FeatureBART: Feature Based Sequence-to-Sequence Pre-Training for Low-Resource NMT

Abhisek Chakrabarty[*], **Raj Dabre**[*], **Chenchen Ding**,
Hideki Tanaka, **Masao Utiyama** and **Eiichiro Sumita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{abhisek.chakra,raj.dabre,chenchen.ding,hideki.tanaka,
mutiyama,eiichiro.sumita}@nict.go.jp

Abstract

In this paper we present FeatureBART, a linguistically motivated sequence-to-sequence monolingual pre-training strategy in which syntactic features such as lemma, part-of-speech and dependency labels are incorporated into the span prediction based pre-training framework (BART). These automatically extracted features are incorporated via approaches such as concatenation and relevance mechanisms, among which the latter is known to be better than the former. When used for low-resource NMT as a downstream task, we show that these feature based models give large improvements in bilingual settings and modest ones in multilingual settings over their counterparts that do not use features.

1 Introduction

Sequence-to-sequence (S2S) pre-training done via denoising objectives on monolingual corpora is known to improve generation quality in low-resource settings (Lewis et al., 2020). This has been extensively explored for neural machine translation, however most works show that the improvements for translation into English are more pronounced than those for translation into a non-English language (Liu et al., 2020; Tang et al., 2020). One reason for this is that pre-training leads to a strong decoder that learns to de-noise masked inputs, whereas the knowledge retained in the encoder is rather limited. Thus far, there has been no explicit effort towards improving the contribution of the encoder during pre-training.

Most pre-training methods rely on the power of large models and large corpora, but ignore the possibility of incorporating linguistic knowledge into the pre-trained model. On the other hand, there are several works which show that incorporating linguistic knowledge in the form of lemma, part-of-speech tags and dependency labels, lead to a significant improvement in translation quality (Sennrich and Haddow, 2016; Hoang et al., 2016; Li et al., 2018; Pan et al., 2020; Chakrabarty et al., 2020) in both low- and high-resource settings. Most recently, in an extremely low-resource setting, Chakrabarty et al. (2020) show the effectiveness of using the aforementioned linguistic features, especially when their influence on the model is controlled via relevance mechanisms that appropriately scale feature embeddings before they augment word embeddings. We hypothesize that incorporating linguistic features, into the denoising based pre-training framework, should improve the quality of pre-training, which should then have a positive impact on the translation quality via fine-tuning. To this end, we propose FeatureBART, a feature based sequence-to-sequence pre-training method.

In FeatureBART, linguistic features are obtained via automatic annotators and then converted into embeddings, which are used to augment the word embeddings of the encoder. Feature embeddings, are incorporated by either naive concatenation with the word embeddings or by first weighing them with a relevance mechanism and then adding them to word embeddings. The model itself is trained using a monolingual corpus via either the text-infilling or the mask prediction approaches, the former used in BART and the latter used in BERT. This FeatureBART model is then fine-tuned for low-resource language pairs, where linguistic features are also used during fine-tuning. Experiments on English to 8 Asian languages from the Asian Language Treebank (ALT) dataset (Riza et al., 2016), using bilingual as well as multilingual fine-tuning, show that our feature based pre-training and fine-tuning leads to significant improvements in translation quality indicating the complementary nature of denoising pre-training and features. Analyses of training curves show that when compared to non-feature based pre-training, feature based pre-

[*] Equal contribution.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5014–5020
October 12–17, 2022.

training leads to significantly lower perplexities during the initial stages of fine-tuning.

2 Methodology

We first give some background knowledge about feature based NMT modeling, followed by an explanation of FeatureBART.

2.1 Background: Use of Source Side Morphological Features into NMT

Sennrich and Haddow (2016) proposed the concatenation of embeddings of features of a token (word or sub-word) to the token embedding. In case a word is split into sub-words, the feature is duplicated for each sub-word. For K features of a source token denoted by $s_i = (s_{i1}, \ldots, s_{iK})$, let V_k, E_k, and d_k denote the vocabulary, embedding matrix and dimension of the k^{th} feature. So, $E_k \in \mathbb{R}^{d_k \times |V_k|}$, s_{i1} is the word or sub-word feature and $s_{i2}, \ldots, s_{iK}$ are the linguistic features. The embedding of s_i, say e_i, is formulated as $e_{ik} = E_k s_{ik}$, and $e_i = \|_{k=1}^{K} e_{ik}$. e_{ik} is the vector embedding of s_{ik} where $\|$ is the concatenation operation.

