ANVITA Machine Translation System for WAT 2021 MultiIndicMT Shared Task

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul,
Chitra Viswanathan, Prasanna Kumar K R
Centre for Artificial Intelligence and Robotics
CV Raman Nagar, Bangalore
{pavanpankaj333@gmail.com, jsbhavani@cair.drdo.in, biswajit@cair.drdo.in,
chitrav@cair.drdo.in, prasanna@cair.drdo.in }

Abstract

This paper describes ANVITA-1.0 MT system, architected for submission to WAT 2021 MultiIndicMT shared task by mcairt team, where the team participated in 20 translation directions: English→Indic and Indic→English; Indic set comprised of 10 Indian languages. ANVITA-1.0 MT system comprised of two multi-lingual NMT models one for the English→Indic directions and other for the Indic→English directions with shared encoder-decoder, catering 10 language pairs and twenty translation directions. The base models were built based on Transformer architecture and trained over MultiIndicMT WAT 2021 corpora and further employed back-translation and transliteration for selective data augmentation, and model ensemble for better generalization. Additionally, MultiIndicMT WAT 2021 corpora was distilled using a series of filtering operations before putting up for training. ANVITA-1.0 achieved highest AM-FM score for English→Bengali, 2nd for English→Tamil and 3rd for English→Hindi, Bengali→English directions on official test set. In general, performance achieved by ANVITA for the Indic→English directions are relatively better than that of English→Indic directions for all the 10 language pairs when evaluated using BLEU and RIBES, although the same trend is not observed consistently when AM-FM based evaluation was carried out. As compared to BLEU, RIBES and AM-FM based scoring placed ANVITA relatively better among all the task participants.

1 Introduction

This paper presents ANVITA-1.0 (A Neural Version of Indic Translation Assistance) MT system, architected for submission to WAT 2021 MultiIndicMT shared task by mcairt team. WAT 2021 MultiIndicMT shared task (Nakazawa et al., 2021) comprised of translation of 10 Indian languages Bengali(bn), Gujarati(gu), Hindi(hi), Kannada(kn), Marathi(mr), Malayalam(ml), Oriya(or), Punjabi(pa), Tamil(ta), Telugu(te) and Engish(en) in 20 translation directions (English→Indic and Indic→English) and our team participated in all 20 translation directions.

Developing quality machine translation system for the Indian languages still remains a major challenge, as large number of Indian languages are individually resource poor which greatly impacts translation quality. However some of the recent developments do show that careful utilization of multilingualism and/or monolingual corpora, translation quality can be boosted (Johnson et al., 2017; Sennrich et al., 2015).The purpose of WAT 2021 MultiIndicMT shared task is to validate the utility of MT techniques that focus on multilingualism and/or monolingual data in the context of Indian languages.

Our ANVITA-1.0 is realized as a Multilingual Neural Machine Translation(MNMT) system based on Transformer architecture (Vaswani et al., 2017). As transformer is sensitive to training noise (Liu et al., 2018), we have rigorously cleaned up the training corpus by applying set of heuristics. For better transfer of translation knowledge among the language pairs, ANVITA-1.0 used multilingual NMT approach and trained two models, one for the English→Indic and one for the Indic→English with shared encoder-decoder similar to MNMT models described by Johnson et.al (Johnson et al., 2017). Additionally, we employed back-translation (Sennrich et al., 2015) and transliteration techniques between related languages (Li et al., 2019) for selective data augmentation followed by model ensemble for better generalization. As Indian languages are morphologically rich, instead of word level tokenization, ANVITA-1.0 employed subword level tokenization, sentence piece (Kudo and Richardson, 2018) before putting up for training.

Proceedings of the 8th Workshop on Asian Translation, pages 244–249
Bangkok, Thailand (online), August 5-6, 2021.

Details are mentioned in the subsequent sections.

ANVITA-1.0 achieved highest AM-FM score for English→Bengali, 2nd for English→Tamil and 3rd for English→Hindi, Bengali→English directions on the official WAT 2021 MultiIndicMT test set. Overall, as compared to BLEU, RIBES and Adequacy-Fluency based scoring relatively placed us better in the ranking chart.

