# NICT-AI4B's Submission to the Indic MT Shared Task in WMT 2023

**Raj Dabre**
NICT, Japan
raj.dabre@nict.go.jp

**Jay Gala**
AI4Bharat, India
jaygala24@gmail.com

**Pranjal A. Chitale**
AI4Bharat, IIT Madras, India
cs21s022@cse.iitm.ac.in

## Abstract

In this paper, we (Team NICT-AI4B) describe the MT systems that we submit to the Indic MT task in WMT 2023. Our primary system consists of 3 stages: Joint denoising and MT training using officially approved monolingual and parallel corpora, backtranslation and, MT training on original and backtranslated parallel corpora. We observe that backtranslation leads to substantial improvements in translation quality up to 4 BLEU points. We also develop 2 contrastive systems on unconstrained settings, where the first system involves fine-tuning of IndicTrans2 Data Augmentation (DA) models on official parallel corpora and seed data used in Gala et al. (2023), and the second system involves a system combination of the primary and the aforementioned system. Overall, we manage to obtain high-quality translation systems for the 4 low-resource North-East Indian languages of focus.

## 1 Introduction

The increasing online presence[1] of the Indian population along with the economic growth[2] of India has necessitated the development of translation systems for Indian languages. There have been substantial efforts towards collecting monolingual and parallel corpora, as well as developing machine translation systems using them (Ramesh et al., 2022; Doddapaneni et al., 2023). Most recently, IndicTrans2 (Gala et al., 2023), an MT system, and its accompanying parallel corpus BPCC, were released. This corpus covers all 22 Indian languages covered in the 8th schedule[3] of the Constitution of India.

While IndicTrans2 has achieved comparable or better results compared to existing systems like NLLB (Costa-jussà et al., 2022), a major limitation is that there is no specific focus on language subgroups. One of such subgroups is the North-East Indian languages, which this shared task focuses on. The task focuses on translation to/from English and the following 4 North-East Indian languages: Assamese, Manipuri, Mizo and Khasi. We submit constrained as well as unconstrained MT systems for the 8 translation directions in this task. For further details on the shared task, kindly refer to (Pal et al., 2023).

We leveraged ideas such as joint multilingual denoising and MT training followed by back-translation at scale. First, due to the small size of the official parallel corpora, we utilized available and permitted monolingual corpora for all languages involved and trained on a combination of text-infilling and MT objectives to train an initial MT system. We used this system to generate large back-translated corpora, which were combined with the official parallel corpora to train the final primary system. The back-translated corpora, due to their scale, led to improvements up to 4 BLEU as measured on the development set. We also submitted two contrastive systems: the first one was obtained via fine-tuning IndicTrans2 DA models (Gala et al., 2023) and the second one was a system combination of the primary and the aforementioned contrastive system. We observed that our first contrastive system outperformed the primary for some language pairs due to the utilization of a strong pretrained MT model as initialization and additional high-quality data being used to fine-tune them. As for our second contrastive system, we observed improvements for directions where there was a small performance gap between the primary and the first contrastive system.

## 2 Related Work

Our submissions leverage ideas from topics such as multilingualism (Dabre et al., 2020), denoising pre-

[1] https://datareportal.com/reports/digital-2023-india

[2] https://www.cnbc.com/2023/07/26/imf-raises-2023-economic-growth-forecast-for-india.html

[3] https://www.mha.gov.in/sites/default/files/EighthSchedule_19052017.pdf

*Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 941–949
December 6–7, 2023. ©2023 Association for Computational Linguistics

training (Lewis et al., 2020; Dabre et al., 2022), backtranslation (Sennrich et al., 2016), transfer learning (Zoph et al., 2016) and system combination (Heafield and Lavie, 2010, 2011).

The North-East Indian languages of focus in this shared task are all low-resource languages, and transfer learning via multilingualism is a reliable solution in this case. Transfer learning can be achieved by fine-tuning a pre-trained model (Zoph et al., 2016) but this involves two stages. On the other hand, multilingual training (Johnson et al., 2017; Dabre et al., 2020) involves implicit transfer via joint training. We explore both strategies when developing our systems.

