Efficient Neural Machine Translation for Low-Resource Languages via Exploiting Related Languages

Vikrant Goyal, Sourav Kumar, Dipti Misra Sharma
Language Technologies Research Center (LTRC)
IIIT Hyderabad, India
{vikrant.goyal, sourav.kumar}@research.iiit.ac.in, dipti@iiit.ac.in

Abstract

A large percentage of the world's population speaks a language of the Indian subcontinent, comprising languages from both Indo-Aryan (e.g. Hindi, Punjabi, Gujarati, etc.) and Dravidian (e.g. Tamil, Telugu, Malayalam, etc.) families. A universal characteristic of Indian languages is their complex morphology, which, when combined with the general lack of sufficient quantities of high-quality parallel data, can make developing machine translation (MT) systems for these languages difficult. Neural Machine Translation (NMT) is a rapidly advancing MT paradigm and has shown promising results for many language pairs, especially in large training data scenarios. Since the condition of large parallel corpora is not met for Indian-English language pairs, we present our efforts towards building efficient NMT systems between Indian languages (specifically Indo-Aryan languages) and English via efficiently exploiting parallel data from the related languages. We propose a technique called Unified Transliteration and Subword Segmentation to leverage language similarity while exploiting parallel data from related language pairs. We also propose a Multilingual Transfer Learning technique to leverage parallel data from multiple related languages to assist translation for low-resource language pair of interest. Our experiments demonstrate an overall average improvement of 5 BLEU points over the standard Transformer-based NMT baselines.

1 Introduction

In recent years, Neural Machine Translation (Luong et al., 2015; Bahdanau et al., 2014; Johnson et al., 2017; Wu et al., 2017; Vaswani et al., 2017) (NMT) has become the most prominent approach to Machine Translation (MT) due to its simplicity, generality and effectiveness. In NMT, a single neural network often consisting of an encoder and a decoder is used to directly maximize the conditional probabilities of target sentences given the source sentences in an end-to-end paradigm. NMT models have been shown to surpass the performance of previously dominant statistical machine translation (SMT) (Koehn, 2009) on many well-established translation tasks.

However, in order to reach high accuracies, NMT systems tend to require very large parallel training corpora (Koehn and Knowles, 2017). As a matter of fact, such corpora are not yet available for many language pairs. Indian languages are not an exception to this; however they are extremely diverse, belonging to different language families, employing various scripts and spanning a multitude of dialects. The majority of Indian languages are morphologically rich and depict unique characteristics, which are significantly different from languages such as English.

Since NMT models learn poorly from small corpora, building effective NMT systems for low-resource languages (e.g. Indian languages) becomes a primary challenge. The bulk of research on low-resource NMT has focused on exploiting monolingual data, or parallel data involving other language pairs. Some of the most well-known methods to improve NMT models with monolingual data range from backtranslation (Sennrich et al., 2016), dual learning (He et al., 2016) to Unsupervised MT (Artetxe et al., 2017; Lample et al., 2017, 2018). Similarly, parallel data from other languages can be exploited to either pretrain the network or jointly learn the representations (Zoph et al., 2016; Firat et al., 2017; Johnson et al., 2017; Kocmi and Bojar, 2018).

Currently, Transfer Learning (TL) is being widely used for low-resource language translation because it is one of the vital directions for addressing the data sparsity problem in low-resource NMT (Zoph et al., 2016; Nguyen and Chiang, 2017; Pass-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 162–168
July 5 - July 10, 2020.

ban et al., 2017; Kocmi and Bojar, 2018). However, most of the existing approaches that take advantage of transfer learning have a major limitation: they do not exploit multiple languages together and in an efficient manner. The idea presented by Zoph et al. (2016) may have the shortcoming of exploiting only one high-resource model (parent) at a time to optimize the low-resource model (child). Actually, the use of highly related multiple language pairs might help to increase the translation quality of the child model. The original Transfer Learning method (Zoph et al., 2016) also makes no assumption about the relatedness of the parent and child languages. Multilingual NMT (Firat et al., 2017; Johnson et al., 2017) approaches which also use parallel data from different languages to improve the translation quality of NMT models does not exploit language relatedness either.

In this paper, we present our efforts towards building efficient NMT systems between Indian languages (specifically Indo-Aryan languages) and English by exploiting parallel data from related languages. We aim to deal with the problem of how to make full use of these corpora of highly related languages, to increase the translation quality of low-resource languages. To this end, we introduce two simple and yet effective approaches:

- Multilingual Transfer Learning: to enable the low-resource languages (child model) to exploit parallel data from multiple related languages which may or may not be high-resourced, and
- Unified Transliteration and Subword Segmentation: to exploit the language similarity between the related language pairs.

