IIIT Hyderabad Submission To WAT 2021: Efficient Multilingual NMT systems for Indian languages

Sourav Kumar, Salil Aggarwal, Dipti Misra Sharma
LTRC, IIIT-Hyderabad
sourav.kumar@research.iiit.ac.in
salil.aggarwal@research.iiit.ac.in
dipti@iiit.ac.in

Abstract

This paper describes the work and the systems submitted by the IIIT-Hyderbad team (Id: IIIT-H) in the WAT 2021 (Nakazawa et al., 2021) MultiIndicMT shared task. The task covers 10 major languages of the Indian subcontinent. For the scope of this task, we have built multilingual systems for 20 translation directions namely English-Indic (one-to-many) and Indic-English (many-to-one). Individually, Indian languages are resource poor which hampers translation quality but by leveraging multilingualism and abundant monolingual corpora, the translation quality can be substantially boosted. But the multilingual systems are highly complex in terms of time as well as computational resources. Therefore, we are training our systems by efficiently selecting data that will actually contribute to most of the learning process. Furthermore, we are also exploiting the language relatedness found in between Indian languages. All the comparisons were made using BLEU score and we found that our final multilingual system significantly outperforms the baselines by an average of **11.3** and **19.6** BLEU points for English-Indic (en-xx) and Indic-English (xx-en) directions, respectively.

1 Introduction

Good translation systems are an important requirement due to substantial government, business and social communication among people speaking different languages. Neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017) is the current state-of-the-art approach for Machine Translation in both academia and industry. The success of NMT heavily relies on substantial amounts of parallel sentences as training data (Koehn and Knowles, 2017) which is again an arduous task for low resource languages like Indian languages (Philip et al., 2021). Many techniques have been devised to improve the translation quality of low resource languages like back translation (Sennrich et al., 2015), dual learning (Xia et al., 2016), transfer learning (Zoph et al., 2016; Kocmi and Bojar, 2018), etc. Also, using the traditional approaches, one would still need to train a separate model for each translation direction. So, building multilingual neural machine translation models by means of sharing parameters with high-resource languages is a common practice to improve the performance of low-resource language pairs (Firat et al., 2017; Johnson et al., 2017; Ha et al., 2016). Low resource language pairs perform better when combined opposed to the case where the models are trained separately due to sharing of parameters. It also enables training a single model that supports translation from multiple source languages to a single target language or from a single source language to multiple target languages. This approach mainly works by combining all the parallel data in hand which makes the training process quite complex in terms of both time and computational resources (Arivazhagan et al., 2019). Therefore, we are training our systems by efficiently selecting data that will actually contribute to most of the learning process. Sometimes, this learning is hindered in case of language pairs that do not show any kind of relatedness among themselves. But on the other hand, Indian languages exhibit a lot of lexical and structural similarities on account of sharing a common ancestry (Kunchukuttan and Bhattacharyya, 2020). Therefore, in this work, we have exploited the lexical similarity of these related languages to build efficient multilingual NMT systems.

This paper describes our work in the WAT 2021 MultiIndicMT shared task (cite). The task

Proceedings of the 8th Workshop on Asian Translation, pages 212–216

Domain	PMI	Cvit	IITB	ocor	m2o	ufal	Wmat	ALT	JW	Osub	Ted	Wtile	nlpc	Tanz	urst	Bible
Vocab Overlap	100	74.14	72.04	70.60	65.30	47.47	42.93	31.12	29.99	22.44	22.15	16.70	16.28	14.86	10.58	10.09

Table 1: Vocab Overlap of domains with PMI

covers 10 Indic Languages (Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu) and English. The objective of this shared task is to build translation models for 20 translation directions (English-Indic and Indic-English). This paper is further organized as follows. Section 2 describes the methodology behind our experiments. Section 3 talks about the experimental details like dataset pre-processing and training details. Results and analysis have been discussed in Section 4, followed by conclusion in Section 5.

2 Methodology

2.1 Exploiting Language Relatedness

India is one of the most linguistically diverse countries of the world but underlying this vast diversity in Indian languages are many commonalities. These languages exhibit lexical and structural similarities on account of sharing a common ancestry or being in contact for a long period of time (Bhattacharyya et al., 2016). These languages share many common cognates and therefore, it is very important to utilize the lexical similarity of these languages to build good quality multilingual NMT systems. To do this, we are using the two different approaches namely **Unified Transliteration** and **Sub-word Segmentation** proposed by (Goyal et al., 2020).

2.1.1 Unified Transliteration

The major Indian languages have a long written tradition and use a variety of scripts but correspondences can be established between equivalent characters across scripts. These scripts are derived from the ancient Brahmi script. In order to achieve this, we transliterated all the Indian languages into a common Devanagari script (which in our case is the script for Hindi) to share the same surface form. This unified transliteration is a string homomorphism, replacing characters in all the languages to a single desired script.

