Addressing Word-order Divergence in Multilingual Neural Machine Translation for Extremely Low Resource Languages

Rudra Murthy V[†], **Anoop Kunchukuttan**[‡], **Pushpak Bhattacharyya**[†]
[†] Center for Indian Language Technology (CFILT)
Department of Computer Science and Engineering
IIT Bombay, India.
[‡]Microsoft AI & Research, Hyderabad, India.
{rudra,pb}@cse.iitb.ac.in, ankunchu@microsoft.com

Abstract

Transfer learning approaches for Neural Machine Translation (NMT) trains a NMT model on an assisting language-target language pair (parent model) which is later fine-tuned for the source language-target language pair of interest (child model), with the target language being the same. In many cases, the assisting language has a different word order from the source language. We show that divergent word order adversely limits the benefits from transfer learning when little to no parallel corpus between the source and target language is available. To bridge this divergence, we propose to pre-order the assisting language sentences to match the word order of the source language and train the parent model. Our experiments on many language pairs show that bridging the word order gap leads to major improvements in the translation quality in extremely low-resource scenarios.

1 Introduction

Transfer learning for multilingual Neural Machine Translation (NMT) (Zoph et al., 2016; Dabre et al., 2017; Nguyen and Chiang, 2017) attempts to improve the NMT performance on the *source* to *target* language pair (child task) using an *assisting source* language (assisting to target language translation is the parent task). Here, the parent model is trained on the assisting and target language parallel corpus and the trained weights are used to initialize the child model. If source-target language pair parallel corpus is available, the child model can further be fine-tuned. The weight initialization reduces the requirement on the training data for the source-target language pair by transferring knowledge from the parent task, thereby improving the performance on the child task.

However, the divergence between the source and the assisting language can adversely impact the benefits obtained from transfer learning. Multiple studies have shown that transfer learning works best when the languages are related (Zoph et al., 2016; Nguyen and Chiang, 2017; Dabre et al., 2017). Zoph et al. (2016) studied the influence of language divergence between languages chosen for training the parent and the child model, and showed that choosing similar languages for training the parent and the child model leads to better improvements from transfer learning.

Several studies have tried to address the *lexical divergence* between the source and the target languages either by using Byte Pair Encoding (BPE) as basic input representation units (Nguyen and Chiang, 2017) or character-level NMT system (Lee et al., 2017) or bilingual embeddings (Gu et al., 2018). However, the effect of *word order divergence* and its mitigation has not been explored. In a practical setting, it is not uncommon to have source and assisting languages with different word order. For instance, it is possible to find parallel corpora between English (SVO word order) and some Indian (SOV word order) languages, but very little parallel corpora between Indian languages. Hence, it is natural to use English as an assisting language for inter-Indian language translation.

To address the word order divergence, we propose to pre-order the assisting language sentences (SVO) to match the word order of the source language (SOV). We consider an extremely resource-constrained scenario, where there is no parallel corpus for the child task. From our experiments, we show that there is a significant increase in the translation accuracy for the unseen source-target language pair.

2 Related Work

To the best of our knowledge, no work has addressed word order divergence in transfer learning

Proceedings of NAACL-HLT 2019, pages 3868–3873
Minneapolis, Minnesota, June 2 - June 7, 2019.

for multilingual NMT. However, some work exists for other NLP tasks in a multilingual setting. For Named Entity Recognition (NER), Xie et al. (2018) use a self-attention layer after the Bi-LSTM layer to address word-order divergence for Named Entity Recognition (NER) task. The approach does not show any significant improvements, possibly because the divergence has to be addressed before/during construction of the contextual embeddings in the Bi-LSTM layer. Joty et al. (2017) use adversarial training for cross-lingual question-question similarity ranking. The adversarial training tries to force the sentence representation generated by the encoder of similar sentences from different input languages to have similar representations.

