Pre-training Synthetic Cross-lingual Decoder for Multilingual Samples Adaptation in E-Commerce Neural Machine Translation

Kamal Kumar Gupta, Soumya Chennabasavaraj,[†] Nikesh Garera,[†] and Asif Ekbal
Department of Computer Science and Engineering,
Indian Institute of Technology Patna, Patna, India
[†]Flipkart, India
kamal.pcs17, asif@iitp.ac.in
[†]soumya.cb, nikesh.garera@flipkart.com

Abstract

Availability of the user reviews in vernacular languages is helpful for the users to get information regarding the products. Since most of the e-commerce websites allow the reviews in English language only, it is important to provide the translated versions of the reviews to the non-English speaking users. Translation of the user reviews from English to vernacular languages is a challenging task, predominantly due to the lack of sufficient in-domain datasets. In this paper, we present a pre-training technique which is used to adapt and improve the single multilingual neural machine translation (NMT) model for the low-resource language pairs. The pre-trained model contains a special synthetic cross-lingual decoder trained over the cross-lingual target samples where the phrases are replaced with their translated counterparts. After pre-training, the model is adapted to multiple samples of the low-resource language pairs using incremental learning. We perform the experiments over eight low-resource and three high resource language pairs from the generic and product review domains. Through our proposed pre-training, we achieve upto 4.35 BLEU improvements compared to the baseline and 2.13 BLEU points compared to the previous code-switched pre-trained models. The review domain outputs are evaluated in human evaluators in the e-commerce company Flipkart.

1 Introduction

Neural machine translation models (Bahdanau et al., 2015; Vaswani et al., 2017) are effective for a specific language pair or domain when trained on a large amount of parallel corpus. It is often difficult to obtain such a large corpus, especially in non-English languages and in specialized domains such as product reviews (Gupta et al., 2021). Currently, in the e-commerce domain, providing the translation of the user reviews in vernacular languages is a need. In a multilingual country like India where English is not a primary language, reviews in local languages will be very helpful for the users as well as e-commerce platforms like Flipkart. As of December 2021, Flipkart leads[1] in the Indian e-commerce market with a market share of 31.9%. In the process of building a one-to-many multilingual translation system to translate the low-resource review domain data on the e-commerce platform Flipkart from English to multiple Indian languages, we propose a synthetic decoder based pre-training approach. To see the impact of the proposed model on translation quality, we perform experiments over the general domain data available publicly. Along with it, we also evaluate our model for review domain data using English-Hindi, English-French and English-Tamil testset.

Recently, pre-training based NMT (Lewis et al., 2019; Devlin et al., 2019) models have attracted attention to improve the translation quality of low as well as high resource language pairs (Yang et al., 2020b; Lin et al., 2020). Pre-training based models first train a parent model over a large dataset and then use the learnt weights to fine-tune for a spe-

[1]https://inc42.com/datalab/amazon-vs-flipkart-who-led-the-indian-//ecommerce-war-in-2021/

cific low-resource language pair or domain (Conneau and Lample, 2019; Song et al., 2019). These approaches have some limitations, e.g. these use some special symbols in the parent models which may not be present in the data during the training of child model. As the samples are taken from the same languages, these approaches fail to capture the cross-lingual information in two languages (Yang et al., 2020b). Fine-tuning also has a limitation that it is not able to remember the information of the parent model's language pairs while training over the child model (new language pair or domain). To resolve this, source side code-switching is used to generate synthetic parallel samples to train the parent model and later use it for fine-tuning over new language pair (Lin et al., 2020; Yang et al., 2020b). These approaches use the parent model's weights to fine-tune for a bi-lingual translation task.

In our work, we perform random phrase substitution at the target side of a parallel sample to capture the shared target context. Our final trained model is a multilingual translation model which can translate the source sentence into multiple languages. Multilingual adaptation helps the incoming pairs to learn from each other because of the shared parameter training. Also, unlike Yang et al., 2020 and Lin et al., 2020 (Yang et al., 2020b; Lin et al., 2020), we use incremental learning instead of fine-tuning where the model can adapt over the incoming input samples from different language pairs without forgetting the information of previously adopted language pairs. Incremental learning allows to update the model even with a small size of available parallel samples without full re-training.

