Exploiting Multilingualism through Multistage Fine-Tuning for Low-Resource Neural Machine Translation

Raj Dabre **Atsushi Fujita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
firstname.lastname@nict.go.jp

Chenhui Chu
Osaka University, Japan
chu@ids.osaka-u.ac.jp

Abstract

This paper highlights the impressive utility of multi-parallel corpora for transfer learning in a one-to-many low-resource neural machine translation (NMT) setting. We report on a systematic comparison of multistage fine-tuning configurations, consisting of (1) pre-training on an external large (209k–440k) parallel corpus for English and a helping target language, (2) mixed pre-training or fine-tuning on a mixture of the external and low-resource (18k) target parallel corpora, and (3) pure fine-tuning on the target parallel corpora. Our experiments confirm that multi-parallel corpora are extremely useful despite their scarcity and content-wise redundancy thus exhibiting the true power of multilingualism. Even when the helping target language is not one of the target languages of our concern, our multistage fine-tuning can give 3–9 BLEU score gains over a simple one-to-one model.

1 Introduction

Encoder-decoder based neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) allows for end-to-end training of a translation model. While NMT is known to perform well for resource-rich language pairs, such as French–English and German–English (Bojar et al., 2018), it performs poorly for resource-poor pairs (Zoph et al., 2016; Riza et al., 2016).

Translating from a resource-poor language into English is typically easier than the other direction, because large parallel corpora between English and other languages can be used for transfer learning (Zoph et al., 2016) and many-to-one modeling.[1] Additionally, large English monolingual data can be back-translated to obtain pseudo-parallel corpora for augmenting the performance of such NMT models.

Translating from English into a resource-poor language is substantially more difficult. Dong et al. (2015) have shown that a one-to-many model trained on middle-sized parallel data (200k sentence pairs) can improve the translation quality over a one-to-one model. However, it is unclear whether this works for much resource-poorer, more distant, and more diverse language pairs. Using pseudo-parallel data is a potential solution, but for most resource-poor languages, the amount of available clean and in-domain monolingual data are limited. It is also unclear what the real reason behind improvements in translation is: increase in the training data or multilingualism.

This paper focuses on (a) improving the performance of NMT for translating English to low-resource languages (b) via exploiting multilingualism (c) through transfer learning based on multistage fine-tuning. The power of multilingualism is verified by using multi-parallel corpora, i.e., the same text in different languages.[2] Unlike previous studies that only apply single-stage fine-tuning for one-to-one translation (Zoph et al., 2016), we systematically compare multistage fine-tuning that exploits one-to-many modeling. We show that our approach can significantly improve the quality of translations from English to seven Asian languages (Bengali, Filipino, Indonesian, Japanese, Khmer, Malay, and Vietnamese) in the Asian Language Treebank (ALT) corpus (Riza et al., 2016),[3] as a result of soft division of labor: training of a strong English encoder, domain adaptation, and tuning of the decoder for each target language.

[1] Note that "many-to-one" does not mean multi-source machine translation (Zoph and Knight, 2016).

[2] Adding a new language does not add any new translation knowledge. Such corpora deserve more attention, since adding a new language to it leads to parallel corpora between the new language and all the other languages in the corpora.

[3] There exist other multi-parallel corpora, such as those for United Nations (Ziemski et al., 2016), Europarl (Koehn, 2005), Ted Talks (Cettolo et al., 2012), ILCI (Jha, 2010), and the Bible (Christodouloupoulos and Steedman, 2015).

2 Related Work

In this paper, we address (a) the importance of multilingualism (b) via one-to-many NMT (c) through robust fine-tuning for transfer learning. Dong et al. (2015) have worked on one-to-many NMT, whereas Firat et al. (2016) and Johnson et al. (2017) have worked on multilingual and multi-way NMT. These studies have focused on training a multilingual model from scratch, but we explore transfer learning for one-to-many NMT.

Fine-tuning based transfer learning has been studied for transferring proper parameters (Zoph et al., 2016; Gu et al., 2018b), lexical (Zoph et al., 2016; Nguyen and Chiang, 2017; Gu et al., 2018a; Lakew et al., 2018), and syntactic (Gu et al., 2018a; Murthy et al., 2018) knowledge from a resource-rich language pair to a resource-poor language pair. On the other hand, Chu et al. (2017) proposed a more robust training approach for domain adaptation, called mixed fine-tuning, which uses a mixture of data from different domains. Imankulova et al. (2019) proposed a multistage fine-tuning method which combines fine-tuning techniques for domain adaptation and back-translation (Sennrich et al., 2016). Unlike us, these approaches do not try to combine multilingualism with multistage fine-tuning using only parallel corpora involving a large number of languages.

