BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese

Nguyen Luong Tran, Duong Minh Le, Dat Quoc Nguyen
VinAI Research, Hanoi, Vietnam
{v.nguyentl12, v.duonglm1, v.datnq9}@vinai.io

Abstract

In this paper, we present BARTpho with two versions BARTpho$_{\text{syllable}}$ and BARTpho$_{\text{word}}$, which are the first public large-scale monolingual sequence-to-sequence models pre-trained for Vietnamese. BARTpho uses the "large" architecture and the pre-training scheme of the sequence-to-sequence denoising autoencoder BART, thus especially suitable for generative NLP tasks. We conduct experiments to compare our BARTpho with its competitor mBART on a downstream task of Vietnamese text summarization and show that: in both automatic and human evaluations, BARTpho outperforms the strong baseline mBART and improves the state-of-the-art. We release BARTpho to facilitate future research and applications of generative Vietnamese NLP tasks. Our BARTpho models are publicly available at: `https://github.com/VinAIResearch/BARTpho`.

1 Introduction

The masked language model BERT (Devlin et al., 2019) and its variants, pre-trained on large-scale corpora, help improve the state-of-the-art (SOTA) performances of various natural language understanding (NLU) tasks. However, due to a bidirectionality nature, it might be difficult to directly apply those pre-trained language models to natural language generation tasks (Wang and Cho, 2019). Therefore, pre-trained sequence-to-sequence (seq2seq) models are proposed to handle this issue (Dong et al., 2019; Lewis et al., 2020; Zhang et al., 2020; Raffel et al., 2020; Qi et al., 2020; Xue et al., 2021a). The success of these pre-trained seq2seq models has largely been limited to the English language. From a societal, cultural, linguistic, cognitive and machine learning perspective (Ruder, 2020), it is worth investigating pre-trained seq2seq models for languages other than English. For other languages, one could employ existing pre-trained multilingual seq2seq models (Liu et al., 2020; Xue et al., 2021b; Qi et al., 2021) or retrain language-specific models using the proposed seq2seq architectures (Eddine et al., 2020; Shao et al., 2021). Note that retraining a language-specific model might be preferable as dedicated language-specific models still outperform multilingual ones (Nguyen and Nguyen, 2020).

Regarding Vietnamese, to the best of our knowledge, there is not an existing monolingual seq2seq model pre-trained for Vietnamese. In addition, another concern is that all publicly available pre-trained multilingual seq2seq models are not aware of the linguistic characteristic difference between Vietnamese syllables and word tokens. This comes from the fact that when written in Vietnamese, in addition to marking word boundaries, the white space is also used to separate syllables that constitute words.[1] For example, 7-syllable written text "chúng tôi là những nghiên cứu viên"$_{\text{we are researchers}}$ forms 4-word text "chúng_tôi$_{\text{we}}$ là$_{\text{are}}$ những nghiên_cứu_viên$_{\text{reseacher}}$". Without applying a Vietnamese word segmenter, those pre-trained multilingual seq2seq models directly apply Byte-Pair encoding models (Sennrich et al., 2016; Kudo and Richardson, 2018) to the syllable-level Vietnamese pre-training data. Therefore, it is interesting to investigate the influence of word segmentation on seq2seq pre-training for Vietnamese.

In this paper, we introduce BARTpho with two versions—BARTpho$_{\text{syllable}}$ and BARTpho$_{\text{word}}$—the first large-scale monolingual seq2seq models pre-trained for Vietnamese, which are based on the seq2seq denoising autoencoder BART (Lewis et al., 2020). The difference between our two BARTpho versions is that they take different types of input texts: a syllable level for BARTpho$_{\text{syllable}}$ vs. a word level for BARTpho$_{\text{word}}$. We compare BARTpho with mBART (Liu et al., 2020)—a multilingual variant of BART—on a downstream task

[1] Note that 85% of Vietnamese word types are composed of at least two syllables (Thang et al., 2008).

of Vietnamese text summarization. We find that our BARTpho models outperform mBART in both automatic and human evaluations, and help produce a new SOTA performance, thus showing the effectiveness of large-scale monolingual seq2seq pre-training for Vietnamese. We also find that BARTpho$_{\text{word}}$ does better than BARTpho$_{\text{syllable}}$, showing the positive influence of Vietnamese word segmentation towards seq2seq pre-training.

