Neural Machine Translation for Similar Languages: The Case of Indo-Aryan Languages

Santanu Pal[1], Marcos Zampieri[2]
[1]Wipro AI Lab, India
[2]Rochester Institute of Technology, USA
santanu.pal2@wipro.com

Abstract

In this paper we present the WIPRO-RIT systems submitted to the Similar Language Translation shared task at WMT 2020. The second edition of this shared task featured parallel data from pairs/groups of similar languages from three different language families: Indo-Aryan languages (Hindi and Marathi), Romance languages (Catalan, Portuguese, and Spanish), and South Slavic Languages (Croatian, Serbian, and Slovene). We report the results obtained by our systems in translating from Hindi to Marathi and from Marathi to Hindi. WIPRO-RIT achieved competitive performance ranking 1st in Marathi to Hindi and 2nd in Hindi to Marathi translation among 22 systems.

1 Introduction

WMT 2020 is the fifth edition of WMT as a conference following a series of well-attended workshops that date back to 2006. WMT became a well-established conference due to its blend of research papers and popular shared tasks on different topics such as translation in various domains (e.g. biomedical, news), translation quality estimation, and automatic post-editing. The competitions co-organized with WMT provide important datasets and benchmarks widely used in the MT community. The vast majority of these tasks so far, however, involved training systems to translate to and from English (Bojar et al., 2016, 2017) while only a few of them addressed the problem of translating between pairs of languages with less resources.

To address this issue, in 2019, the Similar Language Translation (SLT) shared task was introduced at WMT. SLT's purpose was to evaluate the performance of state-of-the-art MT systems on translating between pairs of similar languages without English as a pivot language (Barrault et al., 2019). The organizers provided participants with training, development, and testing parallel data from three pairs of languages from three different language families: Spanish - Portuguese (Romance languages), Czech - Polish (Slavic languages), and Hindi - Nepali (Indo-Aryan languages). Systems were evaluated using automatic metrics, namely BLEU (Papineni et al., 2002) and TER (Snover et al., 2006).

In SLT 2020, the task organizes once again included an Indo-Aryan language track with Hindi and Marathi. Indo-Aryan languages are a sub-family of the Indo-European language family which includes Bengali, Bohjpuri, Hindi, Marathi, and Nepali. These languages are mainly spoken in North and Central India, and some neighbouring countries such as Nepal, Bangladesh, and Pakistan etc. The script used in most of these languages are derived from the ancient Brahmi script and enriched with high grapheme to phoneme correspondence leading to many orthographic similarities across these languages.

In addition to Hindi and Marathi, SLT 2020 features two other tracks with similar languages from the following language families: Romance languages (Catalan, Portuguese, and Spanish) and South Slavic Languages (Croatian, Serbian, and Slovene). In this paper we describe the WIPRO-RIT submission to the SLT 2020 Indo-Aryan track. Our WIPRO-RIT system is based on the model described in Johnson et al. (2017). WIPRO-RIT achieved competitive performance ranking 1st in Marathi to Hindi and 2nd in Hindi to Marathi translation among 22 systems.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 424–429
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

2 Related Work

With the substantial performance improvements brought to MT by neural approaches, a growing interest in translating between pairs of similar languages, language varieties, and dialects has been observed. Recent studies have addressed MT between Arabic dialects (Harrat et al., 2019; Shapiro and Duh, 2019) Catalan and Spanish, Croatian and Serbian (Popović et al., 2020), (Costa-jussà, 2017), Brazilian and European Portuguese (Costa-jussà et al., 2018), and several pairs of languages and language varieties such as Brazilian and European Portuguese, Canadian and European French, and similar languages such as Croatian and Serbian, and Indonesian and Malay (Lakew et al., 2018).

The interest on diatopic language variation is evidenced by the recent iterations of the VarDial workshop in which papers on MT applied to similar languages varieties, and dialects (Shapiro and Duh, 2019; Myint Oo et al., 2019; Popović et al., 2020) have been presented along with evaluation campaigns featuring multiple shared tasks on a number of related topics such as cross-lingual morphological analysis, cross-lingual parsing, dialect identification, and morphosyntactic tagging (Zampieri et al., 2018, 2019; Găman et al., 2020).

