On the Transferability of Massively Multilingual Pretrained Models in the Pretext of the Indo-Aryan and Tibeto-Burman Languages

Salam Michael Singh[1], **Loitongbam Sanayai Meetei**[1], **Alok Singh**[1], **Thoudam Doren Singh**[1], and **Sivaji Bandyopadhyay**[1]

[1]Centre for Natural Language Processing (CNLP) & Dept. of CSE, NIT Silchar, India
{salammichaelcse,loisanayai,alok.rawat478,thoudam.doren,sivaji.cse.ju}@gmail.com

Abstract

In recent times, machine translation models can learn to perform implicit bridging between language pairs never seen explicitly during training and showing that transfer learning helps for languages with constrained resources. This work investigates the low resource machine translation via transfer learning from multilingual pre-trained models i.e. mBART-50 and mT5-base in the pretext of Indo-Aryan (Assamese and Bengali) and Tibeto-Burman (Manipuri) languages via finetuning as a downstream task. Assamese and Manipuri were absent in the pretraining of both mBART-50 and the mT5 models. However, the experimental results attest that the finetuning from these pre-trained models surpasses the multilingual model trained from scratch.

1 Introduction

Recent years have witnessed the growing advances in the field of neural machine translation (NMT) specifically for the resource rich languages. However, NMT requires enormous amount of parallel data in order to have a decent translation system. On the other hand, the low resource languages lacks sufficient amount of parallel data, thus making the translation system far from the production level. Meanwhile, monolingual data is readily available as compared to the parallel data and many works have been done to exploit it, most notably in a semi-supervised approach for data augmentation using self-training (Ueffing, 2006; Zhang and Zong, 2016; He et al., 2020) and back-translation (Sennrich et al., 2013; Edunov et al., 2018). However, these approaches are prone to generate erroneous translations due to the noisy synthetic data and often requires an iterative refinement procedure which is both resource intensive (Hoang et al., 2018) and time consuming process. Unsupervised machine translation (Lample et al., 2018; Artetxe et al., 2018; Lample and Conneau, 2019) on the other hand uses only the monolingual data and do not require any parallel data which appears to be intimidating for a low resource scenario. Additionally, the initial cross-lingual mapping between the two monolingual data requires a maximal amount of vocabulary overlaps which is crucial for a stronger cross-lingual mapping between the source and the target monolingual vector spaces. However, the vocabulary overlaps is maximised only when the two languages are closely related thus making the unsupervised machine translation approach unsuitable for the distant language pairs even if they have large amount of monolingual data (Kim et al., 2020). Moreover, conventional unsupervised systems utilises iterative back-translation for the refinement purpose, thus the unsupervised methods are imposed with the issues of the back-translation (noisy translations and resource intensive). Multilingual neural machine translation (MNMT) (Johnson et al., 2017; Fan et al., 2021) on the other hand supports the translation among multiple languages which has shown to be beneficial for low resource machine translation via the transfer of cross-linguistic information from the higher resource languages (Aharoni et al., 2019; Dabre et al., 2020). This can be facilitated by transferring the trained parameters from a parent model to a child model (Zoph et al., 2016; Nguyen and Chiang, 2017; Kocmi and Bojar, 2018) or through a bridge or pivot language (Dabre et al., 2015; Utiyama and Isahara, 2007; More et al., 2015). However, MNMT can be further simplified by converting it into a single bilingual NMT by jointly training (Firat et al., 2016; Johnson et al., 2017) all the languages. Furthermore, the jointly trained MNMT system is extended with 50 or more languages in a massively multilingual (Aharoni et al., 2019; Fan et al., 2021; Xue et al., 2021) scenario which has shown to im-

Proceedings of the 18th International Conference on Natural Language Processing, pages 64–74
Silchar, India. December 16 - 19, 2021. ©2021 NLP Association of India (NLPAI)

prove the low resource machine translation (Dabre et al., 2020) in the presence of the higher resource languages with the advantage of training a single NMT model instead of training separate bilingual models. However, training these massively multilingual models from scratch for every new languages is not feasible both in terms of time and the resource and has negative impact to the environment for training such enormous models which can be coped up via transfer learning where the downstream translation task can be simply finetuned from a large pre-trained model (Liu et al., 2020; Tang et al., 2020; Conneau et al., 2020; Kakwani et al., 2020; Khanuja et al., 2021; Xue et al., 2021; Dabre et al., 2021). Primitive transfer learning in the NLP flourished with the pretrained word embedding vectors (Mikolov et al., 2013; Pennington et al., 2014), followed by the pretrained encoder (Devlin et al., 2019) or decoders or pretraining the full seq2seq model (Liu et al., 2020). These multilingual pretrained models such as the mBART (Liu et al., 2020) and the mT5 (Xue et al., 2021) has shown to benefit the low resource machine translation during the downstream finetuning step. Additionally, these pretrained models can be extended to even new languages (Tang et al., 2020) which was absent during the pretraining process by simply resuming the training with the new language data with the pretrained model checkpoint as a finetuning step and sometimes increasing the BLEU score also.