Experiments show that our approaches are effective and significantly outperform the state-of-the-art Transformer (Johnson et al., 2017) baseline. Our proposed approach of Multilingual Transfer Learning also significantly outperforms simple Transfer Learning (Zoph et al., 2016) approach, where NMT models are also built using Unified Transliteration and Subword Segmentation approach.

2 Methodology

The core idea of our method is to extend the Multilingual Learning (Johnson et al., 2017) and Transfer Learning (Zoph et al., 2016) approaches to effectively exploit parallel data from multiple related languages. In Section 2.2, we explain our Unified Transliteration and Subword Segmentation technique to exploit language relatedness among the parallel data of related languages. Sections 2.3 and 2.4 describe our modified Multilingual Learning and Transfer Learning techniques for NMT. In Section 2.5, we describe our Multilingual Transfer Learning approach.

2.1 Language Relatedness

In this work, we experiment on Indo-Aryan languages specifically Hindi, Punjabi, Gujarati, Marathi and Bengali. Being from one language family, these languages are closely related to each other and share many features. These languages are morphologically rich and depict unique characteristics, which are significantly different from languages such as English. Some of these characteristics are the relatively free word-order with a tendency towards the Subject-Object-Verb (SOV) construction, a high degree of inflection, usage of reduplication and conjunct verbs. These languages share many common words which have the same root and meaning. They use different Indic scripts derived from the ancient Brahmi script, but correspondences can be established between equivalent characters across scripts.

2.2 Unified Transliteration and Subword Segmentation

Unlike the original Transfer Learning (Zoph et al., 2016) and the Multilingual Neural MT (Johnson et al., 2017) methods which do not exploit any language relatedness, the basic idea of this approach is to exploit the relationship between the related language lexicons while using parallel data from related languages to assist with translation of low-resource languages. To do so, we find a representation of the data that ensures a sufficient overlap between the vocabularies of the related languages.

Since the languages involved in the models have different orthographies, the data processing should help to map them into a common orthography but here we take a minimalist approach; we transliterate all the Indian languages (Hindi, Gujarati, Bengali, Marathi and Punjabi) into a common Devanagari script to share the same surface form. This unified transliteration is a string homomorphism, replacing characters in all the languages mentioned above with Hindi characters (script conversion to Devanagari) or consonant clusters independent of context.

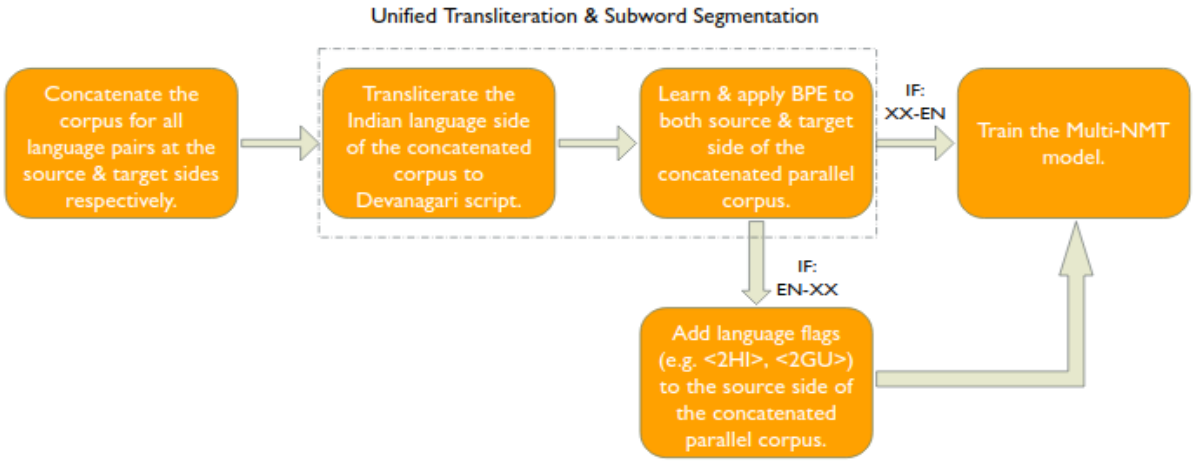

Figure 1: Our pipeline for building Multilingual NMT models for Indian languages.

Now, to increase the overlap between the vocabularies of the languages used in a model, which are already transliterated into a common script and consequently share the same surface form, we use Byte Pair Encoding (BPE) (Sennrich et al., 2015) to break words into subwords. For the BPE merge rules to not only find the common subwords between two related languages but also ensure consistency between source and target segmentation among each language pair, we learn the rules from the union of source and target data of all the language pairs involved in the model construction. The rules are then used to segment the corpora. It is important to note that this results in a single vocabulary, used for both the source and target languages in all the language pairs.