We publicly release our BARTpho models that can be used with popular libraries `fairseq` (Ott et al., 2019) and `transformers` (Wolf et al., 2020). We hope that our BARTpho can serve as a strong baseline for future research and applications of generative natural language processing (NLP) tasks for Vietnamese.

2 Related work

PhoBERT (Nguyen and Nguyen, 2020) is the first public large-scale monolingual language model pre-trained for Vietnamese, which helps obtain state-of-the-art performances on various downstream Vietnamese NLP/NLU tasks (Truong et al., 2021; Nguyen and Nguyen, 2021; Dao et al., 2021; Thin et al., 2021). PhoBERT is pre-trained on a 20GB word-level corpus of Vietnamese texts, using the RoBERTa pre-training approach (Liu et al., 2019) that optimizes BERT for more robust performance. Following PhoBERT, there are also public monolingual language models for Vietnamese such as viBERT and vELECTRA (Bui et al., 2020), which are based on BERT and ELECTRA pre-training approaches (Devlin et al., 2019; Clark et al., 2020) and pre-trained on syllable-level Vietnamese text corpora. Following Rothe et al. (2020) who leverage pre-trained language model checkpoints for sequence generation tasks, Nguyen et al. (2021) conduct an empirical study and show that PhoBERT helps produce better performance results than viBERT for a downstream task of Vietnamese abstractive summarization.

Our BARTpho is based on BART. We employ BART because it helps produce the strongest performances on downstream tasks in comparison to other pre-trained seq2seq models under a comparable setting in terms of the relatively equal numbers of model parameters and pre-training data sizes (Lewis et al., 2020; Raffel et al., 2020; Qi et al., 2020). BART is also used to pre-train monolingual models for other languages such as French (Eddine et al., 2020) and Chinese (Shao et al., 2021).

3 Our BARTpho

This section describes the architecture, the pre-training data and the optimization setup, that we use for BARTpho.

3.1 Architecture

Both BARTpho$_{\text{syllable}}$ and BARTpho$_{\text{word}}$ use the "large" architecture with 12 encoder and decoder layers and pre-training scheme of BART (Lewis et al., 2020). In particular, pre-training BART has two stages: (i) corrupting the input text with an arbitrary noising function, and (ii) learning to reconstruct the original text, i.e. optimizing the cross-entropy between its decoder's output and the original text. Here, BART uses the standard architecture Transformer (Vaswani et al., 2017), but employing the GeLU activation function (Hendrycks and Gimpel, 2016) rather than ReLU and performing parameter initialization from $\mathcal{N}(0, 0.02)$. Following Liu et al. (2020), we add a layer-normalization layer on top of both the encoder and decoder. Following Lewis et al. (2020), we also employ two types of noise in the noising function, including text infilling and sentence permutation. For text infilling, we sample a number of text spans with their lengths drawn from a Poisson distribution (λ = 3.5) and replace each span with a single special <mask> token. For sentence permutation, consecutive sentences are grouped to generate sentence blocks of 512 tokens, and sentences in each block are then shuffled in random order.

3.2 Pre-training data

For BARTpho$_{\text{word}}$, we employ the PhoBERT pre-training corpus (Nguyen and Nguyen, 2020), that contains 20GB of uncompressed texts (about 145M automatically word-segmented sentences). In addition, we also reuse the PhoBERT's tokenizer that applies a vocabulary of 64K subword types and BPE (Sennrich et al., 2016) to segment those word-segmented sentences with subword units. BARTpho$_{\text{word}}$ has about 420M parameters. Pre-training data for BARTpho$_{\text{syllable}}$ is a detokenized variant of the PhoBERT pre-training corpus (i.e. about 4B syllable tokens). We employ the pre-trained SentencePiece model (Kudo and Richardson, 2018) from XLM-RoBERTa (Conneau et al., 2020), used in mBART (Liu et al., 2020), to segment sentences with sub-syllable units and select a vocabulary of the top 40K most frequent types. BARTpho$_{\text{syllable}}$ has about 396M parameters.

