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,
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) train a NMT model on the assisting-target language pair (parent model) which is later fine-tuned for the source-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 sentence 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 significant improvement in the translation quality.

1 Introduction

Deep Learning approaches have achieved impressive results on various NLP tasks (Huang et al., 2015; Luong et al., 2015; Peters et al., 2017) and have become the de facto approach for any NLP task. However, these deep learning techniques have found to be less effective for low-resource languages when the available training data is very less (Zoph et al., 2016). Recently, several approaches like Multi-task learning (Collobert et al., 2011), multilingual learning (Yang et al., 2017), semi-supervised learning (Peters et al., 2017; Rei, 2017) and transfer learning (Pan and Yang, 2010; Zoph et al., 2016) have been explored by the deep learning community to overcome data sparsity in low-resource languages. Transfer learning trains a model for a parent task and fine-tunes the learned parent model weights (features) for a related child task (Pan and Yang, 2010; Ruder and Plank, 2017). This effectively reduces the requirement on training data for the child task as the model would have learned relevant features from the parent task data thereby, improving the performance on the child task.

Transfer learning has also been explored in the multilingual Neural Machine Translation (Zoph et al., 2016; Dabre et al., 2017; Nguyen and Chiang, 2017). The goal is to improve the NMT performance on the source to target language pair (child task) using an assisting source language (assisting to target 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. The child model can now be fine-tuned on the source-target language pairs, if parallel corpus is available. 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). Several studies have tried to address lexical divergence between the source and the target languages (Nguyen and Chiang, 2017; Lee et al., 2017; 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 and some Indian 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 see how word order divergence can be detrimental, let us consider the case of the standard RNN (Bi-LSTM) encoder-attention-decoder architecture (Bahdanau et al., 2015). The encoder generates contextual representations (annotation vectors) for each source word, which are used by
the attention network to match the source words to the current decoder state. The contextual representation is word-order dependent. Hence, if the assisting and the source languages do not have similar word order the generated contextual representations will not be consistent. The attention network (and hence the decoder) sees different contextual representations for similar words in parallel sentences across different languages. This makes it difficult to transfer knowledge learned from the assisting language to the source language.

We illustrate this by visualizing the contextual representations generated by the encoder of an English to Hindi NMT system for two versions of the English input: (a) original word order (SVO) (b) word order of the source language (SOV, for Bengali). Figure 1 shows that the encoder representations obtained are very different. The attention network and the decoder now have to work with very different representations. Note that the plot below does not take into account further lexical and other divergences between source and assisting languages, since we demonstrated word order divergence with the same language on the source side.

![Figure 1: Encoder Representations for English sentences with and without Pre-Ordering](image)

To address this word order divergence, we propose to pre-order the assisting language sentences to match the word order of the source language. We consider an extremely resource constrained scenario, where we do not have any parallel corpus for the child task. We are limited to a bilingual dictionary for transfer information from the assisting to the source language. 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

2.1 Addressing Lexical Divergence

Zoph et al. (2016) explored transfer learning for NMT on low-resource languages. They studied the influence of language divergence between languages chosen for training the parent and child model, and showed that choosing similar languages for training the parent and child model leads to better improvements from transfer learning. A limitation of Zoph et al. (2016) approach is that they ignore the lexical similarity between languages and also the source language embeddings are randomly initialized. Nguyen and Chiang (2017); Lee et al. (2017); Gu et al. (2018) take advantage of lexical similarity between languages in their work. Nguyen and Chiang (2017) proposed to use Byte-Pair Encoding (BPE) to represent the sentences in both the parent and the child language to overcome the above limitation. They show using BPE benefits transfer learning especially when the involved languages are closely-related agglutinative languages. Similarly, Lee et al. (2017) utilize lexical similarity between the source and assisting languages by training a character-level NMT system. Gu et al. (2018) address lexical divergence by using bilingual embeddings and mixture of universal token embeddings. One of the languages’ vocabulary, usually English vocabulary is considered as universal tokens and every word in the other languages is represented as a mixture of universal tokens. They show results on extremely low-resource languages.

