Multiple Pivot Languages and Strategic Decoder Initialization helps Neural Machine Translation

Shivam Mhaskar, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai, India
{shivammhaskar, pb}@cse.iitb.ac.in

Abstract

In machine translation, a pivot language can be used to assist the source to target translation model. In pivot-based transfer learning, the source to pivot and the pivot to target models are used to improve the performance of the source to target model. This technique works best when both source-pivot and pivot-target are high resource language pairs and the source-target is a low resource language pair. But in some cases, such as Indic languages, the pivot to target language pair is not a high resource one. To overcome this limitation, we use multiple related languages as pivot languages to assist the source to target model. We show that using multiple pivot languages gives 2.03 BLEU and 3.05 chrF score improvement over the baseline model. We show that strategic decoder initialization while performing pivot-based transfer learning with multiple pivot languages gives a 3.67 BLEU and 5.94 chrF score improvement over the baseline model.

1 Introduction

Neural Machine Translation (NMT) models have made huge improvements in the performance of machine translation systems. But NMT models are *data hungry*. NMT models require huge amounts of parallel corpus for training. To overcome this limitation and improve the performance of the source to target NMT model, the resources of a pivot language can be used. Zoph et al. (2016) used a parent model trained on a high resource language pair to initialize the parameters of the child model, which is then trained on a low resource language pair. Kim et al. (2019) introduced pivot-based transfer learning techniques to utilize the resources of the pivot language. In pivot-based transfer learning techniques, the source to pivot and the pivot to target models are used to initialize the source to target NMT model.

The pivot-based transfer learning techniques work best when both the source to pivot and the pivot to target language pairs are relatively high resource language pairs. It also helps if the pivot language is related to the source or target language, to utilize language relatedness (Kunchukuttan and Bhattacharyya, 2020). In the task of translation from English to an Indic language, another Indic language can be used as a pivot language, as Indic languages are related. But in such a setting, the pivot to target language pair may not be a high resource language pair. In the task of English to Marathi translation, Hindi can be used as a pivot language, as Hindi is a related language to Marathi. The English-Hindi language pair is a relatively high resource language pair, but the Hindi-Marathi language pair is not a high resource language pair. To overcome this shortcoming, we use multiple Indic languages as pivot languages to assist the source to target NMT model.

Transformer (Vaswani et al., 2017) model has shown state-of-the-art results for various natural language processing tasks, including machine translation. In a Transformer based NMT model, the decoder consists of two modules, self-attention, and cross attention. The self-attention layer works only with the target side language, but the cross attention layer works with the source and target side languages. We experiment with various techniques to initialize the modules of the decoder.

The major contributions of this work are as follows,

- We show that using multiple pivot languages to assist the source to target model helps improve the performance of NMT models.
- We show that strategic decoder initialization while performing pivot language-based transfer learning improves the performance of NMT models.

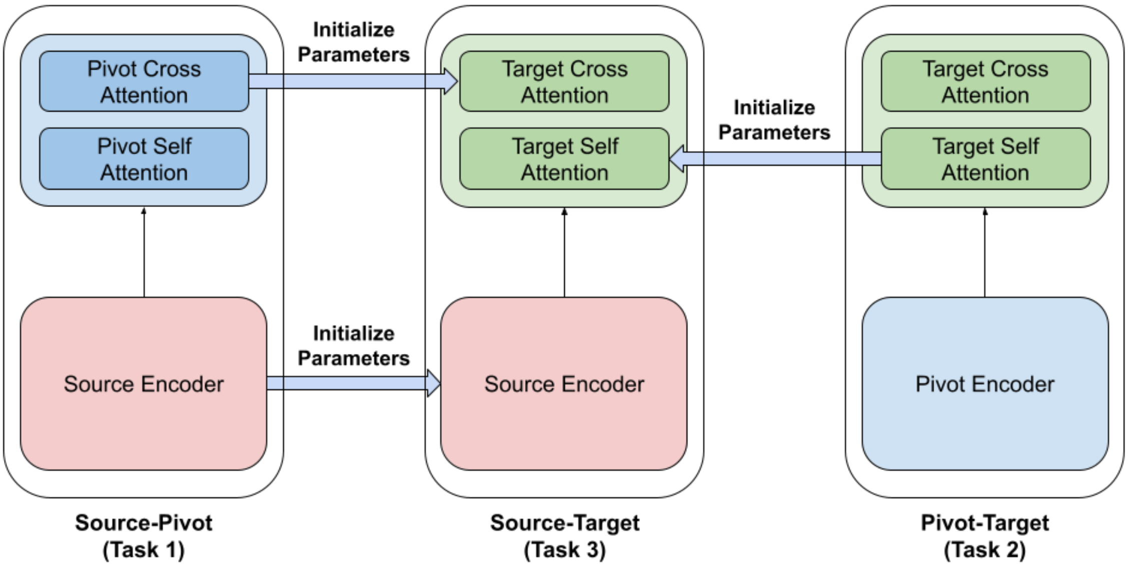

Figure 1: Initializing the *source* ⟶ *target* cross attention module with the cross attention module of the *source* ⟶ *pivot* model in pivot based transfer learning.

