ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation

Long Phan[1,2], Hieu Tran[1], Hieu Nguyen[1,2], Trieu H. Trinh[1,3]

[1]VietAI Research
[2]Case Western Reserve University
[3]New York University
`long.phan@case.edu`

Abstract

We present ViT5, a pretrained Transformer-based encoder-decoder model for the Vietnamese language. With T5-style self-supervised pretraining, ViT5 is trained on a large corpus of high-quality and diverse Vietnamese texts. We benchmark ViT5 on two downstream text generation tasks, Abstractive Text Summarization and Named Entity Recognition. Although Abstractive Text Summarization has been widely studied for the English language thanks to its rich and large source of data, there has been minimal research into the same task in Vietnamese, a much lower resource language. In this work, we perform exhaustive experiments on both Vietnamese Abstractive Summarization and Named Entity Recognition, validating the performance of ViT5 against many other pretrained Transformer-based encoder-decoder models. Our experiments show that ViT5 significantly outperforms existing models and achieves state-of-the-art results on Vietnamese Text Summarization. On the task of Named Entity Recognition, ViT5 is competitive against previous best results from pretrained encoder-based Transformer models. Further analysis shows the importance of context length during the self-supervised pretraining on downstream performance across different settings.

1 Introduction

In recent years, Transformer-based architecture models and pretrained language models (LMs) have played a crucial role in the development of Natural Language Processing (NLP). Large pretrained models such as ELMo (Peters et al., 2018), GPT (Brown et al., 2020), BERT (Devlin et al., 2018) is trained on large corpora and have the ability to derive contextual representation of the language(s) in the training data. After pretraining is complete, these models achieved state-of-the-art results on a broad range of downstream tasks (Devlin et al., 2018). These self-supervised learning methods make use of learning objectives such as Masked Language Modeling (MLM) (Devlin et al., 2018) where random tokens in the input sequence are masked and the model attempts to predict the original tokens. The successes of pretrained models in English have inspired new research efforts to develop pretrained models in other languages such as Vietnamese (i.e., PhoBERT (Nguyen and Nguyen, 2020) and ViBERT (Bui et al., 2020)) and Italian (Sarti and Nissim, 2022). There are also ongoing efforts to develop multilingual pretrained models (mT5 (Xue et al., 2020), mBART (Liu et al., 2020)), in order to improve performance across multiple languages by learning both general and language-specific representations.

A short time ago, BARTpho (Tran et al., 2021), a large pretrained sequence-to-sequence model for Vietnamese inheriting BART style (Lewis et al., 2019), demonstrated the effectiveness of pretrained language models on Vietnamese abstractive summarization. Nevertheless, there are some past works that have shown that T5 architecture (Raffel et al., 2019) might outperform BART in some aspects (i.e., (Phan et al., 2021a)). Inspired by that, we propose ViT5, trained on the Vietnamese monolingual subset of CC100, following the architecture and training methodology in Raffel et al. (2019). We perform exhaustive comparisons on downstream performance to many different pretrained Transformer-based models (Nguyen et al., 2021; Tran et al., 2021; To et al., 2021). Specifically, we finetune the ViT5 on two summarization datasets, Wikilingua (Ladhak et al., 2020) and Vietnews (Nguyen et al., 2019), and one Named Entity Recognition dataset (PhoNER (Truong et al., 2021)).

Text summarization is an important downstream

task whose input is a free-form text paragraph or document(s), and the output sequence is expected to be a short summarization of the input. ViT5 achieves state-of-the-art results on both two of the single-document summarization tasks. We also perform an analysis on the max-length hyperparameter for input and output sequences during self-supervised learning and showed that longer lengths that match the downstream document's length lead to better result.

For NER, we reformulated the per-token classification task into a generation task, where the decoder reconstructs the original input sentence with inserted Named Entity tags following each token (Phan et al., 2021b). This simple and straightforward formulation achieves competitive results in comparison to direct per-token classification done on encoder-only model (Nguyen and Nguyen, 2020).

2 Related Work

There are lots of abstractive summarization studies in English. In an early example, (Gehrmann et al., 2018) employed a bottom-up content selector (BottomUp) to determine which phrases in the source document should be part of the summary, and then a copy mechanism was applied only to pre-select phrases during decoding. Their experiments obtained significant improvements on ROUGE for some canonical summarization datasets.