In our premise, we make use of the mBART-50 (Tang et al., 2020) and the mT5-base (Xue et al., 2021) pretrained models for the English (*en*) to {Assamese (*asm*), Bengali (*bn*) and Manipuri (*mni*)} translation in a one-to-many multilingual setup. All the three languages apart from English are the scheduled languages of India where Assamese and Bengali belong to the Indo-Aryan language family while Manipuri is a Tibeto-Burman language and very few works have been reported in this language most notably (Singh and Bandyopadhyay, 2010; Singh, 2013; Singh and Singh, 2020; Singh et al., 2021; Singh and Singh, 2021; Sanayai Meetei et al., 2020; Rahul et al., 2021; Laitonjam and Ranbir Singh, 2021). Additionally, only the Bengali language is present during the pretraining of both mBART-50 and the mT5-base models while Assamese and Manipuri were absent during the pretraining phase. Hence, the finetuning process involves the transfer learning to totally unseen languages and this work investigates the effect of these pretrained models to the low resource translation task for these unseen languages. We also evaluate our performance on the WAT-2021 MultiIndicMT [1] test set for English to Bengali and Flores-101 test set (Goyal et al., 2021) for the English to (Bengali and Assamese)

2 Multilingual Neural Machine Translation

Multilingual NMT facilitates the translation between multiple languages via pivot based (Dabre et al., 2015), transfer learning (Zoph et al., 2016) or through a jointly trained single NMT model (Johnson et al., 2017). In this work, we utilise the jointly trained single multilingual NMT model. Additionally, this single MNMT can be further divided into three types according to the mapping of the source and the target languages, **Many-to-one** (**m2o**). In this setting, the model is trained to translate multiple source languages into a single target language. **One-to-many** (**o2m**). This MNMT model translates from a single source language to multiple target languages and many-to-many (**m2m**). Here, translation between many source and many target languages is possible. Moreover, as there are several target languages in the **o2m** and **m2m**, a target language tag is typically prepended at the beginning of the source sentence to specify the predicted target language. Given K sentence pairs and L language pairs the training objective of an MNMT model is to maximise the log-likelihood over the whole parallel pairs $\{\mathbf{x}^{(l,k)}, \mathbf{y}^{(l,k)}\}_{k\in(1,\ldots,K_l)}^{l\in(1,\ldots,L)}$ as:

$$\mathcal{L}_\theta = \frac{1}{K}\sum_{l=1}^{L}\sum_{k=1}^{K_l} log\ p(\mathbf{y}^{(l,k)}|\mathbf{x}^{(l,k)};\theta), \qquad (1)$$

where the total parallel sentences $K = \sum_{l=1}^{L} K_l$.

3 Multilingual Pretrained Model

3.1 mBART

The mBART model which follows the sequence-to-sequence (Seq2Seq) pre-training scheme of the BART model and pre-trained on large scale monolingual corpora in 25 languages is used in our work. There are two types of noises used to produce the corrected text by removing the text spans and replacing them with a mask token and secondly by

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual

permuting the order of the sentences within each instance. The large-scale pre-training on multiple diverse languages has shown to be helpful at building low-resource NMT systems by being fine-tuned to the target language pair (Dabre et al., 2021; Xue et al., 2021). This also has shown to possess a powerful generalization ability to languages that do not appear in the pre-training corpora.

3.2 mT5

mT5 is a massively multilingual pretrained model variant of Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020). The T5 is trained on a multitask scenario which is governed by the pre-training on a masked language modeling "span-corruption" objective, in which consecutive input token spans are replaced with a mask token and the model is trained to reconstruct the masked-out tokens.

4 Experimental Setup

4.1 Dataset

The experimentation uses the parallel data from CVIT-PIB (PIB) (Philip et al., 2021) and PMIndia (PMI) (Haddow and Kirefu, 2020) dataset. The Assamese (*asm*) and Manipuri (*mni*) data is curated from PMIndia while Bengali (*bn*) data is taken from both CVIT-PIB and PMIndia dataset. For the development, a small subset of 1000 sentences from the PMI is used for the *mni* and *asm*, while WAT-2021 is used for the *bn* side.