Backtranslation (Sennrich et al., 2016) involves taking intermediate translation systems and translating monolingual corpora into another language. The synthetic-source and original target parallel corpora, can typically be orders of magnitude larger than the parallel corpora used to train the intermediate systems and when used at scale, such backtranslated corpora are known to help improve translation quality (Edunov et al., 2018) and therefore we attempt to use as much monolingual corpora as possible for backtranslation. While there are iterative backtranslation (Hoang et al., 2018) strategies where the process of model training and backtranslation is performed repeatedly, their computational complexity makes them a less attractive solution to us.

An alternative to backtranslation is denoising pretraining using monolingual corpora (Lewis et al., 2020; Dabre et al., 2022) and when combined with MT training as a joint objective (Kamboj et al., 2022) is known to significantly improve MT quality. Since backtranslation and denoising pre-training are known to be orthogonal (Liu et al., 2020), we leverage the joint denoising and MT training approach only for intermediate models which are used for backtranslation.

## 3 Our Systems

We submit 3 systems, one primary (constrained) and two contrastive (unconstrained).

### 3.1 Primary System

To create our primary system, we do the following:

1. Augment official monolingual data with the external monolingual corpora for the 4 North-East Indian languages and English.
2. Train a many-to-many encoder-decoder Transformer (Vaswani et al., 2017) model with the joint text-infilling (denoising) (Lewis et al., 2020; Dabre et al., 2022) and the MT objectives, using the augmented monolingual and official parallel corpora, respectively. To prevent the model from over-adapting to the infilling objective, we oversample the parallel corpora.
3. Use the aforementioned model to backtranslate the monolingual corpora.
4. Combine the backtranslated and official parallel corpora while oversampling the latter, and then train a many-to-many MT model.

### 3.2 Contrastive System #1

For our contrastive system, we investigate the potential of leveraging strong pretrained IndicTrans2 DA En-Indic and Indic-En models (Gala et al., 2023) for adaptation to newer languages and domains. It is important to note that we utilize off-the-shelf IndicTrans2 DA models trained on large-scale general-purpose corpora comprising both original and augmented backtranslated data. We refrain from using final models that are already fine-tuned with the same seed data that we use in this study, making them redundant in this context.

A trivial solution would be to adapt IndicTrans2 DA models to target languages and domains. However, this solution can often lead to catastrophic forgetting of translation ability on existing languages and domains. As a result, we explore approaches that satisfy two-fold objectives: 1) maximize the performance on a specific set of target languages and domains in the context of WMT shared task and 2) retain the overall performance on existing languages supported by the IndicTrans2 DA models. Our experiments involve a comparison of either of the approaches for adaptation of IndicTrans2 DA models to a specific set of few known and few unseen languages. We explore the following approaches:

1. **A1**: Direct fine-tuning of IndicTrans2 DA models on a combination of official parallel corpora and seed data used in Gala et al. (2023) for a set of languages under consideration for WMT Indic MT shared task.
2. **A2**: Direct fine-tuning of IndicTrans2 DA models on a combination of official parallel corpora and seed data used in Gala et al.

| Parallel | | Monolingual | | |
|---|---|---|---|---|
| lang pair | # lines | lang | # lines (Org) | # lines (Aug) |
| as-en | 50K | as | 2.6M | 8.05M |
| mz-en | 50K | mz | 1.9M | 8.8M |
| kha-en | 24K | kha | 0.18M | 0.73M |
| mni-en | 21.6K | mni | 2.14M | 2.20M |
| | | en | 0 | 20M |

Table 1: Parallel and monolingual data statistics. For the primary system, we only use the organizers' provided parallel data. We use monolingual data provided by the organizers as well as from Gala et al. (2023) and indicate the organizers' (Org) and augmented (Aug) sizes.

(2023) for all the languages supported by the IndicTrans2 DA models and WMT Indic MT shared task.

3. **A3**: Two-stage fine-tuning of IndicTrans2 DA models on 1) a combination of official parallel corpora and seed data used in Gala et al. (2023) for a set of languages under consideration for WMT Indic MT shared task, followed by 2) on a combination of official parallel corpora and seed data used in Gala et al. (2023) for all the languages supported by the IndicTrans2 DA models and WMT Indic MT shared task.