3 Multistage Fine-Tuning for NMT

This paper focuses on translation from English (En) to N different languages. In particular, we consider exploiting two types of corpora. One is a small-scale multi-parallel corpus, En-YY_1-$\cdots$-YY_N, consisting of English and N target languages of interest. The other is a relatively larger helping parallel corpus, En-XX, where XX indicates the helping target language which needs not be one of the target languages in the multi-parallel corpus, YY_k $(1 \leq k \leq N)$.

All the fine-tuning techniques that we examine can be generalized as follows.

Pre-training (Pre): An initial NMT model is trained on a parallel corpus for a resource-rich language pair, i.e., En-XX.

Mixed pre-training / fine-tuning (Mix): Training of the NMT model is newly started or continued (Chu et al., 2017) on a mixture of parallel corpora for En-XX and one or more low-resource En-YY pairs.

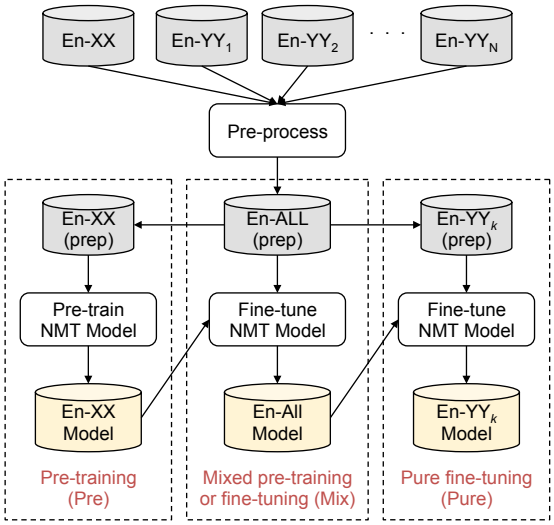

Figure 1: Our multistage fine-tuning for one-to-many NMT. Three dotted sections respectively indicate (a) the pre-training stage, (b) the mixed pre-training / fine-tuning stage, and (c) the pure fine-tuning stage The training data for (a) and (c), i.e., "En-XX (prep)" and "En-YY_k (prep)," are selectively extracted from the pre-processed data labeled "En-ALL (prep)."

Pure fine-tuning (Pure): Training further continues using only a parallel corpus for a particular En-YY pair.

Figure 1 illustrates the training procedure with all of the above three stages. Before starting the training, vocabularies for source and target languages are determined. First, to indicate the target language of each sentence pair, we prepend an artificial token, for instance "2zz" for a target language "zz," onto source sentences (Johnson et al., 2017). All corpora are then concatenated into a single corpus, "En-ALL (prep)," where each of the smaller corpora are enlarged by oversampling so that the size of each corpus matches the largest one. Finally, (sub-word) vocabularies are determined based on "En-ALL (prep)." For simplicity, we keep source and target vocabularies separate, but the target vocabulary is shared by multiple languages. Note that all the NMT models are essentially one-to-many, because they use a shared target vocabulary of XX and one or all of YY_k.

Our training procedure can be viewed as a soft division of labor in NMT training: the first stage learns a strong English encoder by training on a large parallel corpus,[4] whereas the remaining

[4]One may be interested in training an English encoder on a gigantic monolingual corpus. We leave a comparison of the utility of such monolingual corpora and the parallelism in bilingual corpora for our future work.

stages can focus more on training the decoder for the target language(s) of interest and updating the encoder. Whereas we pre-train an encoder for low-resource NMT, so do McCann et al. (2017) to improve several monolingual NLP tasks.

4 Experiments

We conducted a systematic comparison of multistage fine-tuning configurations, specifically focusing on low-resource English-to-Asian language translation.

4.1 Datasets and Pre-processing

As the test-bed, we chose the ALT multilingual multi-parallel corpus[5] (Riza et al., 2016), because it offers the same test sentences in different languages, allowing us to determine whether multilingualism is the true reason behind improved translation quality. We used seven Asian languages[6] as the target languages (i.e., YY): Bengali (Bn), Filipino (Tl), Indonesian (Id), Japanese (Ja), Khmer (Km), Malay (Ms), and Vietnamese (Vi). We randomly[7] split the ALT data into 18,000, 1,000, and 1,106 multi-parallel sentences for training, development, and testing, respectively.