Pre-ordering the source language sentences to match the target language word order has been found useful in addressing word-order divergence for Phrase-Based SMT (Collins et al., 2005; Ramanathan et al., 2008; Navratil et al., 2012; Chatterjee et al., 2014). For NMT, Ponti et al. (2018) and Kawara et al. (2018) have explored pre-ordering. Ponti et al. (2018) demonstrated that by reducing the syntactic divergence between the source and the target languages, consistent improvements in NMT performance can be obtained. On the contrary, Kawara et al. (2018) reported drop in NMT performance due to pre-ordering. Note that these works address source-target divergence, not divergence between source languages in multilingual NMT scenario.

3 Proposed Solution

Consider the task of translating for an extremely low-resource language pair. The parallel corpus between the two languages, if available may be too small to train an NMT model. Similar to Zoph et al. (2016), we use transfer learning to overcome data sparsity between the source and the target languages. We choose *English* as the assisting language in all our experiments. In our resource-scarce scenario, we have no parallel corpus for training the child model. Hence, at test time, the source language sentence is translated using the parent model after performing a word-by-word translation from source to the assisting language using a bilingual dictionary.

Since the source language and the assisting language (English) have different word order, we hypothesize that it leads to inconsistencies in the contextual representations generated by the encoder for the two languages. Specifically, given an English sentence (SVO word order) and its translation in the source language (SOV word order), the encoder representations for words in the two sentences will be different due to different contexts of synonymous words. This could lead to the attention and the decoder layers generating different translations from the same (parallel) sentence in the source or assisting language. This is undesirable as we want the knowledge to be transferred from the parent model (assisting source→ target) to the child model (source→target).

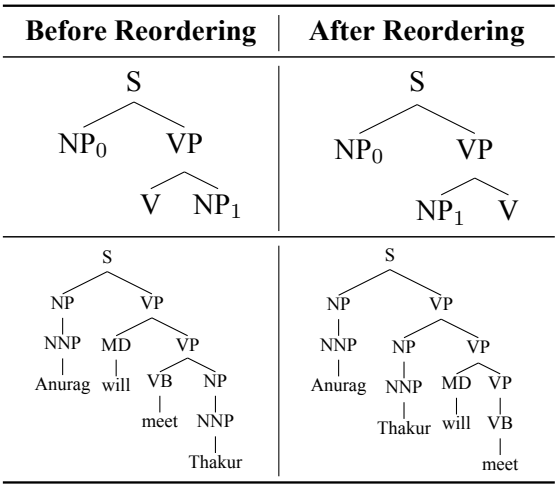

Table 1: Example showing transitive verb before and after reordering (Adapted from Chatterjee et al. (2014))

In this paper, we propose to pre-order English sentences (assisting language sentences) to match the source language word-order and train the parent model on the pre-ordered corpus. Table 1 shows one of the pre-ordering rules (Ramanathan et al., 2008) used along with an example sentence illustrating the effect of pre-ordering. This will ensure that context of words in the parallel source and assisting language sentences are similar, leading to consistent contextual representations across the source languages. Pre-ordering may also be beneficial for other word order divergence scenarios (*e.g.,* SOV to SVO), but we leave verification of these additional scenarios for future work.

4 Experimental Setup

In this section, we describe the languages experimented with, datasets used, the network hyperparameters used in our experiments.

Languages: We experimented with English → Hindi translation as the parent task. English is

the assisting source language. Bengali, Gujarati, Marathi, Malayalam and Tamil are the source languages, and translation from these to Hindi constitute the child tasks. Hindi, Bengali, Gujarati and Marathi are Indo-Aryan languages, while Malayalam and Tamil are Dravidian languages. All these languages have a canonical SOV word order.

Datasets: For training English-Hindi NMT systems, we use the IITB English-Hindi parallel corpus (Kunchukuttan et al., 2018) ($1.46M$ sentences from the training set) and the ILCI English-Hindi parallel corpus ($44.7K$ sentences). The ILCI (Indian Language Corpora Initiative) multilingual parallel corpus (Jha, 2010)[1] spans multiple Indian languages from the health and tourism domains. We use the 520-sentence dev-set of the IITB parallel corpus for validation. For each child task, we use $2K$ sentences from ILCI corpus as test set.