### 3.3 Contrastive System #2

Our second contrastive system combines the primary and first contrastive systems using a system combination approach called Multi-Engine Machine Translation (MEMT) (Heafield and Lavie, 2010, 2011). MEMT involves aligning 1-best outputs from each system using the METEOR aligner (Denkowski and Lavie, 2011), identifying candidate combinations by forming left-to-right paths through the aligned system outputs, and scoring these candidates using a battery of features. MEMT does not leverage any neural networks. We refer the readers to Heafield and Lavie (2010, 2011) for additional details.

## 4 Experiments

In this section, we describe the datasets, implementation and evaluation settings.

### 4.1 Datasets

We use the official parallel corpora and monolingual corpora provided by the organizers. We augment the monolingual corpora with those used in Gala et al. (2023). Particularly, we sample 20M English sentences, since the organizers did not provide any English monolingual data. The parallel and augmented monolingual corpora statistics are described in Table 1. For our first contrastive system, we also use a combination of BPCC seed corpora (Gala et al., 2023) and NLLB-seed corpora (Costa-jussà et al., 2022; Maillard et al., 2023) which was used in Gala et al. (2023) along with the official parallel corpora provided by the organizers for adaptation / fine-tuning IndicTrans2. Table 2 reports the statistics of different subsets used for training contrastive #1 system. For the languages primarily under consideration for the WMT Indic MT shared task, namely Assamese, Manipuri (Bengali), Khasi and Mizo, we use a total of ~196K bitext pairs encompassing seed and official parallel data.

| | # langs / script | # samples |
|---|---|---|
| †BPCC seed | 23 | 654,806 |
| NLLB seed | 3 | 18,579 |
| WMT | 4 | 145,321 |
| Total | 27 | 818,706 |

Table 2: Statistics of the parallel corpora used for training contrastive #1 system. † indicates that the BPCC seed also includes transliterated Sindhi (Arabic) data as released by Gala et al. (2023).

### 4.2 Implementation

Our primary systems are trained using YANMTT (Dabre et al., 2023). We train a single sentencepiece (Kudo and Richardson, 2018) tokenizer of 64K subwords for the Indic languages and English. We use 1M sentences per language, taken from the parallel and monolingual corpora. The model hyperparameters and optimizer details are described in Table 3. We ensure that the ratio of the official parallel and monolingual/backtranslated corpora remains balanced via temperature sampling (T=5.0) (Arivazhagan et al., 2019). We train our models till convergence with early stopping criteria with a patience of 5 and save separate checkpoints for each direction that exhibit best results for that direction based on BLEU (Papineni et al., 2002) metric on the development set. We use a fixed beam size of 4 and a length penalty of 0.8 when doing backtranslation.

For our first contrastive system, we fine-tune IndicTrans2 DA models with the standard fine-

| Hyperparameter | Value |
|---|---|
| #Layers | 12 (6) |
| Hidden size | 1024 (512) |
| FFN hidden size | 4096 (2048) |
| #Heads | 16 (8) |
| Positional Encoding | Embedding |
| Batch size | 1024 (4096) |
| Parameters | 420M (77M) |
| Dropout | 0.1 |
| Label smoothing | 0.1 |
| Optimizer | Adam |
| #GPUs | 64 (8) |
| GPU Type | V100 |
| Learning rate | 0.0005 (0.001) |
| Warmup steps | 16,000 |
| Data sampling temperature | 5.0 |
| #Train steps | ∼380K (225K) |

Table 3: Hyperparameter settings for primary systems. The values in round brackets, if at all, indicate those used for training smaller models, which only leverage organizers' parallel corpora.

tuning hyperparameter settings following Gala et al. (2023). Our first contrastive system is based on fine-tuning of IndicTrans2 DA models (Gala et al., 2023) which uses the fairseq library[4] (Ott et al., 2019). We train our systems till convergence on the development set and use the BLEU metric for early checkpointing. Furthermore, the vocabulary of IndicTrans2 DA models (Gala et al., 2023) lacks coverage for Mizo and Khasi. To address this, we extend the vocabulary and randomly initialize the newly added tokens in the embedding matrix of the IndicTrans2 DA models to incorporate representation for these languages. The expanded models serve as the base for fine-tuning.

