PhoBERT: Pre-trained language models for Vietnamese

Dat Quoc Nguyen and **Anh Tuan Nguyen**
VinAI Research, Vietnam
{v.datnq9, v.anhnt496}@vinai.io

Abstract

We present **PhoBERT** with two versions of "**base**" and "**large**"—the *first* public large-scale monolingual language models pre-trained for Vietnamese. We show that PhoBERT improves the state-of-the-art in multiple Vietnamese-specific NLP tasks including Part-of-speech tagging, Named-entity recognition and Natural language inference. We release PhoBERT to facilitate future research and downstream applications for Vietnamese NLP. Our PhoBERT is released at: `https://github.com/VinAIResearch/PhoBERT`.

1 Introduction

Pre-trained language models, especially BERT—the Bidirectional Encoder Representations from Transformers [Devlin *et al.*, 2019], have recently become extremely popular and helped to produce significant improvement gains for various NLP tasks. The success of pre-trained BERT and its variants has largely been limited to the English language. For other languages, one could retrain a language-specific model using the BERT architecture [Vu *et al.*, 2019; Martin *et al.*, 2019; de Vries *et al.*, 2019] or employ existing pre-trained multilingual BERT-based models [Devlin *et al.*, 2019; Conneau *et al.*, 2019; Conneau and Lample, 2019].

In terms of Vietnamese language modeling, to the best of our knowledge, there are two main concerns: **(i)** The Vietnamese Wikipedia corpus is the only data used to train all monolingual language models [Vu *et al.*, 2019], and it also is the only Vietnamese dataset included in the pre-training data used by all multilingual language models except XLM-R [Conneau *et al.*, 2019]. It is worth noting that Wikipedia data is not representative of a general language use, and the Vietnamese Wikipedia data is relatively small (1GB in size uncompressed), while pre-trained language models can be significantly improved by using more data [Liu *et al.*, 2019]. **(ii)** All monolingual and multilingual models, except ETNLP [Vu *et al.*, 2019], are not aware of the difference between Vietnamese syllables and word tokens (this ambiguity comes from the fact that the white space is also used to separate syllables that constitute words when written in Vietnamese). Without doing a pre-process step of Vietnamese word segmentation, those models directly apply Bype-Pair encoding (BPE) methods [Sennrich *et al.*, 2016] to the syllable-level pre-training Vietnamese data. Also, although performing word segmentation before applying BPE on the Vietnamese Wikipedia corpus, ETNLP in fact does not publicly release any pre-trained BERT-based model.[1] As a result, we find difficulties in applying existing pre-trained language models for word-level Vietnamese NLP tasks.

To handle the two concerns above, we train the *first* large-scale monolingual BERT-based "base" and "large" models using a 20GB word-level Vietnamese corpus. We evaluate our models on three downstream Vietnamese NLP tasks: the two most common ones of Part-of-speech (POS) tagging and Named-entity recognition (NER), and a language understanding task of Natural language inference (NLI). Experimental results show that our models obtain state-of-the-art (SOTA) performances for all three tasks. We release our models under the name PhoBERT in popular open-source libraries, hoping that PhoBERT can serve as a strong baseline for future Vietnamese NLP research and applications.

2 PhoBERT

This section outlines the architecture and describes the pre-training data and optimization setup we use for PhoBERT.

Architecture: PhoBERT has two versions $\text{PhoBERT}_{\text{base}}$ and $\text{PhoBERT}_{\text{large}}$, using the same configuration as $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$, respectively. PhoBERT pre-training approach is based on RoBERTa [Liu *et al.*, 2019] which optimizes the BERT pre-training method for more robust performance.

Data: We use a pre-training dataset of 20GB of uncompressed texts after cleaning. This dataset is a combination of two corpora: (i) the first one is the Vietnamese Wikipedia corpus (~1GB), and (ii) the second corpus (~19GB) is a subset of a 40GB Vietnamese news corpus after filtering out similar news and duplications.[2] We employ RDRSegmenter [Nguyen *et al.*, 2018] from VnCoreNLP [Vu *et al.*, 2018] to perform word and sentence segmentation on the pre-training dataset, resulting in ~145M word-segmented sentences (~3B word tokens). Different from RoBERTa, we then apply `fastBPE` [Sennrich *et al.*, 2016] to segment these sentences with subword units, using a vocabulary size of 64K subword types.

Optimization: We employ the RoBERTa implementation in `fairseq` [Ott *et al.*, 2019]. Each sentence contains at most 256 subword tokens (here, 5K/145M sentences with more

---

[1] `https://github.com/vietnlp/etnlp` – last access on the 28th February 2020.
[2] `https://github.com/binhvq/news-corpus`, crawled from a wide range of websites with 14 different topics.

Table 1: Performance scores (in %) on test sets. "Acc." abbreviates accuracy. [♣], [★], [♦] and [♠] denote results reported by Nguyen *et al.* (2017), Nguyen (2019), Vu *et al.* (2018) and Vu *et al.* (2019), respectively. "mBiLSTM" denotes a BiLSTM-based multilingual embedding method. Note that there are higher NLI results reported for XLM-R when fine-tuning on the concatenation of all 15 training datasets in the XNLI corpus. However, those results are not comparable as we only use the Vietnamese monolingual training data for fine-tuning.