Network: We use OpenNMT-Torch (Klein et al., 2018) to train the NMT system. We use the standard encoder-attention-decoder architecture (Bahdanau et al., 2015) with input-feeding approach (Luong et al., 2015). The encoder has two layers of bidirectional LSTMs with 500 neurons each and the decoder contains two LSTM layers with 500 neurons each. We use a mini-batch of size 50 and a dropout layer. We begin with an initial learning rate of 1.0 and continue training with exponential decay till the learning rate falls below 0.001. The English input is initialized with pre-trained *fastText* embeddings (Grave et al., 2018) [2].

English and Hindi vocabularies consists of $0.27M$ and $50K$ tokens appearing at least 2 and 5 times in the English and Hindi training corpus respectively. For representing English and other source languages into a common space, we translate each word in the source language into English using a bilingual dictionary (we used *Google Translate* to get single word translations). In an end-to-end solution, it would be ideal to use bilingual embeddings or obtain word-by-word translations *via* bilingual embeddings (Xie et al., 2018). However, publicly available bilingual embeddings for English-Indian languages are not good enough for obtaining good-quality, bilingual representations (Smith et al., 2017; Jawanpuria et al., 2019) and publicly available bilingual dictionaries have limited coverage. The focus of our study is the influence of word-order divergence on Multilingual NMT. We do not want bilingual embeddings quality or bilingual dictionary coverage to influence the experiments, rendering our conclusions unreliable. Hence, we use the above mentioned large-coverage bilingual dictionary.

Language	BLEU			LeBLEU		
	No Pre-Order	Pre-Ordered		No Pre-Order	Pre-Ordered	
		HT	G		HT	G
Bengali	6.72	8.83	**9.19**	37.10	41.50	**42.01**
Gujarati	9.81	**14.34**	13.90	43.21	47.36	**47.60**
Marathi	8.77	10.18	**10.30**	40.21	41.49	**42.22**
Malayalam	5.73	6.49	**6.95**	33.27	33.69	**35.09**
Tamil	4.86	**6.04**	6.00	29.38	30.77	**31.33**

Table 2: Transfer learning results for *X-Hindi* pair, trained on *English-Hindi* corpus and sentences from *X* word translated to English.

Language	No Pre-Order	Pre-Ordered	
		HT	G
Bengali	1324	1139	1146
Gujarati	1337	1190	1194
Marathi	1414	1185	1178
Malayalam	1251	1067	1059
Tamil	1488	1280	1252

Table 3: Number of UNK tokens generated by each model on the test set.

Pre-ordering: We use *CFILT-preorder*[3] for pre-reordering English sentences. It contains two pre-ordering configurations: (1) *generic* rules (G) that apply to all Indian languages (Ramanathan et al., 2008), and (2) *hindi-tuned* rules (HT) which improves generic rules by incorporating improvements found through error analysis of English-Hindi reordering (Patel et al., 2013). The Hindi-tuned rules improve translation for other English to Indian language pairs too (Kunchukuttan et al., 2014).

5 Results

We experiment with two scenarios: (a) an extremely resource scarce scenario with no parallel corpus for child tasks, (b) varying amounts of parallel corpora available for child task.

5.1 No Parallel Corpus for Child Task

The results from our experiments are presented in the Table 2. We report BLEU scores and LeBLEU[4]

[1]The corpus is available on request from http://tdil-dc.in/index.php?lang=en

[2]https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md

[3]https://github.com/anoopkunchukuttan/cfilt_preorder

[4]LeBLEU (Levenshtein Edit BLEU) is a variant of BLEU that does a soft-match of reference and output words based

English	the treatment of migraine is done in two ways								
Gujarati (Original)	માઇગ્રેનની	સારવાર	બે	રીતે	કરી	શકાય	છે.		
Gujarati (Word Translate)	migraine	treatment	two	the way	doing be	done	is there .		
Hindi (Reference)	माइग्रेन	का	ट्रीटमेंट	दो	तरह	से	किया	जाता	है ।
(Word Translate)	migraine	of	treatment	two	kind	from	did	go	is .
No Pre-Order	<unk>	उपचार	दो	प्रकार	से	किया	जाता	है	।
		upachAra	do	prakAra	se	kiyA	jAtA	hai	.
	<unk>	treatment	two	kind	from	did	go	is	.
Pre-ordered (HT)	माइग्रेन	का	उपचार	दो	तरह	से	किया	जाता	है।
	mAigrena	kA	upachAra	do	prakAra	se	kiyA	jAtA	hai.
	migraine	of	treatment	two	kind	from	did	go	is .