2.3 Multilingual Learning for NMT

The objective of Multilingual Learning for NMT is to construct a single model for translating to and from multiple languages. Early work in multilingual NMT utilizes a separate encoder, decoder and an attention mechanism to support the translation of either one-to-many or many-to-one language directions. Firat et al. (2017) introduced a many-to-many system, which still relied upon separate encoder-decoder setup with a shared attention mechanism. In a simplified manner and yet delivering better performance, Johnson et al. (2017) introduced a "language flag"-based approach that shares the attention mechanism and a single encoder-decoder network to enable multilingual models. A language flag or token is prepended to the input sequence to indicate which direction to translate in. The decoder learns to generate the target given this input.

However, the Multilingual NMT approaches do not consider the relatedness of the languages or how many shared words there are among the different source and target languages. Mainly, they aim at exploiting many different source and target languages rather than focusing on similarities between many languages that are used in the training and the languages that is used in testing. Accordingly, we modify the Multilingual NMT approach (Johnson et al., 2017) with Unified Transliteration and Subword segmentation technique to exploit the language relatedness. We experiment with this modified approach in our work on efficient NMT for Indian languages.

2.4 Transfer Learning for NMT

Zoph et al. (2016) proposed how Transfer Learning between two NMT models can improve a low-resource NMT task. In their approach, a language pair with a relatively large amount of parallel data is first utilized to train a parent model in a phase known as "pretraining". Then the encoder-decoder parameters are transferred to initialize a child model for a low-resource language pair of interest. After initializing, the model enters the "fine-tuning" stage, where the child model is fine-tuned on the low-resource language pair. This enables the inductive transfer of knowledge from the parent model to the child model. This approach does not make any assumption between the relatedness of the parent and child language pair. However, in our work we use a relatively high-resource lan-

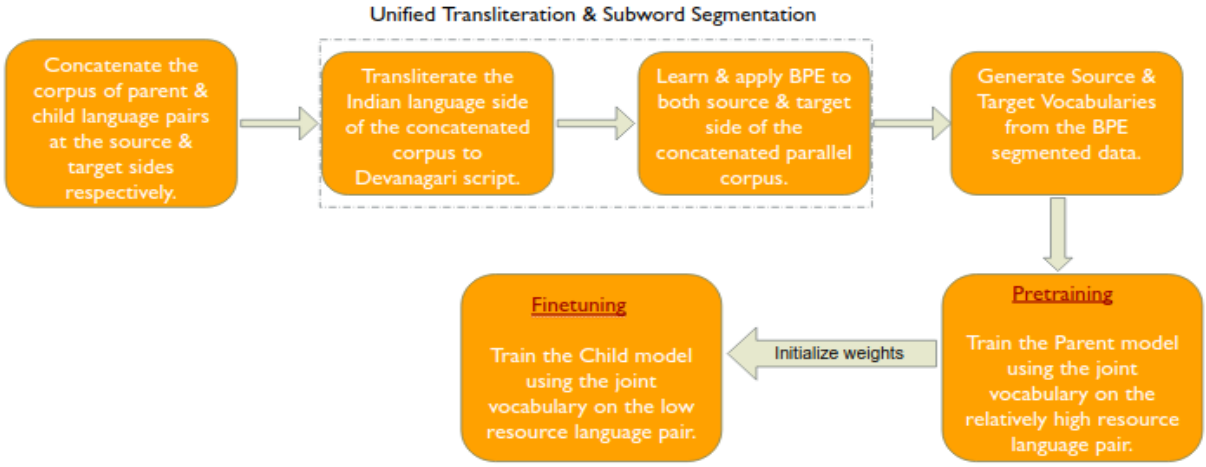

Figure 2: Our pipeline for building Transfer Learning models for Indian languages.

guage pair as our parent model which has similar syntactic and morphological properties as the child language pair. We further exploit the language relatedness of parent and child language pairs via our Unified Transliteration and Subword Segmentation technique. We experiment with this modified Transfer Learning technique and demonstrate huge BLEU improvements over the Transformer NMT baseline for low-resource Indian languages.

2.5 Multilingual Transfer Learning for NMT

In the normal Transfer Learning (Zoph et al., 2016) approach for NMT, the parent model is trained on a single high-resource language pair which may or may not be related to the child language pair of interest. Passban et al. (2017) presented a double transfer learning technique which first trains a parent model on a single high-resource language pair, then initializes the next parent model on the same single high-resource language pair but with different domain and corpus size, and finally fine-tunes it on the child task. To the best of our knowledge, previous Transfer Learning approaches do not exploit parallel data from multiple languages. However, learning from multiple languages can result in better knowledge transfer.