As for the resource-rich En-XX data, we separately used two parallel corpora with two different target languages. One is the IWSLT 2015 English–Chinese corpus (IWSLT En-Zh) (Cettolo et al., 2015), comprising 209,491 parallel sentences. We chose it to examine whether the target language (Chinese) other than the target languages in our setting is helpful. Separately, we also experimented with the Kyoto free translation task English–Japanese corpus (KFTT En-Ja),[8] consisting of Japanese-to-English translations: 440,288 parallel sentences. With this corpus, we could additionally observe how ALT En-Ja can benefit from KFTT En-Ja. Note that the ALT, IWSLT, and KFTT corpora belong to the news, spoken, and Wikipedia domains, respectively.

We tokenized English sentences using the *tokenizer.perl* script in Moses.[9] To tokenize Chinese and Japanese sentences, we used KyotoMorph[10] and JUMAN,[11] respectively. For the other languages in the ALT corpus, we used the data in its raw form, except Khmer for which segmented data were provided by the ALT project.

4.2 Training NMT Models

We compared a total of seven training configurations as shown in Table 1. The model #1 is the simplest baseline model trained only on the En-YY parallel corpus. The models #2 to #4 are three conceivable combinations of fine-tuning stages that exploit an external large parallel corpus for En-XX, showing the limit reachable without a multi-parallel corpus. The models #5 to #7 follow the same training procedure with #2 to #4, respectively, exploiting a small-scale but multi-parallel corpus containing all the seven target languages.

Note that the source side vocabulary is the same across models #2 to #7 and is determined using the English side of the En-XX corpus and the English part of the multi-parallel corpus. On the other hand, the target side vocabulary for models #2 to #4 is different from the vocabulary for models #5 to #7. This is because vocabularies for models #2 to #4 involve the target side of the En-XX corpus and "one" non-English corpus from the multi-parallel corpus. On the other hand, the vocabularies for models #5 to #7 involve the target sides of the En-XX corpus and "all" non-English corpora from the multi-parallel corpus.

We used the open-source implementation of the Transformer model (Vaswani et al., 2017) in the version 1.6 branch of *tensor2tensor*,[12] and used hyper-parameter settings[13] corresponding to *transformer_base_single_gpu* which uses dropout by default. We also used the default sub-word segmentation mechanism of tensor2tensor. The "En-YY" models (#1) used a vocabulary of 8,192 (2^{13}) sub-words due to small corpora sizes, whereas all the other models used a vocabulary of 32,768 (2^{15}) sub-words to account for the larger amount of parallel data in multiple languages. We trained the models for a sufficient number of iterations till

[5] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/

[6] Thai and Burmese (Myanmar) were excluded due to the unavailability of reliable tokenized data. Although some of the YY languages in our experiments, such as Japanese, were relatively resource-rich, we exploited only a small parallel corpus between them and English to highlight the importance of multilingualism.

[7] Our results might not be comparable to those obtained using the official splits mentioned on the website.

[8] http://www.phontron.com/kftt/

[9] https://github.com/moses-smt/mosesdecoder

[10] https://bitbucket.org/msmoshen/kyotomorph-beta

[11] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[12] https://github.com/tensorflow/tensor2tensor

[13] We did not perform any hyper-parameter tuning because we wanted to focus more on simple out-of-the box approaches coupled with multilingualism.

#	XX	N	Model capacity	Training configuration			YY test set						
				Pre	Mix	Pure	Bn	Tl	Id	Ja	Km	Ms	Vi
1.	-	1	1-to-1	-	-	✓	3.99	24.04	24.10	11.03	22.53	29.85	27.39
2.	Zh	1	1-to-2	✓	-	✓	8.86*	27.54*	27.10*	19.07*	28.41*	32.52*	34.63*
3.	Zh	1	1-to-2	-	✓	✓	4.90*	23.07	23.37	13.97*	26.13*	29.24	29.82*
4.	Zh	1	1-to-2	✓	✓	✓	7.99*	26.61*	25.62*	18.39*	27.49*	31.63*	34.22*
5.	Zh	7	1-to-8	✓	-	✓	8.54*	26.88*	26.02*	18.99*	27.07*	32.39*	33.32*
6.	Zh	7	1-to-8	-	✓	✓	9.43*	25.86*	26.33*	19.34*	26.86*	32.39*	33.28*
7.	Zh	7	1-to-8	✓	✓	✓	**10.30**$^{*+\dagger}$	**28.22**$^{*+\dagger}$	**27.24***†	**20.08**$^{*+\dagger}$	**28.66***†	**33.19**$^{*+\dagger}$	**35.34**$^{*+\dagger}$
2.	Ja	1	1-to-2	✓	-	✓	9.16*	28.06*	26.53*	21.55*	27.98*	33.68*	33.93*
3.	Ja	1	1-to-2	-	✓	✓	4.37	22.91	23.37	16.47$*$	23.36$*$	29.28	29.10$*$
4.	Ja	1	1-to-2	✓	✓	✓	8.77*	26.64*	25.88*	21.61*	27.55*	32.45*	34.29*
5.	Ja	7	1-to-8	✓	-	✓	9.43*	27.45*	26.70*	21.79*	27.87*	32.92*	34.28*
6.	Ja	7	1-to-8	-	✓	✓	9.96^{*+}	28.39*	27.22^{*+}	21.03*	28.91^{*+}	33.75*	36.00^{*+}
7.	Ja	7	1-to-8	✓	✓	✓	**10.77**$^{*+\dagger}$	**28.62**$^{*+}$	**28.89**$^{*+\dagger}$	**22.60**$^{*+\dagger}$	**30.03**$^{*+\dagger}$	**34.75**$^{*+\dagger}$	**37.06**$^{*+\dagger}$