2 Related Work

A comprehensive survey covering challenges, design choices and other aspects related to Multilingual Neural Machine Translation(MNMT) was presented by Dabre et.al (Dabre et al., 2020). Siripragada et al. (2020) published a low resource Indian language dataset and trained a Multilingual NMT model on it. Aharoni et al. (2019) presented a massive multilingual neural translation model with 102 languages. Li et al. (2019) has done rigorous filtering of parallel corpora. Liu et al. (2018) and Pinnis (2018) have proposed some heuristics for rigorous filtering of noise from parallel corpora. Li et al. (2019) have proposed combining parallel corpora by transliteration of related languages(grammar similarity) which improves performance. Back translation (Sennrich et al., 2015) is considered by many as one of the effective mechanism for enhancing MT performance.

3 Data sets

ANVITA-1.0 was primarily trained using MultiIndicMT WAT 2021[1] corpora. Additionally AI4Bharat[2] monolingual corpora was used for generating synthetic parallel data by back translation. No other additional corpora or linguistic resources were used in ANVITA-1.0.

MultiIndicMT WAT 2021 corpora (Nakazawa et al., 2021) as shared by the organizer comprises of approximately 10 million parallel sentences covering 10 language pairs (Indic, English) and sourced from the following multiple datasets. CVIT-PIB, PMIndia, IITB 3.0, JW, NLPC, UFAL EnTam, Uka Tarsadia, Wikititles, ALT, OpenSubtitles, Bible-uedin, MTEnglish2Odia, OdiaEncorp2.0, TED, WikiMatrix. MultiIndicMT WAT 2021 training corpora is summarised in Table-1. Hindi↔English has the highest number of sentence pair and Oriya↔English lowest.

[1]http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/
[2]https://indicnlp.ai4bharat.org/corpora/

Sl. No.	Indic↔En	# Sentences	%Share
1	bn-en	1302940	13.52%
2	gu-en	518179	5.37%
3	hi-en	3070239	31.86%
4	kn-en	396882	4.11%
5	ml-en	1142115	11.85%
6	mr-en	621725	6.45%
7	or-en	252160	2.62%
8	pa-en	518520	5.38%
9	ta-en	1354374	14.05%
10	te-en	457523	4.75%

Table 1: Statistics of MultiIndicMT WAT 2021 training corpora (before filtering)

4 System Overview

This section describes ANVITA-1.0 MT system and its subsystems with reasonable details.

4.1 Data Preprocessing

This section presents set of preprocessing steps employed by ANVITA-1.0.

4.1.1 Data Filtering

Like most automatically curated corpora and corpora compiled from such curated corpus, MultiIndicMT WAT 2021 corpora is also not free from noises. A quick glance through the corpora provided with a rough assessment of noises present and aided in employing set of heuristics to filter out many of those noisy sentence pairs. This is all the more critical as transformer based models are sensitive to noises (Liu et al., 2018). Rigorous distillation of training corpora was carried by employing set of heuristics similar to as described by Bei Li (Li et al., 2019). The heuristics applied for filtering out noises from MultiIndicMT WAT 2021 corpora are as given below.

- Filter out sentence pair, in which either source or target sentence is empty.
- Filter out sentence pair, in which either source or target sentence length greater than 800 characters.
- Filter out sentence pair in which length of source and target sentence ratio is greater than 2.5.
- Filter out sentence pair in which length of source and target sentence ratio is less than 0.4.

Indic↔En	# Sentences	%Share	%Filtered
bn-en	1198915	13.73%	7.98%
gu-en	491036	5.62%	5.23%
hi-en	2885632	33.05%	6.01%
kn-en	336967	3.85%	15.09%
ml-en	967909	11.08%	15.25%
mr-en	587576	6.72%	5.49%
or-en	245077	2.80%	2.81%
pa-en	493337	5.65%	4.85%
ta-en	1123269	12.85%	17.06%
te-en	401318	4.60%	12.28%

Table 2: Statistics of MultiIndicMT WAT 2021 training corpora after filtering noisy sentence pairs

- Filter out sentence pair , if source or target sentence contains word having length greater than 10.
- Filter out sentence pair, if source sentence has at least 10 characters of other language.
- Filter out sentence pair, if source sentence has at least 60% characters of other language (used utf-8 ranges for other language character identification).