2 Related Work

Pre-training a NMT model and fine-tuning it for specific translation tasks is one of the popular approaches for dealing with the resource-scarce language scenario. Pre-trained language models (LMs) like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) have been used to improve the NMT models (Yang et al., 2020a; Zhu et al., 2020). Edunov et al. (2019) introduced ELMo to the encoder of the NMT model as a pre-trained LM to improve the performance of the NMT model. Weng et al. (2019) used bi-directional self-attention LM in the NMT by weighted-fusion mechanism and knowledge transfer paradigm to improve the learning of encoder and decoder. Zhu et al. (2020) incorporated the representations from the BERT into the encoder and decoder layers of the NMT model. But such large parameters also added extra overhead and delay in the inference time. The recent studies of Yang et al. (2020b) and Lin et al. (2020) used code-switching at source side to train the parent model. The trained parent model is used for fine-tuning over the specific bi-lingual translation direction. Yang et al. (2020b); Lin et al. (2020) trained a multilingual parent model. Unsupervised pre-training has also been popular in several natural language understanding problems, such as word embedding representation (Pennington et al., 2014), pre-trained context representation (Devlin et al., 2019) and sequence-to-sequence pre-training (Song et al., 2019). In this pre-training, scale of data is found to be a very important attribute.

Multilingual NMT (Dong et al., 2015; Johnson et al., 2017; Lu et al., 2018; Rahimi et al., 2019; Tan et al., 2019) is also a useful paradigm where a model trained in a parameter sharing fashion shares the information among the language pairs. In multilingual NMT, low-resource language pairs leverage the information of the high-resource languages. Johnson et al. (2017) added language specific tags before each source sentence in the parallel corpus, merged all the data from multiple language pairs and trained them in a single NMT model. Firat et al. (2016) used shared attention to transfer information between multiple encoder-decoders in a multilingual NMT. Rahimi et al. (2019) performed the training of massively multilingual NMT models. They showed that training a many-to-many multilingual model is helpful in low-resource scenarios. By keeping this in mind, we also use pre-training to improve a multilingual NMT. Unlike Yang et al. (2020b); Lin et al. (2020), we use the pre-trained NMT model to adapt over multilingual NMT using incremental learning instead of bi-lingual pair using fine-tuning.

3 Dataset

We need two types of corpora *i.e.* parallel and monolingual. For the experiments, we include a total of 11 language pairs; out of which 3 belong to the European language pairs, and the rest 8 are low-resource English-Indian language pairs. The data statistics are shown in Table 1. For the

	Parallel	Mono	Dev	Test
En→Fr	15M	224M	2,000	3,000
En→Fr(R)	36,058	224M	2,000	1,020
En→De	4.5M	622M	2,000	3,000
En→Es	3.9M	122M	2,000	3,000
En→Hi	3M	62.9M	1,000	2,390
En→Hi(R)	19,457	62.9M	1,000	2,539
En→Bn	1.7M	3.5M	1,000	2,390
En→Gu	0.51M	7.8M	1,000	2,390
En→Mr	0.78M	9.9M	1,000	2,390
En→Pa	0.52M	6.5M	1,000	2,390
En→Ta	1.4M	20.9M	1,000	2,390
En→Te	0.68M	15.1M	1,000	2,390
En→Ml	1.2M	11.6M	1,000	2,390

Table 1: Size of parallel and monolingual data used for the experiments in million (M). English monolingual corpus size: 495M. Monolingual column in the table shows the size of the corpus for the non-English language in that row. En→Fr(R) and En→Hi(R) are the user review domain datasets.

parallel and monolingual data of English-{French, German} and English-{Spanish}, we use WMT14 (Bojar et al., 2014) and WMT13 (Bojar et al., 2013) corpus, respectively. For evaluation, we use newstest2014 and newstest2013 test sets, respectively. Size of test and development sets are shown in Table 1. Monolingual corpus for English, French and German are taken from the WMT14, and from WMT13 for Spanish. For English-Indian language pairs, we use the parallel data for training, development and testing from WAT21[2]. The monolingual corpus for the Indian languages are taken from the AI4Bharat-IndicNLP Dataset[3]. We also experiment over two product review dataset i.e. English-French (Michel and Neubig, 2018) and English-Hindi (Gupta et al., 2021). Data statistics are shown in Table 1.