2 Approaches

We first discuss the approach to using multiple pivot languages to assist the source to target model. Then we discuss the various techniques to initialize the decoder of the source to target model in pivot-based transfer learning.

2.1 Multiple Pivot Languages

The task is to improve the performance of the English to Marathi NMT model. Initially, we use Hindi as a pivot language, which is related to Marathi and is a relatively high resource language among Indic languages. The English-Hindi language pair is a high resource language pair, but the Hindi-Marathi language pair is not a high resource. The amount of parallel corpus available for Hindi-Marathi is lower than English-Marathi. In order to bridge this gap, we introduce multiple Indic languages as pivot languages. We use Hindi, Bengali, Gujarati, and Tamil as pivot languages to assist the English-Marathi NMT model.

As we are using four pivot languages, the amount of parallel corpus for source-pivot and pivot-target language pairs increases significantly. This helps train better source-pivot and pivot-target models, which can be used to initialize the source-target model. In this technique, we first train an English to four Indic languages NMT model using the English to Hindi, Bengali, Gujarati, and Tamil parallel corpus. Then we train four Indic languages to the Marathi NMT model using the Hindi, Bengali, Gujarati, and Tamil to Marathi parallel corpora. We use these models to initialize the encoder and decoder modules of the source to target model and train it on the source-target parallel corpus.

2.2 Decoder Initialization

In direct pivot-based transfer learning, the decoder of the source to target model is initialized with the decoder of the pivot to target model. The decoder cross attention layer of the source to target works with the source-target language pair. The decoder cross attention layer of the pivot to target model works with the pivot-target language pair. In order to overcome this mismatch, we experiment with various initialization techniques for the decoder module.

2.2.1 Randomly Initialized Cross Attention Module

In this technique, we first initialize the encoder of the source to target model with the encoder of the source to pivot model. Then we only initialize the decoder self-attention layer of the source to target model with the decoder self-attention of the pivot to target model. The cross attention layer of the source to target model is randomly initialized. In the English-Marathi (source-target) model, the decoder self-attention layer is initialized with the decoder self-attention layer of the Hindi-Marathi (pivot-target) model.

2.2.2 Initializing the Cross Attention Module from *source to pivot* model

In this technique, the encoder of the source to target model is initialized with the encoder of the source to pivot model. The decoder self-attention layer of the source to target model is initialized with the decoder self-attention layer of the pivot to target model. The decoder cross attention layer of the source to target model is initialized with the decoder cross attention of the source to pivot model.

The cross attention layer of a Transformer decoder consists of three types of parameters, the query matrix, the key matrix, and the value matrix. The cross attention module is also called encoder-decoder attention, as it works with the source and target sequence. The query matrix is exposed to the target side sequence, and the key and value matrices are exposed to the source side sequence. The decoder cross attention of the Hindi-Marathi (pivot-target) model works with the Hindi and Marathi sequences. But in English-Marathi (source-target) model, we want the cross-attention module to work with the English and Marathi sequence. So there is a mismatch between, the sequence to which the key and value matrices are exposed during the training of, the pivot to target and the source to target model. During the training of the Hindi-Marathi model, the key and value matrices are exposed to the Hindi language but during the training of the English-Marathi (source-target) model, the key and value matrices are exposed to the English language.

In order to overcome this mismatch, we initialize the cross attention module of the English-Marathi (source-target) model with the cross attention module of the English-Hindi (source-pivot) model. Now there is no mismatch between the sequence exposed to the key and value matrices. But there is a mismatch between the sequence exposed to the query matrix. As in the English-Hindi model, the query matrix is exposed to the Hindi language but in the English-Marathi model, it is exposed to the Marathi language. But the effect of this mismatch is minimized because Hindi and Marathi are related languages.

3 Experimental Setup

In this section, we discuss the setup of the various experiments that we performed. We use byte pair encoding (BPE) (Sennrich et al., 2016) technique to split words into subwords. We use the fairseq (Ott et al., 2019) library to perform all the experiments.

Language Pair	# Sentence Pairs
English-Marathi	3.2M
English-Hindi	8.4M
English-Bengali	8.4M
English-Gujarati	3.0M
English-Tamil	5.0M
Hindi-Marathi	1.9M
Bengali-Marathi	1.8M
Gujarati-Marathi	1.7M
Tamil-Marathi	2.0M

Table 1: Dataset Statistics of Samanantar Parallel Corpus

3.1 Model

We used the Transformer model to implement all the NMT models. The model has 6 encoder layers and 6 decoder layers. The number of encoder attention heads is 8 and the number of decoder attention heads is 8. The Transformer feed-forward layer dimensions are 2048. The encoder and decoder embedding dimensions are 512. We used the same model architecture to implement the bidirectional NMT models and En-Indic multilingual NMT models.

