CVIT's Submissions to WAT-2019

Jerin Philip [*†]
IIIT Hyderabad

Shashank Siripragada[*‡]
IIIT Hyderabad

Upendra Kumar
IIIT Hyderabad

Vinay P. Namboodiri
IIT Kanpur

C.V. Jawahar
IIIT Hyderabad

Abstract

This paper describes the Neural Machine Translation systems used by IIIT Hyderabad (CVIT-MT) for the translation tasks part of WAT-2019. We participated in tasks pertaining to Indian languages and submitted results for English-Hindi, Hindi-English, English-Tamil and Tamil-English language pairs. We employ Transformer architecture experimenting with multilingual models and methods for low-resource languages.

1 Introduction

Neural Machine Translation (NMT) has emerged as the de-facto standard for language translation following the success of deep learning. Recurrent Neural Networks (Sutskever et al., 2014), Convolutional sequence to sequence (Gehring et al., 2017) and pure attention based Transformer (Vaswani et al., 2017) architectures have incrementally improved translation numbers over the years.

Recent works demonstrate success in training multiway among several languages while sharing parameters and learning across languages (Aharoni et al., 2019; Artetxe and Schwenk, 2018). Multiway models enable few-shot learning among several pairs of languages for which parallel data does not exist in training by being able to implicitly pivot (Johnson et al., 2017) through parameter sharing across languages.

Despite the success of NMT and surrounding research in neural methods in other languages around the world, not many successful NMT systems or trained models for Indian languages are available for public use at the time of writing this paper. Indian languages pose a challenge for NMT due to scarcity of parallel corpora across many languages.

In this edition of Workshop on Asian Translation (WAT) (Nakazawa et al., 2019), we explore multiway-models for Indian Languages, improving upon our WAT 2018 submissions in the IIT-Bombay Hindi-English tasks. We pursue two approaches to the UFAL English-Tamil tasks, one training from scratch (cold-start) and fine-tuning an already trained model from a pretrained model on a different dataset (warm-start).

The rest of this document is organized as follows: Section 2 outlines ideas used in the task. Section 3 details the implementation and in Section 4 we summarize our findings.

2 System Components

NMT is commonly formulated in literature within an encoder-decoder framework. An encoder consumes the source-side sequence and provides representations rich in context across the sentence. The decoder along with an attention module looks at the encoded-representations of the source-sequence and generated target-language tokens to predict the token at the current time-step.

In our experiments, we use the Transformer architecture (Vaswani et al., 2017) which is state-of-the-art in several natural language tasks such as Translation, Language modelling (Lample and Conneau, 2019) and Language understanding (Devlin et al., 2019). The transformer is used in both the encoder and decoder.

2.1 Multiway Translation Models

Recent advances and extensive studies (Aharoni et al., 2019; Johnson et al., 2017) suggest using multilingual models to get best results and robust translation systems. A single model is trained here to translate across several languages sharing parameters. We use a shared encoder and decoder for multiway training, switching between target lan-

[*] Equal contribution.
[†] jerin.philip@research.iiit.ac.in
[‡] shashank.siripragada@alumni.iiit.ac.in

Proceedings of the 6th Workshop on Asian Translation, pages 131–136
Hong Kong, China, November 4, 2019.

guages by use of a special token (`__t2xx__`) following Johnson et al. (2017).

2.2 Backtranslation

One widely successful method to exploit monolingual data to improve the NMT systems is backtranslation proposed by Sennrich et al. (2016) wherein an NMT system trained from target to source is used to translate the monolingual data. The synthetic parallel data thus obtained is used to augment the source to target NMT system. We employ backtranslation in both the multiway model and the model trained from scratch.

2.3 Low-Resource settings

It has been shown that the performance of neural machine translation (NMT) drops in low-resource conditions, underperforming statistical machine translation (SMT). Sennrich and Zhang (2019) argue that this is due to lack of system adaptation to low-resource settings. They demonstrate that with suitable choice of parameters in low-data setting NMT systems can outperform Phrase Based SMT (PBSMT). To this end they propose reduction of subword vocabulary size, aggressive dropout, label smoothing and some more set of best practices. Following their settings for our English Tamil model, we restrict the subword vocabulary size of English and Tamil to 2000 each. We also use layer normalization after every encoder and decoder layers and label smoothing.