4 Methodology

Our proposed approach has four modules: *i.* Training cross-lingual word mapping, *ii.* Generation of synthetic phrase table for source-target phrase pairs, *iii.* Generation of synthetic cross-lingual target samples and training the parent model and *iv.* adapting new input samples from multiple language pairs using incremental learning.

[2]http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/indic_wat_2021.tar.gz
[3]https://github.com/AI4Bharat/indicnlp_corpus

4.1 Word level substitution

Artetxe et al. (2017); Lample et al. (2017) introduced the strategies to learn translation pairs from the lexicons of two monolingual corpora using a shared semantic space. This strategy provides the mapping between the tokens from two languages which can be considered as the translations of each other. Based on the word mapping procedure of Artetxe et al. (2017), we use the unsupervised word mapping based on their embeddings to extract the probabilistic translation lexicons. These translation lexicon pairs are considered as the one-to-many source and target token translations. For example, given independent word embeddings of source and target languages, X_i and Y_j trained on source and target monolingual corpus X and Y, respectively, the unsupervised word mapping exploits self training in third semantic space (Artetxe et al., 2017) or adversarial training in the available semantic space (Conneau et al., 2018) to learn a mapping function f(X) = WX, which provides the source and target word representations in a common embedding space. Here, W is a mapping matrix that is learnt during training. With the word embeddings in the common semantic space, the cosine similarity is measured between the source and target tokens. After that, the probabilistic translation lexicons are selected based on the top-k nearest neighbours in the common embedding space. We can say that for the source language word x_i, its top-k nearest neighbour tokens y_{i1}, y_{i2},..., y_{ik} in the counter target language are extracted as its translation counterparts. The normalized similarities s_{i1}, s_{i2},..., s_{ik} for the word pairs are also given and defined as the translation probabilities. This word mapping is used to randomly replace the target side tokens of one language with another.

4.2 Phrase level substitution

For the phrase substitution, we use the unsupervised phrase table generation technique (Lample et al., 2017). Lample et al. (2017) uses the shared latent semantic space shown in the section above (ref. Section 4.1) and back-translation approach for the unsupervised phrase table generation. Each source and target phrase are considered as a translation candidate and using the shared semantic embedding and back-translation, the translations of the source and target phrase (n-gram sequences) are generated. Each source phrase can be paired

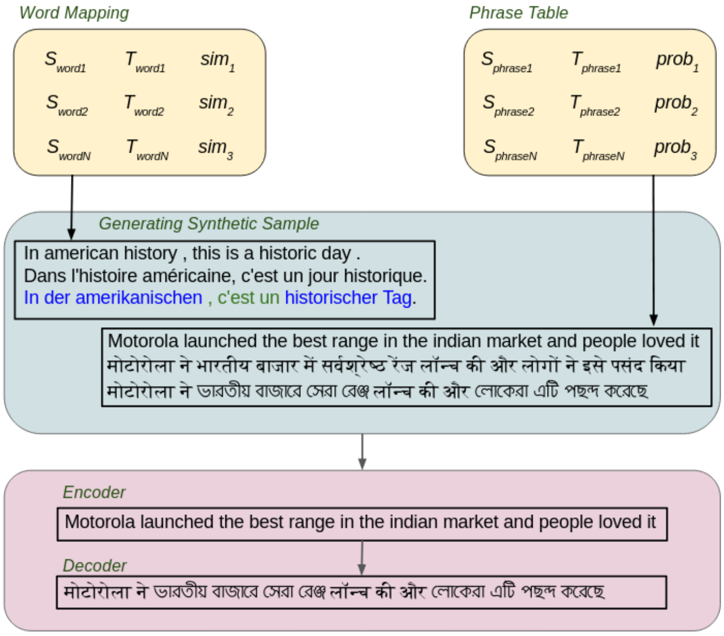

Figure 1: Representation of mapped phrase table, bi-lingual word mapping, target side synthetic sample generation and training of parent model using the synthetic parallel pairs.

