Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 29–32
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Multilingual Neural Machine Translation involving Indian Languages

Pulkit Madaan, Fatiha Sadat
IIIT Delhi, UQAM
Delhi India, Montreal Canada
pulkit16257@iiitd.ac.in, sadat.fatiha@uqam.ca

Abstract

Neural Machine Translations (NMT) models are capable of translating a single bilingual pair and require a new model for each new language pair. Multilingual Neural Machine Translation models are capable of translating multiple language pairs, even pairs which it hasn't seen before in training. Availability of parallel sentences is a known problem in machine translation. Multilingual NMT model leverages information from all the languages to improve itself and performs better. We propose a data augmentation technique that further improves this model profoundly. The technique helps achieve a jump of more than 15 points in BLEU score from the Multilingual NMT Model. A BLEU score of 36.2 was achieved for Sindhi–English translation, which is higher than any score on the leaderboard of the LoResMT SharedTask at MT Summit 2019, which provided the data for the experiments.

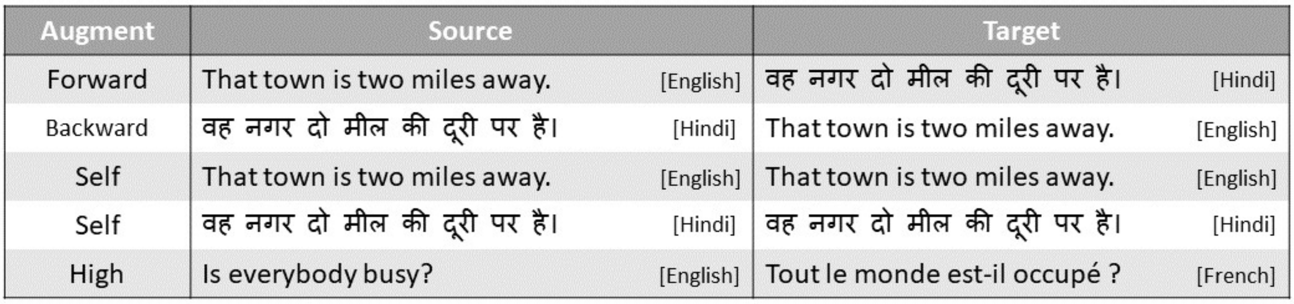
Keywords: Neural Machine Translation, Low Resource Languages, Multilingual Transformer, Deep Learning, End-to-end Learning, Data Augmentation, Transfer Learning

1. Introduction

A lot of the models for end-to-end NMT are trained for single language pairs. Google's Multilingual NMT (Johnson et al., 2017) is a single model capable of translating to and from many languages. The model is fed a token identifying a target language uniquely along with the source language sentence. This allows the model to translate between pairs for which the model hasn't seen parallel data, essentially zero-shot translations. The model is also able to improve upon individual translation qualities too by the help of other languages. NMT suffers from the lack of data. And as Arivazhagan et al.(2019b) and Koehn et al.(2017) too recognize, lack of data makes NMT a non-trivial challenge for low-resource languages. Multilingual NMT is a step towrds solving this problem which leverages data from other language pairs and does an implicit transfer learning. We propose to improve this quality further with a data-augmentation technique that was able to improve the BLEU scores two fold in our experiments. The technique is simple and can work with any model. We show that increasing the amount of data available for training artificially with our technique in a way as simple as just swapping the source with target sentences and using the same sentence as source and target can improve the BLEU scores significantly. Also, we show that since all language pairs share the same encoder and the same decoder, in a case of transfer learning, the model is able to leverage data from rich resource language pairs for learning better translations for low-resource pairs. Using Hindi–English data in training improved the BLEU scores for {Bhojpuri, Sindhi, Magahi}<>English. The structure of the present paper is described as follows: Section 2 presents the state of the art. Section 3 presents our proposed methodology. Section 4 describes the corpora used in this research. In section 5, we put forward our experiments and evaluations, perform an ablative analysis and compare our system's performance with other Google's Neural Machine Translation(Johnson et al., 2017). Section 6, compares our results with other methods that participated in the LoResMT Shared Task(Karakanta et al., 2019) at the MT Summit 2019. Finally in Section 7, we state our conclusions and perspectives for future research.