3 Experimental Setup

In this section, we describe our setup in detail. In 3.1, we describe the multiway system which gave the best numbers for the Hindi-English tasks provided by the IIT-Bombay Hindi-English corpus, followed by the setup for UFAL English-Tamil task in 3.2. 3.3 discusses evaluation metrics common to both tasks.

3.1 Indian Language Multiway System

We use The IIT-Bombay English-Hindi (IITB-hi-en) (Kunchukuttan et al., 2018) corpus provided by the organizers. This dataset supplies parallel corpus for English-Hindi as well as monolingual Hindi corpus. We use noisy backtranslated Hindi-English corpus obtained through our previous models for the same task translating Hindi monolingual data provided by IITB-hi-en to English. In addition to this, we use the Indian Language Corpora Initiative Corpus (ILCI) (Jha, 2010) and the Indian Language Multi Parallel Corpus (WAT-ILMPC) (Nakazawa et al., 2018) consisting of subtitles provided as training data for WAT-2018.

Source	#pairs	type
IITB-hi-en	1.5M	en-hi
Backtranslated-Hindi	2.5M	en-hi
WAT-ILMPC	188K	xx-en
ILCI	50K	xx-yy
Backtranslated-wiki	10.4M	mono

Table 1: Training dataset used for ilmulti model. xx-yy indicates parallel sentences aligned across multiple languages. xx-en indicates bilingual corpora with English in one direction.

We use pairs obtained among Hindi (hi), English (en), Tamil (ta), Malayalam (ml), Telugu (te) and Urdu (ur) from the datasets mentioned in Table 1 in training our model hereafter referred to as `ilmulti` .

We use sentences extracted from Wikipedia dumps of the respective languages, monolingual data provided by WAT-ILMPC and some additionally crawled news-articles for further backtranslation to obtain more training samples across languages. We backtranslate only to Hindi and English from other low-resource languages since the BLEU scores for the other directions were not promising. We refer the reader to Philip et al. (2019) for comprehensive information on the data used in training this model and multilingual comparisons on other test-sets.

Preprocessing and Filtering We use trained SentencePiece (Kudo, 2018)[1] models to tokenize the sentences in all languages and source to target token count ratio to filter sentences. We chose sentences whose source to target ratio is between 0.8 and 1.2. In addition to this, we use a threshold of 98% language match through `langid.py` (Lui and Baldwin, 2012) to remove sentences that did not belong to the language the parallel corpus was provided for. These methods are applied on both the original training data and the backtranslated corpus added to augment training data.

Training and Inference We use the default configuration provided by `transformer` model in `fairseq` (Ott et al., 2019).[2] Embedding layers of

[1] https://github.com/google/sentencepiece
[2] https://github.com/pytorch/fairseq

No	Model	BLEU		RIBES		AM-FM		Human	
		en-hi	hi-en	en-hi	hi-en	en-hi	hi-en	en-hi	hi-en
1	`ilmulti`	20.17	22.62	0.761061	0.766180	0.701670	0.637230	-	-
2	1 + backtranslation	20.46	**22.91**	0.765422	**0.768324**	0.702380	**0.641730**	-	-
		en-ta	ta-en	en-ta	ta-en	en-ta	ta-en	en-ta	ta-en
3	2 (out of the box inference)	0.80	4.68						
4	2 + UFAL warm-start	10.91	27.14	0.671850	0.770024	0.795160	0.693750	-	-
5	UFAL cold-start	**13.05**	**30.04**	**0.698482**	**0.788588**	**0.801570**	**0.707060**	-	-

Table 2: Translation evaluation scores on IIT-Bombay Hindi-English and UFAL English-Tamil test sets. 3 and 4 indicate BLEU obtained during ilmulti inference out-of-box and warm-start respectively. Bold indicates best values among all submissions at the time of writing this paper.

dimension 512 are in place and are shared among the encoder and decoder (also known in literature as tied embeddings) along with the parameters. Stacked 6 Multi-Head-Attention layers were used to realize both the encoder and decoder. The model is trained with Adam optimizer with the token-wise negative log-likelihood objective. We trained on 4 nodes with 4 NVIDIA 1080Ti GPUs. We used beam-search with beam-size of 10 for generating the translations at test time.