In recent years, pretrained language models have been used to enhance performance on language generation tasks. (Liu and Lapata, 2019) developed a Transformer-based encoder-decoder model so that pretrained language models like BERT can be adopted for abstractive summarization. Here, the authors proposed a novel document-level BERT-based encoder (*BERTSum*) and a general framework encompassing both extractive and abstractive summarization tasks. Based on *BERTSum*, Dou et al. (2021) introduced *GSum* that effectively used different types of guidance signals as input in order to generate more suitable words and more accurate summaries. This model accomplished state-of-the-art performance on four popular English summarization datasets.

Meanwhile, there are a small number of studies on Vietnamese text summarization. Most of these focus on inspecting extractive summarization. The researchers (Nguyen et al., 2018) com-

pared a wide range of extractive methods, including unsupervised ranking methods (e.g., LexRank, LSA, KL-divergence), supervised learning methods using TF-IDF and classifiers (e.g., Support Vector Machine, AdaBoost, Learning-2-rank), and deep learning methods (e.g., Convolutional Neural Network, Long-Short Term Memory). Similarly, the authors (Nguyen et al., 2019) also evaluated the extractive methods on their own dataset, which was released publicly as a benchmark for future studies.

Recent work (Quoc et al., 2021) investigated the combination of a pretrained BERT model and an unsupervised K-means clustering algorithm on extractive text summarization. The authors utilized multilingual and monolingual BERT models to encode sentence-level contextual information and then ranked this information using the K-means algorithm. Their report showed that monolingual models achieved better results compared when to multilingual models performing the same extractive summarization tasks. However, due to the lack of studies on Vietnamese abstractive summarization, we compare both multilingual and monolingual encoder-decoder models.

3 ViT5

In this section, we will explain our newly released ViT5 models, the vocabulary generation steps, the pretraining data, and the training setup.
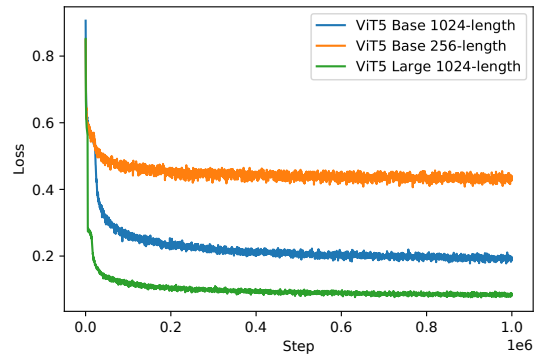

Figure 1: Loss curves for the masked span prediction task were used to pretrain the ViT5 models. Larger model with larger context optimizes much better, which leads to better downstream performance.

3.1 Model

ViT5 follows the encoder-decoder architecture proposed by Vaswani et al. (2017) and the T5
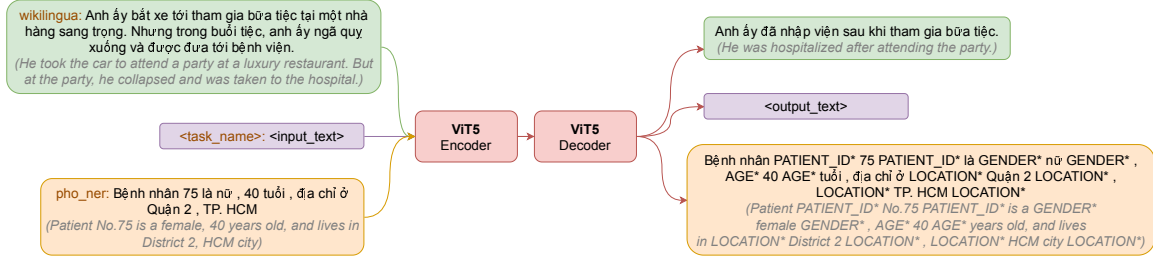
Figure 2: An overview of ViT5 encoder-decoder architecture, with input-output examples of two downstream tasks. For Named Entity Recognition, the decoder reconstructs the sentence with inserted Entity tags.

framework proposed by (Raffel et al., 2019). The original works of T5 proposed five different configs of model size: small, base, large, 3B, and 11B. For the purpose of practical study, we adapt the base (310M parameters) and large (866M parameters) models for ViT5 models and leave bigger models for future works.