3 Data

For our experiments, we use the Hindi–Marathi and Marathi–Hindi WMT 2020 SLT data. The released parallel dataset was collected from news (Siripragada et al., 2020), PMIndia (Haddow and Kirefu, 2020) and Indic 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 2 million parallel sentences, which is available as an additional resource. We use a subset of 5 million segments of Hindi monolingual news crawled from ca. 32 million data. We also use a subset 5 million Marathi monolingual data. We performed similar cleaning and pre-processing methods as we described in case of parallel data.

The five million Hindi monolingual sentences were first back-translated to English using a Hindi–English NMT system. The Hindi–English NMT system was trained on English–Hindi parallel data released in WMT 2014 (Bojar et al., 2014), IITB parallel corpus (Kunchukuttan et al., 2018), the parallel dataset was collected from news (Siripragada et al., 2020) and the PMIndia (Haddow and Kirefu, 2020) parallel corpus (see Table 1).

Data Sources	#sentences
WMT	273,885
News	156,344
IITB	1,561,840
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 also back-translated 5 million Marathi monolingual sehments using our WIPRO-RIT CONTRASTIVE 1 system described in more detail Section 6. For Marathi–Hindi we did not use any back translation data in our CONTRASTIVE 2 and PRIMARY submissions. In the both cases 5 million English–Hindi back-translation data provide significant ($p < 0.01$) improvements over CONTRASTIVE 1 (detailed in Section 6).

The released WMT 2014 EN-HI data and the WMT SLT 2020 data were noisy for our purposes, so we apply methods for cleaning (see data statistics in Table 2).

Parallel	#sentences
News	12,349
PM India	25,897
Indic WordNet	11,188
Total	49,434
Filtered*	33923

Table 2: Data statistics of released SLT Data; *Filtration methods: (i) remove duplicates and (ii) filtering noisy mixed language sentences.

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 re-

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.
MR→HI	Raw data	देश आता सामाईक प्रयत्न करत आहेत.	देश एकल प्रयासों से आगे बढ़ चुके हैं।
	Processed data	TO_HI देश आता सामाईक प्रयत्न करत आहेत.	देश एकल प्रयासों से आगे बढ़ चुके हैं।
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 3: Multilingual **Processed data**, indicating TO_XX as target language:

moving 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. For the case of Hindi and Marathi, we use Indic NLP Library[1] (Kunchukuttan, 2020b) for tokenizaton.

4 Model Architecture

Our model is based on a transformer architecture (Vaswani et al., 2017) 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 the transformer architecture that can translate between multiple languages. To make use of multilingual data within a single NMT model, we perform one simple modification to the source side of 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 as shown in Table 3.

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.

[1] https://github.com/anoopkunchukuttan/indic_nlp_library/

5 Experiments

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

5.1 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., 2016) by using byte-pair encoding (BPE). In the preprocessing step, instead of learning an explicit mapping between BPEs in the English (EN), Hindi (HI) and Marathi (MR), we define BPE tokens by jointly processing all parallel data. Thus, all derive a single BPE vocabulary. Since HI and MR belong to the similar languages, they naturally share a good fraction of BPE tokens, which reduces the vocabulary size.

We report evaluation results (evaluated by the shared task organizers) of our approach with the released Test data. BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and TER (Snover et al., 2006) are used to evaluate the performance of all participating systems in the shared task.

Parallel Data	#sentences	C1	C2	P
Filtered SLT	33,923	✓	✓	✓
Filtered EN–HI	961,036	✓	✓	✓
BT EN–HI	5 million	✓	✓	✓
BT HI–MR	5 million		✓	✓

Table 4: The training criteria data statistics of our submitted systems (C1 = Contrastive 1, C2 = Contrastive 2, P = Primary, and BT = Back-translated data).

5.2 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.

6 Results

We present the results obtained by our systems for Hindi–Marathi in Table 5 and for Marathi–Hindi in Table 6 in terms of BLEU, RIBES, and TER. We apply our proposed method to train multilingual models in three different configurations. Table 4 shows different training data used to train our CONTRASTIVE 1 (C1), CONTRASTIVE 2 (C2) and Primary (P) submissions.