Approximately 15% of the total sentence pairs, amounting to 1.5 million sentence pairs were tagged as noisy after applying the above heuristics and were filtered out from the MultiIndicMT WAT 2021 training corpora. Detailed corpus statistics after filtering operation is given in Table-2. Final training data size after filtering turned out to be 8731036 sentence pairs. Data filtering improved both translation performance and convergence rate.

4.1.2 Tokenization at Sub-word Level

To effectively make use of the morphological richness property of Indian languages, sub-word level tokenization is employed instead of word or character level tokenization.

English→Indic: Sentence piece tokenizer (Kudo and Richardson, 2018) was used with 80K joint vocabulary of 10 target Indic languages, 16K vocabulary of English and character coverage of 1.0.

Indic→English: Sentence piece tokenizer (Kudo and Richardson, 2018) was used with 48K joint vocabulary of 10 Indic source languages, 16K vocabulary of English and character coverage of 1.0.

Indic↔English	Special Token
bn-en	@%+@
gu-en	{%-}
hi-en	—_-^—
kn-en	&*—&
ml-en	?:/?
mr-en	#_+#
or-en	=&-=
pa-en	~&[~
ta-en	:*&:
te-en	*]*

Table 3: Special tokens used for tagging language pairs at the source side

4.1.3 Tagging of Source Sentences

To guide the input-output sequence mapping task better under multilingual setting, all sentences at the source side were tagged with language pair information using special tokens and placed at the beginning of each source sentence (Johnson et al., 2017). Special language tokens consisted of 4 characters and all having special symbols. Special symbols were used to avoid overlapping of language tokens with data tokens and token lengths were decided based on minimum number of characters required to tag 10 language pairs distinctly. Language tokens were used only at the source side during training of both the models i.e Indic→English and English→Indic models. Table-3 lists out the language tokens used.

4.2 Data Augmentation

Data augmentation has become a de-facto step for low resource MT. Following strategies were applied for augmenting data in ANVITA-1.0.

4.2.1 Related Language Transliteration

As most of the languages fall under low resource category, we employed related-language transliteration strategy for the top three low resource languages. Relatedness is decided based on similarities between languages (Li et al., 2019). Top three low resource languages as found in MultiIndicMT WAT 2021 corpora are Oriya(or), Kannada(kn), and Gujarati(gu). To the best of our knowledge, related languages of these three low resource Indian languages are listed in Table-4. Relatively high resource related language training data were transliterated into low resource language using transliterated method as described by Ahmad Bhat et

Low Resource Language	Related Language
Oriya	Bengali
Kannada	Telugu
Gujarati	Hindi

Table 4: Related languages of top three low resource languages

Language Pair	# Sentence (%Share)	
Indic↔English	Indic→En	En→Indic
bn-en	1198915 (7.67%)	1198915 (9.07%)
gu-en	3976668 (25.46%)	3376668 (25.54%)
hi-en	2885632 (18.47%)	2885632 (21.83%)
kn-en	1338285 (8.56%)	738285 (5.58%)
ml-en	967909 (6.19%)	967909 (7.32%)
mr-en	587576 (3.76%)	587576 (4.44%)
or-en	2043992 (13.08%)	1443992 (10.92%)
pa-en	1093337 (7.00%)	493337 (3.73%)
ta-en	1123269 (7.19%)	1123269 (8.49%)
te-en	401318 (2.56%)	401318 (3.03%)

Table 5: Statistics of final training data after applying transliteration and back translation

al. (Bhat et al., 2014) and added to the low resource language training data. For instance, Bengali sentences were transliterated into Oriya and augmented with Oriya training data.

