Enhancing Low-Resource NMT with a Multilingual Encoder and Knowledge Distillation: A Case Study

Aniruddha Roy[1], Pretam Ray[1], Ayush Maheshwari[2], Sudeshna Sarkar[1], Pawan Goyal[1]
[1] Indian Institute of Technology Kharagpur [2] Vizzhy Inc, Bengaluru
{aniruddha.roy, pretam.ray}@iitkgp.ac.in, ayush.maheshwari@vizzhy.com,
{sudeshna, pawang}@cse.iitkgp.ac.in

Abstract

Neural Machine Translation (NMT) remains a formidable challenge, especially when dealing with low-resource languages. Pre-trained sequence-to-sequence (seq2seq) multi-lingual models, such as mBART-50, have demonstrated impressive performance in various low-resource NMT tasks. However, their pre-training has been confined to 50 languages, leaving out support for numerous low-resource languages, particularly those spoken in the Indian subcontinent. Expanding mBART-50's language support requires complex pre-training, risking performance decline due to catastrophic forgetting. Considering these expanding challenges, this paper explores a framework that leverages the benefits of a pre-trained language model along with knowledge distillation in a seq2seq architecture to facilitate translation for low-resource languages, including those not covered by mBART-50. The proposed framework employs a multilingual encoder-based seq2seq model as the foundational architecture and subsequently uses complementary knowledge distillation techniques to mitigate the impact of imbalanced training. Our framework is evaluated on three low-resource Indic languages in four Indic-to-Indic directions, yielding significant BLEU-4 and chrF improvements over baselines[1]. Further, we conduct human evaluation to confirm effectiveness of our approach.

1 Introduction

Neural Machine Translation (NMT) models (Bahdanau et al., 2016; Vaswani et al., 2017; Liu et al., 2020a; Khandelwal et al.) have shown impressive results on benchmark datasets, mainly containing large amounts of parallel data. However, these models face challenges when applied to low-resource languages or languages with rich and diverse morphology. Previous approaches have leveraged pre-trained models trained on extensive corpora (Weng et al., 2019; Wang et al., 2022; Liu et al., 2020b, 2021; Haddow et al., 2022; Roy et al., 2023, 2022) to address these limitations.

Pre-trained multilingual seq2seq-based models based on an encoder-decoder framework such as mBART-50 (Liu et al., 2020b) have been successfully used for various low-resource NMT tasks. Despite being pre-trained with 50 languages, it needs more support for numerous low-resource languages. Expanding the capabilities of mBART-50 to encompass new languages entails a cumbersome process involving the collection of substantial amounts of monolingual data and the execution of pre-training with denoising objectives after initializing mBART-50. This process is time-consuming and may decrease performance on the initial 50 languages when incorporating new ones, a phenomenon known as catastrophic forgetting (French, 1999).

In contrast, encoder-based pretrained model XLM-R (Conneau et al., 2020) is designed to accommodate 100 languages, making it suitable for a wide range of low-resource cross-lingual Natural Language Understanding (NLU) tasks. Both cross-lingual and Machine Translation (MT) functionalities share certain similarities. In cross-lingual scenarios, training and evaluation occur across different languages, while MT systems process input in one language and produce output in another. This distinction prompts several experimental research questions, including: 1) How does an XLM-R based NMT model perform on low-resource morphologically rich languages, particularly those not covered by mBART-50? 2) Given that low-resource NMT may be affected by training imbalances leading to performance degradation, can the application of knowledge distillation further enhance the results?

To address the two aforementioned experimen-

[1]Our code is publicly available at https://github.com/raypretam/Two-step-low-res-NMT

Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 64–73

tal research questions, we utilize our base model, which follows a seq2seq framework. Here, we initialize the encoder with the multilingual pretrained model XLM-R large, while decoder layers are initialized from scratch, we call this base approach as XLM-MT. Similar frameworks have been explored in previous studies (Zhu et al., 2020; Li et al., 2023), with our approach sharing similarities with (Chen et al., 2022), who employed it for zero-shot cross-lingual NMT tasks and froze the embedding layers. However, our base approach differs in considering only decoder training. Thereafter, we apply complementary knowledge distillation (CKD) (Shao and Feng, 2022) to the base XLM-MT model to address training imbalances. The objective of this complementary knowledge distillation is to train the student model with knowledge which complements the teacher model and avoid knowledge forgetting, and we refer to this as XLM-MT+CKD. We empirically evaluate our model across three Indic languages and observe significant improvement in BLEU and chrF scores. Finally, we use human evaluation to assess the fluency, relatedness, and correctness of our output. Our contributions are as follows:

1. We repurpose the XLM-based seq2seq framework in conjunction with a complementary knowledge distillation approach to effectively design an NMT model for low-resource MT tasks. To the best of our knowledge, we are the first to integrate these two approaches effectively for NMT tasks.