2.2 Relevance of Features

Chakrabarty et al. (2020) proved the effectiveness of the following two feature weighting strategies to be applied to features prior to concatenation:

Self-Relevance: The relevance of a feature embedding is evaluated w.r.t itself. For $k \in \{1, \ldots, K\}$, the self relevance is calculated as $mask_{ik} = sigmoid(W_k e_{ik})$, and then $e'_{ik} = mask_{ik} \odot e_{ik}$. $W_k \in R^{d_k \times d_k}$ is the learnable weight matrix for the k^{th} feature, and $\odot$ is the element-wise multiplication operation. The vector $mask_{ik}$ signifies the self relevance of e_{ik} and is multiplied element-wise with e_{ik} to produce the modified feature embedding e'_{ik}. Finally, $e'_{i1}, \ldots, e'_{iK}$ are concatenated to make the final embedding e'_i for the source token s_i. Thus, $e'_i = \|_{k=1}^{K} e'_{ik}$.

Word-Relevance: It uses the word/sub-word embedding (e_{i1}) to determine the relevance of the remaining feature embeddings. Formally, for $k \in \{2, 3, \ldots, K\}$, $mask_{ik} = sigmoid(W_k(e_{i1} \| e_{ik}))$, and then $e'_{ik} = mask_{ik} \odot e_{ik}$. $W_k \in R^{d_k \times (d_1 + d_k)}$ is the learnable weight matrix and the final embedding e'_i is obtained by concatenating $e'_{i2}, \ldots, e'_{iK}$ with e_{i1}.

2.3 FeatureBART

FeatureBART, is a feature based encoder-decoder pre-trained model trained using linguistic features which are used to augment the word embeddings of the encoder. First, a large monolingual corpus is annotated via morphological and syntactic annotators to obtain different types of features for each token (word or sub-word). This feature annotated monolingual corpus is used for self-supervised training where during training, some tokens (token masking) or token spans (text infilling) in a sentence are replaced with the "MASK" token just like in Lewis et al. (2020). We use dummy mask features for masked tokens or spans. The sentence containing masked content is fed to the encoder and the model is trained so that the decoder can predict the original sentence. We hypothesize that features help in better pre-training as they provide the model with additional information for denoising.

3 Experiments

We describe the datasets for pre-training and fine-tuning, features used and model training details.

Datasets: For pre-training of our BART and FeatureBART models, we choose the English monolingual News Crawl articles[1] of 2007, 2010, and 2013 from WMT-2017 with varying sizes that contain 3.8, 6.8, and 21.7 million sentences respectively. For fine-tuning, we experiment with the multilingual, multi-parallel Asian language treebank (ALT) (Riza et al., 2016)[2] for English to Asian language translation. Following (Chakrabarty et al., 2020), eight Asian languages - Bengali (bg), Filipino (fi), Hindi (hi), Indonesian (id), Khmer (khm), Malay (ms), Myanmar (my) and Vietnamese (vi) are set as the targets. So, bilingual experiments cover eight language pairs from en-bg to en-vi. For our multilingual experiments, we explore one-to-many multilingual translation setup keeping the source side fixed to English (en) and the target side to the eight Asian languages mentioned above. The source side is fixed to English throughout because as the initial attempt on feature-based pre-training, we can rely on high quality automatic morphological analyzers available for English. Given the potential of this work established by empirical results, other languages with moderate quality morphological annotation can also be tried in the future. We use the official train/dev/test splits containing 18088, 1000, and 1018 sentences, respectively.