As Marathi and Hindi languages both share the same script, so in order to avoid script overlapping, we mapped characters of Marathi sentences to Unicode Block 0D80- 0DFF. This seems to have reduced sharing of translation knowledge and impacted results. However this needs to be verified further through experimentation.

4.2.2 Back Translation

Back translation (Sennrich et al., 2015) is considered as one of the effective mechanism for enhancing MT performance, specially involving low resource languages. As most of languages in the task involved are low resource, back translation was applied for the top four low resource languages observed in the MultiIndicMT WAT 2021 corpora namely Oriya, Kannada, Punjabi, and Gujarati. We extracted monolingual corpora of 6 lakh sentences for each of the four low resource language pair from the AI4Bharat (Kakwani et al., 2020) corpora for the purpose. Statistics of the final training corpora after data augmentation is shown in Table-5.

4.3 Model Training

ANVITA-1.0 was trained based on Transformer architecture and for better sharing of knowledge among Indian languages, specially for resource poor languages, two multilingual models were trained in (a) One-to-Many fashion for English→Indic and (b) Many-to-One fashion for Indic→English with shared encoder-decoder, similar to as described by Johnson et.al (Johnson et al., 2017).

Ensembling of multiple models, which are diverse in nature, have shown improvement of translation performance and better generalization (Li et al., 2019). Due to time and resource limitations, we could not work out on diverse models. However, we ensemble last 5 checkpoints i.e (560000-600000 iterations).

5 Experimental Details

ANVITA-1.0 used OpenNMT-py 2.0 (Klein et al., 2017) toolkit for training. Training configuration are 600000 steps for Indic→English, 440000 steps for English→Indic, with batch size of 4096, dropout 0.1, batch type tokens, adam optimizer, warmup steps 8000, word embedding size 512, encoder layers 6, decoder layers 6, heads 8,feed forward dimension of 2048, rnn size 512 and noam as learning rate decay method. ANVITA-1.0 was trained on NVIDIA DGX machine having 4 V100 GPU cards, each having 32GB of GPU memory. Training of Indic→English took approximately 96 hours and English→Indic took approximately 72 hours.

6 Evaluation and Results

Translation quality of ANVITA-1.0 was assessed by the organizer (Nakazawa et al., 2021) on the official WAT 2021 MultiIndicMT test set using BLEU, RIBES(Isozaki et al., 2010) and Adequacy-Frequency(Banchs et al., 2015) based metrics. The official evaluation results as declared by the organizer for all the 20 translation directions are shown in Table-6 and Table-7.

Performance of Indic→English 10 translation directions ranges from 27.29 to 40.05 BLEU points, where Marathi→English happens to be the lowest and Hindi→English highest scorers respectively. For English→Indic 10 translation directions performance ranges from 35.85 to 6.17 BLEU points, in which English→Malayalam scored lowest and English→Hindi highest. We believe that, because of the relatively high resource nature of Hindi↔English language pair, this particular pair outperformed all other pairs.

Indic→English	BLEU	RIBES	AM-FM
bn→en	29.96	0.798326	0.786717
gu→en	36.77	0.829389	0.819546
hi→en	40.05	0.850322	0.832119
kn→en	31.16	0.803525	0.799216
ml→en	28.07	0.792884	0.794932
mr→en	27.29	0.785579	0.780231
or→en	29.96	0.798326	0.795586
pa→en	38.42	0.840360	0.818332
ta→en	28.04	0.793839	0.790184
te→en	29.26	0.790319	0.786396

Table 6: Performance of ANVITA-1.0 for Indic→English directions on the official WAT 2021 MultiIndicMT test set.