2. We conduct comprehensive experiments on three Indian languages in four directions that are not included in mBART-50 and demonstrate the significance of our approach in enhancing translation results.

3. We also perform a detailed analysis of the results, including human evaluation and error analysis, for our proposed model.

2 Methodology

Given a source language sentence $X = (x_1, x_2, \ldots, x_S)$, and its corresponding target language translation $Y = (y_1, y_2, \ldots, y_T)$, an NMT model is trained to predict the translated sequence Y' using the maximum log-likelihood estimation (MLE) objective. The probability of predicting the target sequence Y' is computed as $p(Y'|X;\theta) = \prod_{t=1}^{T} p(y_t|y_{0:t-1}, x_{1:S}, \theta)$, where θ represents the model parameters.

2.1 Base Model (XLM-MT)

We initialize encoder layers and encoder embeddings with an unsupervised pre-trained multilingual model, XLM-R large (Conneau et al., 2020) which is trained using masked language model objective. Then, we train the decoder from scratch while freezing the encoder parameters. During training, decoder parameters are learned with an MLE objective. The underlying assumption is that the pre-trained encoder parameters have already learnt a multilingual representation of the source language. As a result, only the decoder is trained using MLE objective while leveraging the encoder embeddings learned by the pre-trained model. $\mathcal{L}_{\theta_{dec}} = \sum_{(X,Y)\in D} \log P(Y|X;\theta_{dec})$ where X and Y represents the source target sentence respectively from the dataset D. The parameter θ_{dec} refers to the parameters of the decoder layers and embedding.

Algorithm 1 Complementary Knowledge Distillation

Input: Training data D, the number of teachers n.
Output: Student model S.
Initialize S and teacher models ($T_{1:n}$) with the base model, XLM-MT.
while not converge **do**
    randomly divides the training data D in mutually exclusive $n+1$ subsets $D_1, D_2, ...D_{n+1}$
    for $t = 1$ to $n+1$ **do**
        for $i = 1$ to n **do**
            Train T_i on $D_{O(i,t)}$
        end for
        Train S on D_t using Eq 3
    end for
    for $i = 1$ to n **do**
        $T_i \leftarrow S$ (At the end of each epoch, reinitialize teacher models with the student model:)
    end for
end while
return student model S

2.2 Complementary Knowledge Distillation

Imbalances in training data lead to performance degradation in low-resource NMT due to catastrophic knowledge forgetting (LeCun et al., 2002; Shao and Feng, 2022). We leverage complementary knowledge distillation (CKD) technique (Shao and Feng, 2022) to overcome this problem in low-

resource MT. In CKD, n teacher models and a student model S are trained in a complementary manner such that S learns from new training samples while teacher models dynamically provide complementary early samples knowledge to the S. In our case, both teacher and student models are initialized with the parameters of our base model, XLM-MT.

We divide the training set D into $n+1$ mutually exclusive subsets for each epoch. The student model S sequentially learns from D_1 to D_{n+1} while the teacher models learn from all data splits except D_t. To determine the training data for the teacher models at timestep t, we utilize an ordering function, as shown in Eq 1 (Shao and Feng, 2022). This ordering function covers all data splits except D_t, ensuring that the teacher models complement the student model.

$$O(i,t) = \begin{Bmatrix} i+t, & i+t \leq n+1 \\ i+t-n-1, & i+t > n+1 \end{Bmatrix} \quad (1)$$

where, $i \in \{1,2,\ldots,n\}$ and $t \in \{1,2,\ldots,n+1\}$

In the process of word-level knowledge distillation, the student model S benefits from an additional supervision signal, aligning its outputs with the probability outputs of the teacher model T.