3.2 UFAL English-Tamil Tasks

For UFAL English-Tamil tasks, we explore training single direction models from scratch and fine-tuning our ilmulti model.

Source	#pairs	type
UFAL EnTam	160K	en-ta
Leipzig Newscrawl	300K	ta mono
Kaggle Indian Politics News	300K	en mono

Table 3: Training dataset used for UFAL English-Tamil task.

Dataset For the UFAL English-Tamil translation task we used the EnTam v2.0 dataset (Ramasamy et al., 2012). This parallel corpora covers texts from bible, cinema and news domains. Additional Tamil monolingual data was obtained by sampling a subset of 300K sentences from Leipzig Tamil Newscrawl[3] data to avoid deterioration from noise per Edunov et al. (2018). For English monolingual data, we used a subset of 300K sentences randomly sampled from Kaggle Indian Politics News data[4] which contains 15346 news articles along with their headlines. We have restricted to use of only 300K additional English and Tamil monolingual sentences in order to maintain a appropriate ratio of original and synthetic parallel data after back-translation. Adding too much synthetic parallel data introduces more than feasible noise in already brittle model trained in low-resource settings.

Preprocessing and Filtering We used SentencePiece to restrict the vocabulary size while being able to cover the full text. For the UFAL English-Tamil task we have trained a SentencePiece model separately on English and Tamil corpus restricting the Vocabulary size to 2000 tokens in each language. Pairs with length ratio of target to source sentences less than 0.7 were filtered out from both original as well as backtranslated data.

Backtranslation For backtranslation experiments, we augmented training corpus with additional data comprising of 300K sentences. We obtained the noisy synthetic data for augmentation by translating monolingual data in both en→ta and ta→en directions, using the data described in Table 3. For obtaining synthetic data, beam search with beam size of 5 was used. Edunov et al. (2018) demonstrate that the original parallel data provides much richer training signal as compared to synthetic data generated by beam search. Hence we upsample the original data by a factor of 2 which results in the ratio of UFAL EnTam(∼150K) to synthetic data(∼300K) being 1:1.

Training We used the Transformer-Base implementation available in `fairseq`. The encoder and decoder have 5 layers each with and embedding dimension of 512 and 8 attention heads. The inner-layer dimension is 2048. We apply layer normalization (Ba et al., 2016) before each encoder and decoder layer. We use dropout, weight decay and label smoothing to regularize the model. The model is trained to minimize the label smoothed cross entropy loss using Adam optimizer with label smoothing of 0.2. We run the training on 4

[3] `http://cls.corpora.uni-leipzig.de/en/tam_newscrawl_2011`

[4] `https://www.kaggle.com/xenomorph/indian-politics-news-2018`

Source	A room was arranged for him at Sun Towers Lodge.
Hypothesis	இவருக்கு சன் டோர்ஸ் லோட்ஜில் ஒரு அறை ஏற்பாடு செய்யப்பட்டிருந்தது.
Target	அங்குள்ள சன் டவர்ஸ் லாட்ஜில் அவருக்கு அறை ஏற்பாடாகியிருந்தது.
Source	His administration, however, has been regarded as untenable in the eyes of substantial sections of the ruling class.
Hypothesis	ஆனால் அவருடைய நிர்வாகம் ஆளும் வர்க்கத்தின் கணிசமான பிரிவுகளின் பார்வையில் தக்கவைத்துக் கொள்ள முடியாதது என்று கருதப்படுகிறது.
Target	எவ்வாறாயினும், அவரது நிர்வாகம், ஆளும் வர்க்கத்தின் கணிசமான பிரிவினரது கண்களுக்கு ஏற்புடையதாகத் தோன்றவில்லை.
Source	பிரெஞ்சு முதலாளித்துவத்திற்கு மற்றொரு முண்டுகோல் தேவை
Hypothesis	French capitalism needs another prop.
Target	French capitalism needs another prop
Source	இந்த சூழ்நிலையில் Lufthansa விமானிகளுக்கு சலுகைகளைக் கொடுக்கத் தயாராக இருக்காது.
Hypothesis	Under these conditions, Lufthansa would not be prepared to make concessions to pilots.
Target	Under these circumstances, Lufthansa will hardly be prepared to make any concessions to the pilots.