Model	Validation set			Test set			
	R-1	R-2	R-L	R-1	R-2	R-L	Human
mBART	60.06	28.69	38.85	60.03	28.51	38.74	21/100
BARTpho$_{\text{syllable}}$	60.29	29.07	39.02	60.41	29.20	39.22	37/100
BARTpho$_{\text{word}}$	**60.55**	**29.89**	**39.73**	**60.51**	**29.65**	**39.75**	**42/100**

Table 1: Detokenized and case-sensitive ROUGE scores (in %) w.r.t. duplicate article removal. R-1, R-2 and R-L abbreviate ROUGE-1, ROUGE-2 and ROUGE-L, respectively. Every score difference between mBART and each BARTpho version is statistically significant with p-value < 0.05.

3.3 Optimization

We utilize the BART implementation with the denoising task from `fairseq` (Ott et al., 2019). We use Adam (Kingma and Ba, 2015) for optimization, and use a batch size of 512 sequence blocks across 8 A100 GPUs (40GB each) and a peak learning rate of 0.0001. Note that we initialize parameter weights of BARTpho$_{\text{syllable}}$ by those from mBART. For each BARTpho model, we run for 15 training epochs in about 6 days (here, the learning rate is warmed up for 1.5 epochs).

4 Experiments

4.1 Experimental setup

We evaluate and compare the performance of BARTpho with the strong baseline mBART on a downstream generative task of Vietnamese text summarization. Here, mBART is pre-trained on a Common Crawl dataset of 25 languages, which includes 137 GB of syllable-level Vietnamese texts.

We employ the single-document summarization dataset VNDS (Nguyen et al., 2019), consisting of 150704 news articles each including a news abstract (i.e. gold summary) and body content (i.e. input text). In particular, 105418, 22642 and 22644 articles are used for training, validation and test, respectively. However, we find that there are duplicate articles in this dataset. Therefore, we filter the duplicates, resulting in 99134, 22184 and 22498 articles for training, validation and test, respectively.[2] When fine-tuning BARTpho$_{\text{syllable}}$ and mBART, we use a detokenized version of the filtered dataset, while its automatically word-segmented version is used for fine-tuning BARTpho$_{\text{word}}$.

We formulate this task as a monolingual translation problem and fine-tune our BARTpho and the baseline mBART using the same hyper-parameter tuning strategy. We fix the maximum number of tokens in a batch at 4096. We use Adam and run for 20 training epochs. We also perform grid search to select the Adam initial learning rate from {1e-5, 2e-5, 3e-5, 5e-5}. We employ beam search with a beam size of 4 for decoding. We evaluate each model 4 times in every epoch. We select the model checkpoint that produces the highest ROUGE-L score (Lin, 2004) on the validation set, and we then apply the selected one to the test set.

Note that we compute the detokenized and case-sensitive ROUGE scores for all models (here, we detokenize the fine-tuned BARTpho$_{\text{word}}$'s output before computing the scores).

4.2 Main results

Table 1 presents our obtained ROUGE scores on the validation and test sets for the baseline mBART and our two BARTpho versions w.r.t. the setting of duplicate article removal. Clearly, both BARTpho versions achieve significantly better ROUGE scores than mBART on both validation and test sets.

