances our understanding of NMT in LRL contexts and emphasises context-aware linguistic variation. The baseline and backtranslated models had similar BLEU values for longer sentences, indicating a moderate translation quality improvement. Based on this conclusion, this study will use a sentence length independent hybrid deep learning model for NMT. This strategy aims to address the issues of fluctuating sentence durations and achieve more consistent improvements across phrase patterns in translation.

REFERENCES

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.

[2] K. Revanuru, K. Turlapaty, and S. Rao, “Neural machine translation of indian languages,” in *Proceedings of the 10th Annual ACM India Compute Conference*, New York, USA: ACM, Nov. 2017, pp. 11–20, doi: 10.1145/3140107.3140111.

[3] S. Dewangan, S. Alva, N. Joshi, and P. Bhattacharyya, “Experience of neural machine translation between Indian languages,” *Machine Translation*, vol. 35, no. 1, pp. 71–99, 2021, doi: 10.1007/s10590-021-09263-3.

[4] S. Sen, M. Hasanuzzaman, A. Ekbal, P. Bhattacharyya, and A. Way, “Neural machine translation of low-resource languages using SMT phrase pair injection,” *Natural Language Engineering*, vol. 27, no. 3, pp. 271–292, 2021, doi: 10.1017/S1351324920000303.

[5] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 86–96, doi: 10.18653/v1/P16-1009.

[6] R. Jiao, Z. Yang, M. Sun, and Y. Liu, “Alternated training with synthetic and authentic data for neural machine translation,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 1828–1834, doi: 10.18653/v1/2021.findings-acl.160.

[7] D. R. Beddiar, M. S. Jahan, and M. Oussalah, “Data expansion using back translation and paraphrasing for hate speech detection,” *Online Social Networks and Media*, vol. 24, 2021, doi: 10.1016/j.osnem.2021.100153.

[8] Z. Z. Hlaing, Y. K. Thu, T. Supnithi, and P. Netisopakul, “Improving neural machine translation with POS-tag features for low-

resource language pairs," *Heliyon*, vol. 8, no. 8, p. e10375, Aug. 2022, doi: 10.1016/j.heliyon.2022.e10375.

[9] H. Bąk, "Issues in the translation equivalence of basic emotion terms," *Ampersand*, vol. 11, Dec. 2023, doi: 10.1016/j.amper.2023.100128.

[10] S. Lei and Y. Li, "English machine translation system based on neural network algorithm," *Procedia Computer Science*, vol. 228, pp. 409–420, 2023, doi: 10.1016/j.procs.2023.11.047.

[11] J. Lun, J. Zhu, Y. Tang, and M. Yang, "Multiple data augmentation strategies for improving performance on automatic short answer scoring," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, pp. 13389–13396, Apr. 2020, doi: 10.1609/aaai.v34i09.7062.

[12] B. Marie, R. Rubino, and A. Fujita, "Tagged back-translation revisited: why does it really work?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, Sep. 2020, pp. 5990–5997, doi: 10.18653/v1/2020.acl-main.532.

[13] S. Shleifer, "Low resource text classification with ULMFit and backtranslation," *Computer Science*, pp. 1–9, 2019.

[14] G. Moro, N. Piscaglia, L. Ragazzi, and P. Italiani, "Multi-language transfer learning for low-resource legal case summarization," *Artificial Intelligence and Law*, Sep. 2023, doi: 10.1007/s10506-023-09373-8.

[15] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?," in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, Nov. 2016, pp. 1–6, doi: 10.1109/DICTA.2016.7797091.

[16] A. Banerjee, A. Jain, S. Mhaskar, S. D. Deoghare, A. Sehgal, and P. Bhattacharyya, "Neural machine translation in low-resource setting: a case study in English-Marathi pair," *Proceedings of Machine Translation Summit XVIII: Research Track*, vol. 1, pp. 35–47, 2021.