$$\mathcal{L}_{KD}(\theta) = -\sum_{t=1}^{T}\sum_{k=1}^{|V|}\sum_{i=1}^{n}\frac{q_i(y_t = k|y_{<t}, X)}{n} \times \log\, p(y_t = k|y_{<t}, X, \theta) \quad (2)$$

where $|V|$ denotes the number of classes, p denotes the prediction of student and q_i is the prediction of teacher model T_i. To balance the distillation loss and the cross-entropy loss, we introduce a hyperparameter α for interpolation. Finally, the overall objective function is

$$\mathcal{L}(\theta) = \alpha \cdot \mathcal{L}_{KD}(\theta) + (1-\alpha) \cdot \mathcal{L}_{NLL}(\theta) \quad (3)$$

We employ a reinitialization technique (Zhang et al., 2018; Zhu et al., 2018) to facilitate two-way knowledge transfer. After each epoch, we reset the parameters of the teacher models using those of the student model. This reinitialization ensures that the student and teachers begin each epoch with identical settings. We present the training procedure for CKD in Algorithm 1. We apply CKD to our base model in the following 2-step process.

Step 1 - Initialization: In this step, we initialize both the student and teacher models with the model obtained after the first step training (*c.f.*, Section 2.1). This initialization ensures that the student model benefits from the knowledge acquired during the initial decoder training.

Step 2 - CKD: In this step, we apply the complementary KD technique (*c.f.*, Section 2.2) which enables the model to benefit from the transfer of complementary knowledge.

3 Experimental Setup

Dataset: For our experiments, we specifically select three Indic languages, namely Kannada, and Punjabi that are not included in mBART-50, to assess the effectiveness of our approach. We use the Samanantar dataset (Ramesh et al., 2022) for training all our NMT models which contains parallel sentences for 11 Indic language pairs. We consider three languages in 4 directions, namely Hindi-Kannada, Kannada-Hindi, Kannada-Punjabi, and Punjabi-Kannada, containing 2.1 million and 1.1 million parallel sentences respectively. We use the FLORES-200 (Team, 2022) containing 997 and 1012 sentences as our validation and test set respectively.

Implementation Details: We implement our approach using the Fairseq Toolkit (Ott et al., 2019). We use Adam optimizer (Kingma and Ba, 2017) with $\beta_1 = 0.9$ and $\beta = 0.98$. Following the work by Chen et al. (2021), we use learning rates $5e-3$ and $1e-3$ for the base model and CKD, respectively. We set maximum updates of 200K for the base model training and 40K for the CKD. We use 12 layers with 16 attention heads in the decoder. We use the 'Large' variant of XLM-R that has 550 million parameters for our experiments. We set the number of teachers to 1 and $\alpha = 0.95$. We set batch size = 32k, and used beam size = 5 throughout our experiments, and following Shao and Feng (2022) we averaged the last five checkpoints. We use BLEU-4 (Papineni et al., 2002) and chrF (Popović, 2015) score to evaluate our approach. All the models have been trained on single A100 GPUs. None of the training methods consumed more than 96 hours.

Baselines We employ various baseline models for comparison with our approach. To ensure a fair assessment, we train all baseline models using identical training data and assess their performance on the Flores dataset.

Transformer (Vaswani et al., 2017): We uti-

Model	hi-kn	kn-hi	kn-pa	pa-kn	hi-kn	kn-hi	kn-pa	pa-kn
	BLEU				**chrF**			
Transformer	3.60	7.61	1.39	1.04	37.40	34.09	24.19	25.75
Sequence-KD (Kim and Rush, 2016)	4.23	7.88	1.71	1.08	37.43	34.23	24.31	25.89
mBERT-KD (Chen et al., 2020)	4.73	8.67	2.01	1.31	37.47	34.67	24.43	26.22
Selective KD (Wang et al., 2021)	5.35	8.08	2.24	1.19	39.23	35.02	24.57	26.78
Transformer+CKD	4.51	8.89	3.23	1.98	38.54	35.13	24.54	27.01
mBERT-MT (Zhu et al., 2020)	4.98	10.23	3.78	4.17	38.98	35.56	25.01	29.68
SixTp (Chen et al., 2022)	7.01	10.80	6.14	5.45	40.98	35.74	27.62	32.47
XLM-MT (base)	6.08	8.75	6.01	2.98	40.38	35.38	27.12	28.11
XLM-MT + CKD (ours)	**9.15**	**11.46**	**7.23**	**6.43**	**41.11**	**35.88**	**29.12**	**33.98**