The WAT-2021 test set is in-domain with the PMI and PIB data which are mostly news domain and we also investigate the domain adaptability of these pretrained models on a general domain test set FLORES-101. For this, the *en*-{*asm*, *bn*} translations are finetuned in a multilingual way with the FLORES development data.

4.2 Dataset Preprocessing

The text preprocessing step initially tokenizes the raw texts. English side data is tokenized using the *moses-scripts*[2] while the Indic data are normalized and tokenized using the `IndicNLP` toolkit[3]. Additionally, we do not perform any sort of script conversion for the orthogonality matching as *bn*, *asm* and *mni* all use the same script.

Furthermore, foreign language text are identified and removed using *langid*[4] and their dataset is de-duplicated and ensured that the training data excludes any instances of the development and test sets. Following the work of (Philip et al., 2021), a sentencepiece (Kudo and Richardson, 2018) BPE of 3K subword merges is learnt for each language separately over the normalized and the tokenized text data. However, the vocabulary for *en* is learnt over the combined *en* data. Finally, the union of all the unique tokens is taken to make a common dictionary.

4.3 Training setup

1. **One-to-Many multilingual model trained from scratch (O2M-S)**: A one-to-many multilingual NMT is trained from scratch using transformer with 6 layers of encoders and decoders, 4 attention heads, 512 embedding dimension and a feedforward dimension of 1024. The encoder and decoder are shared and optimised using adam with the betas (0.9, 0.98) with an initial learning rate of 0.0005 which is scheduled using inverse square root with 4000 warmup updates. The training is done using fairseq (Ott et al., 2019) toolkit for 100,000 update steps with a token based batch of batchsize 4000.

2. **mBART+O2M**: We finetune the mBART-50 model in a one-to-many multilingual setup for the *en* to (*asm*, *bn* and *mni*) translation. Furthermore, the fairseq toolkit is used and in particular the multi-simple-epoch task of the fairseq to finetune from mBART-50 pretrained model. The system is an mbart-large architecture and uses the default parameters as in this setup[5] and finetuned for 80,000 update steps.

3. **mT5+O2M**: The mT5-base model is used for the finetuning using the simpletransformers library[6] with the default setup and finetuned for 80,000 update steps.

Furthermore, all the systems are finetuned for another 15,000 update steps upon the FLORES development set after resetting the training optimizers for the domain adaptation as all the systems are

[2] `https://github.com/moses-smt/mosesdecoder/tree/master/scripts`
[3] `https://github.com/anoopkunchukuttan/indic_nlp_library`
[4] `https://github.com/saffsd/langid.py`
[5] `https://github.com/pytorch/fairseq/tree/main/examples/multilingual`
[6] `https://github.com/ThilinaRajapakse/simpletransformers`

trained only on the PMI and PIB data which is a news domain whereas the FLORES-101 test set is a general domain data.

4.4 Comparison with Other Works

This work is compared with the following work evaluated upon the WAT-2021 and FLORES-101 test sets:

1. Ramesh et al. (2021): A multilingual model trained on the largest publicly available parallel corpora.

2. IndicBART (Dabre et al., 2021): A multilingual pretrained model trained on 11 Indic languages trained using mBART objective.

4.5 Evaluation Metrics

1. **Automatic Evaluation**: The automatic evaluation is done using BLEU which is reported over the geometric mean of the 4-gram precision or BLEU-4, ranging from 0-100, with 100 being the highest. The hypothesis for the *en* to {*asm*, *bn*, *mni*} translation evaluation is detokenized and then retokenized using the IndicNLP tokenizer and then evaluated without using any tokenizer in SacreBLEU[7].

2. **Human Evaluation**: Human evaluation is carried out by considering the fluency and adequacy of the translated output. In this pretext, three human translators fluent in English-Manipuri, English-Assamese and English-Bengali are assigned to separately rate each sentence from 1-5 for the fluency and the adequacy criteria. Finally, the sentence wise scores are averaged to get the corpus level score for both the criteria.

5 Experimental Results

Table 1 reports the automatic evaluation scores of the systems based on the BLEU score for the *en* to {*asm*, *bn* and *mni*} one-to-many translations. Both the pretrained models outperforms the multilingual system trained from the scratch (**O2M-S**) across all the translation directions suggesting a successful transfer of information from the pretrained models to the downstream finetuning task.