Therefore, in this work, we propose a new Transfer Learning approach called as Multilingual Transfer Learning to enable the low-resource languages to efficiently learn from multiple related languages which may or may not be high-resourced. In this approach, the parent model is a Multilingual NMT model of related languages and also the child language pair. This Multilingual parent NMT model also uses the Unified Transliteration and Subword Segmentation technique to exploit language relatedness more efficiently as discussed in Section 2.2. After pretraining the parent model, the child model is initialyzed with parent model parameters and is then fine-tuned on the low-resource language pair of interest.

The proposed approach may deliver better results than Multilingual NMT and Transfer Learning because adding more languages into one model may result in better knowledge transfer (i.e multilingual NMT) but it can also result in ambiguities between languages at the inference time. Accordingly, a multilingual NMT model fine-tuned on the language pair of interest can potentially remove all the inconsistencies at the inference time.

3 Experimental Settings

3.1 Dataset

In our experiments, we use the IIT-Bombay (Kunchukuttan et al., 2017) parallel data for Hindi-English. The training corpus consists of data from mixed domains. We use the multilingual ILCI (Indian Language Corpora Initiative) corpus (Jha, 2010), which contains roughly 50,000 parallel sentences for each of the Indian languages (Gujarati, Punjabi, Marathi, Bengali) and also for English. The ILCI data is from tourism and health domains. For every XX-EN language pair (where XX is Gujarati, Marathi, Bengali or Punjabi), the English side of the data is same because of the multilingual nature of the corpus. We check and clean the ILCI corpus manually as it contains a lot of misalign-

ments and mistranslations.

Table 1: Statistics of our cleaned and processed parallel data, where XX is Gujarati, Marathi, Bengali or Punjabi

Dataset	Sentences
IITB HI-EN Train	1,528,631
ILCI XX-EN Train	46,490
ILCI XX-EN Test	2,000
ILCI XX-EN Dev	500

3.2 Data Processing

We use the Moses (Koehn et al., 2007) toolkit[1] for tokenization and cleaning the English side of the data. All the Indian language data is first normalized with the Indic NLP library[2] followed by tokenization with the same library. As our preprocessing step, we remove all sentences of length greater than 80 words from our training corpus and lowercase the English side of the data. In all cases, we use BPE segmentation with 16k merge operations as described in Section 2.2.

3.3 Training Details

For all of our experiments, we use the OpenNMT-py (Klein et al., 2018) toolkit[3]. We use the Transformer model with 6 layers in both the encoder and decoder, each with 512 hidden units. The word embedding size is set to 512 with 8 heads. The training is done in batches of maximum 4096 tokens at a time with dropout set to 0.3. We use the Adam (Kingma and Ba, 2014) optimizer to optimize model parameters. We validate the model every 5,000 steps via BLEU (Papineni et al., 2002) and perplexity on the development set. We train all our NMT models for 150k steps except for fine-tuning which is done for 10k steps. After translation at the test time, we rejoin the translated BPE segments and convert the translated sentences back to their original language scripts. Finally, we evaluate the accuracy of our translation models using BLEU.

4 Results

We report the results of Multilingual Learning, Transfer Learning and Multilingual Transfer Learning for Gujarati-English, Bengali-English, Marathi-English and Punjabi-English language pairs for both translation directions (XX-EN and EN-XX). Table 2 shows our main results for the Indian language to English (XX-EN) translation direction. Multilingual models for XX-EN language direction do not show any improvements. The reason for this might be the multiparallel nature of the ILCI data where each English sentence on the target side appears 4 times in the model, thereby creating ambiguities in the model. The transfer learning model built using Unified Transliteration and Subword Segmentation that was trained on the IITB HI-EN data and then fine-tuned on XX-EN data (see model no. 8 in Table 2) resulted in an average improvement of 5 BLEU points.

Table 3 shows our main results for the English to Indian language (EN-XX) translation direction. In this case, the multilingual model using all ILCI data shows significant improvements over the baseline, unlike in the XX-EN translation direction. The reason for this is that in the EN-XX direction, language flags are used on the source side which guides the decoder to which language the model translate in, whereas the same is not possible for the XX-EN direction as verified by our preliminary experiments. The other two multilingual models containing the IITB EN-HI data show performance degradation, potentially due to the mismatch between the size of the IITB EN-HI (1.5M sentences) and ILCI data (47k sentences). The transfer learning model that was trained on IITB EN-HI data and then fine-tuned on EN-XX data (see model no. 8 in Table 3) also resulted in an average improvement of 5 BLEU points.