Table 1: Performance (BLEU-4 and chrF scores) of our model along with seven baseline models on the FLORES-200 dataset on 3 languages in 4 directions between Indic languages: Hindi ('hi'), Kannada ('kn'), Punjabi ('pa'). We provide additional human evaluation results in Table 4.

lize a standard transformer-based encoder-decoder model, employing six layers for both the encoder and decoder.

Word-level Knowledge Distillation (Kim and Rush, 2016) is a conventional method applied to enhance NMT results by distilling knowledge at the word level.

Sequence-level Knowledge Distillation (Kim and Rush, 2016) is a conventional knowledge distillation technique applied to enhance NMT results by distilling knowledge at the sequence level.

BERT-KD (Chen et al., 2020) is a knowledge extracted from a fine-tuned BERT model is transferred to NMT models.

Selective KD (Wang et al., 2021) refers to the process of distilling and transferring specific, relevant knowledge from a teacher model to a student model. Instead of transferring all the knowledge indiscriminately, this approach involves selecting and distilling the most valuable and informative aspects of the teacher model's knowledge.

mBERT-MT(Zhu et al., 2020), integrates BERT into the NMT process. Initially, BERT is employed to extract representations for an input sequence. Subsequently, these representations are fused with each layer of the NMT model's encoder and decoder using attention mechanisms.

sixTp (Chen et al., 2022) is a sequence-to-sequence (seq-to-seq) model. In its initialization, the encoders are initialized with the XLM-R large model, while the decoder is initialized randomly. The model undergoes a two-stage fine-tuning process. In the initial stage, the encoder layers are frozen, and fine-tuning is performed on the decoders. Subsequently, in the second stage, the model is trained in an end-to-end fashion.

4 Results

Table 1 presents the BLEU-4 and chrF results for Hindi to Kannada, Kannada to Punjabi in both directions. It is noteworthy that Hindi, and Punjabi belong to the Indo-Aryan language family, while Kannada belongs to the Dravidian family. We compare our results against seven competitive baselines, namely, vanilla transformer, knowledge distillation techniques, transformer with CKD, two step training techniques using mBERT and SixTp, which is XLM-R based model. We observe that XLM-MT + CKD achieves BLEU scores within the range of (6.43 to 11.46) consistently surpassing the baselines. We observe an average improvements of 1.22-5.15 BLEU scores across all language pairs. We also present chrF scores in Table 1. Notably, XLM-MT+CKD consistently demonstrates its superiority, outperforming all the baselines with averages of 0.82-4.66 chrF score, across all language pairs. Further, we conduct human evaluation to assess the fluency, relatedness and correctness of the generated text. We present human evaluation results of sixTp and our model, XLM-MT+CKD in Table 4.

We also investigate various variants of our model to validate the effectiveness of our architecture and present results in Table 2. Additionally, we conduct comprehensive error analysis in Section 7.

5 Analysis

How would the method perform with the languages that mBART-50 supported? In addition to the language pairs outlined in Section 3, we extend our exploration to include language pairs supported by mBART-50, facilitating effec-

Model	hi-kn	kn-hi	kn-pa	pa-kn
Transformer	3.60	7.61	2.39	1.04
$\text{Enc}^{\text{XLM-R}}_{\text{train}}$ + Dec	4.72	8.32	5.75	2.11
$\text{Enc}^{\text{XLM-R}}_{\text{no-train}}$ + Dec	6.08	8.75	6.01	2.98
SixTp	7.01	10.80	6.14	5.45
XLM-MT + CKD	**9.15**	**11.46**	**7.23**	**6.43**

Table 2: Performances (BLEU-4 scores) of our model along with its variants. The score in **bold** shows the best scores for the corresponding language pair. Enc + Dec refers to the transformer model without XLM initialization. $\text{Enc}^{\text{XLM-R}}_{\text{train}}$ + Dec refers to joint training of XLM-R based encoder and decoder. $\text{Enc}^{\text{XLM-R}}_{\text{no-train}}$+ + Dec refers to only decoder training.