Additionally, the significant improvement in BLEU score after the finetuning is observed for both the *asm* and *mni* languages which were absent during the pretraining step revealing that these multilingual pretrained models are language independent up to an extent and can be extended to any new languages irrespective of their relatedness from the pretrained languages and thus ideal for a low resource machine translation.

System	asm	bn	mni
O2M-S	11	16.2	19.5
mBART+O2M	15.9	19.8	26.3
mT5+O2M	15.4	18.6	29.2

Table 1: BLEU score evaluated using PMI test set for the en to (asm, bn, mni) translation.

5.1 Comparison With Other Works

Table 2 reports the BLEU score of the trained systems i.e. O2M-S, mBART+O2M and mT5 which is compared with Ramesh et al. (2021) and IndicBART (Dabre et al., 2021) evaluated upon the WAT-2021 and PMI test sets. O2M-S performs the worst amongst all the systems for both the test sets across all the translation directions. For the WAT-2021 test set, mT5+O2M has the best performance followed by Ramesh et al. (2021). Ramesh et al. (2021) is trained using the largest available training data for the Indian languages thus giving an extra edge. On the other hand FLORES test is a general domain data thus making the task more challenging as our systems are trained using only the news domain from PMI and PIB which is reflected in the low BLEU scores of our trained systems for the FLORES test set.

However, IndicBART trained their systems using Samanantar dataset (Ramesh et al., 2021) thus making their system more adaptive to the FLORES domain and surpassing both the mBART+O2M and mT5+O2M models with a

System	Test Set		
	WAT-2021	FLORES	
	bn	asm	bn
Ramesh et al. (2021)	16.0	-	-
IndicBART	11.1	-	30.7
O2M-S	10.7	1.2	2.3
mBART+O2M	14.7	3.5	5.6
mT5+O2M	16.2	2.3	4.8

Table 2: BLEU score of the systems for the en to (asm and bn) evaluated on WAT-2021 and FLORES TEST set.

[7] BLEU+case.mixed+numrefs.1+smooth.exp+tok.none+version.1.5.1

Systems	asm	bn
mBART+O2M w/o FT	2.9	4.6
+5K steps FT	3.1	5.2
+10K steps FT	3.4	5.5
+15K steps FT	3.5	5.6
mT5+O2M w/o FT	0.1	3.3
+5K steps FT	0.3	3.9
+10K steps FT	1.3	4.2
+15K steps FT	1.8	4.8

Table 3: Effect of the BLEU score on the finetuning steps (FT) which is finetuned using FLORES development set for the *en* to (*asm* and *bn*) directions.

whooping 30.7 BLEU score in comparison to the 5.6 and 4.8 BLEU scores for the mBART+O2M and mT5+O2M respectively. Additionally, for the WAT-2021 *en-bn* task, IndicBART performed poorly even though they pretrain an mBART model from the Indic languages and finetune upon it. Furthermore, the low performance of IndicBART on WAT-2021 test reveals two possibilities, i) the finetuning of IndicBART involved more number of languages than our setting, which in turn induced a negative transfer (Dabre et al., 2020) due to the incompatibility of the languages involved thus the degradation in the performance, ii) transfer learning from a massively multilingual pretrained model followed by the multilingual finetuning as in our case is more beneficial than transfer learning from a limited language pretrained model as in the case of IndicBART and we put forward these as a future work.

5.2 Domain Adaptation via Few Shot Learning

The systems in our experimentation are trained on a narrow domain data, thus these systems choke when evaluated on a general domain data. Hence, the systems are further finetuned using the FLORES development set for another 15,000 update steps by resetting the optimisers. The results are reported in Table 3.

It is observed that this domain adaptation using incremental finetuning upon the FLORES development set improves the BLEU score across all the directions for both mBART+O2M and mT5+O2M models. However, this increment is still insignificant in comparison to IndicBART (Dabre et al., 2021) as presented in Table 2.

5.3 Human Evaluation Score

Table 4 reports the human evaluation score of the **O2M-S**, **mBART+O2M** and **mT5+O2M** for the *en* to (*asm*, *bn* and *mni*) translations based on the adequacy and fluency criteria which is evaluated upon the PMI test set. For the *en-mni* translation direction presented in Table 4, the multilingual finetuning over both the pretrained models (**mBART+O2M**) and (**mT5+O2M**) is superior to the multilingual model trained from scratch (**O2M-S**) qualitatively. Additionally, in terms of the adequacy score, **mT5+O2M** performs better than the **mBART+O2M**. However, **mBART+O2M** gives a competitive performance to the **mT5+O2M** in terms of the fluency score.