For our second contrastive system using MEMT, we train 5-gram language models using KenLM (Heafield, 2011) and use default settings for system combination. Instead of taking only the best beam search output of each system being combined, we take the top 2 best translations in the beam, which simulates a combination of 4 systems.

For local evaluation, we use BLEU score (Papineni et al., 2002) measured using sacrebleu (Post, 2018), however, organizers additionally report chrF2 (Popović, 2017), RIBES (Isozaki et al., 2010) and TER (Snover et al., 2006). Human evaluation is not performed, but the organizers release COMET (Rei et al., 2022) scores as an approximation. For test set decoding, we identify optimal decoding hyperparameters (beam size and length penalty) by grid searching on the development set and list said hyperparameters in Table 4 for our primary and first contrastive system.

| Pair | Primary | | Contrastive | |
|---|---|---|---|---|
| | beam | penalty | beam | penalty |
| as-en | 16 | 1.4 | 16 | 1.2 |
| kha-en | 16 | 0.6 | 8 | 0.6 |
| mz-en | 16 | 1.4 | 16 | 1.4 |
| mni-en | 8 | 1.4 | 16 | 1.2 |
| en-as | 8 | 1.4 | 8 | 1.4 |
| en-kha | 16 | 1.4 | 8 | 1.4 |
| en-mz | 16 | 1.4 | 8 | 1.4 |
| en-mni | 16 | 1.2 | 8 | 0.8 |

Table 4: Optimal decoding hyperparameters settings (beam size and length penalty) obtained by performing grid search on the development set for both primary and contrastive #1 systems.

## 5 Results

In this section, we describe the results we obtained on the test sets.

### 5.1 Primary

**Main result.** Table 5 shows the results of our primary many-to-many system. For the Indic-En direction, Manipuri and Mizo to English exhibit reasonably high translation quality, at BLEU/chrF2 scores of 39.40/64.70 and 32.47/51.33 respectively. Assamese to English translation is the next best at 27.02/50.71. However, Khasi to English has the lowest translation quality among all. A critical observation is that there is no particular correlation between the sizes of the corpora and the MT quality. For example, Manipuri-English has the smallest parallel corpus (21,687 lines) and the second-smallest monolingual corpus (2.2M lines) but still exhibits the best translation quality for Manipuri to English. This could mean that the evaluation set is either easier for this pair or that it is easier to translate the pair compared to others.

For the reverse direction, once again Mizo and Manipuri exhibit the best translation quality, followed by Khasi and Assamese. Despite Assamese having more than 8 million monolingual sentences that were used for backtranslation, its translation quality is at 17.03 and 45.31 (BLEU and chrF2) which is not particularly high. The same decorrela-