Model	hi-bn	mr-hi	hi-te
mBART-50	**9.25**	**17.06**	9.35
XLM-MT	8.13	15.78	11.56
XLM-MT + CKD	8.67	15.81	**12.01**

Table 3: Performances BLEU-4 of our model along with mBART-50 model on the FLORES-200 dataset for the translation between Indic languages: Hindi ('hi'), Telugu ('te'), and Bengali ('bn').

tive comparisons with the mBART-50 model. We extracted three language pairs from the Samantar dataset—namely, Hindi-Bengali, Telugu-Hindi, and Marathi-Hindi—and compared our approach with mBART-50. We present the results in Table 3. mBART-50 achieves BLEU-4 scores of 9.35 and 17.06 for the language pairs of Hindi-Bengali and Marathi-Hindi respectively, surpassing the performance of XLM-MT+CKD model. For the Hindi-Telugu pair, our model XLM-MT+CKD achieves better performance than mBART-50.

5.1 Analysis of different model variants

The aim of this analysis is to assess the effectiveness of our model with different approaches in addressing the challenges of machine translation, particularly for low-resource and morphologically rich languages. The obtained BLEU scores are presented in Table 2.

Enc + Dec: To assess the importance of pre-training initialization in the encoder, we compare the performance of XLM-MT, which is initialized with XLM-R large, against a randomly initialized model. We observe that the encoder initialized with XLM-R large produces better performance than the randomly initialized encoder.

$\text{Enc}^{\text{XLM-R}}_{\text{train}}$ + Dec, $\text{Enc}^{\text{XLM-R}}_{\text{no-train}}$ + Dec: To analyze the effectiveness of the two-stage training process employed in XLM-MT, we experiment with two different settings: (a) training encoder and decoder jointly (the second stage), denoted as $\text{Enc}^{\text{XLM-R}}_{\text{train}}$ + Dec, and (b) only training the decoder (the first stage), denoted as $\text{Enc}^{\text{XLM-R}}_{\text{no-train}}$ + Dec. From Table 2, we clearly see the effectiveness of two-stage training compared to only one of these stages across all language pairs.

6 Human Evaluation

We follow a procedure similar to previous studies (Chi et al., 2019; Maurya et al., 2021) to assess the quality of translated sentences in three Indic languages in four Indic-to-Indic language pairs. We randomly selected 50 test data-points for each language pair for evaluation. Three key metrics are used to evaluate the translated sentences: fluency, relatedness, and correctness. Fluency refers to the smoothness and coherence of the generated text, evaluating how well the sentences flow and adhere to grammatical rules. Relatedness measures how well the translated sentences are connected to the given ground truth sentences and capture its key information. Correctness assesses the accuracy and appropriateness of the translated sentences in terms of their meaning and semantics. We present the translated sentences (randomly shuffled) from two models XLM-MT and XLM-MT+CKD to three language experts for each language pair. The selected 12 experts are well versed in the corresponding target language including English. The experts attained a minimum of graduate degree in English and have native proficiency in the target language. The experts are informed about the task and were renumerated as per industry standard norms. The experts rated the sentences on a 5-point scale, with 1 indicating very bad and 5 indicating very good, for each of the three metrics. The final numbers are in Table 4. These are calculated by averaging all

Metric	hi-kn	kn-hi	kn-pa	pa-kn
SixTp				
Fluency	3.13	3.71	2.47	1.97
Correctness	3.04	3.68	2.40	**1.87**
Relatedness	3.18	3.75	2.33	1.92
XLM-MT + CKD				
Fluency	**3.23**	**3.75**	**2.53**	**2.01**
Correctness	**3.71**	**3.92**	**2.47**	1.85
Relatedness	**3.43**	**3.78**	**2.51**	**1.97**

Table 4: Human evaluation results of our approach sixTp and XLM-MT+ CKD for three languages in four directions. The three metrics are Fluency, Relatedness, and Correctness, respectively.

the experts' responses for each parameter. The annotation experts received compensation according to industry standards for their work. We briefed them on the objectives and explicit usage of their annotations

7 Case Study

Table 5 presents several example sentences and their translations by our proposed approach. Notably, there are specific issues with reference 1 in the XLM-MT translations. In reference 1, XLM-MT incorrectly translates the sentence using a wrong gender concept, whereas XLM-MT+CKD translates correctly. Regarding the Kannada sentence in reference 2, the XLM-MT and XLM-MT+CKD approaches provide a correct and meaningful translation, albeit with some paraphrasing.