We train ViT5 models with two different input and output lengths: 256 and 1024-length. We thoroughly experimented with these two models to have an insight into the importance of pretraining data length for summarization tasks. For the self-supervised training learning objectives, we use the span-corruption objective with a corruption rate of 15%. Figure 1 shows the computed loss during the self-supervised training stage for the three models.

3.2 Vocabulary

Different from some other current Vietnamese Transformer-based language models, we find that an effective vocabulary can contribute a significant improvement to our model performance. Therefore, we did pre-process on a 5GB subset of our pretraining corpus with care like normalizing punctuation and capitalization, splitting numbers. We fixed the size of vocabulary to 36K sub-words and trained SentencePiece (Kudo and Richardson, 2018) model on that dataset.

3.3 Pretraining Data

We use the CC100 Dataset (Monolingual Datasets from Web Crawl Data) (Wenzek et al., 2020; Conneau et al., 2020). The corpus contains monolingual data for over 100 languages. The corpus was constructed using the pipeline provided by (Wenzek et al., 2020) through processing January-December 2018 Commoncrawl snapshots. The total size for the Vietnamese Corpus is 138GB of raw text. We process and filter out 69GB of short paragraphs for 256-length model and 71GB of long paragraphs for 1024-length model.

Table 1: Input and Output Length of Finetuned Datasets

	Wikilingua	Vietnews
Train	13707	99134
Test	3916	22498
#avg body length	521	519
#avg abstract length	44	38

4 Abstractive Summarization

4.1 Wikilingua

Wikilingua (Ladhak et al., 2020) is a large-scale multilingual corpus for abstractive summarization tasks. The corpus consists of 18 languages, including Vietnamese. These article and summary pairs are extracted from WikiHow[1]. These articles have been reviewed by human authors to ensure quality. The Vietnamese articles are translated from the original English articles and have been reviewed by WikiHow's international translation team.

4.2 Vietnews

Vietnews (Nguyen et al., 2019) is a single-document abstractive summarization dataset including news data from reputable Vietnamese news website (*tuoitre.vn*, *vnexpress.net*, and *nguoiduatin.vn*). The authors of this work removed all articles related to questionnaires, analytical comments, and weather forecasts to ensure the quality of document summarization. The final released dataset only includes long document news events. The data consists of 150704 word-level news articles with a summary abstract and body text pairs. We follow the filtering pipeline by Tran et al. (2021) to deduplicate the train/dev/test dataset. The statistics after filtering are shown in Table 1.

---

[1]https://www.wikihow.com

Table 2: Test result on Wikilingua and Vietnews Summarization

Models	WikiLingua			Vietnews		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Transformer (RND2RND)	46.25	16.57	29.82	57.56	24.25	35.53
PhoBERT2PhoBERT	50.4	19.88	32.49	60.37	29.12	39.44
mBERT2mBERT	52.82	20.57	31.55	59.67	27.36	36.73
mBART	55.21	25.69	37.33	59.81	28.28	38.71
mT5	55.27	27.63	38.30	58.05	26.76	37.38
BARTpho	57.16	31.18	40.89	61.14	30.31	40.15
ViT5$_{base\ 256\text{-length}}$	57.86	29.98	40.23	61.85	31.70	41.70
ViT5$_{base\ 1024\text{-length}}$	<u>58.61</u>	<u>31.46</u>	<u>41.45</u>	<u>62.77</u>	<u>33.16</u>	<u>42.75</u>
ViT5$_{large\ 1024\text{-length}}$	**60.22**	**33.12**	**43.08**	**63.37**	**34.24**	**43.55**

Notes: The best scores are in bold and second best scores are underlined. The scores in gray color are our experiments. Code and models for reproducing our experiments: https://github.com/vietai/ViT5

4.3 Baselines

In order to verify the effectiveness of our proposed methods, we compare ViT5 models with the Transformer models based on (Vaswani et al., 2017), the ViSum BERT2BERT models (Nguyen et al., 2021), multilingual encoder-decoder model (Xue et al., 2020; Liu et al., 2020), and Vietnamese encoder-decoder BARTpho model (Tran et al., 2021). The baseline transformer models (labeled RND) have a multi-head self-attention and a feed-forward network. RND models are initialized with random weights. For the BARTpho models, we follow the models set up and results released by (Tran et al., 2021). All finetuned ViT5 models are conducted with a sequence length of 1024.