We also conduct a human-based manual comparison between the outputs produced by the baseline mBART and our two BARTpho versions. In particular, we randomly sample 100 input text examples from the test set; and for each input example, we anonymously shuffle the summary outputs from three fine-tuned models (here, each input sampled example satisfies that any two out of three summary outputs are not exactly the same). We then ask two external Vietnamese annotators to choose which summary they think is the best. We obtain a Cohen's kappa coefficient at 0.61 for the inter-annotator agreement between the two annotators. Our second co-author then hosts and participates in a discussion session with the two annotators to resolve annotation conflicts (here, he does not know which model produces which summary). Table 1 shows final scores where BARTpho obtains a better human evaluation result than mBART.

[2] Firstly, we remove duplicates inside each of the training, validation and test sets. Secondly, if an article appears in both training and validation/test sets, then the article is filtered out of the training set. Lastly, if an article appears in both validation and test sets, then the article is filtered out of the validation set.

```python
from transformers import AutoModel, AutoTokenizer
# BARTpho_syllable
tokenizer = AutoTokenizer.from_pretrained("vinai/bartpho-syllable")
bartpho_syllable = AutoModel.from_pretrained("vinai/bartpho-syllable")
input_text = 'Chúng tôi là những nghiên cứu viên.'
input_ids = tokenizer(input_text, return_tensors='pt')
features = bartpho_syllable(**input_ids)
# BARTpho_word
tokenizer = AutoTokenizer.from_pretrained("vinai/bartpho-word")
bartpho_word = AutoModel.from_pretrained("vinai/bartpho-word")
input_text = 'Chúng_tôi là những nghiên_cứu_viên .'
input_ids = tokenizer(input_text, return_tensors='pt')
features = bartpho_word(**input_ids)
```

Figure 1: An example code using BARTpho for feature extraction with `transformers` in Python.

Model	Original test set		
	R-1	R-2	R-L
fastAbs [⋆]	54.52	23.01	37.64
viBERT2viBERT [∗]	59.75	27.29	36.79
PhoBERT2PhoBERT [∗]	60.37	29.12	39.44
mT5 [∗]	58.05	26.76	37.38
mBART	60.35	29.13	39.21
BARTpho$_{\text{syllable}}$	60.88	29.90	39.64
BARTpho$_{\text{word}}$	**61.14**	**30.31**	**40.15**

Table 2: ROUGE scores (in %) w.r.t. the original dataset setting (i.e. without duplicate article removal). [⋆] denotes the best performing model among different models experimented by Nguyen et al. (2019). [∗] denotes scores reported by Nguyen et al. (2021).

For comparison with previously published results (Nguyen et al., 2019, 2021), we also fine-tune our BARTpho models and baseline mBART on the original training set (i.e. without duplicate article removal),[3] using the same hyper-parameter tuning strategy as presented in Section 4.1. We report ROUGE scores on the original test set in Table 2. The previous highest ROUGE-L score obtained among different models experimented by Nguyen et al. (2019, 2021) is 39.44 accounted for PhoBERT2PhoBERT, that is 0.2 and 0.7 points lower than BARTpho$_{\text{syllable}}$ and BARTpho$_{\text{word}}$, respectively. Tables 1 and 2 show that BARTpho helps attain a new SOTA performance for this task.

Our automatic and human evaluation results from tables 1 and 2 demonstrate the effectiveness of large-scale BART-based monolingual seq2seq models for Vietnamese. Note that mBART uses 137 / 20 ≈ 7 times bigger Vietnamese pre-training data than BARTpho. In addition, the multilingual seq2seq mT5 (Xue et al., 2021b) is pre-trained on the multilingual dataset mC4 that includes 79M Common Crawl Vietnamese pages consisting of 116B syllable tokens, i.e. mT5 uses 116 / 4 = 29 times bigger Vietnamese pre-training data than BARTpho. However, BARTpho surpasses both mBART and mT5, reconfirming that the dedicated language-specific model still performs better than the multilingual one (Nguyen and Nguyen, 2020). Tables 1 and 2 also show that BARTpho$_{\text{word}}$ outperforms BARTpho$_{\text{syllable}}$, thus demonstrating the positive influence of word segmentation for seq2seq pre-training and fine-tuning in Vietnamese.