[4] https://github.com/facebookresearch/fairseq

| Pair | BLEU | chrF2 | RIBES | TER | COMET |
|---|---|---|---|---|---|
| | | | **Primary** | | |
| **as-en** | 27.02 | 50.71 | 0.71 | 62.46 | 0.76 |
| **mz-en** | 32.47 | 51.33 | 0.69 | 60.56 | 0.67 |
| **kha-en** | 17.80 | 39.22 | 0.66 | 74.10 | 0.60 |
| **mni-en** | 39.40 | 64.70 | 0.77 | 51.27 | 0.79 |
| **en-as** | 17.03 | 45.31 | 0.58 | 76.57 | 0.78 |
| **en-mz** | 33.18 | 56.73 | 0.73 | 55.68 | 0.70 |
| **en-kha** | 19.95 | 43.30 | 0.68 | 66.47 | 0.67 |
| **en-mni** | 27.36 | 61.60 | 0.74 | 58.28 | 0.76 |
| | | | **Contrastive #1** | | |
| **as-en** | **37.28** | 59.97 | 0.72 | 58.81 | 0.81 |
| **mz-en** | 28.47 | 47.93 | 0.61 | 67.54 | 0.69 |
| **kha-en** | **20.06** | 40.33 | 0.58 | 78.44 | 0.60 |
| **mni-en** | **46.06** | 69.96 | 0.80 | 47.44 | 0.83 |
| **en-as** | 18.09 | 51.98 | 0.57 | 73.41 | 0.82 |
| **en-mz** | 26.47 | 50.60 | 0.66 | 65.97 | 0.69 |
| **en-kha** | 20.77 | 43.82 | 0.65 | 69.51 | 0.68 |
| **en-mni** | 24.17 | 62.95 | 0.70 | 62.85 | 0.76 |
| | | | **Contrastive #2** | | |
| **as-en** | 36.97 | 59.82 | 0.72 | 58.53 | 0.81 |
| **mz-en** | **33.30** | 52.74 | 0.70 | 60.87 | 0.68 |
| **kha-en** | 20.02 | 39.82 | 0.59 | 77.50 | 0.59 |
| **mni-en** | 43.35 | 69.27 | 0.80 | 47.43 | 0.82 |
| **en-as** | **21.07** | 51.71 | 0.58 | 73.03 | 0.81 |
| **en-mz** | **33.64** | 56.88 | 0.72 | 57.71 | 0.71 |
| **en-kha** | **21.05** | 46.06 | 0.65 | 73.80 | 0.68 |
| **en-mni** | **27.40** | 61.55 | 0.74 | 58.16 | 0.76 |

Table 5: Our primary and contrastive system results for Indic-En and En-Indic translation on the test set. These scores for all the metrics are directly reported as provided by organizers.

| Pair | WMT PC | Stage | |
|---|---|---|---|
| | | Intermediate | Final |
| as-en | 17.63 | 24.06 | **26.11** |
| mz-en | 22.36 | 25.98 | **28.34** |
| kha-en | 11.03 | 13.22 | **14.68** |
| mni-en | 31.70 | 36.73 | **40.43** |
| en-as | 13.23 | 16.62 | **17.51** |
| en-mz | 21.54 | 24.25 | **26.12** |
| en-kha | 14.72 | 15.99 | **17.60** |
| en-mni | 20.35 | 23.72 | **24.62** |

Table 6: Greedy search BLEU scores on the development set for Indic-En and En-Indic direction for the various models we trained in the process of getting to our final model. The "WMT PC" model uses only the parallel corpus for training. The "Intermediate" model is trained using the joint text infilling and MT objective and the "Final" model is trained with the backtranslated and organizers' parallel data. All models are many-to-many. Please note that we use the IndicNLP tokenizer (Kunchukuttan, 2020) instead of standard tokenizer provided in sacrebleu (Post, 2018) for computing scores locally.

tion between corpora sizes and translation quality that existed for translation into English holds for the reverse direction. In addition, we report the BLEU scores for NLLB 54B MoE model on test set in Table 8.

**Ablations.** Although we report test set results only using the final system, we also report the BLEU scores on the organizer's official dev set of the intermediate and final models in Table 6. Additionally, we report the results of a model that is trained only using the organizers' official parallel corpora. It is clear that the intermediate model using joint denoising and MT training leads to a vast improvement in translation quality, indicating that the monolingual corpus brings substantial benefits. This is especially the case for Indic-En direction since we use around 20M monolingual English sentences. We observe that the En-Indic direction also has some performance gains (around 3 BLEU) but not as much as compared to the gains in the Indic-En direction (around 6 BLEU). This implies that the scale of monolingual data is an important factor, however, we are limited by the scale of monolingual data available for the Indic languages.

Furthermore, the final model, which uses backtranslated data from the intermediate model further shows improvements of approximately 4 BLEU. This indicates that in low-resource settings similar to ours, leveraging monolingual corpora first via denoising followed by backtranslation leads to the best models. Iterative backtranslation (Hoang et al., 2018) would be the ideal next step, but we chose to not pursue it because of compute constraints.