Moreover, the human evaluation scores correlates well with the automatic scores as reported in Table 1 suggesting the effectiveness of the transfer learning for this unseen language during the pretraining time. On the other hand, **mBART+O2M** has higher human evaluation scores than **mT5+O2M** for the *en-asm* and *en-bn* translations as reported in Table 4. However,

Models	en-mni		en-asm		en-bn	
	Adequacy	Fluency	Adequacy	Fluency	Adequacy	Fluency
O2M-S	3.25	3.07	2.91	3.17	2.75	2.823
mBART+O2M	4.15	4.31	3.82	3.782	3.9122	3.782
mT5+O2M	4.42	4.37	3.801	3.775	3.8622	3.688

Table 4: Human evaluation score evaluated on PMI test set based on the adequacy and fluency criteria.

Source-1	In particular, he mentioned the Buddha and the Ramayana.
Ref	বুদ্ধা অমসুং রামায়নবু মহাক্না অকক্ননা পনখি।
TT	buddhaa amasung raamaayanbu mahakna akaknanaa pankhi.
Gloss	the buddha and the ramayan he particularly mentioned.
O2M-S	মহাক্না পনখি মদুদি মহাক্না বুদ্ধ অমসুং রামাধন বুদ্ধগী মতাংদা পনখি।
TT	mahakna pankhi madudi mahakna buddha amasung raamaadhan buddhagi matanga pankhi.
Gloss	he mentioned that he the buddha and the ramadan buddha's about mentioned.
ET	He mentioned that he mentioned about the Buddha and the Ramadan Buddha.
mBART+O2M	মরুওইনা মহাক্না বুদ্ধ অমসুং রামায়ণগী মতাংদা পনখি।
TT	maruoina mahakna buddha amasung raamaayanagi mataangda pankhi.
Gloss	importantly he the buddha and the ramayan's about mentioned.
ET	Importantly he mentioned about the Buddha and the Ramayan.
mT5+O2M	অকক্ননা মহাক্না বুদ্ধ অমসুং রামায়নগী মতাংদা পনখি।
TT	akaknanaa mahakna buddha amasung raamaayangi mataandga pankhi.
Gloss	in particular he the buddha and the raamaayana about mentioned.
ET	In particular he about mentioned about the buddha and the raamaayana.
Source-2	The Officer Trainees belong to 17 Civil Services, and 3 Services from the Royal Bhutan Civil Service.
Ref	ওফিসর ত্রেনীশিং অদুদা সিবিল সর্বিসকী ১৭ অমসুং রোয়েল ভুতান সিবিল সর্বিসকী অহুম য়াওরি।
TT	ophisar trenishing aduda sibil sarbiski 17 amasung royel bhutan sibil sarbiski ahum yaori.
Gloss	officer trainees in civil services 17 and royal bhutan civil service's three belong to.
O2M-S	ওফিসর ১৭, সিবিল সর্বিসশিং, সিভিল সর্বিসশিং অমসুং রোয়েল সর্বিসশিং অসি ভুতানগী সিভিল সর্বিসশিংদগীনি।
TT	ophisar 17, sibil sarbis-shing, sibhil sarbis-shing amasung royel sarbis-shing asi bhutangi sibhil sarbis-shingdagini.
Gloss	officer 17, civil services, civil services and royal services is bhutan's civil services from.
ET	17 officers, Civil Services, Civil Services and the Royal Services are from Bhutan's Civil Services.
mBART+O2M	ওফিসর ত্রেনীশিং অসি সিবিল সর্বিস ১৭ অমসুং রোয়েল ভুতান সিবিল সর্বিসতগী সর্বিস ৩নি।
TT	ophisar trenishing asi sibil sarbis 17 amasung royel bhutan sibil sarbis-tagi sarbis 3ni.
Gloss	officer trainees these civil service 17 and royal bhutan civil service from service is 3.
ET	These officer trainees are from 17 Civil Services and 3 Services from the Royal Bhutan Civil Service.
mT5+O2M	ওফিসর ত্রেনীশিং অদুদা সিবিল সর্বিসকী ১৭ অমসুং রোয়েল ভুতান সিবিল সর্বিসকী সর্বিস অহুম য়াওরি।
TT	ophisar trenishing aduda sibil sarbiski 17 amasung royel bhutan sibil sarbiski sarbis ahum yaori.
Gloss	officer trainees in civil services 17 and royal bhutan civil service's service three belong to.
ET	The Officer Trainees belong to 17 Civil Services and 3 Services from the Royal Bhutan Civil Service.
Source-3	PMSSY has two components
Ref	পি. এম. এস. এস. ৱাই .গী মশা অনি লৈ
TT	pi. em. ess. ess. yai. gi masa ani lei
Gloss	PMSSY's components two has
O2M-S	PMSSYগী কম্পোনেন্ট অনি লৈ
TT	PMSSYgi kamponenṭ ani lei
Gloss	PMSSY's components two has
ET	PMSSY has two components
mBART+O2M	PMSSYগী কম্পোনেন্ট অনি লৈ
TT	PMSSYgi kamponenṭ ani lei
Gloss	PMSSY's components two has
ET	PMSSY has two components
mT5+O2M	পি এম এস এস এস এস এস হায়বসিগী কম্পোনেন্ট অনি লৈ
TT	pi em ess ess ess ess ess haibasigi kamponent ani lei
Gloss	PMSSSSS so called component two has
ET	The so called PMSSSSS has two components