[17] V. Karyukin, D. Rakhimova, A. Karibayeva, A. Turganbayeva, and A. Turarbek, "The neural machine translation models for the low-resource Kazakh–English language pair," *PeerJ Computer Science*, vol. 9, Feb. 2023, doi: 10.7717/peerj-cs.1224.

[18] A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Low-resource neural machine translation improvement using source-side monolingual data," *Applied Sciences*, vol. 13, no. 2, Jan. 2023, doi: 10.3390/app13021201.

[19] P. K. Nagaraj, K. S. Ravikumar, M. S. Kasyap, M. H. S. Murthy, and J. Paul, "Kannada to English machine translation using deep neural network," *Ingenierie des Systemes d'Information*, vol. 26, no. 1, pp. 123–127, Feb. 2021, doi: 10.18280/isi.260113.

[20] T. I. Ramadhan, N. G. Ramadhan, and A. Supriatman, "Implementation of neural machine translation for English-Sundanese language using long short term memory (LSTM)," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2614.

[21] S. A. Mohamed, A. A. Elsayed, Y. F. Hassan, and M. A. Abdou, "Neural machine translation: past, present, and future," *Neural Computing and Applications*, vol. 33, no. 23, pp. 15919–15931, Dec. 2021, doi: 10.1007/s00521-021-06268-0.

[22] Y. Pan, X. Li, Y. Yang, and R. Dong, "Multi-source neural model for machine translation of agglutinative language," *Future Internet*, vol. 12, no. 6, Jun. 2020, doi: 10.3390/FI12060096.

[23] D. V Sindhu and B. M. Sagar, "Dictionary based machine translation from Kannada to Telugu," *IOP Conference Series: Materials Science and Engineering*, vol. 225, Aug. 2017, doi: 10.1088/1757-899x/225/1/012182.

[24] Y. Li, J. Li, J. Jiang, S. Tao, H. Yang, and M. Zhang, "P-transformer: towards better document-to-document neural machine translation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 31, pp. 3859–3870, 2023, doi: 10.1109/TASLP.2023.3313445.

[25] N. Goyal *et al.*, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 2022, doi: 10.1162/tacl_a_00474.

[26] A. K. Madasamy *et al.*, "Overview of the shared task on machine translation in dravidian languages," *DravidianLangTech 2022 - 2nd Workshop on Speech and Language Technologies for Dravidian Languages, Proceedings of the Workshop*, pp. 271–278, 2022, doi: 10.18653/v1/2022.dravidianlangtech-1.41.

[27] J. Waldendorf, A. Birch, B. Haddow, and A. V. M. Barone, "Improving translation of out of vocabulary words using bilingual lexicon induction in low-resource machine translation," *AMTA 2022 - 15th Conference of the Association for Machine Translation in the Americas, Proceedings*, vol. 1, pp. 144–156, 2022.

[28] B. C. Melinamath, "Handling of Auxiliaries in Kannada Morphology," in *Techno-Societal 2020*, Cham: Springer International Publishing, 2021, pp. 439–447, doi: 10.1007/978-3-030-69921-5_44.

[29] X. Liu *et al.*, "On the complementarity between pre-training and back-translation for neural machine translation," *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2900–2907, Oct. 2021.

[30] N. K. Ghanghor, P. Krishnamurthy, S. Thavareesan, R. Priyadarshini, and B. R. Chakravarthi, "IIITK@DravidianLangTech-EACL2021: offensive language identification and meme classification in Tamil, Malayalam and Kannada," *Proceedings of the 1st Workshop on Speech and Language Technologies for Dravidian Languages, DravidianLangTech 2021 at 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021*, pp. 222–229, 2021.

[31] Z. Li and L. Specia, "Improving neural machine translation robustness via data augmentation: beyond back translation," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 328–336, doi: 10.18653/v1/D19-5543.

[32] Q. Liu, M. Kusner, and P. Blunsom, "Counterfactual data augmentation for neural machine translation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 187–197, doi: 10.18653/v1/2021.naacl-main.18.

