Low-Resource Machine Translation Systems for Indic Languages

Ivana Kvapilíková and **Ondřej Bojar**
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University,
Prague, Czech Republic,
kvapilikova@ufal.mff.cuni.cz, bojar@ufal.mff.cuni.cz

Abstract

We present the submission of the CUNI team to the WMT23 shared task in translation between English and Assamese, Khasi, Mizo, and Manipuri. All our systems were pretrained on the task of multilingual masked language modelling and denoising auto-encoding. Our primary systems for translation into English were further pretrained for multilingual MT in all four language directions and fine-tuned on the limited parallel data available for each language pair separately. We used online back-translation for data augmentation. The same systems were submitted as contrastive for translation out of English where the multilingual MT pretraining step seemed to harm the translation performance. Other contrastive systems used additional pseudo-parallel data mined from monolingual corpora.

1 Introduction

We present our submission to the Indic MT shared task of the WMT23 workshop. We trained constrained systems in all evaluated language directions: English-Assamese (en-as), English-Manipuri (en-mni), English-Mizo (en-mz) and English-Khasi (en-kha).

A majority of languages in the world have a very limited amounts of translation resources to be used for training machine translation (MT) systems. Unsupervised learning techniques have been proposed to leverage monolingual texts in MT training, either in the pretraining phase (Liu et al., 2020; Conneau and Lample, 2019) or during fine-tuning by means of back-translation (Sennrich et al., 2016). This shared task is proposed as a realistic scenario where for each Indic language, the participants have access to several thousand parallel sentences paired with English and up to 2.6M additional unaligned sentences in each language. The texts are mixed from the religious domain and the general domain. In addition to the provided data, participants were allowed to use any monolingual texts and any pretrained models trained on monolingual texts.

In our other research, we focus primarily on unsupervised MT and we participated in this shared task to measure the impact of adding at least a small number of parallel sentences into the training. Therefore, we also evaluated our fully unsupervised systems in the conditions of Indic MT, where the languages are linguistically very different from English and some also have a different script.

A major obstacle, especially for our unsupervised models, is the domain mismatch in our training data. While monolingual English data we used come from NewsCrawl, the Indic training data includes texts from the religious domain. The issue is especially pronounced when we struggle with finding equivalent sentences in the monolingual corpora section 2.5, but it is problematic for the entire unsupervised training as the domain mismatch interferes with the underlying assumption of isomorphism of embedding spaces.

In this paper, we first introduce our training methodology (section 2), describe the data (section 3.1) and comment on the results (sections 4 and 5)

2 Methodology

2.1 Model Architecture

The architecture of all our NMT models is a 6-layer Transformer with 6 attention heads, GELU activations, and 0.1 dropout. In addition to token embeddings and trained positional embeddings, the model features language embeddings to pass the information which language direction is being used. Both input token embeddings and the final softmax layer have tied weights.

2.2 Pretraining on Monolingual Texts

We pretrain a Transformer encoder on the task of masked language modelling (MLM) on all available corpora in all languages. The details of

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 954–958
December 6–7, 2023.

	as	kha	mni	mz	en
train (mono)	2.6M	183k	2.1M	1.9M	33M
train (para)	50k	24k	22k	50k	-
train (pseudo-para)	81k	95k	150k	66k	-
dev	2k	1k	1k	1.5k	-
test	2k	1k	1k	2k	-

Table 1: The number of sentences in the training, dev and test sets. Monolingual (mono) and parallel (para) data were provided by the organizers, pseudo-parallel data was created as described in Section 2.3.

the MLM task are given in Conneau and Lample (2019). We copy its weights into both the encoder and the decoder of the Transformer model and we continue the pretraining phase by training a multilingual denoising autoencoder (DAE). The noise function applied to the input sentence has the following components: word deletion with probability $p_{del} = 0.1$, word masking with probability $p_{mask} = 0.1$ and word shuffling within the window of length $l_{shuf} = 4$.

All our systems are pretrained on both MLM and DAE.

2.3 Pretraining on Multilingual Parallel Texts

In the second pretraining stage, we train a multilingual neural machine translation model (MNMT) on all available parallel data. In each training step, the model sees a mini-batch of parallel sentences for all language pairs. It uses language embeddings to detect the right translation direction.

2.4 Fine-tuning for Machine Translation

In the fine-tuning stage, we train a bidirectional model for each language pair in a semi-supervised fashion, using a cross-entropy loss on a small authentic parallel corpus. We augment the data with online back-translation (OBT) (Lample et al., 2018; Artetxe et al., 2018) to avoid over-fitting. In every OBT training step, the model is switched into an inference mode to create a mini-batch of training data by translating a portion of monolingual sentences. This operation is performed in both translation directions and the resulting mini-batch (with the synthetic sentences placed on the source side) is directly used for training.