Table 4: Sample Hindi translation generated by the Gujarati-Hindi NMT model. Text in red indicates phrase dropped by the no pre-ordered model.

scores. We observe that both the pre-ordering models significantly improve the translation quality over the no-preordering models for all the language pairs. The results support our hypothesis that word-order divergence can limit the benefits of multilingual translation. Thus, reducing the word order divergence improves translation in extremely low-resource scenarios.

An analysis of the outputs revealed that pre-ordering significantly reduced the number of UNK tokens (placeholder for unknown words) in the test output (Table 3). We hypothesize that due to word order divergence between English and Indian languages, the encoder representation generated is not consistent leading to decoder generating unknown words. However, the pre-ordered models generate better encoder representations leading to lesser number of UNK tokens and better translation, which is also reflected in the BLEU scores and Table 4.

5.2 Parallel Corpus for Child Task

We study the impact of child task parallel corpus on pre-ordering. To this end, we fine-tune the parent task model with the child task parallel corpus. Table 5 shows the results for *Bengali-Hindi*, *Gujarati-Hindi*, *Marathi-Hindi*, *Malayalam-Hindi*, and *Tamil-Hindi* translation. We observe that pre-ordering is beneficial when almost no child task corpus is available. As the child task corpus increases, the model learns the word order of the source language; hence, the non pre-ordering models perform almost as good as or sometimes better than the pre-ordered ones. The non pre-ordering model is able to forget the word-order of English and learn the word order of Indian languages. We attribute this behavior of the non pre-ordered model to the phenomenon of catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999) which enables the model to learn the word-order of the source language when sufficient child task parallel corpus is available.

We also compare the performance of the fine-tuned model with the model trained only on the available source-target parallel corpus with randomly initialized weights (No Transfer Learning). Transfer learning, with and without pre-ordering, is better compared to training only on the small source-target parallel corpus.

6 Conclusion

In this paper, we show that handling word-order divergence between the source and assisting languages is crucial for the success of multilingual NMT in an extremely low-resource setting. We show that pre-ordering the assisting language to match the word order of the source language significantly improves translation quality in an extremely low-resource setting. If pre-ordering is not possible, fine-tuning on a small source-target parallel corpus is sufficient to overcome word order divergence. While the current work focused on Indian languages, we would like to validate the hypothesis on a more diverse set of languages. We

on edit distance, hence it can handle morphological variations and cognates (Virpioja and Grönroos, 2015).

Corpus Size	No Transfer Learning	No Pre-Order	Pre-Ordered	
			HT	G
Bengali				
-	-	6.72	8.83	**9.19**
500	0.0	11.40	**11.49**	11.00
1000	0.0	13.71	**13.84**	13.62
2000	0.0	16.41	**16.79**	16.01
3000	0.0	17.44	**18.42**†	17.82
4000	0.0	18.86	**19.17**	18.66
5000	0.07	19.58	**20.15**†	19.82
10000	1.87	22.50	**22.92**	22.53
Gujarati				
-	-	9.81	**14.34**	13.90
500	0.0	17.27	17.11	**17.75**
1000	0.0	21.68	**22.12**	21.45
2000	0.0	25.34	**25.73**	25.63
3000	0.29	27.48	27.77	**27.83**
4000	0.82	29.20	29.49	**29.51**
5000	0.0	29.87	**31.09**†	30.58†
10000	1.52	33.97	**34.25**	34.08
Marathi				
-	-	8.77	10.18	**10.30**
500	0.0	12.84	**13.61**†	12.97
1000	0.0	15.62	15.75	**16.10**†
2000	0.0	18.59	**19.10**	18.67
3000	0.0	20.51	**20.76**	20.29
4000	0.24	**21.78**	21.77	21.39
5000	0.29	22.21	22.41	**22.73**†
10000	7.90	25.16	**25.88**	25.36
Malayalam				
-	-	5.73	6.49	**6.95**
500	0.0	5.40	5.54	**6.17**†
1000	0.0	7.34	7.36	**7.63**
2000	0.0	8.24	**8.66**†	8.31
3000	0.0	9.11	9.30	**9.31**
4000	0.0	9.65	**9.91**	9.87
5000	0.03	10.26	**10.47**	10.28
10000	0.0	**11.96**	11.85	11.63
Tamil				
-	-	4.86	**6.04**	6.00
500	0.0	5.49	**5.85**†	5.59
1000	0.0	7.04	7.23	**7.44**†
2000	0.0	8.83	8.84	**9.24**
3000	0.0	9.80	**10.04**	9.56
4000	0.0	9.69	**10.59**†	10.25†
5000	0.03	10.84	**10.93**	10.69
10000	0.0	12.71	**13.05**	12.69