Table 1: BLEU scores for all the tested configurations. The "XX" column indicates the resource-rich external (helping) parallel corpus if used, where "Zh" and "Ja" stand for using the IWSLT English–Chinese and the KFTT English–Japanese corpora, respectively. The columns under "YY test set" indicate the the target languages in our multi-parallel corpus, where **bold** marks the best scores for each target language with each external parallel corpus. The "$*$," "$+$," and "$\dagger$" respectively mark the scores significantly (bootstrap re-sampling with $p < 0.05$) better than the 1-to-1 "En-YY" model (#1), the best model without the multi-parallel corpus (#2–#4), and the strongest baselines (#5–#6).

the BLEU score on the development set (simply concatenated unlike training data) did not vary by more than 0.1 BLEU over 10 checkpoints.

Instead of choosing the model with the best development set BLEU, we averaged the last 10 checkpoints saved every after 1,000 updates, following Vaswani et al. (2017), and decoded the test sets with a beam size of 4 and a length penalty, α, of 0.6 consistently across all the models.

4.3 Results

Table 1 gives the BLEU scores (Papineni et al., 2002) for all the configurations. Among the seven configurations, irrespective of the external parallel corpus for En-XX, the three-stage fine-tuned model (#7) achieved the highest BLEU scores for all the seven target languages.

Results for #1 demonstrate that NMT systems trained on 18k parallel sentences can achieve only poor results for Bn and Ja, whereas reasonably high BLEU scores (> 20) are achieved for the other target languages.

Introducing a large external En-XX parallel corpus improved the translation quality consistently and significantly for all the seven target languages,[14] irrespective of the way of its use: simple pre-training (#2), mixed pre-training (#3), and their combination (#4). Among these three training configurations, in most cases, the simple pre-training (#2) attained the best performance. The BLEU gains over #1 were 2.67 (Ms) to 8.04 (Ja) points with the IWSLT En-Zh corpus and 2.43 (Id) to 10.52 (Ja) points with the KFTT En-Ja corpus.

Simple fine-tuning (#2 and #5) gave great improvements over the baseline. However, #5 was worse than #2, despite being trained in the same way, because #5 involves up to eight target languages, inevitably reducing the allowance for each target language. In contrast, exploiting the multi-parallel corpus through mixed pre-training (#6) or mixed fine-tuning (#7) brought consistent improvements over the corresponding 1-to-2 models (#3 and #4). Three-stage fine-tuning (#4 and #7) was always better than two-stage fine-tuning (#3 and #6), because the pre-final model in #4 and #7, obtained using mixed-fine tuning, was superior to and more robust than the one in #3 and #6, obtained using naive multilingual training. These comparisons also justifies the utility of external large parallel corpus for En-XX. #7 was substantially better than #5 simply because the pre-final model in #7 was also trained on multilingual data, unlike the one in #5.

Irrespective of the external parallel corpus for En-XX, the three-stage fine-tuned model (#7) achieved the highest BLEU scores for all the seven target languages, consistently outperforming the other models, with gains over the best scores of 0.14 (Id) to 0.87 (Bn) BLEU points with the

[14] We suspected that the gain is partially owing to the use of the larger vocabularies of these models. However, we only obtained translations of degraded quality with a larger vocabulary size of #1, presumably due to the scarce parallel data for En-YY language pair.

IWSLT En-Zh corpus and 0.23 (Tl) to 1.67 (Id) BLEU points with the KFTT En-Ja corpus. As mentioned in Section 4.2, these models do not introduce any new source sentences compared to the models #2 to #4. Therefore, we can safely conclude that the gain is due to multilingualism.

5 Discussion

In this section, we discuss the effect of the two types of corpora, i.e., external large parallel corpora and small multi-parallel corpora.