2. Related Work

Significant progress has been made in end-to-end NMT (Cho et al., 2014; Sutskeveret al., 2014; Bahdanau et al., 2015) and some work has been done to adapt it to a multilingual setting. But, before the mulitilingual approach of Johnson et al., 2017, none of the approaches have a single model capable of dealing with multiple language pairs in a many-to-many setting. Dong et al.(2015) use different decoders and attention layers for different target languages. Firat et al.(2016) use a shared attention layer but an encoder per source language and a decoder per target language. Lee et al.(2017) use a single model with the whole model shared across all pairs but it can only be used for a single target language. The model proposed by Johnson et al.(2017) has a single model for a many-to-many task and is able to perform in zero-shot setting too wher it can translate sentences between pairs whose parallel data wasn't seen by the model during training. Arivazhagan et al.(2019a) also propose a model for zero-shot translation that improves upon Google's Multilingual NMT Model (Johnson et al., 2017) and achieves results on par with pivoting. They propose English as the pivot language and feed the target language token to the decoder instead of the encoder. In order to improve the independence of encoder on source language they maximise the similarity between all sentence vectors and their English parallel sentence embeddings and minimize the translation cross-entropy loss. They use a discriminator and train the encoder adversarially for similarity maximisation. Artetxe et al.(2018) and Yang et al.(2018) also train the encoder adversarially to learn a shared latent space. There has been a lot of work done to improve NMT models using data augmentation. Sennrich et al (2016a) proposed automatic back-translation to augment the dataset. But, as mentioned in SwitchOut (Wang et al., 2018) faces challenges in initial models. Fadaee et al.(2017) propose



Augment	Source		Target	
Forward	That town is two miles away.	[English]	वह नगर दो मील की दूरी पर है।	[Hindi]
Backward	वह नगर दो मील की दूरी पर है।	[Hindi]	That town is two miles away.	[English]
Self	That town is two miles away.	[English]	That town is two miles away.	[English]
Self	वह नगर दो मील की दूरी पर है।	[Hindi]	वह नगर दो मील की दूरी पर है।	[Hindi]
High	Is everybody busy?	[English]	Tout le monde est-il occupé ?	[French]

Figure 1: An example of different augments. Here the low resource pair of languages is English–Hindi, and the high resource pair language set is English–French

a data augmentation technique where they synthesise new data by replacing a common word in the source sentence with a rare word and the corresponding word in the target sentence with its translation. And to maintain the syntactic validity of the sentence, they use an LSTM language model. Zhu et al.(2019) propose a method in which they obtain parallel sentences from multilingual websites. They scrape the websites to get monolingual data on which they learn word embeddings. These embeddings are used to induce a bilingual lexicon and then use a trained model to identify parallel sentences. Ours is a much simpler way, which does not require an additional model, is end-to-ed trainable and is still at par with some Statistical Machine Translation methods submitted at the SharedTask.

3. The Proposed Methodology

The technique we propose is simple consists of four components named **Forward**, **Backward**, **Self** and **High**. **Forward** augmentation is the given data itself. **Backward** augmentation is generated by switching the source and target label in the **Forward** Data, so the source sentence becomes the target sentence and vice versa in parallel sentence pair. **Self** augmentation is generated by using only the required language from the parallel sentences and cloning them as their own target sentences, so the source and target sentence are the same. An example of the augmentations is shown in Figure 1. We know that translation models improve with increase in data and since we also have the same encoder for every language, we can use a language pair that is similar to the language pairs of the task and is a high resource pair to further improve the encoder in encoding source independent embeddings, for transfer learning through the Multilingual architecture of Johnson et al.(2017) . So we propose **Multilingual+** which uses the above mentioned three augmentations (Forward, Backward, Self) along with **High** augmentation; **High** augmentation consists of high-resource language pairs, like Hindi–English parallel data, in Forward, Backward and Self augmentations. This helps in improving the translation models of low resource pairs; {Bhojpuri, Sindhi, Magahi}<>English.