Table 5: Transfer learning results (BLEU) for *Indian Language-Hindi* pair, fine-tuned with varying number of *Indian Language-Hindi* parallel sentences. †Indicates statistically significant difference between *Pre-ordered* and *No Pre-ordered* results using paired bootstrap resampling (Koehn, 2004) for a p-value less than 0.05. *No Transfer Learning* model refers to training the model on varying number of *Indian Language-Hindi* parallel sentences with randomly initialized weights.

would also like to explore alternative methods to address word-order divergence which do not require expensive parsing of the assisting language corpus. Further, use of pre-ordering to address word-order divergence for multilingual training of other NLP tasks can be explored.

Acknowledgements

We would like to thank Raj Dabre for his helpful suggestions and comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations, ICLR 2015*.

Rajen Chatterjee, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2014. Supertag Based Pre-ordering in Machine Translation. In *Proceedings of the 11th International Conference on Natural Language Processing, ICON 2014*.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, ACL 2005*.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An Empirical Study of Language Relatedness for Transfer Learning in Neural Machine Translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128 – 135.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), NAACL 2018*.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning Multilingual Word Embeddings in Latent Metric Space: A Geometric Approach. *Transactions of the Association for Computational Linguistics, TACL*.

Girish Nath Jha. 2010. The TDIL program and the Indian Langauge Corpora Intitiative (ILCI). In *Proceedings of the Seventh conference on International Language Resources and Evaluation, LREC 2010*.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language Learning with Adversarial Neural Networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning, CoNLL 2017*.

Yuki Kawara, Chenhui Chu, and Yuki Arase. 2018. Recursive Neural Network Based Preordering for English-to-Japanese Machine Translation. In *Proceedings of ACL 2018, Student Research Workshop*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural Machine Translation Toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi Parallel Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*.

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Shata-Anuvadak: Tackling Multiway Translation of Indian Languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics, TACL*, 5.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. volume 24 of *Psychology of Learning and Motivation*, pages 109 – 165. Academic Press.

Jiri Navratil, Karthik Visweswariah, and Ananthakrishnan Ramanathan. 2012. A Comparison of Syntactic Reordering Methods for English-German Machine Translation. In *Proceedings of COLING 2012, The 24th International Conference on Computational Linguistics, COLING 2012*.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), IJCNLP 2017*.

Raj Nath Patel, Rohit Gupta, Prakash B. Pimpale, and Sasikumar M. 2013. Reordering rules for English-Hindi SMT. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2018*.

Ananthakrishnan Ramanathan, Jayprasad Hegde, Ritesh M. Shah, Pushpak Bhattacharyya, and Sasikumar M. 2008. Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Aligning the fastText vectors of 78 languages.

Sami Virpioja and Stig-Arne Grönroos. 2015. LeBLEU: N-gram-based Translation Evaluation Score for Morphologically Complex Languages. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell. 2018. Neural Cross-Lingual Named Entity Recognition with Minimal Resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*.