2.5 Data Augmentation with Pseudo-Parallel Texts

We also measure whether we can earn some benefits by incorporating pseudo-parallel (PP) sentences into the MT training. We use the methodology of Kvapilíková et al. (2020) and search for parallel sentences in the training corpora. We search for the nearest neighbors in the multilingual sentence embedding space created by a multilingual sentence encoder. The encoder is the modified XLM-100 (Conneau and Lample, 2019) pretrained model fine-tuned on the MLM task for Assamese, Khasi, Mizo, Manipuri and English. The search metric is the modified cosine similarity $xsim$ (Artetxe and Schwenk, 2019) between the sentence embeddings which is required to be higher than 1:

	en-as	**en-kha**	**en-mni**	**en-mz**
Precision	35.03	9.67	7.92	22.54
Recall	18.55	10.50	5.70	18.00
F1 Score	24.26	10.07	6.63	20.01
Threshold	1.022	1.027	1.022	1.022

Table 2: Precision, Recall and F1 score on the Parallel Sentence Matching Task.

$$\text{xsim}(x, y) = \frac{\cos(x, y)}{\text{avgcos}(x) + \text{avgcos}(y)} > 1 \quad (1)$$

where

$$\text{avgcos}(\cdot) = \sum_{z \in \text{NN}_k(\cdot)} \frac{\cos(\cdot, z)}{2k} \quad (2)$$

where $\text{NN}_k(x)$ is the set of k nearest neighbors of x. We augment the existing authentic parallel corpora with the pseudo-parallel sentence pairs and train on the resulting corpus. The number of retrieved pseudo-parallel sentence pairs is indicated in table 1. The performance of the sentence encoder at the task of parallel corpus mining for the languages in question was measured by an auxiliary task where it was asked to find 1-2k parallel sentences (dev set) among 200k monolingual sentences from the train set in both languages. The results are summarized in section 2.4 where we see that the precision of correctly matched parallel sentences for Khasi and Manipuri is very low.

	en-as	en-kha	en-mni	en-mz	as-en	kha-en	mni-en	mz-en
MT+OBT	**14.1**	**16.6**	29.5	31.2	17.6	12.8	33.9	28.3
MNMT+MT+OBT	13.9	16.4	**29.9**	**31.5**	**20.7**	**13.8**	**36.1**	**29.5**
PP+MT+OBT	13.3	15.9	29.8	30.8	16.8	12.1	30.2	28.7
OBT (unsup)	0.2	-	-	0.8	0.3	-	-	1.3
PP+OBT (unsup)	2.9	-	-	6.1	3.1	-	-	5.5

Table 3: BLEU score of our MT systems on the WMT23 test set.

In our experiments we evaluate the impact of MNMT and PP pretraining on the final translation quality.

3 Experiments

We train several models for each language pair. All models are pre-trained as described in 2.2. For our shared task submission, we train three kinds of semi-supervised models using all available parallel data:

- MNMT+MT+OBT models were trained for multilingual MT and fine-tuned for each language pair separately on a combination of authentic parallel data and synthetic parallel data created by OBT;
- MT+OBT models skip the multilingual MT pre-training step;
- PP+MT+OBT models are trained on pseudo-paralle data in addition to authentic and synthetic data. The pseudo-parallel corpus is removed after 5 epochs of training.

We compare the results of the semi-supervised models to unsupervised models trained without the authentic parallel data to measure the effect of limited amounts of parallel data. We experiment with gradually adding parallel sentences into the training and evaluate the performance of a model trained on 1k, 2k, 5k, 10k and 25k parallel sentences.

3.1 Data

In addition to the data provided by the organizers (Pal et al., 2023), we used 33M English sentences from NewsCrawl2022. The summary of the data is in table 1. We trained a BPE model on the concatenation of all Indic corpora and a downsampled Englih corpus. The BPE vocabulary size is 52k. During pre-processing, we first tokenized the texts using the Moses tokenizer which created a problem with the Assamese script as it decomposed several compound Unicode characters which had impact on the segmentation of texts with Assamese script (as, mni). The decomposed accents form a separate BPE unit which lead to a high segmentation of the Assamese and Manipuri texts. During post-processing we managed to compose the segmented text by running a special substitution on top of the standard detokenization. The unnecessary step of Moses tokenization likely cost us some final translation performance due to the sub-optimal BPE segmentation.