[1] https://statmt.org/wmt17/translation-task.html

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

Bilingual Results											
Pre-training	Noise	Config	en-bg	en-fi	en-hi	en-id	en-khm	en-ms	en-my	en-vi	Avg.
None	-	Base	7.5	26.98	23.62	30.88	26.24	35.78	16.48	29.05	24.57
	-	+Self-Rel	8.4†	28.22†	26.13†	32.65†	27.33†	37.22†	18.13†	29.91†	26
NC-07	I	Base	9.77	31.28	25.63	33.86	29.0	38.59	18.83	32.76	27.46
	I	+Self-Rel	10	31.53	26.93†	34.47†	29.02	39.01	19.21	33.5†	27.96
NC-10	I	Base	9.78	32.07	26.21	35.25	27.5	38.79	19.5	32.32	27.68
	I	+Word-Rel	9.7	32.57†	26.44	**36.07**†	29.4†	40.34†	18.98	34.78†	28.53
NC-13	I	Base	**10.09**	32.62	26.3	35.56	29.5	40.09	19.72	33.95	28.48
	I	+Concat	9.75	**33.12**†	**26.96**†	35.32	**30.21**†	**40.64**†	**20.01**	**35.26**†	28.91
Multilingual Results											
Pre-training	Noise	Config	en-bg	en-fi	en-hi	en-id	en-khm	en-ms	en-my	en-vi	Avg.
None	-	Base	11.55	31.04	27.29	34.78	30.27	39.37	20.93	34.58	28.73
	-	+Self-Rel	11.40	31.14	**27.94**†	34.42	30.09	39.84†	20.99	33.85	28.71
NC-07	M	Base	11.78	31.90	27.03	35.77	30.47	40.39	20.65	34.94	29.12
	M	+Self-Rel	11.51	32.17	27.68†	36.11	30.94†	40.80	21.13†	35.28	29.45
NC-10	I	Base	11.84	32.66	26.77	35.71	30.72	40.11	20.65	34.56	29.13
	I	+Self-Rel	11.54	32.57	27.76†	36.25†	**31.16**†	40.48	**21.26**†	35.44†	29.58
NC-13	I	Base	**11.98**	**32.67**	26.69	36.11	30.41	40.20	20.60	35.24	29.24
	I	+Self-Rel	11.65	32.38	27.77†	**36.39**	30.98†	**41.35**†	20.95†	**35.75**†	29.65

Table 1: BLEU scores of the bilingual and multilingual models. For a given pre-training corpus size, we only show the results of the best feature and pre-training configuration due to lack of space. Highest scores are BOLD. † marks scores significantly better ($p < 0.05$) than the corresponding non-feature (base) counterparts. The "Noise" column indicates the pre-training approach, "I" and "M" for text-infilling and token masking. Self-Rel, Word-Rel, and Concat denote self-relevance, word-relevance, and concatenation of feature embedding configurations respectively.

Pre-Processing: The monolingual corpora for pre-training and the source side of fine-tuning corpora are tokenized and true-cased by Moses tokenizer (Papineni et al., 2002) and the target languages are tokenized to separate the delimiters and the punctuation symbols. Following Johnson et al. (2017), each source language sentence in multilingual translation setup is appended with a token like <*tgt-id*> which indicates the target language. Byte-pair-encoding (BPE) (Sennrich et al., 2016) is performed to obtain subword vocabularies. We train a single BPE model of vocabulary size $32K$ on the combined training corpora of all 9 languages and use it during pre-training as well as fine-tuning.

Features Used: Morphological annotation of the English datasets is done using Stanford CoreNLP toolkit (Manning et al., 2014). There are three word-level linguistic features - lemma, part-of-speech (POS), and dependency labels. All subwords of a word take the features of that word. We use subword tags (Sennrich et al., 2016) to denote beginning, middle and ending of a subword unit.

Hyperparameters and Training Details: We use variations of the Transformer-base model (Vaswani et al., 2017) for our experiments available from OpenNMT PyTorch (Klein et al., 2017), which we modify for feature based experiments. Wherever possible, we perform hyperparameter tuning of layers, hidden sizes, number of attention heads, dropouts, number of training epochs etc. (See A for details). All training is done on a single 32 GB V-100 GPU. Pre-training is done for 3 epochs for each monolingual corpora. During fine-tuning, validation is done after every 10000 steps and training stops if validation accuracy does not improve for consecutive 5 evaluations. Test set decoding is done using beam search with a beam size of 5 and length penalty of 1.0. Translation performance is measured by BLEU score calculated using *multi-bleu.perl*.

4 Results

Table 1 contains our results for bilingual and multilingual fine-tuning of FeatureBART along with their non-feature counterparts. Bilingual and multilingual results are divided into 4 groups: no pre-training, pre-training using News Crawl articles of 2007 (NC-07), 2010 (NC-10), and 2013 (NC-13) from WMT-2017. We analyze the results as follows:

Bilingual vs. Multilingual Translation without and with Pre-Training: Comparing corresponding rows between the bilingual and multilingual blocks, shows that multilingual models are significantly better than bilingual ones, observations which are in accordance with Arivazhagan et al. (2019); Johnson et al. (2017); Dabre et al. (2020); Zhang et al. (2020). Without

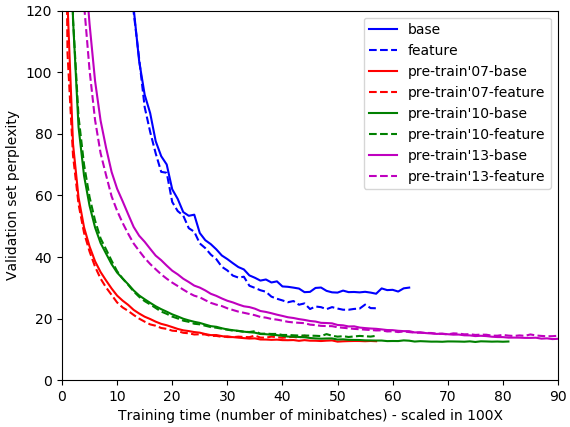

Figure 1: Perplexity plots for English–Bengali models.

