WT: Wipro AI Submissions to the WAT 2020

Santanu Pal
Wipro AI Lab, India
santanu.pal2@wipro.com

Abstract

In this paper we present an English–Hindi and Hindi–English neural machine translation (NMT) system, submitted to the Translation shared Task organized at WAT 2020. We trained a multilingual NMT system based on transformer architecture. In this paper we show: (i) how effective pre-processing helps to improve performance, (ii) how synthetic data through back-translation from available monolingual data can help in overall translation performance, (iii) how language similarity can aid more onto it. Our submissions ranked 1st in both English to Hindi and Hindi to English translation achieving BLEU 20.80 and 29.59 respectively.

1 Introduction

Now-a-days Neural Machine Translation (NMT) has evolved promising Machine Translation (MT) paradigm to an established state-of-the-art technology. The SOTA MT models are following the popular encoder-decoder framework, which encodes the source sentences and decodes the target sentences with a Transformer network (Vaswani et al., 2017). Transformer networks are trained in a supervised manner, relying on paired source-target datasets.

High quality translation from NMT can be achieved by using large amounts of sentence aligned parallel corpora and an efficient modeling of an NMT architecture. However, the upstream precess i.e., data scarcity can be a challenge for low resource language pairs. Therefore, existing SOTA NMT architecture like Transformer fails to produce quality translation output for low resource scenario. However, NMT systems have constantly ranked in the top positions in WMT (Bojar et al., 2016, 2017) and WAT (Nakazawa et al., 2016). Given the youth of the paradigm and while the main structure of encoder-decoder is still maintained. The research in NMT goes in many directions, including subword unit (Sennrich et al., 2016b) for translation of rare Words, back translation (Sennrich et al., 2016a)or transfer learning (Zoph et al., 2016) for low resource settings and recently unsupervised training for less or no resources (Artetxe et al., 2018).

In this paper we describe the WIPRO-NMT submission to the WAT 2020 translation track (Nakazawa et al., 2020). Our WIPRO-NMT system is inspired from the model described in Johnson et al. (2017) and trained using transformer network. The system achieved best performance ranking 1st in English to Hindi and Hindi to English translation among all participants. The paper poses the following contributions:

- How effective pre-processing help to improve performance.
- How synthetic data through back-translation from available monolingual data could help in overall translation performance.
- How language similarity can aid more onto it.

2 Data

For our experiments, we use the Hindi–English and English–Hindi workshop of Asian translation (WAT) 2020 translation data. We used a subset of the released parallel dataset, was collected from news (Siripragada et al., 2020), PMIndia (Haddow and Kirefu, 2020) and Indic

Proceedings of the 7th Workshop on Asian Translation, pages 122–126

Wordnet (Bhattacharyya, 2010; Kunchukuttan, 2020a) datasets. To augment our dataset, we use English–Hindi parallel data released in WMT 2014 (Bojar et al., 2014), consisting of more than 2M parallel sentences, is available as an additional resource. All dataset used to train our system are detailed in Table 1.

Data Sources	#sentences
IITB	1,561,840
WMT	273,885
News	156,344
PM India	56,831
Total	2,048,900
Remove duplicates	1,464,419
Cleaning*	961,036

Table 1: English–Hindi parallel data statistics. *Removing noisy mixed language sentences.

We use a subset of 5 million segments of Hindi monolingual news crawled from approximately 32 million data. We performed similar cleaning and pre-processing methods as we described in case of parallel data. Five million Hindi monolingual sentences were first back translated to English using a Hindi-English NMT system.

The released WMT 2014 EN-HI data and the WAT 2020 data were noisy for our purposes, so we apply methods for cleaning. We performed the following two steps: (i) we use the cleaning process described in Pal et al. (2015), and (ii) we execute the Moses (Koehn et al., 2007) corpus cleaning scripts with minimum and maximum number of tokens set to 1 and 100, respectively. After cleaning and removing duplicates, we have 1M EN-HI parallel sentences. Next, we perform punctuation normalization, and then we use the Moses tokenizer to tokenize the English side of the parallel corpus with 'no-escape' option. Finally, we apply true-casing. We submitted a multilingual system, which additionally use Hindi–Marathi parallel data from WMT 2020 Similar Language Task[1]. For tokenizaton, we use Indic NLP Library (Kunchukuttan, 2020b).