2.2 Addressing Word Order Divergence

To the best of our knowledge, no work has addressed word order divergence in transfer learning for multilingual NMT. However, some work exists for other NLP tasks that could potentially address word order. 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 over multiple languages. A possible reason is that the divergence has to be addressed before/during construction of the contextual embeddings in the Bi-LSTM layer, and the subsequent self-attention layer does not ad-
dress word-order divergence. Joty et al. (2017) use adversarial training for cross-lingual question-question similarity ranking in community question answering. The adversarial training tries to force the encoder representations of similar sentences from different input languages to have similar representations.

2.3 Use of Pre-ordering

Pre-ordering the source language sentences to match the target language word order has been 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). Recently, Ponti et al. (2018) proposed a way to measure and reduce the divergence between the source and target languages based on morphological and syntactic properties, also termed as anisomorphism. They demonstrated that by reducing the anisomorphism between the source and target languages, consistent improvements in NMT performance were obtained. The NMT system used additional features like word forms, POS tags and dependency relations in addition to parallel corpora. On the other hand, Kawara et al. (2018) observed a drop in performance due to pre-ordering for NMT. Unlike Ponti et al. (2018), the NMT system was trained on pre-ordered sentences and no additional features were provided to the system. Note that all these works address source-target divergence, not divergence between source languages in multilingual NMT.

3 Proposed Solution

Consider the task of translating from an extremely low-resource language (source) to a target language. The parallel corpus between the two languages if available may be too small to train a NMT model. Similar to existing works (Zoph et al., 2016; Nguyen and Chiang, 2017; Gu et al., 2018), we use transfer learning to overcome data sparsity and train a NMT model between the source and the target languages. Specifically, the NMT model (parent model) is trained on the assisting language and target language pairs. We choose English as the assisting language in all our experiments. In our resource-scarce scenario, we have no parallel corpus for the child task. Hence, at test time, the source language sentence is translated using the parent model after performing a word-by-word translation into the assisting language.

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. In this paper, we propose to pre-order English sentences (assisting language sentences) to match the word-order of the source language and train the parent model on this pre-ordered corpus. In our experiments, we look at scenarios where the assisting language has SVO word order and the source language has SOV word order.

For instance, consider the English sentence *Anurag will meet Thakur*. One of the pre-ordering rule swaps the position of the noun phrase followed by a transitive verb with the transitive verb. The original and the resulting re-ordered parse tree will be as shown in the Table 1. Applying this reordering rule to the above sentence *Anurag will meet Thakur* will yield the reordered sentence *Anurag Thakur will meet*. Additionally, the Table 1 shows the parse trees for the above sentence with and without pre-ordering.

Pre-ordering should 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 hyper-parameters used in our experiments.
**Table 2:** Transfer learning results for X-Hindi pair, trained on English-Hindi corpus and sentences from X word translated to English. NP: No pre-ordering, HT: Hindi-tuned pre-ordering rules, G: Generic pre-ordering rules.

| Language    | BLEU NP | BLEU HT | BLEU G | LeBLEU NP | LeBLEU HT | LeBLEU G |
|-------------|---------|---------|--------|-----------|-----------|----------|
| Bengali     | 6.72    | 8.83    | **9.19** | 0.3710    | 0.4150    | **0.4201** |
| Gujarati    | 9.81    | 14.34   | 13.90  | 0.4321    | 0.4736    | **0.4760** |
| Marathi     | 8.77    | 10.18   | **10.30** | 0.4021    | 0.4149    | 0.4222    |
| Malayalam   | 5.73    | 6.49    | **6.95** | 0.3327    | 0.3369    | **0.3509** |
| Tamil       | 4.86    | 6.04    | 6.00   | 0.2938    | 0.3077    | **0.3133** |

4.1 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 primary 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.

4.2 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 the test set.

4.3 Network

We use OpenNMT-Torch (Klein et al., 2017) to train the NMT system. We use the standard sequence-to-sequence architecture with attention (Bahdanau et al., 2015). We use an encoder which contains two layers of bidirectional LSTMs with 500 neurons each. The decoder contains two LSTM layers with 500 neurons each. Input feeding approach (Luong et al., 2015) is used where the previous attention hidden state is fed as input to the decoder LSTM. We use a mini-batch of size 50 and use a dropout layer. We begin with an initial learning rate of 1.0 and decay the learning rate by a factor of 0.7 when the perplexity on validation set increases. The training is stopped when the learning rate falls below 0.001 or number of epochs=22. The English input is initialized with pre-trained embeddings trained using fastText (Grave et al., 2018).\(^2\)