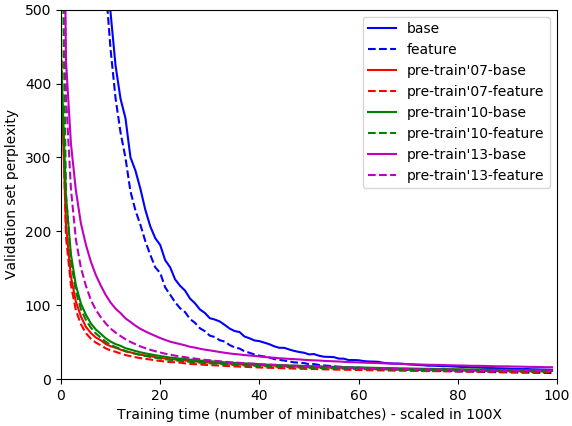

Figure 2: Perplexity plots for multilingual models.

pre-training, improvements in maximum BLEU points comparing bilingual vs. multilingual setups are $(3.15, 2.92, 1.81, 2.13, 2.94, 2.62, 2.86, 4.67)$ for (en $\rightarrow$ bg, fil, hi, id, khm, ms, my, vi) respectively. Pre-training, although it improves performance in both cases, naturally reduces the gap between bilingual and multilingual models. Nevertheless, we get maximum improvements of $(1.89, 0.98, 0.95, 0.71, 1.25, 0.49)$ BLEU points for (en $\rightarrow$ bg, hi, khm, ms, my, vi) when comparing bilingual and multilingual models.

Impact of Features: From Table 1 the following observations can be made: (1) In case of no pre-training, features are very useful in bilingual settings but not in multilingual settings as multilingual systems can utilize a multi-parallel corpora efficiently by acquiring supplementary knowledge from other languages, thus making linguistic information redundant. (2) However, when pre-training is used, multilingual models tend to benefit more from features. As an example, consider en-khm and en-vi multilingual scores in Table 1. Under no pre-training setup, adding features deteriorates the scores, whereas feature based pre-training gives consistent improvements, indicating that feature based modeling and pre-training are complementary.

Optimal Pre-Training and Feature Configurations: From Table 1, looking at the "Noise" column, it is clear that text infilling is the best pre-training objective in most cases. With regard to feature incorporation mechanism, self-relevance predominantly gives the best results. We therefore recommend the use of self-relevance and text-infilling based FeatureBART in low-resource settings.

Studying Model Perplexities: Figure 1 shows the perplexities[3] of en-bg bilingual models. The perplexities with pre-training and features are lower compared to when features are not used, during initial stages of training, but this changes towards convergence. This explains why en-bg does not show performance improvements from features. Figure 2 contains plots of cumulative perplexity of all language pairs for multilingual models. Here, features seem to have larger impact than in bilingual settings, both at the beginning and later stages of training. However, BLEU gains are not always observed, indicating that it may not always be reliable motivating future multi-metric and human evaluation (Marie et al., 2021).

5 Related Work

Pre-Training: Pre-trained models reduce the need for large fine-tuning data for a given downstream task, and in this paper we focus on extremely low-resource settings. In this context, T5 (Raffel et al., 2020), mT5 (Xue et al., 2021), BART (Lewis et al., 2020), mBART-25 (Liu et al., 2020), mBART-50 (Tang et al., 2020) and, most recently, IndicBART (Dabre et al., 2021) are most commonly used for fine-tuning. None of these works focus on linguistic features, which is the key focus of our paper.