English→Indic	BLEU	RIBES	AM-FM
en→bn	13.02	0.715490	0.779592
en→gu	23.21	0.809389	0.816739
en→hi	35.85	0.846656	0.822626
en→kn	14.58	0.726259	0.805963
en→ml	6.17	0.622598	0.793308
en→mr	14.90	0.740079	0.791850
en→or	17.71	0.743984	0.763064
en→pa	30.56	0.830405	0.810106
en→ta	11.98	0.707054	0.801632
en→te	11.17	0.702337	0.783647

Table 7: Performance of ANVITA-1.0 for English→Indic directions on the official WAT 2021 MultiIndicMT test set

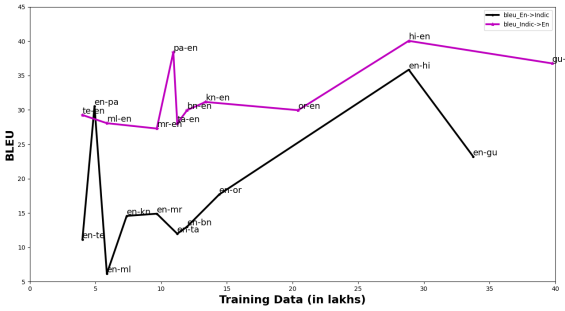

Figure 1: Performance of ANVITA-1.0 wrt size of training data, when evaluated using BLEU for Indic→English and English→Indic directions on the official WAT 2021 MultiIndicMT test set.

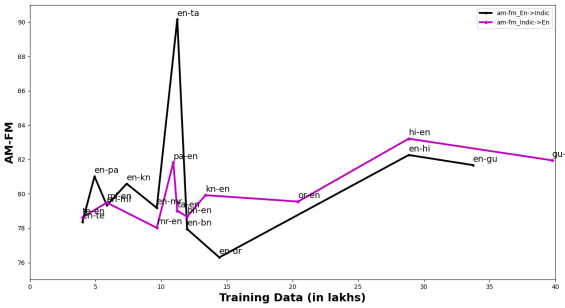

Figure 2: Performance of ANVITA-1.0 wrt size of training data, when evaluated using AM-FM scores for Indic→English and English→Indic directions on the official WAT 2021 MultiIndicMT test set.

Figure-1 and Figure-2 show how performance of ANVITA-1.0 changes as a parameter of training data size. This evaluated was carried out on the official WAT 2021 MultiIndicMT test set using BLEU and AM-FM metrics. Barring few exceptions, training data size seems to be positively correlated with the translation performance. The exceptions are possibly due to implicit transfer of translation knowledge among the related languages.

7 Conclusion and Future Directions

The overall translation performance achieved by ANVITA-1.0 for the Indic→English directions are encouraging. Data augmentation largely aided the relatively lower resource languages well. Transfer of translation knowledge through shared encoder-decoder seems to be aided the related language better and data filtering improved the overall performance. RIBES and AM-FM based scoring placed us relatively better than BLEU.

Translation performance figures for the Indic→English directions achieved by ANVITA-1.0 are relatively better than that of English→Indic directions for all language pairs, when evaluated using BLEU and RIBES, though the same trend is not observed consistently when AM-FM based evaluation was carried out. Potential reasons could be One to Many mapping is relatively harder to learn as compared to Many to One mapping with shared decoder. One of the future direction would be to closely investigate whether having shared encoder but separate decoders helps for One-to-Many models in the Indic context. Though we have applied a large number of data filtering heuristics, we noticed that training data was still not free from noises. So another potential future direction would be to explore more effective data filtering techniques and its impacts on MT performance. Exploration of additional data augmentation strategies and effective transfer of

translation knowledge, their shares in improving MT performance would be a critical direction when it comes to handling low resource languages. Having more diverse parallel corpora for the Indian languages will help Indic MT tasks and automated methods for compilation of large and diverse Indic corpus is a much needed one.

8 Acknowledgments

The authors would like to thank Director, CAIR for his constant encouragement, guidance and enablement.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Rafael E. Banchs, Luis F. DHaro, and Haizhou Li. 2015. Adequacyfluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2014. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, pages 48–53.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A comprehensive survey of multilingual neural machine translation. *arXiv preprint arXiv:2001.01115*.

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

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Googles 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, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4948–4961.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, et al. 2019. The niutrans machine translation systems for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2018. Robust neural machine translation with joint textual and phonetic embedding. *arXiv preprint arXiv:1810.06729*.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondej 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.

Mārcis Pinnis. 2018. Tildes parallel corpus filtering methods for wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 939–945.

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*.

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*.