In both translation directions, the multilingual models do not prove to be effective. Fine-tuning the multilingual models (multilingual transfer learning) on XX-EN or EN-XX data removes some ambiguities in the model and shows significant improvements compared to their simple multilingual model counterparts. The best performance (almost 5-6 BLEU improvements over the baseline) is achieved by fine-tuning the multilingual model (trained on IITB HI-EN or EN-HI data and all the ILCI data) on EN-XX or XX-EN outperforming all the NMT, Multilingual NMT and Transfer Learning baselines thus demonstrating the effectiveness of our technique.

[1] https://github.com/moses-smt/mosesdecoder
[2] https://anoopkunchukuttan.github.io/indic_nlp_library/
[3] https://github.com/OpenNMT/OpenNMT-py/

Table 2: BLEU scores of the contrastive experiments for Indian Language to English translation (XX to EN).

Model No.	Model Description	Gujarati	Bengali	Marathi	Punjabi
1	Baseline	28.37	22.40	25.29	30.51
2	Multilingual Model of all ILCI data	25.14	21.47	23.56	25.43
3	Multilingual Model of IITB HI-EN data & all ILCI data	28.62	22.71	26.90	29.46
4	Multilingual Model of IITB HI-EN data & ILCI data of XX-EN	29.18	23.93	27.15	30.54
5	Fine-tuning model no. 2 on XX-EN	26.83	22.72	25.36	27.12
6	Fine-tuning model no. 3 on XX-EN	**33.78 (+5.41)**	**27.55 (+5.15)**	**31.79 (+6.5)**	**34.70 (+4.19)**
7	Fine-tuning model no. 4 on XX-EN	33.72	27.40	31.80	34.68
8	Fine-tuning model pretrained on IITB HI-EN data on XX-EN	33.13	27.06	31.27	34.54

Table 3: BLEU scores of the contrastive experiments for English to Indian Language translation (EN to XX).

Model No.	Model Description	Gujarati	Bengali	Marathi	Punjabi
1	Baseline	20.67	16.59	15.13	25.20
2	Multilingual Model of all ILCI data	24.61	19.81	17.92	28.02
3	Multilingual Model of IITB EN-HI data & all ILCI data	20.63	16.51	15.05	21.76
4	Multilingual Model of IITB EN-HI data & ILCI data of EN-XX	14.30	6.38	8.88	14.54
5	Fine-tuning model no. 2 on EN-XX	24.75	20.25	18.75	28.16
6	Fine-tuning model no. 3 on EN-XX	**26.22 (+5.55)**	**21.62 (+5.03)**	**19.90 (+4.77)**	**30.27 (+5.07)**
7	Fine-tuning model no. 4 on EN-XX	25.52	20.45	19.77	29.53
8	Fine-tuning model pretrained on IITB EN-HI data on EN-XX	25.35	21.77	19.58	29.54

5 Conclusion & Future Work

In this paper, we explore effective methods to exploit parallel data from multiple related languages to improve the translation between Indian languages and English. Our results show that Multilingual Learning for translation between Indian Languages and English is not very effective given the set of data we have. However, the performance of multilingual models can easily be enhanced by fine-tuning them on the low-resource language pairs of interest. Our experiments show that using a Multilingual NMT model as a parent model (consisting of multiple language pairs with related languages either on the source side or on the target side) and fine-tuning it on the low-resource language pair of interest yields an overall average improvement of 5 BLEU points over a standard Transformer-based NMT baseline. Our proposed Multilingual Transfer Learning approach also outperforms the simple Transfer Learning approach by a significant amount. In future, we would like to work on effective techniques to exploit monolingual data and parallel data from other languages together to improve the translation of low-resource languages.

Acknowledgements

We would like to thank Prof. Andy Way for his valuable and detailed comments in improving the paper.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Computer Speech and Language*, 45(C):236–252.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

Girish Nath Jha. 2010. The tdil program and the indian langauge corpora intitiative (ilci). In *LREC*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 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 Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M Rush. 2018. Opennmt: Neural machine translation toolkit. *arXiv preprint arXiv:1805.11462*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

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)*, pages 296–301.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Peyman Passban, Qun Liu, and Andy Way. 2017. Translating low-resource languages by vocabulary adaptation from close counterparts. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):1–14.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Lijun Wu, Yingce Xia, Li Zhao, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2017. Adversarial neural machine translation. *arXiv preprint arXiv:1704.06933*.

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*, pages 1568–1575.