Table 5: Sample *en-mni* translations by the MT systems

mT5+O2M gives a competitive score in terms of fluency for the *en-asm*. Based on the quantitative and qualitative findings from Table 1 and Table 4 respectively, **mT5+O2M** is beneficial for the *en-mni* translation while for the *en* to (*asm* and *bn*), **mBART+O2M** is found to be effective and we plan to explore these discrepancies in our future work.

6 Qualitative and Error Analysis

6.1 Qualitative Analysis

A qualitative analysis in the form of sample input and output is also presented in Table 5 in addition to the qualitative scores reported in Section 5.3 to compare the translation qualities of the **O2M-S**, **mBART+O2M** and **mT5** for the *en* to *mni* translation of the PMI test set. In doing so, we randomly select three *en* test sentences (Source-1, Source-2 and Source-3) and present the respective translated outputs by the systems. Table 5 contains the following abbreviations: The Roman transliterated *mni* sentence is denoted by TT, Gloss is the *en* word-for-word translation, and the *en* translation for the *mni* sentence is ET.

In the first source sentence (Source-1), **O2M-S** the phrase "*mahakna pankhi*" (he mentioned) twice thus degrading the fluency and the term "*raamaayan* has been wrongly generated as "raamaadhan" (ramadan) which in turn detoriates the adequacy. Similarly, there are several instances where **O2M-S** has generated erroneous words. On the other hand, **mBART+O2M** and **mT5+O2M** made a better translation as compared to the **O2M-S** in terms of both adequacy and fluency. However, **mBART+O2M** translated the source word *In particular* to "*maruoina*" (importantly) while **mT5+O2M** translated into the accurate word "*akaknanaa*" (in particular). Although, the word order has been displaced even after generating the correct word hence the automatic scores which depends upon the exact word overlapping gets penalised. The second (Source-2) and the third source (Source-3) sentences are challenging ones. The Source-2 has complex contextual dependencies which is evident with the struggle to establish the correct dependency relations in the translations of the **O2M-S** and **mBART+O2M** while, **mT5+O2M** is the only system which can successfully establish the meaning of the source sentence along with a fluent translation. Apart from this, the Source-2 contains numerical values *17* and *3* which is successfully translated by all the three systems.

Another challenging instance is the presence of abbreviations in the source sentence and the valid English terms which exists as in the target language. This phenomenon is illustrated in Source-3 translation where all the three systems generated the source word *components* as "*kamponent*" (component) instead of "*masa*" (branch; part; component). Thus, even though the **O2M-S** and **mBART+O2M** generated the correct translation due to token mismatch between the reference and the translations, the BLEU score is penalised. In the same Source-3 sentence, the abbreviation of *PMSSY* is directly copied in the outputs of **O2M-S** and **mBART+O2M** which exists as "*pi. em. ess. ess. yai.*" (PMSSY) in the reference thus degrading the BLEU score. **mT5+O2M** on the other hand generated the extra three extra *S* in the abbreviations and excluded *Y*.

6.2 Error Analysis

The error analysis of the systems are conducted based on the sentence length. Figure 1A displays the distribution of the difference between the length of the translated output from the reference sentence length of the three systems. Here, the value of "0" at the X-axis signifies that the translated output and the reference sentence are of equal length. In this regard, **mBART+O2M** has the highest count for "0" length difference than both the **mT5+O2M** and **O2M-S** systems across all the translation directions, thus providing the heuristics that the reference and the outputs match word by word which contradicts the superior automatic and human evaluation scores of the **mT5+O2M** than the other two systems for *en* to *mni* translation.