## 5.2 Contrastive

**Contrastive #1: Main Result.** Table 5 shows the results of our contrastive #1 system. We observe superior performance for the languages that are already covered in the off-the-shelf IndicTrans2 models (Assamese, Manipuri (Bengali)) as compared to the primary system. For Indic-En direction, Assamese and Manipuri to English exhibit reasonably high translation quality, achieving BLEU scores of 37.28 and 46.06 respectively. Furthermore, we also find mixed results between both systems for newly introduced languages such as Mizo and Khasi. For Khasi, the contrastive #1 system outperforms the primary system on both directions, whereas for

| Model variants | FLORES-200 (18 lang) | | WMT (all langs) | | WMT (new langs) | |
|---|---|---|---|---|---|---|
| | En-Indic | Indic-En | En-Indic | Indic-En | En-Indic | Indic-En |
| IT2-DA | 19.03 | 37.25 | - | - | - | - |
| A1 | 3.85 | 35.81 | 24.70 | 32.50 | 23.20 | 24.40 |
| A2 | 19.46 | 37.62 | 20.90 | 24.60 | 18.95 | 13.00 |
| A3 | 18.68 | 38.07 | 25.80 | 32.30 | 24.50 | 24.20 |

Table 7: BLEU scores of different ablations described in Section 3.2 explored under contrastive #1 system on FLORES-200 devtest set (covers 18 languages) and WMT dev set (4 languages). Please note that we use the IndicNLP tokenizer (Kunchukuttan, 2020) instead of standard tokenizer provided in sacrebleu (Post, 2018) for computing scores locally.

| | Primary | | Contrastive #1 | | Contrastive #2 | | NLLB 54B MoE | |
|---|---|---|---|---|---|---|---|---|
| | en-xx | xx-en | en-xx | xx-en | en-xx | xx-en | en-xx | xx-en |
| as | 17.03 | 27.02 | 18.09 | **37.28** | **21.07** | 36.97 | 19.60 | 26.8 |
| mz | 33.18 | 32.47 | 26.47 | 28.47 | **33.64** | 33.30 | 27.50 | **38.50** |
| mni | 27.36 | 39.40 | 24.17 | **46.06** | **27.40** | 43.35 | 14.90 | 31.50 |

Table 8: Comparison of BLEU scores of all our systems - Primary, Contrastive #1, Contrastive #2 with massively multilingual NLLB 54B MoE model (Costa-jussà et al., 2022).

Mizo, the primary system outperforms the contrastive #1 system across both directions.

**Contrastive #1: Ablations.** In order to identify the optimal configuration for training the Contrastive #1 system, a series of ablations were conducted, involving a comparative analysis of three distinct approaches for fine-tuning the IndicTrans2 model (Gala et al., 2023), as detailed in Section 3.2. The baseline approach, denoted as A1, focuses solely on optimizing performance across 4 languages under consideration for WMT languages. However, this approach exhibits catastrophic forgetting on the existing supported languages. This is evident in the significant drop in average BLEU scores on the FLORES-200 test set (Goyal et al., 2022; Costa-jussà et al., 2022). Specifically, when fine-tuning IndicTrans2 DA model (Gala et al., 2023) to obtain A1 for the En-Indic language direction, the average BLEU score significantly dropped from 19.03 to 3.85. However, for the Indic-En direction, the drop is relatively modest, shifting from 37.25 to 35.81, although this drop can be made even lower.

To prevent catastrophic forgetting on existing supported languages, an alternative approach, labeled as A2, was experimented. This approach involves a joint fine-tuning on a combined set involving all the existing supported languages along with the newly introduced ones. Notably, this approach averts the issue of catastrophic forgetting. On the FLORES-200 benchmark (Goyal et al., 2022; Costa-jussà et al., 2022), the models resulting from this joint fine-tuning slightly surpass the performance of the IndicTrans2 DA models in both translation directions, showing an improvement of approximately 0.4 points. However, despite this improvement, the performance on the newly added languages such as Khasi and Mizo is suboptimal, significantly trailing behind the scores obtained using the A1 approach. We observe respective drops of 3.8 and 7.9 points in the En-Indic and Indic-En directions over A1.