For training the model we used label smoothed cross entropy criterion with label smoothing of 0.1. We used the Adam optimizer with beta values of 0.9 and 0.98. We used the inverse square root learning rate scheduler with 4000 warmup updates. We used a dropout value of 0.3. The batch size was 4096 tokens. We trained the model for 300,000 iterations and chose the model that gave the best loss value on the validation set.

3.2 Datasets

For all the experiments, we used the Samanantar (Ramesh et al., 2022) parallel corpus. We used the parallel corpora for the English to Hindi, Marathi, Gujarati, Bengali, and Tamil language pairs. We also used the Hindi, Gujarati, Bengali, and Tamil to Marathi parallel corpora. The dataset statistics of the parallel corpora used are mentioned in Table 1. We evaluate our models on the Facebook Low Resource (FLORES) MT Benchmark (Guzmán et al., 2019) which consists of 1012 sentence pairs from various domains.

Technique	English→Marathi			
	Pivot=Hi		Pivot=Hi,Bn,Gu,Ta	
	BLEU	chrF	BLEU	chrF
Baseline	9.02	38.58	9.02	38.58
Direct Pivoting	10.49	40.47	11.95	43.82
+ Randomly Initialized Cross Attention Module	10.82	40.90	11.99	43.69
+ Cross Attention Module Initialized from source → pivot model	**11.05**	**41.63**	**12.69**	**44.52**

Table 2: Results (BLEU and chrF Scores) of the English→Marathi NMT model. The table shows a comparison of models using only one pivot language, Hindi (Hi), and using multiple pivot languages, Hindi (Hi), Bengali (Bn), Gujarati (Gu), and Tamil (Ta). The table also shows the comparison between different decoder initialization techniques in pivot-based transfer learning. The Baseline model score is the score of the English-Marathi model trained on the English-Marathi parallel corpus

Pivot Language	English→Marathi
	BLEU
Hi	10.49
Bn	9.95
Gu	10.17
Ta	9.15
Hi, Bn, Gu, Ta	**11.95**

Table 3: Results (BLEU scores) of English→Marathi model trained by using different pivot languages as the single pivot language. The single pivot languages used are Hindi (Hi), Bengali (Bn), Gujarati (Gu), and Tamil (Ta). The last row shows the results of the English→Marathi model trained with multiple pivot languages.

3.3 Baseline

The baseline model is an English to Marathi NMT model which is trained on English-Marathi parallel corpus.

3.4 Direct Pivoting

In the Direct Pivoting model, we first train an English-Hindi and Hindi-Marathi NMT model. Then we initialize the encoder and decoder of the English-Marathi model using the encoder and decoder of the English-Hindi and Hindi-Marathi model, respectively. Finally, we train the English-Marathi model on English-Marathi parallel corpus.

3.5 Multiple Language Pivoting

In Multiple-Language Pivoting models, we use Hindi, Gujarati, Bengali, and Tamil as pivot languages. The source to pivot model is now an English to Indic NMT model, and the pivot to target model is an Indic to Marathi NMT model. For all the experiments with multiple pivoting languages, we use the four Indic languages as pivot languages instead of using only Hindi as the pivot language.

3.6 Randomly Initialized Cross Attention Module

In this experiment, we first train an English-Hindi and Hindi-Marathi NMT model. We initialize the encoder of the English-Marathi model with the encoder of the English-Hindi model. The decoder self-attention layer of the English-Marathi model is initialized with the decoder self-attention layer of the Hindi-Marathi model. The decoder cross attention layer of the English-Marathi model is randomly initialized. Finally, the model is trained on English-Marathi parallel corpus.

3.7 Initializing Cross Attention module from *source to pivot* model

In this experiment, an English-Hindi and a Hindi-Marathi model are trained. The encoder of the English-Marathi model is initialized using the encoder of the English-Hindi model. The decoder self-attention layer of the English-Hindi model is initialized using the decoder self-attention layer of the Hindi-Marathi model. The decoder cross attention layer of the English-Marathi model is initialized using the decoder cross attention layer of the English-Hindi model. Finally, the model is trained on English-Marathi parallel corpus.