Additionally, for the *en-asm* direction in Figure 1A(i) **O2M-S** and **mT5+O2M** have similar counts for the "0" difference. Furthermore, **mT5+O2M** tends to generate more shorter length sentences than the reference sentence in comparison to the other two systems for all directions, while **O2M-S** generates more longer sentences. Hence, **mBART+O2M** produces more equivalent length to that of the reference than the other two systems.

Figure 1B depicts the change in the BLEU score with the varying sentence length. For this, the test sentences are grouped together in buckets based on the sentence length of the reference sentences. For the *en-mni* direction in Figure 1B(iii), **mT5+O2M**

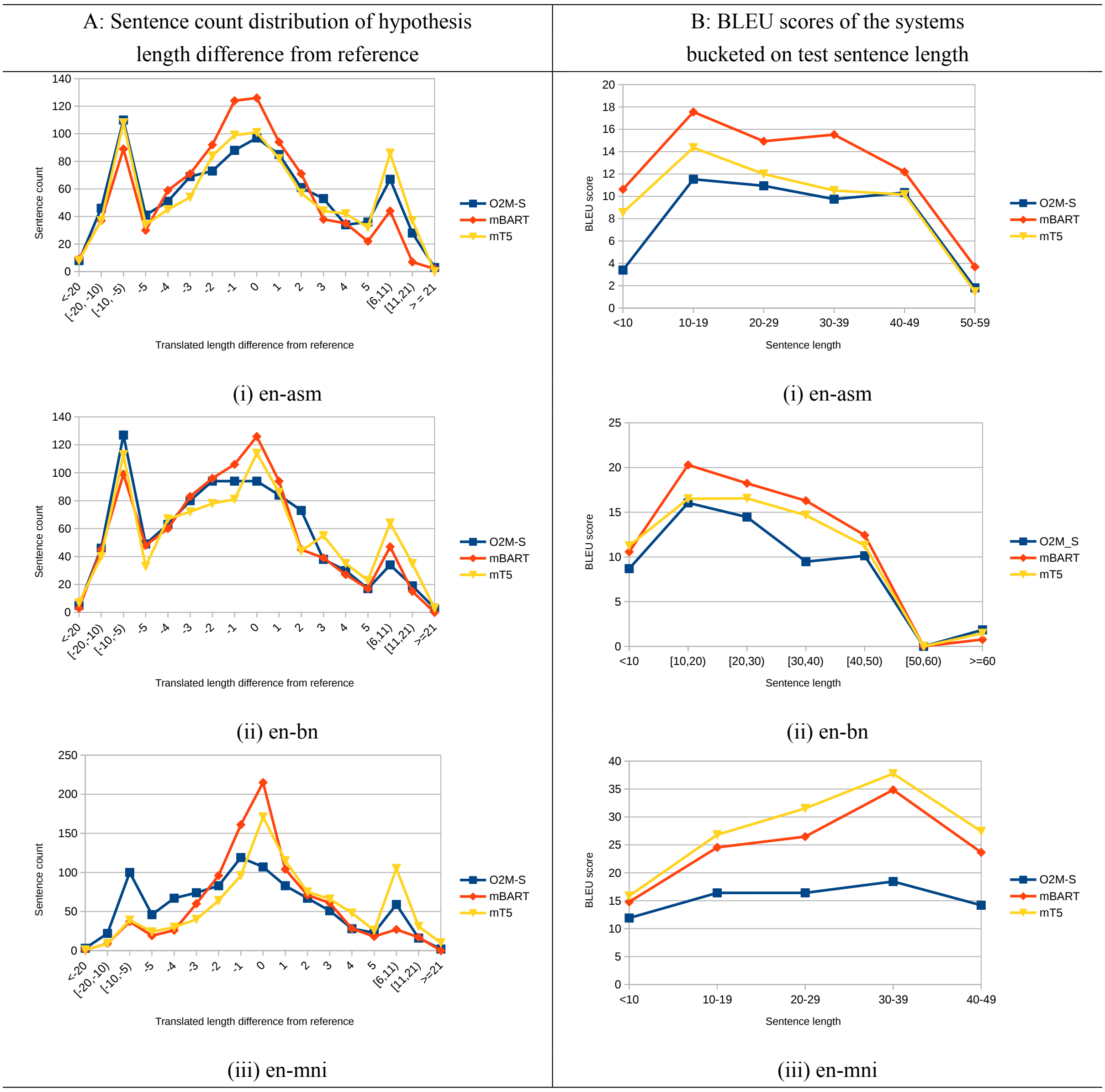

Figure 1: Error analysis of the systems based on the sentence length.

supersedes the other two systems across all the sentence length, followed by the **mBART+O2M**. Meanwhile, **mBART+O2M** is robust to longer sentence length for the *en-asm* (Figure 1B(i)) and similar trend exists in the *en-bn* direction (Figure 1B(ii)) although, **O2M+S** and **mT5+O2M** has higher BLEU scores than **mBART+O2M** for sentences longer than 60 tokens.