Although approach A2 successfully resolved the issue of catastrophic forgetting, it did not fully meet our objective of optimizing for the newly introduced languages. As a result, we explored approach A3 which involves a two-stage fine-tuning procedure, wherein A1 is initially employed, followed by A2. As previously discussed, A1 resulted in a sharp decline in performance across the existing languages, but optimized to the newly introduced languages. However, we observe that this performance drop can be rectified by introducing an additional stage of fine-tuning involving a combined set of all languages, as seen in approach A2. A3 results in a fair retention of performance in terms of BLEU scores across the existing languages for both translation directions, Indic-En (with scores of 37.24 for IndicTrans2 DA and 36.68

for A3) and En-Indic (with scores of 19.03 for IndicTrans2 DA and 18.68 for A3). Moreover, on the newly introduced languages, models trained using the A3 approach demonstrate an improvement of nearly 8 points in the Indic-En direction and 4.9 points in the En-Indic direction on average, when compared to A2, as observed on the WMT 2023 IndicMT dev set. Notably, A3 achieves performance on par with A1 (optimized for four languages) in the Indic-En direction and even outperforms A1 by a margin of +1.1 in the En-Indic direction. Therefore, A3 achieves both the outcomes: performance retention on existing languages as well as optimization in performance for newer languages.

**Contrastive #2: Main Result.** Having obtained the best primary and contrastive systems, we combine them via MEMT. Table 5 contains the result of the system combination on the test set. For Indic-En direction, only Mizo to English benefits from system combination, where the best BLEU score improves from 32.47 to 33.30. For the En-Indic direction, we see improvements for all directions. Most notable is the improvement for English to Assamese, whose best BLEU score improves from 18.09 to 21.07. For other directions, the improvements are relatively smaller. One important observation is that when the performance gaps between the primary and contrastive #1 system is larger, the gains are smaller or are negative. Overall, it is important to note that such word level system combination still works despite the idea being over a decade old, however, the use of n-gram based LMs might be a limitation and replacing said LMs with neural LLMs might bring large benefits. We leave this for future work.

### 5.3 Lessons Learned

- In low-resource settings, leverage monolingual data first via denoising and then via backtranslation.
- A two-stage fine-tuning approach (introducing new languages first, followed by a combination of new and existing languages) is an effective approach when considering extending a pre-trained translation model to newer languages without catastrophic forgetting.
- System combination is still effective despite working at a word level.

## 6 Conclusion

In this paper, we have described our systems submitted to the WMT 2023 Indic translation task. We leveraged ideas ranging from joint denoising and MT training, backtranslation, fine-tuning pretrained models, and system combination. We reported our results, which show the benefits of the various ideas we explored. Finally, we recommend best practices.

## 7 Limitations

We identify the following limitations of our submissions:

- We did not perform ensembling or checkpoint averaging, which could boost our results by another 1-2 BLEU.
- Iterative backtranslation (Hoang et al., 2018) was not adopted due to compute constraints and can potentially boost quality even further.
- Although we reached the monolingual corpora limit for the Indic languages of focus, we could have used much larger English monolingual corpora but opted not to, once again, due to compute constraints. This would also require us to increase model sizes which was also not feasible.
- We have not leveraged any LLMs for our experiments, mainly because we are not sure if they have been trained on any of the test data, a common concern in recent times.
- MEMT is an old idea and does not use any neural language models, especially LLMs, which could enhance its performance.

## Acknowledgement

We graciously thank Varun Gumma (SCAI Fellow, Microsoft Research) for his kind review of this draft, which helped us improve its readability and quality.

## References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Diptesh Kanojia, Chinmay Sawant, and Eiichiro Sumita. 2023. YANMTT: Yet another neural machine translation toolkit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 257–263, Toronto, Canada. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv: 2305.16307*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Kenneth Heafield and Alon Lavie. 2010. CMU multi-engine machine translation for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 301–306, Uppsala, Sweden. Association for Computational Linguistics.

Kenneth Heafield and Alon Lavie. 2011. CMU system combination in WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 145–151, Edinburgh, Scotland. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Samta Kamboj, Sunil Kumar Sahu, and Neha Sengupta. 2022. DENTRA: Denoising and translation pre-training for multilingual machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1057–1067, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In

*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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*, 8:726–742.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Sandeep Kumar Dash, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, and Pankaj Kundan Dadure. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

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. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. 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.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.