[33] M. Graça, Y. Kim, J. Schamper, S. Khadivi, and H. Ney, "Generalizing back-translation in neural machine translation," in *WMT 2019 - 4th Conference on Machine Translation, Proceedings of the Conference*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 45–52, doi: 10.18653/v1/w19-5205.

[34] M. Przystupa and M. A. -Mageed, "Neural machine translation of low-resource and similar languages with backtranslation," in *WMT 2019 - 4th Conference on Machine Translation, Proceedings of the Conference*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 224–235, doi: 10.18653/v1/w19-5431.

[35] S. Ranathunga, E. S. A. Lee, M. P. Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–37, Nov. 2023, doi: 10.1145/3567592.

[36] C. Mi, L. Xie, and Y. Zhang, "Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing," *Neural Networks*, vol. 148, pp. 194–205, Apr. 2022, doi: 10.1016/j.neunet.2022.01.016.

[37] T. T. Hailu, J. Yu, and T. G. Fantaye, "Pre-trained word embedding based parallel text augmentation technique for low-resource NMT in favor of morphologically rich languages," in *Proceedings of the 3rd International Conference on Computer Science and Application Engineering*, New York, USA: ACM, Oct. 2019, pp. 1–5, doi: 10.1145/3331453.3361309.

[38] L. Sun, C. Xia, W. Yin, T. Liang, P. S. Yu, and L. He, "Mixup-transformer: dynamic data augmentation for NLP tasks,"

Proceedings of the 28th International Conference on Computational Linguistics, pp. 3436–3440, Oct. 2020.

[39] V. Goyle, P. Krishnaswamy, K. G. Ravikumar, U. Chattopadhyay, and K. Goyle, “Neural machine translation for low resource languages,” *arXiv-Computer Science*, pp. 1-9, Apr. 2023.

[40] K. D. Chowdhury, M. Hasanuzzaman, and Q. Liu, “Multimodal neural machine translation for low-resource language pairs using synthetic data,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 33–42, doi: 10.18653/v1/w18-3405.

[41] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016,* Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 86–96, doi: 10.18653/v1/p16-1009.

[42] P. Prakash and R. M. Joshi, “Orthography and reading in Kannada: a Dravidian language,” in *Scripts and Literacy. Neuropsychology and Cognition*, Dordrecht: Springer, 1995, pp. 95–108, doi: 10.1007/978-94-011-1162-1_7.

BIOGRAPHIES OF AUTHORS

Padma Prasada obtained a Master of Technology in VLSI Design and Embedded Systems from Visvesvaraya Technological University, Belagavi, in 2013. Presently pursuing his Ph.D. in natural language processing topic with an objective of improving the NMT efficiency of low resource language at Jain University, Bangalore, India. Embedded system design and applications of machine learning are the focus of his research. A professional corporate trainer with 10+ years of teaching and training experience. 5+ research publications in international journals; attended many domestic and international conference and served on the technical program committees of 12+ international conferences. Filed an Indian patent with titled “Cognitively managing water utilization based on user's context" with application number: 202241057792. He can be contacted at email: ppjain15@gmail.com.

Dr. Malode Vishwanatha Panduranga Rao obtained his Ph.D. degree in computer science from National Institute of Technology Karnataka, Mangalore, India. He has completed a Master of Technology in computer science and Bachelor of Engineering in Electronics and Communication Engineering. He is currently working as Professor in Jain (Deemed to be University) Bengaluru, India. His research interests are in the field of real-time and embedded systems - internet of things. He has published various research papers in journal and conferences across India, also in the IEEE international conference in Okinawa, Japan (visited) 2008. He has authored two reference books on Linux Internals. He is the life member of Indian Society for Technical Education and IAENG. Now from past three years, he has published 12 indian patents and three patents are stepping towards grant status. One research scholar under his guidance was awarded a Ph.D. degree. He can be contacted at email: r.panduranga@jainuniversity.ac.in.