English vocabulary consists of 0.27M tokens appearing at least 2 times in the English training corpus. For constructing the Hindi vocabulary we considered only those tokens appearing at least 5 times in the training split resulting in a vocabulary size of 50K tokens. 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 (Google Translate word translation in our case). In an end-to-end solution, it would have been ideal to use bilingual embeddings or obtain word-by-word translations via bilingual embeddings (Xie et al., 2018). But, the quality of publicly available bilingual embeddings for English-Indian languages is very low for obtaining good-quality, bilingual representations (Smith et al., 2017; Jawanpuria et al., 2018). We also found that these embeddings were not useful for transfer learning.

We use the CFILT-preorder\(^3\) system for reordering English sentences to match the Indian language word order. It contains two re-ordering systems: (1) generic rules that apply to all Indian languages (Ramanathan et al., 2008), and (2) hindi-tuned rules which improve the generic rules by incorporating improvements found through an error analysis of English-Hindi reordering (Patel et al., 2013). These Hindi-tuned rules have been

---

\(^1\)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
found to improve reordering for many English to Indian language pairs (Kunchukuttan et al., 2014).

5 Results

In this section, we describe the results from our experiments on NMT task. We report the results on X-Hindi pair, where X is one of Bengali, Gujarati, Marathi, Tamil, and Malayalam. The results are presented in the Table 2. We report BLEU scores and LeBLEU scores (Virpioja and Grönroos, 2015). We observe that both the pre-ordering configurations significantly improve the BLEU scores over the baseline scores. We observe larger gains when generic pre-ordering rules are used compared to the Hindi-tuned pre-ordering rules.

These results support our hypothesis that word-order divergence can limit the benefits of multilingual translation. Reducing the word order divergence can improve translation in extremely low-resource scenarios.

| Language  | NP  | 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. NP: No pre-ordering, HT: Hindi-tuned pre-ordering rules, G: Generic pre-ordering rules

An analysis of the outputs revealed that pre-ordering significantly reducing 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 contextual representations leading to less number of unknown tokens and better translation which is also reflected in the BLEU scores.

LeBLEU (Levenshtein Edit BLEU) is a variant of BLEU that does a soft-match of reference and output words based on edit distance, hence it can handle morphological variations and cognates.

6 Conclusion

In this paper, we show that handling word-order divergence between 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. While the current work focused on Indian languages, we would like to validate the hypothesis on a more diverse set of languages.

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.

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’05). Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12.

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. The National University (Phillippines).

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). European Language Resources Association (ELRA).

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). Association for Computational Linguistics.
Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. CoRR, abs/1508.01991.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2018. Learning multilingual word embeddings in latent metric space: A geometric approach. arXiv preprint arXiv:1808.08773.

Girish Nath Jha. 2010. The tdl program and the indian language corpora initiative (ilci). In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10). European Languages Resources Association (ELRA).

Shafiq Joty, Preslav Nakov, Lluís Marquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Association for Computational Linguistics.

Yuki Kawara, Chenhui Chu, and Yuki Arase. 2018. Recursive neural network based reordering for english-to-japanese machine translation. In Proceedings of ACL 2018, Student Research Workshop. Association for Computational Linguistics.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. ArXiv e-prints.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The it bombay english-hindi parallel corpus. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). European Language Resources Association (ELRA).

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Shata-anuvadak: Tackling multilingual translation of indian languages. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). European Language Resources Association (ELRA).

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. Transactions of the Association for Computational Linguistics, 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.

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 COLING 2012 Organizing Committee.

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). Asian Federation of Natural Language Processing.

S. J. Pan and Q. Yang. 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10).

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. Association for Computational Linguistics.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.

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). Association for Computational Linguistics.

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-1.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with bayesian optimization. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

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 Gönroos. 2015. Lebleu: N-gram-based translation evaluation score for morphologically complex languages. In Proceedings of the Tenth Workshop on Statistical Machine Translation. Association for Computational Linguistics.

J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell. 2018. Neural Cross-Lingual Named Entity Recognition with Minimal Resources. In
Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2017. Multi-task cross-lingual sequence tagging from scratch. In International Conference on Learning Representations.

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. Association for Computational Linguistics.