2.1.2 Subword Segmentation

Despite sharing a lot of cognates, Indian languages do not share many words at their non-root level. Therefore, the more efficient approach is to exploit Indian languages at their sub-word level which will ensure more vocabulary overlap. Therefore, we are converting every word to sub-word level using the very well known technique **Byte Pair Encoding (BPE)** (Sennrich et al., 2015). This technique is applied after the unified transliteration in order to ensure that languages share same surface form (script). BPE units are variable length units which provide appropriate context for translation systems involving related languages. Since their vocabularies are much smaller than the morpheme and word-level models, data sparsity is also not a problem. In a multilingual scenario, learning BPE merge rules will not only find the common sub-words between multiple languages but it also ensures consistency of segmentation among each considered language pair.

2.2 Data Selection Strategy

Since the traditional approaches of training a multilingual system simply work by combining all the parallel dataset in hand, making it infeasible in terms of both time as well as computational resources. Therefore, in order to select only the relevant domains, we are incrementally adding all the domains in decreasing order of their vocab overlap with the PMI domain (Haddow and Kirefu, 2020). Detection of dip in the BLEU score (Papineni et al., 2002) is considered as the stopping criteria for our strategy. The vocab overlap between any two domains is calculated using the formula shown below:

$$\text{Vocab Overlap} = \frac{|Vocab_{d1} \cap Vocab_{d2}|}{max(|Vocab_{d1}|, |Vocab_{d2}|)} * 100$$

Here, $Vocab_{d1}$ & $Vocab_{d2}$ represents vocabulary of domain 1 and domain 2 respectively. Vocab overlap of each domain with PMI is shown in **Table 1**.

2.3 Back Translation

Back translation (Sennrich et al., 2015)is a widely used data augmentation method where the reverse direction is used to translate sentences from target side monolingual data into the source language. This synthetic parallel data is combined with the actual parallel data to re-train the model leading to better language modelling on the target side, regularization and target domain adaptation. Back

Dataset	**En-hi**	**En-pa**	**En-gu**	**En-mr**	**En-bn**	**En-or**	**En-kn**	**En-ml**	**En-ta**	**En-te**	
Parallel corpus											
PMI	50349	28294	41578	28974	23306	31966	28901	26916	32638	33380	
CVIT	266545	101092	58264	114220	91985	94494	-	43087	115968	44720	
IITB	1603080	-	-	-	-	-	-	-	-	-	
Monolingual corpus											
	En	**Hi**	**Pa**	**Gu**	**Mr**	**Bn**	**Or**	**Kn**	**Ml**	**Ta**	**Te**
PMI	89269	151792	87804	123008	118848	116835	103331	79024	81786	90912	111325

Table 2: Training dataset statistics

translation is particularly useful for low resource languages. We use back translation to augment our multilingual models. The back translation data is generated by multilingual models in the reverse direction, hence some implicit multilingual transfer is incorporated in the back translated data also. For the scope of this paper, we have used monolingual data of the PMI given on the WAT website.

2.4 Multilingual NMT and Fine-tuning

Multilingual model enables us to translate to and from multiple languages using a shared word piece vocabulary, which is significantly simpler than training a different model for each language pair. We used the technique proposed by Johnson et al. (2017) where he introduced a "language flag" based approach that shares the attention mechanism and a single encoder-decoder network to enable multilingual models. A language flag or token is part of the input sequence to indicate which direction to translate to. The decoder learns to generate the target given this input. This approach has been shown to be simple, effective and forces the model to generalize across language boundaries during training. It is also observed that when language pairs with little available data and language pairs with abundant data are mixed into a single model, translation quality on the low resource language pair is significantly improved. Furthermore, We are also fine tuning our multilingual system on PMI (multilingual) domain by the means of transfer learning b/w the parent and the child model.

3 Experimental Details

3.1 Dataset and Preprocessing

We are using the dataset provided in WAT 2021 shared task. Our experiments mainly use PMI (Haddow and Kirefu, 2020), CVIT (Siripragada et al., 2020) and IIT-B (Kunchukuttan et al., 2017) parallel dataset, along with monolingual data of PMI for further improvements **Table 2**. We used Moses (Koehn et al., 2007) toolkit for tokenization and cleaning of English and Indic NLP library (Kunchukuttan, 2020) for normalizing, tokenization and transliteration of all Indian languages. For our bilingual model we used BPE segmentation with 16K merge operation and for multilingual models we learned the Joint-BPE on source and target side with 16K merges (Sennrich et al., 2015).