3 Model Architecture

Our model is based on transformer architecture. The *transformer* architecture (Vaswani et al., 2017) is built solely upon such attention mechanisms completely replacing recurrence and convolutions. The transformer uses positional encoding to encode the input and output sequences, and computes both self- and cross-attention through so-called multi-head attentions, which are facilitated by parallelization. We use multi-head attention to jointly attend to information at different positions from different representation subspaces.

We present a single multilingual NMT system based on transformer architecture that can able to translate between multiple languages. To make use of multilingual data within a single NMT model, we perform one simple modification to the source side on the multilingual data, we use an additional token at the beginning of the each source sentence to indicate the target language by the NMT model would be translated. Examples are shown in Table 2.

We train the model with all the processed multilingual data consisting of sentence aligned multiple language pairs at once, During inference, we also need to add the aforementioned additional token to each input source sentence of the source data to specify the desired target language.

4 Experiments

In the next sub-sections we describe the experiments we carried out for translating from Hindi to English and from English to Hindi for WIPRO's WAT 2020 shared task submission.

4.1 Experiment Setup

To handle out-of-vocabulary words and to reduce the vocabulary size, instead of considering words, we consider subword units (Sennrich et al., 2016b) by using byte-pair encoding (BPE). In the preprocessing step, instead of learning an explicit mapping between BPEs in the English (EN) and Hindi (HI), we define BPE tokens by jointly processing all parallel data. Thus, all derive a single BPE vocabulary. We train our system using transformer architecture for NMT available in Marian NMT implementation[2].

We report evaluation results (evaluated by the shared task organizers) of our approach

[1] http://www.statmt.org/wmt20/similar.html

[2] https://marian-nmt.github.io/

L1 → L2		Parallel Sentences	
		Source	Target
HI→MR	Raw data	देश एकल प्रयासों से आगे बढ़ चुके हैं।	देश आता सामाईक प्रयत्न करत आहेत.
	Processed data	TO_MR देश एकल प्रयासों से आगे बढ़ चुके हैं।	देश आता सामाईक प्रयत्न करत आहेत.
HI→EN	Raw data	इस एमओयू पर फरवरी, 2016 में हस्ताक्षर किए हेत.	The MoU was signed in February, 2016.
	Processed data	TO_EN इस एमओयू पर फरवरी, 2016 में हस्ताक्षर किए हेत.	The MoU was signed in February, 2016.
EN→HI	Raw data	The MoU was signed in February, 2016.	इस एमओयू पर फरवरी, 2016 में हस्ताक्षर किए गए थे।
	Processed data	TO_HI The MoU was signed in February, 2016.	इस एमओयू पर फरवरी, 2016 में हस्ताक्षर किए गए थे।

Table 2: Multilingual **Processed data**, indicating TO_XX as target language:

with the released Test data. BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010) are used to evaluate the performance of our systems in the shared task.

4.2 Hyper-parameter Setup

We follow a similar hyper-parameter setup for all reported systems. All encoders, and the decoder, are composed of a stack of $N_X = 6$ identical layers followed by layer normalization. Each layer again consists of two sub-layers and a residual connection (He et al., 2016) around each of the two sub-layers. We apply dropout (Srivastava et al., 2014) to the output of each sub-layer, before it is added to the sub-layer input and normalized. Furthermore, dropout is applied to the sums of the word embeddings and the corresponding positional encodings in both encoders as well as the decoder stacks.

We set all dropout values in the network to 0.1. During training, we employ label smoothing with value $\epsilon_{ls} = 0.1$. The output dimension produced by all sub-layers and embedding layers is $d_{model} = 512$. Each encoder and decoder layer contains a fully connected feed-forward network (FFN) having dimensionality of $d_{model} = 512$ for the input and output and dimensionality of $d_{ff} = 2048$ for the inner layers. For the scaled dot-product attention, the input consists of queries and keys of dimension d_k, and values of dimension d_v. As multi-head attention parameters, we employ $h = 8$ for parallel attention layers, or heads. For each of these we use a dimensionality of $d_k = d_v = d_{model}/h = 64$. For optimization, we use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$.

The learning rate is varied throughout the training process, and increasing for the first training steps $warmup_{steps} = 16000$ and afterwards decreasing as described in (Vaswani et al., 2017). All remaining hyper-parameters are set analogously to those of the transformer's *base* model. At training time, the batch size is set to 25K tokens, with a maximum sentence length of 256 subwords, and a vocabulary size of 32K. After each epoch, the training data is shuffled. During decoding, we perform beam search with a beam size of 4. We use 32K BPE operations to train our BPE models. We use shared embeddings in all our experiments.