5.1 Does the Nature of Corpora Matter?

When using only one En-YY language pair (#2 to #4), only for the three-stage fine-tuned model (#4), the KFTT En-Ja corpus showed slight but consistent superiority (for Tl, Id, Km, and Vi) over the IWSLT En-Zh corpus. In contrast, when all the target languages were exploited (#5–#7), the KFTT En-Ja corpus consistently helped achieve significantly better results than the IWSLT En-Zh corpus. The reason could be three-fold: corpus size, overlap or similarity of helping target language with the target language of interest, and proximity of domain.

En-Ja translation using KFTT En-Ja corpus as a helping data was better by 2.52 BLEU points than when IWSLT En-Zh corpus was used. The difference in BLEU is significant, but it could be because the En-Ja corpus is more than twice the size of the En-Zh corpus. Whereas it is possible that En-Ja translation improves from the overlapping vocabularies of Chinese and Japanese, the performance of other languages also improves despite no vocabulary overlap. Consequently, we believe that translation into a low-resource target language might benefit when the language is also target side of the helping corpus but multistage fine-tuning does not require such overlap and instead shows optimal performance when it leverages multilingualism during stage-wise tuning.

In the future, we will test the above hypotheses thoroughly via some controlled experiments by using, for instance, larger scale multi-parallel corpora (Imamura and Sumita, 2018) and varieties of helping corpora.

5.2 How Does Multilingualism Help?

The ALT corpus is multi-parallel and merely comprises of the same sentences in multiple languages. Although we used seven target languages, we did not introduce new English sentences to the source side. Table 1 shows that some languages are better translated (En-Vi) than others (En-Bn). We speculate that the representations for En-Vi might be better learned than those for En-Bn. When learning all language pairs jointly, the representations of sentences with the same meaning tend to be similar. As such, the representations for En-Bn will be similar to those of En-Vi and hence the former might benefit from the potentially better representations of the latter, leading to improvements in translation quality for the former. This might be how multilingualism is responsible for the improvement in translation quality.

The results also shed light on the value of multi-parallel corpora despite their resource-scarce nature. In a practical situation, a collection of existing multiple bilingual corpora may potentially improve translation quality. Nevertheless, when we create new parallel data, multi-parallel fashion can be a less expensive option, since adding new language leads to parallel corpora between that language and all the other languages in the corpora.

6 Conclusion

We explored the use of small multi-parallel corpora and large helping parallel corpora for one-to-many NMT, which translates English to multiple languages. We proposed multistage fine-tuning and confirmed that it works significantly better than training models via fewer fine-tuning stages despite the resource-scarce and content-redundant nature of the multi-parallel corpora, thereby highlighting the usefulness of multilingualism behind the improvement in translation quality.

In the future, we will explore the utility of multistage fine-tuning for many-to-one and many-to-many NMT. We will also try to explicitly determine the impact of corpora sizes, language similarity, and domain on our approach, and propose further improvements according to our findings.

Acknowledgments

This work was partially conducted under the program “Research and Development of Enhanced Multilingual and Multipurpose Speech Translation Systems” of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, USA.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. In *Proceedings of the Twelfth International Workshop on Spoken Language Translation (IWSLT)*, pages 2–14, Da Nang, Vietnam.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 385–391, Vancouver, Canada.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, USA.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018a. 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)*, pages 344–354.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018b. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631. Association for Computational Linguistics.

Kenji Imamura and Eiichiro Sumita. 2018. Multilingual parallel corpus for Global Communication Plan. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 3453–3458, Miyazaki, Japan.

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139, Dublin, Ireland.

Girish Nath Jha. 2010. The TDIL program and the Indian Language Corpora Initiative (ILCI). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 982–985, Valletta, Malta.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Surafel Melaku Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 54–61, Bruges, Belgium.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Proceedings of the 30th Neural Information Processing Systems Conference (NIPS)*, pages 6294–6305, Long Beach, USA.

V. Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2018. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. *CoRR*, abs/1811.00383.

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, Taipei, Taiwan.

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 (ACL)*, pages 311–318, Philadelphia, USA.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian Language Treebank. In *Proceedings of the 2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)*, pages 1–6, Bali, Indonesia.

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 (ACL), Volume 1: Long Papers*, pages 86–96, Berlin, Germany.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th Neural Information Processing Systems Conference (NIPS)*, pages 3104–3112, Montréal, Canada.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 30th Neural Information Processing Systems Conference (NIPS)*, pages 5998–6008, Long Beach, USA.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 3530–3534, Portorož, Slovenia.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, USA.

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)*, pages 1568–1575, Austin, USA.