with multiple target phrase along with their source-target n-gram probability. The source-target phrase pair having the highest probability is taken as the parallel phrase pair. For the synthetic phrase substitution, the source phrases of length 3 to 5 tokens are considered as the ideal candidates and replaced with their target counterparts. Monolingual sentences (ref. Table 1) are used to generate the phrase table of two languages.

4.3 Training Parent NMT Model with Synthetic Decoder

To train the parent NMT model, we use two methods to generate the synthetic cross-lingual target sequence: using phrase substitution and using word substitution. After following the processes as mentioned in Sections 4.1 and 4.2, we have now a phrase table and bi-lingual word mapping. In the phrase table, each source phrase is aligned with its target phrase pair. In bi-lingual word mapping, we have mapped cross-lingual tokens. Now, we move towards the generation of synthetic parallel pairs for training the multilingual parent NMT model. For each original parallel sample, we randomly mark the target side n-gram sequence (3 to 5 gram) for the substitution. For each such target side phrase, we find the cross-lingual phrase from the phrase table. As shown in Figure 1, an original English-Bengali parallel sample is transformed into a synthetic parallel pair by substituting the Hindi phrase with its counter Bengali phrase. Now, the source is having a monolingual English sentence while the target is a combination of Hindi and Bengali tokens. As shown in Figure 1, an English-French synthetic sample is generated by replacing French phrases with German phrases. Similar to the phrase substitution method, we also use word mapping to substitute tokens instead of phrases. Similarly, we generate such kinds of multiple synthetic samples for other languages (ref. Table 1) too. These synthetic samples are used to train the parent NMT model where the decoder has a cross-lingual sequence knowledge and is capable of capturing the context between the tokens from different languages.

4.4 Adapting Low-Resource Samples through Incremental Learning

After training the parent model using synthetic source and cross-lingual target samples, we use this to adapt over the multiple parallel samples from the low-resource language pairs or domains. We use incremental learning to adapt the parent model over the new samples to obtain a multilingual NMT model which can translate for inputs from the low-resource language pairs. The parent model is updated using the new bi-lingual parallel samples. In order to differentiate the new bilingual parallel samples from the synthetic samples

	Baseline	Proposed (Word)	Proposed (Phrase)	CSP	mRASP
En→Fr	38.24	39.27	40.86	38.80	38.64
En→De	27.38	29.48	30.60	28.90	29.08
En→Es	30.44	32.06	32.74	30.92	31.77
En→Hi	30.42	31.72	32.89	31.08	31.69
En→Bn	12.85	16.45	17.20	14.52	15.61
En→Gu	26.18	29.11	30.09	27.73	28.60
En→Mr	24.08	25.13	26.02	24.13	24.82
En→Pa	25.93	27.86	28.52	26.68	27.34
En→Ta	17.96	19.82	20.77	18.96	19.51
En→Te	16.08	19.14	20.51	17.93	18.38
En→Ml	16.71	18.63	19.50	17.54	18.04
En→Fr(R)	20.72	22.41	22.79	21.16	21.73
En→Hi(R)	34.36	35.84	36.27	34.82	35.38

Table 2: Performance of the proposed models in terms of BLEU score. En→Fr(R) (Michel and Neubig, 2018) and En→Hi(R) (Gupta et al., 2021) are user review domain datasets.

already used, we include language specific tags before each source sentence (Johnson et al., 2017). For example, for English-Spanish, English-Hindi and English-Bengali pairs, we use ES, HI and BN tags. Similarly, we use unique tags for all the language pairs. Instead of fine-tuning, incremental learning allows the model to learn the new input samples without losing the knowledge of previous samples.