8 Related Work

Neural Machine Translation (NMT) aims to translate a given source sentence into a target sentence. Typically, an NMT model comprises an encoder, a decoder, and an attention mechanism. The encoder transforms the input sequence into hidden representations while the decoder maps these representations to the target sequence. The attention mechanism, pioneered by (Bahdanau et al., 2016), enhances alignment between words in the source and target languages. Different architectures can be employed for the encoder and decoder, including LSTM (Long Short-Term Memory), CNN (Convolutional Neural Network), and Transformer. The Transformer architecture, introduced by (Vaswani et al., 2017), consists of three sublayers. Transformer has demonstrated state-of-the-art performance in NMT tasks (Barrault et al., 2019).

Prior studies (Imamura and Sumita, 2019; Conneau and Lample, 2019; Yang et al., 2022; Weng et al., 2019; Ma et al., 2020; Zhu et al., 2020) have investigated the integration of pre-trained language encoders into NMT models to bolster supervised translation performance. (Zhu et al., 2020) introduce a BERT-fused model that extracts representations from input sentences and integrates them into the encoder and decoder using attention mechanisms. Recent research (Song et al., 2019) focuses on developing and refining encoder-decoder-based multilingual trained language models for NMT. (Liu et al., 2020c) present mBART, a Transformer-based encoder-decoder model explicitly tailored for NMT applications. Wei et al. finetune the multilingual encoder-based model for low-resource NMT, and they focus on improving the MPE for a more universal representation across languages. (Chen et al., 2021, 2022) have examined a two-stage framework utilizing an encoder-based multilingual language model for zero-shot neural machine translation.

Numerous studies in NMT have incorporated the Knowledge Distillation (KD) framework. (Kim and Rush, 2016) introduced word-level KD for NMT and later proposed sequence-level KD to enhance overall performance. Investigating the efficacy of various token types in KD, (Wang et al., 2021) suggested strategies for selective KD. (Wu et al., 2020) successfully transferred internal hidden states from teacher models to students, achieving positive results. Various KD approaches have also been employed in non-auto-regressive Machine Translation tasks to enhance outcomes. (Gu et al., 2018) improved non-autoregressive model performance by distilling information from an autoregressive model. (Zhou et al., 2021) conducted systematic experiments highlighting the importance of knowledge distillation in training non-auto-

1. Source (Kannada):	Udaharanage, obbaru, motaru karugale rastegala abhivrd'dhige mula karana endu helabahudu. **Translation:** For example, one could say that motor cars were the root cause of the development of roads.
Reference (Punjabi):	Udaharana vajon, koi kahi sakada hai ki motara kara sarakan nu zaruri taura'te vikas vala lai jandi hai. **Translation:** For example, one could say that the motor car essentially leads to development of roads.
XLM-MT+CKD:	Udaharana vajon, koi kahi sakada hai ki motara kara sarakan nu zaruri taura'te vikas vala lai jandi hai. **Translation:** For example, one could say that the motor car essentially leads to development of roads.
2. Source (Hindi) :	kuchh any visheshagyon kee tarah, unhen is baat par sandeh hai ki kya madhumeh ko theek kiya ja sakata hai, yah dekhate hue ki in nishkarshon kee un logon ke lie koee praasangikata nahin hai jinhen pahale se hee taip 1 madhumeh hai. **Translation:** Like some other experts, he is skeptical about whether diabetes can be cured, noting that these findings have no relevance to people who already have type 1 diabetes.
Reference (Kannada):	Madhumehavannu gunapadisalu sadhyave emba bagge itare itara kelavu tajñarante avaru kuda sansaya vyaktapadisuttare, ī sansodhanegaḷu īgagale ṭaip 1 madhumeha hondiruva janarige yavude prayojanagaḷannu nīdilla. **Translation:** Like some other experts, he doubts whether diabetes can be cured, because these conclusions are not practical for people who have previously had type 1 diabetes.
XLM-MT+CKD:	Itara kelavu tajñarante, avaru madhumehavannu gunapadisabahude endu sansayapaduttare, ekendare ī tīrmanagaḷu ī hinde ṭaip 1 madhumeha hondiruva vyaktigaḷige prayogikavagiruvudilla. **Translation:** Like some other experts, he doubts whether diabetes can be cured, because these conclusions are not practical for people who have previously had type 1 diabetes.