3.2 Training Details

We use the XLM[1] toolkit for training. For language model pretraining, we use mini-batches of 64 text streams (256 tokens per stream) per GPU and Adam (Kingma and Ba, 2015) optimization with `lr=0.0001`. For denoising and MT fine-tuning, we use mini-batches of 3,400 tokens per GPU and Adam optimization with a linear warm-up `(beta1=0.9,beta2=0.98,lr=0.0001)`. The models are trained on 8 GPUs.

4 Shared Task Results

For our shared task submission, we compared the performance of our experiments on the dev set and concluded that the fine-tuned multilingual NMT system (MNMT+MT+OBT) performs better than individual systems (MT+OBT) when translating into English and on par with individual systems when translating from English. Therefore, for our PRIMARY submission, we submitted the output of the multilingual model when translating into English and the output of the individual models when translating from English. The opposite results were submitted as CONTRASTIVE-1. The PP+MT+OBT systems were submitted as CONTRASTIVE-2. The final test set results are summarized in table 3.

The winning system for all language directions was a system called TRANSSION-MT which outperformed other systems with almost double the

[1] https://github.com/facebookresearch/XLM

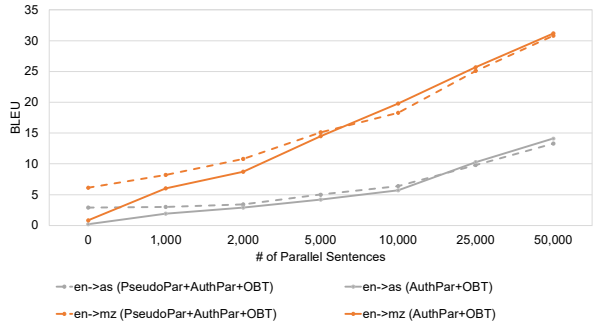

Figure 1: Relationship between the translation quality and the number of parallel sentences used for training.

BLEU score of the second best candidate (Pal et al., 2023). In general, our systems performed relatively better in translation from English which suggests that the translation to English may have been harmed by the bidirectional nature of our systems. Our en→mni system ranked second after TRANSSION-MT out of 14 participants. Our en→mz system ranked fourth out of 11 participants. The the remaining systems finished on the 5th-7th places.

5 Discussion

Asides from the shared task submission, we were interested in the following phenomena which we measured in our experiments:

- The gap between unsupervised and semi-supervised translation systems;
- The impact of pseudo-parallel data augmentation on the final translation quality;
- The development of translation quality in relation to the number of parallel sentences used during training.

Outside of the scope of the shared task, we trained unsupervised MT systems for Mizo and Assamese. For each of the two language pairs, we trained two systems, with and without pseudo-parallel sentences. table 3 shows that the unsupervised systems reach between 3 and 6 BLEU which is not a sufficient quality for practical use of the systems. The poor unsupervised results are most likely the consequence of the domain mismatch between English and Indic data as well as a mismatch between the English train set and the test sets. Our conclusions support the claims of other researchers (Marchisio et al., 2020; Vulić et al., 2019) that unsupervised MT models often fail in truly low-resource scenarios where it is not possible to obtain enough clean and domain-balanced monolingual training data and the underlying assumption of language isomorphism is challenged.

Data augmentation with pseudo-parallel sentences has zero or even a negative impact on the performance of our semi-supervised systems. For the unsupervised systems, on the other hand, it increases BLEU score by up to 5.3 BLEU points. We trained several other systems, gradually adding more parallel sentences, to measure the threshold where pseudo-parallel sentences stop helping. fig. 1 illustrates the relationship between translation quality and reveals that when we have more than 10k parallel-sentences, the unsupervised data augmentation techniques of adding pseudo-parallel sentence pairs is not beneficial anymore.

6 Conclusion

We trained several MT systems for translation between English and four Indic languages. The most promising outcomes were achieved by initially pre-training a multilingual NMT system, followed by fine-tuning using bilingual parallel data along with online back-translation. Our systems ranked between the 2th and 7th place among 10-14 participating teams, depending on the language pair and translation direction. Our systems performed relatively better at translation out of English.

Data augmentation with pseudo-parallel data does not bring any further benefits in the context of the shared task. Our experiments show that their positive effect disappears when we have access to more than 10k authentic parallel sentences.

We compared the results to completely unsupervised systems and we conclude that the domain mismatch between our English and Indic training data and the linguistic dissimilarity of the languages do not allow the unsupervised MT systems to learn to translate without seeing parallel sentences. Incorporating pseudo-parallel sentences into the training helps, but the translation quality remains low.

Acknowledgements

This research was partially supported by the grant 19-26934X (NEUREM3) of the Czech Science Foundation and by the SVV project number 260 698 of the Charles University.

References

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

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*.

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.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. 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.

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.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.