3.2 Training

For all of our experiments, we use the **OpenNMT-py** (Klein et al., 2017) toolkit for training the NMT systems. We used the Transformer model with 6 layers in both the encoder and decoder, each with 512 hidden units. The word embedding size is set to 512 with 8 heads. The training is done in batches of maximum 4096 tokens at a time with dropout set to 0.3. We use Adam (Kingma and Ba, 2014) optimizer to optimize model parameters. We validate the model every 5,000 steps via BLEU (Papineni et al., 2002) and perplexity on the development set. We are training all of our models with early stopping criteria based on validation set accuracy. During testing, we rejoin translated BPE segments and convert the translated sentences back to their original language scripts. Finally, we evaluate the accuracy of our translation models using BLEU.

4 Results and Analysis

We report the Bleu score on the test set provided in the WAT 2021 MultiIndic shared task. **Table 3** and **Table 4** represents the results for different experiments we have performed for En-XX and XX-En directions respectively. The rows corresponding to *PMI + CVIT + Back Translation + Fine tuning on PMI multilingual* is our final system submitted for this shared task (Bleu scores shown in the table for this task are from automatic evaluation system). We observe that Multilingual system of PMI outperforms the bilingual baseline model of PMI by significant margins. The reason for this is the abil-

En-XX	en-hi	en-pa	en-gu	en-mr	en-bn	en-or	en-kn	en-ml	en-ta	en-te
PMI Baselines	23.21	18.26	15.46	7.07	5.25	8.32	8.67	4.63	5.32	6.12
PMI Multilingual	28.22	26.00	21.19	13.37	10.53	14.78	15.39	8.99	9.38	8.57
PMI + CVIT Multilingual	32.86	28.29	23.85	16.74	11.71	16.79	15.63	10.71	11.85	9.18
PMI + CVIT + IITB Multilingual	32.68	23.55	22.36	15.74	8.66	13.88	13.71	8.03	9.23	7.31
PMI + CVIT + Back Translation	35.81	30.15	25.84	18.47	12.50	18.52	17.98	11.99	12.31	12.89
PMI + CVIT + Back Translation + Fine Tuning on PMI Multilingual	**38.25**	**33.35**	**26.97**	**19.48**	**14.73**	**20.15**	**19.57**	**12.76**	**14.43**	**15.61**

Table 3: Results for En-XX direction

XX-En	hi-en	pa-en	gu-en	mr-en	bn-en	or-en	kn-en	ml-en	ta-en	te-en
PMI Baselines	24.69	19.80	20.16	11.70	10.25	13.80	13.32	11.30	9.82	13.39
PMI Multilingual	26.91	24.26	23.91	19.66	17.44	19.65	21.08	18.99	18.95	19.94
PMI + CVIT Multilingual	39.40	37.35	35.12	29.59	25.35	30.38	29.56	27.69	28.12	28.97
PMI + CVIT + IITB Multilingual	37.93	36.08	35.03	28.71	24.18	29.04	28.95	27.24	27.61	28.41
PMI + CVIT + Back Translation	41.41	39.15	37.84	32.17	26.90	32.52	32.58	28.99	29.31	30.29
PMI + CVIT + Back Translation+ Fine Tuning on PMI Multilingual	**43.23**	**41.24**	**39.39**	**34.02**	**28.28**	**34.11**	**34.69**	**29.19**	**29.61**	**30.44**

Table 4: Results for XX-En direction

ity to induce learning from multiple languages; also there is increase in vocab overlap using our technique of exploiting language relatedness. Further we tried to improve the performance of system using the relevant domains by incrementally adding different domains based on vocab overlap to the already existing system. We observed a decrease in Bleu score after adding the IIT-B corpus and therefore we stopped our incremental training at that point. Further we can see that our final multilingual model using back translation and fine tuning outperforms all other systems. Our submission also got evaluated with AMFM scores which can be found in the WAT 2021 evaluation website.

5 Conclusion

This paper presents the submissions by IIIT Hyderabd on the WAT 2021 MultiIndicMT shared Task. We performed experiments by combining different pre-processing and training techniques in series to achieve competitive results. The effectiveness of each technique is demonstrated. Our final submission able to achieve the second rank in this task according to automatic evaluation.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Pushpak Bhattacharyya, Mitesh M Khapra, and Anoop Kunchukuttan. 2016. Statistical machine translation between related languages. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 17–20.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

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

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*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand. 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.

Jerin Philip, Shashank Siripragada, Vinay P Namboodiri, and CV Jawahar. 2021. Revisiting low resource status of indian languages in machine translation. In *8th ACM IKDD CODS and 26th COMAD*, pages 178–187.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Shashank Siripragada, Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2020. A multilingual parallel corpora collection effort for indian languages. *arXiv preprint arXiv:2007.07691*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *arXiv preprint arXiv:1611.00179*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.