4.4 Results

We report the results of the ViT5 models on two datasets: Wikilingua and Vietnews. We do experiments with two versions of pretraining ViT5: 256-length and 1024-length to have an insight into the importance of pretraining data's paragraph length for summarization in Vietnamese. We also compare the results of ViT5$_{base}$ and ViT5$_{large}$ models.

We use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) as our benchmark metrics for both single document summarization datasets. The metric measures the overlap of n-grams and word sequences between two candidate and reference sequences. ROUGE-1, ROUGE-2, and ROUGE-L mean the overlap between unigram, bigram, and longest matching sequence, respectively.

4.4.1 Wikilingua

The results of our models on Wikilingua summarization dataset are shown in Table 2. ViT5 models outperform all of the experimented pretrained models, achieving state-of-the-art on all ROUGE metrics. There is also a significant increase in ROUGE scores when the models are pretrained on a longer input and output sequence (1024 compared to 256).

Both versions of ViT5$_{1024\text{-length}}$ achieve the highest results on Wikilingua summarization tasks across all ROUGE metrics with ViT5$_{large\ 1024\text{-length}}$ achieving state-of-the-art. There is a significant improvement in score between the base and large ViT5$_{1024\text{-length}}$ architectures (approximately 2% for ROUGE-1, ROUGE-2, and ROUGE-L). This is predictable as the number of parameters of ViT5$_{large}$ (866M) is approximately 2.8 times larger than ViT5$_{base}$ (310M).

There are interesting results when comparing the results of 256-length and 1024-length versions of ViT5$_{base}$. Although the finetuning settings are 1024-length for both ViT5$_{base}$ models, ViT5$_{base\ 1024\text{-length}}$ performs slightly better with 1% higher score for ROUGE-1, ROUGE-2, and ROUGE-L. These results are attributed to the longer sequences during self-supervised training. As reported in Table 1, the average words in an input body of Wikilingua corpus are more than 256 tokens, which can be considered long documents. For this reason, pretraining ViT5 on a 1024 sequence length corpus achieves better results on Wikilingua summarization task.

Two-out-of-three ViT5 models perform better

than the published BARTpho model in summarizing Wikilingua corpus. This can be the result of the quality of pretraining data. While BARTpho (and PhoBERT) was trained on 20GB of news data, ViT5 models are trained using CC100, which is a subset of Common Crawl data. CC100 corpus contains more diverse and general representation of the Vietnamese language than news data. Meanwhile, Wikilingua is more of an academic or instruction representation than news-like text.

4.4.2 Vietnews

The size of Vietnews corpus is much larger than Wikilingua corpus (with 7.7% for train and 5.8% for test set). The result of Vietnews abstractive summarization is in Table 2. Following the discussion of the need for an effective large pretrained encoder-decoder model in Section 1, we can see that there is a minimum increase in performance for the existing Vietnamese encoder-only model compared to the Transformer baseline. Pretraining on a large corpus of Vietnamese news, BARTpho still showed its limitation in the Vietnews summarization task with slightly better ROUGE scores than multilingual models (mBART and mT5).

Our ViT5 models still achieve state-of-the-art on Vietnews task for both 256 and 1024-length. For a more specific news-domain corpus, ViT5 models achieve notable results on the news domain although being trained on a more general Vietnamese natural language domain (CC100). This supports the assumption that our ViT5 models learn a better representation of the Vietnamese language even for more domain-specific summarization problems.

Similar to the results discussed in Section 4.4, ViT5$_{base}$ models when pretrained on a longer sequence corpus (1024-length) achieve better performance in summarizing compared to a short sequence corpus (256-length) across all ROUGE metrics. The average input length for Vietnews documents is approximately the same as in the Wikilingua task (more than 500 words). Therefore, the quality of long sequences during self-supervised training data also leads to a better summarizing in downstream Vietnews finetuned tasks.

5   Named Entity Recognition (NER)

Table 3: Test results on PhoNER_COVID19

Models	Micro-F1
XLM-R$_{large}$	93.8
PhoBERT$_{base}$	94.2
PhoBERT$_{large}$	**94.5**
ViT5$_{base\ 256\text{-}length}$	93.19
ViT5$_{base\ 1024\text{-}length}$	**94.5**
ViT5$_{large\ 1024\text{-}length}$	93.8

Notes: The best scores are in bold.