5 Results

We present the released results obtained by our systems for Hindi–English and English to Hindi in Table 5 in terms of BLEU and RIBES. We apply our proposed method to train multilingual models.

Table 3 shows different experiment setting of out WIPRO NMT system. The 'base' model achieve 12.23 BLEU for English–Hindi and 12.83 BLEU for Hindi–English. The 'preprocessed' system includes our preprocessing methods described in Section 2. BT (no MR) system is similar 'preprocessed' system but additionally used back translation data derived from monolingual Hindi and English data.

Table 4 shows how BLEU is increasing/decreasing based on the sentence length. The results are quite surprising based on the target languages. For Hindi–English length > 20 achieve better performance, while for the case of English–Hindi the sentence length between 10 and 20 achieves significant best per-

Systems	L1 → L2	BLEU ↑
base	HI–EN	12.83
preprocessed	HI–EN	17.30
BT (no MR)	HI–EN	19.51
base	EN–HI	12.23
preprocessed	EN–HI	17.03
BT (no MR)	EN–HI	19.63

Table 3: Results for Hindi to English and English to Hindi translation. BT (no MR): No back-translation data for monolingual Marathi were used.

formance over others.

Table 5 presents our submission results released in WAT 2020 evaluation suite. The system 'X1' is using preprocessed data (see 'preprocessed' in Table 3), however, additionally, 5M back translated Hindi–English and English–Hindi, 5M back-translated Marathi–Hindi and 5M back-translated Hindi–Marathi corpus. Source back-translated sentences begin with an additional token indicating the target language. Note that we use the multilingual system presented in Table 3 for back translation.

6 Conclusion and Future Work

This paper presented the WIPRO–NMT system submitted to the Translation shared task at WAT 2020. We presented the results obtained by our system in translating from Hindi to English and English to Hindi. Our system ranked first among all participated teams in terms of BLEU score. This paper also shows How effective pre-processing, back-translation and language similarity help in improving performance.

In future work, we would like to further explore the similarity between languages in translating to other Indo-Aryan languages (e.g. Bengali, Magadhi, and Nepali) and expect that the method presented here to perform well for other languages provided that sufficient training data is available. Furthermore, we would like to apply and evaluate our method on other language families. Finally, we will be incorporating the translation models to CATaLog, an open-source online CAT tool (Nayek et al., 2015; Pal et al., 2016).

References

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

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of LREC*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of WMT*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of WMT*.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of WMT*.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *Proceedings of CVPR*.

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, Macduff Hughes, and Jeffrey Dean. 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 Lei Ba. 2015. Adam: A Method for Stochastic Optimization. *Proceedings of ICLR*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar,

Testset	#Sentence	EN-HI	HI-EN
length (<=10)	359	14.42	15.34
length (>10, <=20)	1058	**17.80**	17.21
length (>20)	996	16.77	**17.32**
Overall		17.03	17.30

Table 4: BLEU scores based on length

System	Desc.	L1 → L2	BLEU ↑	RIBES ↑
WIPRO NMT	X1	HI–EN	29.58	79.20
WIPRO NMT	X1	EN–HI	22.08	76.53
WIPRO NMT	X3	EN–HI	22.80	76.91

Table 5: Results for Hindi to English and English to Hindi translation. X1 = Single system; X3 = ensemble of 3 systems initialized on three different random seeds. Note that we did not tested ensemble model for HI–EN as ensembling does not providing much impact on the performance for EN–HI.

Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL*.

Anoop Kunchukuttan. 2020a. Indowordnet parallel corpus.

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

Toshiaki Nakazawa, Hideya Mino, Chenchen Ding, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2016. Overview of the 3rd Workshop on Asian Translation. *Proceedings of WAT*.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, Suzhou, China. Association for Computational Linguistics.

Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2015. CATaLog: New approaches to TM and post editing interfaces. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 36–42, Hissar, Bulgaria. Association for Computational Linguistics.

Santanu Pal, Sudip Naskar, and Josef van Genabith. 2015. UdS-sant: English–German hybrid machine translation system. In *Proceedings of WMT*.

Santanu Pal, Sudip Kumar Naskar, Marcos Zampieri, Tapas Nayak, and Josef van Genabith. 2016. CATaLog online: A web-based CAT tool for distributed translation with data capture for APE and translation process research. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 98–102, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL*.

Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of LREC*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

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 *Proceedings of NIPS*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of EMNLP*.