System	**BLEU ↑**	**RIBES ↑**	**TER ↓**
P	16.62	62.45	72.23
C2	15.42	61.02	73.59
C1	13.25	58.51	76.17

Table 5: Results for Hindi to Marathi translation ranked by BLEU score.

System	**BLEU ↑**	**RIBES ↑**	**TER ↓**
P	24.53	66.23	66.39
C2	22.93	65.89	68.11
C1	22.69	65.01	68.13

Table 6: Results for Marathi to Hindi Translation ranked by BLEU score.

CONTRASTIVE 1 (C1) Our CONTRASTIVE 1 submission is a multilingual single system and does not use any monolingual back translation data. The system is trained on the released HI-MR and MR-HI parallel data. In addition to we also use EN-HI parallel data.

CONTRASTIVE 2 (C2) This submission is similar to CONTRASTIVE 1, however in this case we used 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.

PRIMARY (P) Our primary submission is trained using the same setting as we described in CONTRASTIVE 2 system. The difference is our primary system is an ensemble of three different CONTRASTIVE 2 systems initiated with three different random seeds.

7 Conclusion and Future Work

This paper presented the WIPRO–RIT system submitted to the Similar Language Translation shared task at WMT 2020. We presented the results obtained by our system in translating from Hindi to Marathi and Marathi to Hindi. Our primary system achieved competitive performance ranking first in Marathi to Hindi and second in Hindi to Marathi among 22 teams in terms of BLEU score.

In future work, we would like to further explore the similarity between these two languages in translating to other Indo-Aryan languages (e.g. Bengali, Bhojpuri, and Nepali). We expect the models presented in this paper to perform well for other Indo-Aryan language provided that suitable training data is available. Furthermore, we would like to apply and evaluate our method on the two other groups of languages in the WMT SLT 2020 shared task, Romance languages: Catalan, Portuguese, and Spanish, and South Slavic languages: Croatian, Serbian, and Slovene. Finally, we will be incorporating the translation models presented in this paper to CATaLog, an open-source online CAT tool that provides users with both MT and TM outputs (Nayek et al., 2015; Pal et al., 2016).

Acknowledgments

We would like to thank the WMT 2020 SLT shared task organizers for making the Hindi - Marathi data available. We further thank the anonymous WMT reviewers for their insightful feedback and suggestions.

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of WMT*.

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*.

Marta R. Costa-jussà. 2017. Why Catalan-Spanish Neural Machine Translation? Analysis, Comparison and Combination with Standard Rule and Phrase-based Technologies. In *Proceedings of VarDial*.

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A Neural Approach to Language Variety Translation. In *Proceedings of VarDial*.

Mihaela Găman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A Report on the VarDial Evaluation Campaign 2020. In *Proceedings of VarDial*.

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

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine Translation for Arabic Dialects (Survey). *Information Processing & Management*, 56(2):262–273.

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 EMNLP*.

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

Anoop Kunchukuttan. 2020a. Indowordnet parallel corpus. https://github.com/anoopkunchukuttan/indowordnet_parallel.

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

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of LREC*.

Surafel M Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural Machine Translation into Language Varieties. *arXiv preprint arXiv:1811.01064*.

Thazin Myint Oo, Ye Kyaw Thu, and Khin Mar Soe. 2019. Neural machine translation between Myanmar (Burmese) and rakhine (arakanese). In *Proceedings of VarDial*.

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 NLP4TM*.

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

Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, Tapas Nayak, Mihaela Vela, and Josef van Genabith. 2016. CATaLog online: Porting a post-editing tool to the web. In *Proceedings of LREC*.

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

Maja Popović, Alberto Poncelas, Marija Brkic, and Andy Way. 2020. Neural Machine Translation for Translating into Croatian and Serbian. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

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

Pamela Shapiro and Kevin Duh. 2019. Comparing pipelined and integrated approaches to dialectal Arabic neural machine translation. In *Proceedings of VarDial*.

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*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*.

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*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign. In *Proceedings of VarDial*.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei Butnaru, and Tommi Jauhiainen. 2019. A Report on the Third VarDial Evaluation Campaign. In *Proceedings of VarDial*.