Feature Based NMT: Most works focusing on linguistic features, experiment on low-resource settings, and a majority of them focus on how to exploit syntactic/dependency structures of the source language (Eriguchi et al., 2016; Shi et al., 2016; Chen et al., 2017; Li et al., 2017; Wu et al., 2018; Zhang et al., 2019; Bugliarello and Okazaki, 2020). These works rely on various sophisticated

[3] Note that, adding features tends to make the training curves smoother, especially in bilingual settings which are comparatively lower resource than multilingual settings.

approaches but, Sennrich and Haddow (2016) show that enriching encoder word embeddings with morphological features is a simple but nice technique to exploit the features. Chakrabarty et al. (2020) improve upon this further by relevance mechanisms on top of morphological and syntactic feature embeddings, to enable effective use of features, an insight we adopt in this paper for feature based pre-training.

Multilingual NMT: Where feature based NMT focuses on utilizing linguistic information, multilingual NMT (Johnson et al., 2017; Dabre et al., 2020) focuses on leveraging training data for other languages to improve translation quality. Incorporating linguistic features into multilingual models has been neglected to the best of our knowledge, and our work aims to fill in this gap.

6 Conclusion

We have presented FeatureBART, an encoder-decoder pre-trained model that augments the encoder's embeddings with linguistic feature embeddings. Our experiments on English to Asian language translation in an extremely low-resource setting show that FeatureBART leads to better translation quality compares to its counterpart that does not use features. Analyses of training curves reveal that compared to pre-training without features, feature based pre-training leads to significantly lower perplexities during the initial stages of fine-tuning, which we think is responsible for improvement in translation quality. Future work will focus on: (1) exploring the impact of individual feature category on feature based pre-training and (2) multilingual version of FeatureBART which uses features in the encoder for languages other than English.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

Emanuele Bugliarello and Naoaki Okazaki. 2020. Enhancing machine translation with dependency-aware self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1618–1627, Online. Association for Computational Linguistics.

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2020. Improving low-resource NMT through relevance based linguistic features incorporation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4263–4274, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Huadong Chen, Shujian Huang, David Chiang, and Jiajun Chen. 2017. Improved neural machine translation with a syntax-aware encoder and decoder. *CoRR*, abs/1707.05436.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for natural language generation of indic languages. *CoRR*, abs/2109.02903.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 823–833, Berlin, Germany. Association for Computational Linguistics.

Cong Duy Vu Hoang, Reza Haffari, and Trevor Cohn. 2016. Improving neural translation models with linguistic factors. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 7–14, Melbourne, Australia.

Melvin Johnson, Mike Schuster, Quoc V. 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.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junhui Li, Deyi Xiong, Zhaopeng Tu, Muhua Zhu, Min Zhang, and Guodong Zhou. 2017. Modeling source syntax for neural machine translation. *CoRR*, abs/1705.01020.

Qiang Li, Derek F. Wong, Lidia S. Chao, Muhua Zhu, Tong Xiao, Jingbo Zhu, and Min Zhang. 2018. Linguistic knowledge-aware neural machine translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(12):2341–2354.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Dual-source transformer model for neural machine translation with linguistic knowledge.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Rapid Sun, Vichet Chea, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian language treebank. In *Proc. of O-COCOSDA*, pages 1–6.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

S. Wu, D. Zhang, Z. Zhang, N. Yang, M. Li, and M. Zhou. 2018. Dependency-to-dependency neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2132–2141.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Meishan Zhang, Zhenghua Li, Guohong Fu, and Min Zhang. 2019. Syntax-enhanced neural machine translation with syntax-aware word representations. 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 1151–1161, Minneapolis, Minnesota. Association for Computational Linguistics.

A Hyperparameters and Training Details

We perform extensive hyperparameter tuning to determine optimial settings for pre-training and

fine-tuning. We use 6 encoder and decoder layers, 8 multi-attention heads, $2,048$ dimension size of fully-connected network. For models without features, the token embedding and model hidden dimension is set to 512. For feature based models, following Sennrich et al. (2016) the embedding and hidden dimensions are 536 $(250, 250, 15, 15, 6$ corresponding to subword, lemma, POS, dependency label, and subword-tag), in order to make the number of parameters comparable. We use batch-sizes of $4,096$ tokens. The dropout rate is set to 0.3 for fine tuning experiments.

We pre-train models using token masking as well as text infilling, where span lengths are drawn from a Poisson distribution ($\lambda = 3$). We investigate the effect of different noise percentage from 10% to 80% and find that $50\% - 60\%$ noising is the optimum to get the best performance in downstream translation task. Pre-training is done for 3 epochs for each monolingual corpora of 2007, 2010, and 2013, covering 43200, 75600, and 237600 training steps respectively. During fine-tuning, maximum training steps are set as 200000 with validation accuracy performed after every 10000 steps.