English-Source	The smaller the Rossby number, the less active the star with respect to magnetic reversals.
Marathi-Reference	रॉस्बी संख्या जितकी लहान असेल तितकाच तो तारा चुंबकीय परावर्तनाच्या बाबतीत कमी सक्रिय असेल.
Marathi-Reference Gloss	Rossby number as-much small will-be that-much that star magnetic changes in-case less active will-be.
Marathi-Single	रॉसबी संख्या जितकी कमी असेल, तितकेच चुंबकीय मागे पडण्याच्या बाबतीत स्टार कमी सक्रिय आहे.
Marathi-Single Gloss	Rossby number as-much small will-be, that-much magnetic behind to-fall in-case star less active is.
Marathi-Multiple	रोस्बी संख्या जितकी लहान तितकीच चुंबकीय उलथापालथींच्या बाबतीत तारा कमी सक्रिय असतो.
Marathi-Multiple Gloss	Rossby number as-much small will-be magnetic of-upheavals in-case star less active is.

Table 4: Illustrative examples of improvement of the English→Marathi model trained with a multiple pivot language over the model trained with a single pivot language on a sentence from the test set. 'English-Source' is the input English sentence. 'Marathi-Reference' is the reference Marathi translation in the test set and 'Marathi-Reference-Gloss' is the word-to-word translation of the Marathi sentence in English which is done manually. 'Marathi-Single' is the output translation of the English→Marathi model trained with single pivot language Hindi and 'Marathi-Multiple' is the output translation of the English→Marathi model trained with multiple pivot languages. 'Marathi-Single Gloss' and 'Marathi-Multiple Gloss' are the word-to-word translations of the outputs 'Marathi-Single' and 'Marathi-Multiple', respectively, in English which is done manually.

4 Results And Analysis

We use BLEU (Papineni et al., 2002) and chrF (Popović, 2015) scores to evaluate the performance of all the models. We used the sacreblue[1] implementation for computing the BLEU scores and the NLTK[2] implementation for computing the chrF scores. Table 2 shows the results of various strategies to initialize the decoder module in pivot language-based transfer learning. The table also shows the results of experiments performed by using a single pivot language and using multiple pivot languages.

From the results, we can observe that models using multiple pivot languages outperform models using only Hindi as a pivot language. The best model using only a single pivot language achieves a BLEU score of 11.05 and chrF score of 41.63. The model using multiple pivot languages improves the BLEU score by 1.64 points to 12.69 and chrF score by 2.89 points to 44.52. This shows that using multiple pivot languages improves the performance of the source to target NMT models.

We can observe that randomly initializing the decoder cross attention module of the source to target model gives better or comparable performance over direct pivoting. Initializing the decoder cross attention module of the source to target model with the decoder cross attention module of the source to pivot model gives the best performance. In multi pivot languages setting, the direct pivoting technique achieves a BLEU score of 11.95 and chrF score of 43.82 and the strategic decoder initialization technique improves the BLEU score by 0.74 BLEU points to 12.69 and the chrF score by 0.7 points to 44.52.

Table 3 shows the results of the English-Marathi model trained using different pivot languages as the single pivot language and the model trained with multiple pivot languages. From the results, we can observe that using Hindi as single a pivot language performs better than using other languages such as Bengali, Gujarati, and Tamil as single pivot languages. We can also observe that a model trained using multiple pivot languages performs better than any model trained with only a single pivot language.

[1] https://github.com/mjpost/sacrebleu

[2] https://www.nltk.org

5 Illustrative examples of improvement

In this section, we show some examples of improvement in translation with the model with multiple pivot languages over the model with a single pivot language. Table 4 shows an English sentence, its reference Marathi translation (Marathi Reference), the output of the model trained with a single pivot language (Marathi-Single), and the output of the model trained with multiple pivot languages (Marathi-Multiple). The model with a single pivot language does not translate the word ’reversals’ properly but the model with multiple pivot languages is able to translate the word properly. The model with single pivot language translated the word ’reversals’ as ’मागे पडण्याच्या’ which means ’to fall behind’. The model with multiple pivot languages correctly translated the word ’reversals as ’उलथापालथींच्या’ which means ’of-upheavals’.

The model with a single pivot language transliterated the word ’star’ to ’स्टार’ whereas the model with multiple pivot languages correctly translated the word ’star’ to ’तारा’.

6 Conclusion and Future Work

In this work, we show that using multiple pivot languages to assist the source-target NMT model improves its performance. We show using various metrics such as BLEU and chrF, that using multiple Indic languages as pivot languages and utilizing language relatedness improves the performance of the English-Marathi NMT model. We also show that strategic decoder initialization techniques while performing pivot language-based transfer learning improves the performance of the source-target NMT models. In the future, we plan to perform experiments by adding more pivot languages to assist the source to target the NMT model and see the performance of the system.

References

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. 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 866–876, Hong Kong, China. Association for Computational Linguistics.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2020. Utilizing language relatedness to improve machine translation: A case study on languages of the indian subcontinent. *arXiv preprint arXiv:2003.08925*.

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. *arXiv preprint arXiv:1904.01038*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.