Table 4: Examples from the test set of correctly translated samples.

NVIDIA 1080Ti GPUs with mini-batches of maximum size of 4K tokens. The model described above is referred to hereafter as Transformer-base.

We further extend the existing `ilmulti` + backtranslation model to UFAL English-Tamil training data domain by warm-starting and training for a few epochs.

Inference and decoding Decoding was performed with beam size of 5 for generation of hypotheses for both en→ta and ta→en tasks. For UFAL-3 and UFAL-5, ensembles of models were used in inference by test time averaging outputs from last 5 checkpoints saved at interval of 10 epochs. In experiment UFAL-6, for generating hypotheses, length penalty of 1.5 for en→ta task and 2.0 for ta→en task was enforced.

3.3 Evaluation

We primarily use Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) scores for comparisons. BLEU is an automatic evaluation metric widely use for translation and is based on precision. N-grams of sizes 1-4 are used to compute precision and the geometric mean of the same is multiplied by a brevity-penalty (BP) to obtain the final score. For aggregate value over a corpus, micro averaging is performed. In addition to BLEU, we report AM-FM, RIBES (Isozaki et al., 2010) and Human Evaluation scores from the submission site, when available.

4 Results and Discussion

Since IITB-hi-en has been widely discussed in the past, we focus on UFAL English Tamil in this paper. We provide both qualitative and quantitative analyses of the results obtained below.

4.1 Quantitative Results

IITB-en-hi The automated evaluation scores for both directions in IITB-hi-en are reported in Table 2. For hi→en, the `ilmulti` model provides BLEU scores higher than past submissions, and the additional augmentation through backtranslation gives an extra +0.39 increase in BLEU. A similar increase in en→hi direction with respect to the `ilmulti` model was observed through addition of backtranslated data. Both provide competitive numbers, although not the best in the category[5].

UFAL English-Tamil With no further training on the `ilmulti` model with backtranslation, we evaluate for BLEU scores on the test-set of UFAL English Tamil task. However, the non-adapted model leads to poor BLEU scores. On warm-starting and training with UFAL English-Tamil dataset further for a few epochs, we obtain better scores in both directions. These numbers are reported in Table 2.

However, the warm-started multiway model underperforms compared to model trained from scratch described below. Table 6 indicates the incremental improvements along with the numbers which got us to the best scores on the test set, training from scratch using only UFAL English Tamil training data to begin with. We refer to BLEU scores obtained in UFAL-1 as base-

[5]The same model performs reasonably well for the WAT-ILMPC tasks from WAT-2018.

Source	Or you could leave and return to your families as men instead of murderers.
Hypothesis	அல்லது கொலைகாரர்களுக்குப் பதிலாக உங்களது குடும்பங்களுக்குத் திரும்பிப் போகலாம்.
Target	அல்லது நீங்கள் வெளியேறி, கொலைகாரர்களுக்குப் பதிலாக ஆண்களாக உங்கள் குடும்பங்களுக்குத் திரும்பலாம்.
Source	Srinivasan has his hand in the original of 'Vellithirai' currently under production in Tamil.
Hypothesis	'வெள்ளித்திரை'யின் ஒரிஜினலில் ஸ்ரீனிவாசன் கைவசம் வைத்துள்ளார்.
Target	தற்போது தமிழில் தயாராகிக் கொண்டிருக்கும் 'வெள்ளித்திரை'யின் ஒரிஜினல், 'உதயனாறு தாரம்' படத்திலும் ஸ்ரீனிவாசனின் பங்களிப்பு உண்டு.
Source	அங்குள்ள சன் டவர்ஸ் லாட்ஜில் அவருக்கு அறை ஏற்பாடாகியிருந்தது.
Hypothesis	In the Sun Dawers lodged there, he had a slap.**
Target	A room was arranged for him at Sun Towers Lodge.
Source	முதன்முறையாக காதல் படமொன்றை இயக்குகிறேன்.
Hypothesis	I am directing a romantic film for the first time.**
Target	Vikraman is confident that this love story will appeal to the youth.