Table 5: Sample outputs generated from our proposed approach, where the target languages' source language and translations are specified for each reference.

regressive models, showing its ability to reduce dataset complexity and help model variations in output data. In the realm of multilingual NMT, (Baziotis et al., 2020) used language models as instructors for low-resource NMT models. (Chen et al., 2020) extracted knowledge from fine-tuned BERT and transferred it to NMT models. Furthermore, (Feng et al., 2021) and (Zhou et al., 2021) employed KD to introduce forward-looking information into the teacher-forcing training of NMT models.

9 Conclusion

In this paper, we empirically explored the methods for improving low-resource NMT, particularly for Indic languages. We investigated several strategies for initialization of encoder and decoder, along with the knowledge distillation techniques. We conducted experiment on three low-resource Indic languages in four Indic-to-Indic directions belonging to two language families, specifically focusing on those not covered by mBART-50. Further, we perform additional analysis on languages supported by mBART-50 and high-resource language pairs.

Limitations

A limitation of this study is the increased training time required for the XLM-MT+CKD model due to its addition of complementary knowledge distillation. Furthermore, our validation is limited to low-resource machine translation tasks, although seq2seq models have the potential to be utilized

for a wide range of generation tasks, including Question Generation and Summarization in both monolingual and cross-lingual contexts.

Acknowledgement

This work was supported in part by the National Language Translation Mission (NLTM): Bhashini project by the Government of India.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online. Association for Computational Linguistics.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2022. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–157, Dublin, Ireland. Association for Computational Linguistics.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7893–7905, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2019. Cross-lingual natural language generation via pre-training.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Yang Feng, Shuhao Gu, Dengji Guo, Zhengxin Yang, and Chenze Shao. 2021. Guiding teacher forcing with seer forcing for 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 (Volume 1: Long Papers)*, pages 2862–2872, Online. Association for Computational Linguistics.

Robert French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3:128–135.

J Gu, J Bradbury, C Xiong, VOK Li, and R Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations (ICLR)*.

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

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. 2002. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer.

Bryan Li, Mohammad Sadegh Rasooli, Ajay Patel, and Chris Callison-Burch. 2023. Multilingual bidirectional unsupervised translation through multilingual finetuning and back-translation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 16–31.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020a. Norm-based curriculum learning for neural machine translation. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 427–436, Online. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the copying behaviors of pre-training for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4265–4275, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020c. Multilingual denoising pre-training for neural machine translation.

Shuming Ma, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, Xia Song, Arul Menezes, and Furu Wei. 2020. Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. Zmbart: An unsupervised cross-lingual transfer framework for language generation.

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. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Aniruddha Roy, Isha Sharma, Sudeshna Sarkar, and Pawan Goyal. 2023. Meta-ed: Cross-lingual event detection using meta-learning for indian languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

Aniruddha Roy, Rupak Kumar Thakur, Isha Sharma, Ashim Gupta, Amrith Krishna, Sudeshna Sarkar, and Pawan Goyal. 2022. Does meta-learning help mBERT for few-shot question generation in a cross-lingual transfer setting for indic languages? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4251–4257, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chenze Shao and Yang Feng. 2022. Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2023–2036.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for 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 (Volume 1: Long Papers)*, pages 6456–6466.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael Lyu. 2022. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2591–2600, Dublin, Ireland. Association for Computational Linguistics.

Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. 2019. Acquiring knowledge from pre-trained model to neural machine translation.

Yimeng Wu, Peyman Passban, Mehdi Rezagholizade, and Qun Liu. 2020. Why skip if you can combine: A simple knowledge distillation technique for intermediate layers.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang, and Lei Li. 2022. Towards making the most of bert in neural machine translation.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4328.

Chunting Zhou, Graham Neubig, and Jiatao Gu. 2021. Understanding knowledge distillation in non-autoregressive machine translation.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation.

Xiatian Zhu, Shaogang Gong, et al. 2018. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31.