4. Dataset

Parallel data from four different language pairs are used in the experiments. Following are the language pairs along with the number of parallel sentences of each pair:

1. Sindhi–English (29,014)
2. Magahi–English (3,710)
3. Bhojpuri–English (28,999)
4. Hindi–English (1,561,840)

Data for pairs 1–3 were made available at the Shared Task at MT Summit 2019. While data for pair 4 was obtained from the IIT Bombay English–Hindi Corpus (Kunchukuttan et al., 2018). The Train-Val-Test splits were used as given by the respective data providers.

5. Experiments

We performed experiments on the Multingual+ model and showed how the addition of each of augmentations we proposed improves the performance by an ablative analysis. After augmentation, the source sentences get a target language token prepended. Joint Byte-Pair Encoding is learnt for subword segmentation (Sennrich et al., 2016b) to address the problem of rare words. Byte-Pair encoding was learnt over the training data and was used to segment subwords for both the training and the test data. A Joint dictionary was learnt over all the languages. This is the only pre-processing that we do besides the augmentation. The basic architecture is the same as in Johnson et al.(2017). A single encoder and decoder shared over all the languages. Adam (Kingma and Ba, 2015) optimizer was use, with initial beta values of 0.9 and 0.98 along with label smoothing and dropout(0.3). Following are the augmentations included in Multinlingual+

- **Forward**
  Sindhi-to-English, Bhojpuri-to-English, Magahi-to-English
- **Backward**
  English-to-Sindhi, English-to-Bhojpuri, English-to-Magahi
- **Self**
  Sindhi-to-Sindhi, Bhojpuri-to-Bhojpuri, Magahi-to-Magahi, English-to-English
- **High**
  Hindi-to-English, English-to-Hindi, Hindi-to-Hindi

	Sin-to-Eng	**Eng-to-Sin**	**Bho-to-Eng**	**Eng-to-Bho**	**Mag-to-Eng**	**Eng-to-Mag**
Base	15.74*	–	6.11*	–	2.46*	–
Base + Back	18.09*	11.38*	5.01*	0.2	2.55*	0.2
Base + Back + Self	30.77*	18.98*	7.38*	0.6	4.61*	1.2
†Multilingual+	**36.2**	**28.8**	**15.6**	**3.7**	**13.3**	**3.5**

Table 1: BLEU scores of different language pairs and directions in the different experiments.
*Results on test data evaluated by the Shared Task at MT Summit 2019 committee.
† Not submitted for the SharedTask

To understand how each augmentation improves the BLEU score, we create 4 methods:

- **Base**
  This is the standard model as used in (Johnson et al., 2017), hence it uses only **Forward** and forms our baseline.

- **Base + Back**
  We add **Backward** augmentation to the baseline model

- **Base + Back + Self**
  We add **Self** & **Backward** augmentation to the baseline.

- **Multilingual+**
  This uses all the augmentations:**High** along with **Forward**, **Backward** & **Self**.

Parameters and training procedures are set as in Johnson et al.(2017). PyTorch Sequence-to-Sequence library, **fairseq** (Ott et al., 2019), was used to run the experiments. Table 1 shows that **Multilingual+** consistently outperforms the others. The table also confirms that the more augmentations you add to the Multilingual NMT model (Johnson et al., 2017), the more it improves. Adding **Backward**, then **Self** and then a new language pair improved the results at each level. All the BLEU scores reported, except star (*) marked, are calculated using SacreBLEU (Post, 2018) on the development set provided.

6. Comparisons

We compared our results with other models submitted at the LoResMT Shared Task at the MT Summit 2019. The submission to the Shared Task followed a naming convention to distinguish between different types of corpora used, which we will follow too. The different types of corpora and their abbreviations are as follows:

- Only the provided parallel corpora [-a]

- Only the provided parallel and monolingual corpora [-b]

Using these abbreviations the methods were named in the following manner”