5 Experimental Setting

Parent model is trained using the Transformer (Vaswani et al., 2017) based encoder-decoder NMT model. Our training setup is described as follows: the tokens of training, evaluation and validation sets are segmented into the subword units using the BPE technique (Gage, 1994) proposed by (Sennrich et al., 2016). We perform 30,000 and 10,000 join operations for high and low-resource languages, respectively. We learn the BPE vocabulary using the monolingual data and apply it over the parallel samples. We use 6 layers at encoder and decoder sides each, 8-head attention, hidden layer of size 512, embedding vector of size 512, learning rate of 0.0002, and the minimum batch size of 3800 tokens. The evaluation results are based on the BLEU metric (Papineni et al., 2002).

The new samples from the low-resource child pairs are tokenized and true-cased. Here, we also apply the subword operation using the learned vocabulary from the monolingual data as mentioned above. Here, before adding the new parallel samples to the parent models using incremental learning, we add language specific tags before the source sentence of each parallel sample. Adding a tag before the sample (Johnson et al., 2017) is for differentiating between parent samples and new incoming samples as well as highlighting the difference between the parallel samples from different languages too. For example, we append ##HI before source sentence of each English-Hindi parallel sample. Similar to this, we use the tags like ##FR, ##DE, ##ES, ##BN, ##GU, ##MR, ##BN, ##GU, ##MR, ##PA, ##ML, ##TA and ##TE for French, German, Spanish, Bengali, Gujarati, Marathi, Punjabi, Malayalam, Tamil and Telugu languages, respectively.

	Baseline	100%	30%	50%
En–Fr	38.24	40.86	38.82	39.65
En–De	27.38	30.60	28.81	29.02
En-Es	30.44	32.74	30.62	31.15
En-Hi	30.42	32.89	30.78	31.64
En-Ta	17.96	20.77	18.84	19.91
En-Bn	12.85	17.20	14.29	16.26

Table 3: Performance of the proposed models in terms of BLEU score when the parent model is trained on fractions of synthetic parallel data.

6 Results and Analysis

We evaluate our proposed models and compare with the multilingual models for Indian languages as the baseline. We also compare our method with existing two pre-trained models, i.e. CSP (Yang et al., 2020b) and mRASP (Lin et al., 2020). For the low-resource Indian languages, we fine tune

CSP and mRASP models for multilingual child model. For the experiments over Indian languages using WAT21 dataset, we augmented it with European languages dataset. We report the evaluation results of both word based and phrase based models. From Table 2, we can see that both the models *i.e.* word and phrase based outperforms the respective multilingual models. Pre-trained models using phrase substitution perform significantly better than the word based models. By comparing CSP and mRASP, we can observe that both the versions of the proposed model significantly outperform the CSP and mRASP. The behaviour is consistent for the high-resource as well as low-resource language pairs. It is seen that the cross-lingual context captured by the proposed decoder helps the adapted low-resource pairs that result in statistically significant (Koehn, 2004) ($p \leq 0.05$) and consistent improvements over the multilingual models, CSP and mRASP.

To see the impact of synthetic data used to train the parent model, we also perform the experiments by training the parent model over multiple fractions of synthetic data samples. We split the parent data in 30%, 50% and 100% of total size. In Table 3, we can see that the BLEU scores for En→Fr, En→De and En→Es are reported with the parent model trained over different sizes of dataset. We can see that performance of the NMT model improves consistently as the data to train the parent model increases.