5 Usage example

Figure 1 presents a basic usage of BARTpho for feature extraction with `transformers` to show its potential use for other downstream tasks.[4] More usage examples of BARTpho with both `fairseq` and `transformers` can be found at the BARTpho's GitHub repository: `https://github.com/VinAIResearch/BARTpho`.

6 Conclusion

In this paper, we have presented BARTpho$_{\text{syllable}}$ and BARTpho$_{\text{word}}$—the first pre-trained and large-scale monolingual seq2seq models for Vietnamese. We demonstrate the usefulness of our BARTpho by showing that BARTpho performs better than its competitor mBART and helps produce the SOTA performance for the Vietnamese text summarization task. We hope that our public BARTpho models can foster future research and applications of generative Vietnamese NLP tasks.

[3] This is not a proper experimental setup because of data leakage, e.g. 1466 training articles appear in the test set.

[4] `https://huggingface.co/docs/transformers/model_doc/bartpho`

References

The Viet Bui, Thi Oanh Tran, and Phuong Le-Hong. 2020. Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models. In *Proceedings of PACLIC*, pages 13–20.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*, pages 8440–8451.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent Detection and Slot Filling for Vietnamese. In *Proceedings of InterSpeech*, pages 4698–4702.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *Proceedings of NeurIPS*, volume 32.

Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. 2020. BARThez: a Skilled Pretrained French Sequence-to-Sequence Model. *arXiv preprint*, arXiv:2010.12321.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). *arXiv preprint*, arXiv:1606.08415.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of EMNLP: System Demonstrations*, pages 66–71.

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 ACL*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81.

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 ACL*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*, arXiv:1907.11692.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of EMNLP*, pages 1037–1042.

Hieu Nguyen, Long Phan, James Anibal, Alec Peltekian, and Hieu Tran. 2021. VieSum: How Robust Are Transformer-based Models on Vietnamese Summarization? *arXiv preprint*, arXiv:2110.04257v1.

Linh The Nguyen and Dat Quoc Nguyen. 2021. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. In *Proceedings of NAACL: Demonstrations*, pages 1–7.

Van-Hau Nguyen, Thanh-Chinh Nguyen, Minh-Tien Nguyen, and Nguyen Xuan Hoai. 2019. VNDS: A Vietnamese Dataset for Summarization. In *Proceedings of NICS*, pages 375–380.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, pages 48–53.

Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, Houqiang Li, and Nan Duan. 2021. ProphetNet-X: Large-Scale Pre-training Models for English, Chinese, Multi-lingual, Dialog, and Code Generation. In *Proceedings of ACL: System Demonstrations*, pages 232–239.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting Future N-gram for Sequence-to-SequencePre-training. In *Findings of EMNLP*, pages 2401–2410.

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.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the ACL*, 8:264–280.

Sebastian Ruder. 2020. Why You Should Do NLP Beyond English. https://ruder.io/nlp-beyond-english/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL*, pages 1715–1725.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. CPT: A Pre-Trained Unbalanced Transformer for Both Chinese Language Understanding and Generation. *arXiv preprint*, arXiv:2109.05729.

Dinh Quang Thang, Le Hong Phuong, Nguyen Thi Minh Huyen, Nguyen Cam Tu, Mathias Rossignol, and Vu Xuan Luong. 2008. Word segmentation of Vietnamese texts: a comparison of approaches. In *Proceedings of LREC*, pages 1933–1936.

Dang Van Thin, Lac Si Le, Vu Xuan Hoang, and Ngan Luu-Thuy Nguyen. 2021. Investigating Monolingual and Multilingual BERT Models for Vietnamese Aspect Category Detection. *arXiv preprint*, arXiv:2103.09519.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of NAACL-HLT*, pages 2146–2153.

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 *Proceedings of NIPS*, volume 30.

Alex Wang and Kyunghyun Cho. 2019. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. *arXiv preprint*, arXiv:1902.04094.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021a. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv preprint*, arXiv:2105.13626.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of NAACL*, pages 483–498.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of ICML*, pages 11328–11339.