Table 5: Failure cases among translated samples. Red colored words in source text do not have corresponding translation in generated hypothesis. Generated hypotheses marked with ** are fluent but don't preserve meaning of source sentence.

Id	Model	BLEU	
		en-ta	ta-en
UFAL-1	Transformer-base	11.59	27.31
UFAL-2	UFAL-1 + filtered	11.73	27.58
UFAL-3	UFAL-2 + ensemble	11.96	28.05
UFAL-4	UFAL-3 + backtranslation	12.63	29.21
UFAL-5	UFAL-4 + ensemble	12.87	29.75
UFAL-6	UFAL-5 + length penalty	13.14	30.10
	UFAL-5 + length penalty	**13.05**†	**30.04**†

Table 6: Automated evaluation scores on the UFAL EnTam v2.0 test set. This table demonstrates incremental improvements which got us to the final submission in Table 2. † indicates numbers from the submission site, others were computed locally and have minor differences.

line BLEU scores for English-Tamil and Tamil-English tasks. Using filtered data to warm-start the UFAL-1 model provided only marginal increments in BLEU for translation in both directions. In UFAL-4, significant improvements in BLEU scores were obtained by doing warm-start of English to Tamil and Tamil to English model on filtered UFAL EnTam train data augmented with additional back-translated data. Further, based on observation that length ratio of generated hypotheses to reference sentence in UFAL-5 was less than 1.0 on validation data for both tasks, we found that enforcing appropriate length penalty for both tasks gave better BLEU scores on validation data. These settings of length penalty parameters were used for obtaining best evaluation BLEU scores in UFAL-6.

4.2 Qualitative Samples

The qualitative samples from Table 4 indicate en→ta comparable to ta→en, despite the imbalance in BLEU scores. We attribute this to be due to the tokenization in place while determining n-grams for BLEU computation. Whitespace and punctuation based tokenization fails to recognize multiple words conjoined to obtain newer words in Tamil, being an agglutinative language.

Table 5 indicates failure cases, many of which shows under-translation phenomena, when all source tokens do not have corresponding translated tokens in generated translation.

5 Conclusion and Future Work

In this paper, we built and demonstrated that a practical translation system is feasible in low-resource settings with improvements in performance of models obtained from pre-processing and filtering, augmentation with additional training corpus using back-translation and simple intuitive tuning of hyper-parameters like length-penalty. Along with this system description paper, we release the trained models and associated code for tokenization and inference[6]. A live web-interface is hosted on the web and available at preon.iiit.ac.in/babel.

There is an increasing interest in unsupervised methods for NMT (Lample et al., 2017; Artetxe et al., 2018) and also to obtain parallel-pairs from sources which provide same content in different

[6] github.com/jerinphilip/ilmulti/

languages (Schwenk et al., 2019; Schwenk, 2018). We intend to tap into increasing monolingual data online across major languages of the country to collectively improve multilingual models in the future.

Acknowledgements

We thank the multilingual milieu at our lab which enable us to worry less about the challenges in interpreting the results which comes with the many languages involved – special thanks to Prajwal Renukanand, Kiran Devraj, Rudrabha Mukhapodhyay.

We thank the organizers for hosting the tasks including Indian languages curating the platform and providing compiled resources.

References

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *EMNLP*.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *EMNLP*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *EMNLP*.

Girish Nath Jha. 2010. The TDIL Program and the Indian Language Corpora Intitiative (ILCI). In *LREC*.

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 ACL*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of ACL*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi Parallel Corpus. In *LREC*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *ACL (System Demonstrations)*.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. ACL.

Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on asian translation. In *WAT*.

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 *NAACL (Demonstrations)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2019. A baseline neural machine translation system for indian languages. *arXiv preprint arXiv:1907.12437*.

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological processing for english-tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *ACL (Short Papers)*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *ACL*.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. *ACL*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.