<TeamCode>-<Language-and-Direction>-
<MethodName>-<Used-Corpora-Abbreviation>

Our Team Code was L19T3 and we submitted Base (as Method_1), Base+Back (as Method_2) and Base+Back+Self (as Method_3) all under -a category. Multilingual+ was developed later. Table 2 shows the top 3 performers in different translation directions along with Multilingual+. Method3-b from team L19T2 is a Phrase Based Statistical Machine Translation model. While their Method2-a is an NMT model that uses a sequence-to-sequence approach along with self-attention. pbmt-a model from team L19T5 is again a Phrase Based Statistical Machine Translation model. While their xform-a model is an NMT model. Both of the NMT models of the other teams train a different model for different language pairs, one for each, while ours is a one for all model. Multilingual+ is the best performer in Sin-to-Eng and Mag-to-Eng task, second best performer in Eng-to-Sin and Bho-to-Eng tasks. These results show the superiority of our simple approach. Our data augmentation technique is comparable or better than the best of the methods on the leaderboard of the SharedTask.

In Eng-to-Sin task L19T2-Eng2Sin-Method3-b scores the best while the second best is Multinlingual+. This could be because the former is a Statistical Machine Translation Model. Though, it surpasses the L19T2's NMT model. For Bho-to-Eng it is able to surpass pbmt-a of team L19T5 it still lags behind their NMT model. This can be explained as we have more data for Sindhi than Bhojpuri and though we were able to improve the performance by augmenting data, it still remains behind statistical machine translation approach of L19T2. The success of our simple approach can be attributed to its conjunction with Multilingual NMT. Multilingual NMT is able to use data of all langugaes to improve them all together, and by even further increasing this data, we improve the model greatly.

7. Conclusion and Future Work

We have presented a simple data augmentation technique coupled with a multilingual transformer that gives a jump of 15 points in BLEU score without any new data and 20 points in BLEU score if a rich resource language pair is introduced, over a standard multilingual transformer. It performs at par or better than best models submitted at the Shared Task. This demonstrates that a multilingual transformer is sensitive to the amount of data used and a simple augmentation technique like ours can provide a significant boost in BLEU scores. Back-translation (Sennrich et al., 2016a) can be coupled with our approach to experiment and analyse the effectiveness of this amalgam.

Rank	Sin-to-Eng		Eng-to-Sin	
1	L19T2-Sin2Eng-Method3-b	31.32	L19T2-Eng2Sin-Method3-b	**37.58**
2	Base+Back+Self	30.77	L19T2-Eng2Sin-Method2-a	25.17
3	L19T5-sin2eng-xform-a	28.85	Base+Back+Self	18.98
	Multilingual+	**36.2**	Multilingual+	28.8

Rank	Bho-to-Eng		Mag-to-Eng	
1	L19T2-Bho2Eng-Method3-b	**17.03**	L19T2-Mag2Eng-Method3-b	9.71
2	L19T5-bho2eng-xform-a	15.19	L19T5-mag2eng-pbmt-a	5.64
3	L19T5-bho2eng-pbmt-a	14.2	Base+Back+Self	4.61
	Multilingual+	15.6	Multilingual+	**13.3**

Table 2: Top 3 performers in LoResMT Shared Task in different translation directions along with Multilingual+

8. Bibliographical References

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., and Macherey, W. (2019a). The missing ingredient in zero-shot neural machine translation. *CoRR*, abs/1903.07091.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. (2019b). Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *CoRR*, abs/1710.11041.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Cho, K., van Merrienboer, B., Çaglar Gülçehre, Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *ArXiv*, abs/1406.1078.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *ACL*.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada, July. Association for Computational Linguistics.

Firat, O., Cho, K., and Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. *ArXiv*, abs/1601.01073.

Johnson, M., Schuster, M., Le, Q. V., Krikuna, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G. S., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Alina Karakanta, et al., editors. (2019). *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, Dublin, Ireland, August. European Association for Machine Translation.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT bombay english-hindi parallel corpus. *Language Resources and Evaluation Conference*, 10.

Lee, J. D., Cho, K., and Hofmann, T. (2017). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. *ArXiv*, abs/1511.06709.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*.

Wang, X., Pham, H., Dai, Z., and Neubig, G. (2018). SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium, October-November. Association for Computational Linguistics.

Yang, Z., Chen, W., Wang, F., and Xu, B. (2018). Unsupervised neural machine translation with weight sharing. *CoRR*, abs/1804.09057.