To verify the effectiveness of ViT5 on classification tasks, we test our models on PhoNER_COVID19 dataset (Truong et al., 2021). PhoNER is a dataset for recognizing named entities related to the COVID19 domain in Vietnamese. The dataset consists of 35,000 entities in over 10,000 sentences. The goal is to recognize 10 entity types related to the domain of COVID19 and epidemics topics. The dataset was released and benchmarked with PhoBERT (Nguyen and Nguyen, 2020).

We treat the NER classifications tasks as text-to-text generating tasks with tags of labels before and after an entity token (Phan et al., 2021b). An example of NER in text-to-text format is shown in Figure 2. The results are shown in Table 3.

The ViT5$_{large\ 1024\text{-}length}$ model, although effective in generating Vietnamese abstractive summarization, shows its limitation in classification tasks with lower F1 scores on NER task. On the other hand, our ViT5$_{base\ 1024\text{-}length}$ model still performs slightly better than PhoBERT$_{base}$ and competitively the same as the current state-of-the-art PhoBERT$_{large}$ on the PhoNER corpus.

6   Discussion

According to the results on both Wikilingua and Vietnews summarization tasks (Table 2 and Table 4.4.2), there is a steady increase in ROUGE scores going from the baseline Transformer, BERT2BERT related models (PhoBERT2PhoBERT and mBERT2mBERT), multilingual encoder-decoder models (mBART, mT5), to pretrained monolingual models (BARTpho and ViT5). For Vietnamese summarization tasks, monolingual encoder-decoder models noticeably outperform multilingual models, most likely thanks to their more focused and narrower pretraining stage.

Interestingly, a more general domain of pre-training texts can lead to a better domain-specific summarization performance. In Section 4.4.1, our ViT5 models while being trained on a more general corpus (CC100), outperform current models that are trained on news-related corpus. More technical domains such as laws, medicals, or engineering are not tested as we leave these domain-specific summarization tasks for future studies.

The slightly better performance of ViT5$_{\text{base 1024-length}}$ compared to ViT5$_{\text{base 256-length}}$ suggests that longer document summarization (more than 512 tokens) need a comparatively longer context length during the pretraining stage.

7   Conclusion

We introduce ViT5, a pretrained sequence-to-sequence Transformer model for the Vietnamese language. Leveraging the T5 self-supervised pre-training formulation on massive and high-quality Vietnamese corpora, we showed that finetuned ViT5 models are performant on both generation and classification tasks. We exhaustively compare ViT5 with other pretrained formulations on both multilingual and monolingual corpora. Our experiments show that ViT5 achieves state-of-the-art results on summarization in both Wikilingua and Vietnews corpus, and competitive results in generating Named Entity Recognition (NER) on the PhoNER_COVID19 dataset. We also analyze and discuss the importance of context length during the self-supervised pretraining stage, which strongly influences and positively leads to better downstream performance.

8   Acknowledgements

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mc-Candlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

The Viet Bui, Oanh Thi Tran, and Phuong Le-Hong. 2020. Improving sequence tagging for vietnamese text using transformer-based neural models. *CoRR*, abs/2006.15994.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. Gsum: A general framework for guided neural abstractive summarization. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen R. McKeown. 2020. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. *CoRR*, abs/2010.03093.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP/IJCNLP*.

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. *CoRR*, abs/2001.08210.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 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?

Minh-Tien Nguyen, Hoang-Diep Nguyen, Thi-Hai-Nang Nguyen, and Van-Hau Nguyen. 2018. Towards state-of-the-art baselines for vietnamese multi-document summarization. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 85–90.

Van-Hau Nguyen, Thanh-Chinh Nguyen, Minh-Tien Nguyen, and Nguyen Hoai. 2019. Vnds: A vietnamese dataset for summarization. pages 375–380.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Long Phan, Hieu Tran, Daniel Le, Hieu Nguyen, James Annibal, Alec Peltekian, and Yanfang Ye. 2021a. CoTexT: Multi-task learning with code-text transformer. In *Proceedings of the 1st Workshop on Natural Language Processing for Programming (NLP4Prog 2021)*, pages 40–47, Online. Association for Computational Linguistics.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021b. Scifive: a text-to-text transformer model for biomedical literature. *CoRR*, abs/2106.03598.

Huy To Quoc, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2021. Monolingual versus multilingual bertology for vietnamese extractive multi-document summarization.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation.

Huy Quoc To, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2021. Monolingual versus multilingual bertology for vietnamese extractive multi-document summarization.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2021. Bartpho: Pre-trained sequence-to-sequence models for vietnamese.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.