7 Conclusion

In this work, we report the findings of the investigation of low resource machine translation via transfer learning from multilingual pretrained models i.e. mBART-50 and mT5-base in the pretext of Indo-Aryan (Assamese and Bengali) and Tibeto-Burman (Manipuri) languages. It is found that the transfer learning from these pretrained multilingual models outperforms the one-to-many model trained from the scratch across all the translation directions in all the test sets thus suggesting the strong transfer of interliguistic information to the downstream finetuning tasks even for the languages absent during the pretraining step. Furthermore, the superiority of finetuning from these pretrained models than the IndicBART for the English to Bengali translation using the WAT-2021 test set suggests that a stronger transfer learning is possible even without linguistic relatedness during the pretraining step or due to the negative transfer of information between the incompatible languages during the multilingual finetuning of IndicBART. Finally, we plan to explore more on the negative transfer and the linguistic relatedness avenue in future focusing on Indian languages.

Acknowledgments

We acknowledge CNLP (Centre for Natural Language Processing) at NIT Silchar for giving access to the lab and the necessary resources. The authors would like to express their gratitude to the anonymous reviewers for their insightful comments. The authors also want to express their gratitude to the volunteers who assisted them with human evaluation tasks.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, Fabien Cromieres, Toshiaki Nakazawa, and Sadao Kurohashi. 2015. Large-scale dictionary construction via pivot-based statistical machine translation with significance pruning and neural network features. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 289–297, Shanghai, China.

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

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for natural language generation of indic languages. *CoRR*, abs/2109.02903.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

Barry Haddow and Faheem Kirefu. 2020. Pmindia - A collection of parallel corpora of languages of india. *CoRR*, abs/2001.09907.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *Proceedings of ICLR*.

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.

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.

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

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Belgium, Brussels. 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.

Lenin Laitonjam and Sanasam Ranbir Singh. 2021. Manipuri-English machine translation using comparable corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Rohit More, Anoop Kunchukuttan, Pushpak Bhattacharyya, and Raj Dabre. 2015. Augmenting pivot based SMT with word segmentation. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 303–307, Trivandrum, India. NLP Association of India.

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

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jerin Philip, Shashank Siripragada, Vinay P. Namboodiri, and C. V. Jawahar. 2021. Revisiting low resource status of indian languages in machine translation. In *8th ACM IKDD CODS and 26th COMAD*, CODS COMAD 2021, page 178–187, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Laishram Rahul, Loitongbam Sanayai Meetei, and H. S. Jayanna. 2021. Statistical and neural machine translation for manipuri-english on intelligence domain. In *Advances in Computing and Network Communications*, pages 249–257, Singapore. Springer Singapore.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *CoRR*, abs/2104.05596.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, Sivaji Bandyopadhyay, Mihaela Vela, and Josef van Genabith. 2020. English to Manipuri and mizo post-editing effort and its impact on low resource machine translation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 50–59, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting synergies between open resources for German dependency parsing, POS-tagging, and morphological analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 601–609, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Salam Michael Singh, Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021. Multiple captions embellished multilingual multi-modal neural machine translation. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 2–11, Online (Virtual Mode). INCOMA Ltd.

Salam Michael Singh and Thoudam Doren Singh. 2020. Unsupervised neural machine translation for English and Manipuri. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 69–78, Suzhou, China. Association for Computational Linguistics.

Salam Michael Singh and Thoudam Doren Singh. 2021. Statistical and neural machine translation systems of english to manipuri: A preliminary study. In *Soft Computing and Signal Processing*, pages 203–211, Singapore. Springer Singapore.

Thoudam Doren Singh. 2013. Taste of two different flavours: Which Manipuri script works better for English-Manipuri language pair SMT systems? In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 11–18, Atlanta, Georgia. Association for Computational Linguistics.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010. Manipuri-English bidirectional statistical machine translation systems using morphology and dependency relations. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 83–91, Beijing, China. Coling 2010 Organizing Committee.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Nicola Ueffing. 2006. Using monolingual source-language data to improve MT performance. In *Proceedings of the Third International Workshop on Spoken Language Translation: Papers*, Kyoto, Japan.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

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.