6.1 Human Evaluation

The proposed model is evaluated at Flipkart (https://www.flipkart.com/) with the help of the real time human evaluators. The models for Hindi and Tamil are evaluated with the help of English–to–Hindi (Gupta et al., 2021) and English–to–Tamil testset from the review domain. The English-Tamil testset is an in-house testset of Flipkart. For evaluation, 1,000 output samples are taken and tagged with three labels i.e *Good*, *Can be better* and *Bad*. The labels are assigned to the output samples based on their semantic and syntactic accuracy. During the evaluation, while assigning the labels to the output samples, 'tense preservation', 'syntax of output sentence', 'choice of in-domain output tokens' are some important factors which are kept in mind. Table 4 shows the results for quality evaluation. We can see that the proposed model significantly reduces the outputs from *Can be better* and *Bad* category, and increases the *Good* label output sentences.

	Good	**Can be better**	**Bad**
En–Ta (base)	45.6%	28.1%	26.3%
En–Ta (phrase)	60.4%	24.7%	14.9%
En-Hi (base)	52.6%	21.7%	25.7%
En-Hi (phrase)	64.0%	26.3%	9.7%

Table 4: Real time quality evaluation between baseline and proposed phrase based pre-training models.

7 Conclusion

In this paper, we have devised a pre-training based learning where the parent model is trained on the source and cross-lingual target samples. We pre-train a one-to-many multilingual parent model with synthetic decoder and use incremental learning to adapt over new incoming bi-lingual parallel samples from multiple language pairs. Our objective to train such a pre-training model is to capture a cross-lingual context at the target side and use it to adapt the new multilingual parallel samples from the low-resource language pairs.
We have performed experiments over 8 low-resource and 3 high-resource language pairs. We also perform experiments over two product review domain datasets from English-French and English-Hindi language pairs. Through our synthetic multilingual decoder based pre-training, we achieve upto 3.22 and 4.35 BLEU points improvements for high and low-resource language pairs, respectively over the baseline.
From the perspective of the e-commerce platforms, our proposed parent model is able to adapt new samples for multiple language pairs and provide us a single translation model which can translate the English sentence into multiple languages. The proposed model is evaluated by real time evaluators at Flipkart for English–to–Tamil and English–to-Hindi review domain testsets. The human evaluation results show the increment of upto 6% output samples with the *Good* label.

In the future, we aim to utilize language relatedness in the multilingual setting. We believe that language relatedness in terms of vocabulary overlap, syntax sharing and subword learning can help to improve the translation quality in a multilingual model.

Acknowledgement

Authors gratefully acknowledge the unrestricted research grant received from the Flipkart Internet Private Limited to carry out the research. Authors thank Muthusamy Chelliah for his continuous feedback and suggestions to improve the quality of work.

References

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR 2015)*.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT 2013)*, pages 1–44.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word Translation Without Parallel Data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (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. Association for Computational Linguistics.

Edunov, S., Baevski, A., and Auli, M. (2019). Pre-trained language model representations for language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4052–4059, Minneapolis, Minnesota. Association for Computational Linguistics.

Firat, O., Sankaran, B., Al-Onaizan, Y., Vural, F. T. Y., and Cho, K. (2016). Zero-resource translation with multi-lingual neural machine translation. *arXiv preprint arXiv:1606.04164*.

Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Gupta, K., Chennabasavaraj, S., Garera, N., and Ekbal, A. (2021). Product review translation using phrase replacement and attention guided noise augmentation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 243–255, Virtual. Association for Machine Translation in the Americas.

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

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lin, Z., Pan, X., Wang, M., Qiu, X., Feng, J., Zhou, H., and Li, L. (2020). Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., and Sun, J. (2018). A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Michel, P. and Neubig, G. (2018). MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rahimi, A., Li, Y., and Cohn, T. (2019). Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). 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.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Tan, X., Ren, Y., He, D., Qin, T., and Liu, T.-Y. (2019). Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Weng, R., Yu, H., Huang, S., Luo, W., and Chen, J. (2019). Improving neural machine translation with pre-trained representation. *arXiv preprint arXiv:1908.07688*.

Yang, J., Wang, M., Zhou, H., Zhao, C., Zhang, W., Yu, Y., and Li, L. (2020a). Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9378–9385.

Yang, Z., Hu, B., Han, A., Huang, S., and Ju, Q. (2020b). CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online. Association for Computational Linguistics.

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T. (2020). Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.