POS tagging		NER		NLI	
Model	Acc.	Model	F_1	Model	Acc.
RDRPOSTagger [Nguyen *et al.*, 2014] [♣]	95.1	BiLSTM-CNN-CRF [♦]	88.3	mBiLSTM [Artetxe and Schwenk, 2019]	72.0
BiLSTM-CNN-CRF [Ma and Hovy, 2016] [♣]	95.4	VnCoreNLP-NER [Vu *et al.*, 2018]	88.6	multilingual BERT [Wu and Dredze, 2019]	69.5
VnCoreNLP-POS [Nguyen *et al.*, 2017]	95.9	VNER [Nguyen *et al.*, 2019b]	89.6	$XLM_{MLM+TLM}$ [Conneau and Lample, 2019]	76.6
jPTDP-v2 [Nguyen and Verspoor, 2018] [★]	95.7	BiLSTM-CNN-CRF + ETNLP [♠]	91.1	XLM-R$_{base}$ [Conneau *et al.*, 2019]	75.4
jointWPD [Nguyen, 2019]	96.0	VnCoreNLP-NER + ETNLP [♠]	91.3	XLM-R$_{large}$ [Conneau *et al.*, 2019]	79.7
PhoBERT$_{base}$	96.7	PhoBERT$_{base}$	93.6	PhoBERT$_{base}$	78.5
PhoBERT$_{large}$	**96.8**	PhoBERT$_{large}$	**94.7**	PhoBERT$_{large}$	**80.0**

than 256 subword tokens are skipped). Following Liu *et al.* [2019], we optimize the models using Adam [Kingma and Ba, 2014]. We use a batch size of 1024 and a peak learning rate of 0.0004 for PhoBERT$_{base}$, and a batch size of 512 and a peak learning rate of 0.0002 for PhoBERT$_{large}$. We run for 40 epochs (here, the learning rate is warmed up for 2 epochs). We use 4 Nvidia V100 GPUs (16GB each), resulting in about 540K training steps for PhoBERT$_{base}$ and 1.08M steps for PhoBERT$_{large}$. We pretrain PhoBERT$_{base}$ during 3 weeks, and then PhoBERT$_{large}$ during 5 weeks.

3 Experiments

We evaluate the performance of PhoBERT on three downstream Vietnamese NLP tasks: POS tagging, NER and NLI.
Experimental setup: For the two most common Vietnamese POS tagging and NER tasks, we follow the VnCoreNLP setup [Vu *et al.*, 2018], using standard benchmarks of the VLSP 2013 POS tagging dataset and the VLSP 2016 NER dataset [Nguyen *et al.*, 2019a]. For NLI, we use the Vietnamese validation and test sets from the XNLI corpus v1.0 [Conneau *et al.*, 2018] where the Vietnamese training data is machine-translated from English. Unlike the 2013 POS tagging and 2016 NER datasets which provide the gold word segmentation, for NLI, we use RDRSegmenter to segment the text into words before applying `fastBPE` to produce subwords from word tokens.

Following Devlin *et al.* [2019], for POS tagging and NER, we append a linear prediction layer on top of the PhoBERT architecture w.r.t. the first subword token of each word. We fine-tune PhoBERT for each task and each dataset independently, employing the Hugging Face `transformers` for POS tagging and NER and the RoBERTa implementation in `fairseq` for NLI. We use AdamW [Loshchilov and Hutter, 2019] with a fixed learning rate of 1.e-5 and a batch size of 32. We fine-tune in 30 training epochs, evaluate the task performance after each epoch on the validation set (here, early stopping is applied when there is no improvement after 5 continuous epochs), and then select the best model to report the final result on the test set.
Main results: Table 1 compares our PhoBERT scores with the previous highest reported results, using the same experimental setup. PhoBERT helps produce new SOTA results for all the three tasks, where unsurprisingly PhoBERT$_{large}$ obtains higher performances than PhoBERT$_{base}$.

For POS tagging, PhoBERT obtains about 0.8% absolute higher accuracy than the feature- and neural network-based models VnCoreNLP-POS (i.e. VnMarMoT) and jointWPD. For NER, PhoBERT$_{large}$ is 1.1 points higher F_1 than PhoBERT$_{base}$ which is 2+ points higher than the feature- and neural network-based models VnCoreNLP-NER and BiLSTM-CNN-CRF trained with the BERT-based ETNLP word embeddings [Vu *et al.*, 2019]. For NLI, PhoBERT outperforms the multilingual BERT and the BERT-based crosslingual model with a new translation language modeling objective $XLM_{MLM+TLM}$ by large margins. PhoBERT also performs slightly better than the cross-lingual model XLM-R, but using far fewer parameters than XLM-R (base: 135M vs. 250M; large: 370M vs. 560M).

Discussion: Using more pre-training data can help significantly improve the quality of the pre-trained language models [Liu *et al.*, 2019]. Thus it is not surprising that PhoBERT helps produce better performance than ETNLP on NER, and the multilingual BERT and $XLM_{MLM+TLM}$ on NLI (here, PhoBERT employs 20GB of Vietnamese texts while those models employ the 1GB Vietnamese Wikipedia data).

Our PhoBERT also does better than XLM-R which uses a 2.5TB pre-training corpus containing 137GB of Vietnamese texts (i.e. about $137/20 \approx 7$ times bigger than our pre-training corpus). Recall that PhoBERT performs segmentation into subword units after performing a Vietnamese word segmentation, while XLM-R directly applies a BPE method to the syllable-level pre-training Vietnamese data. Clearly, word-level information plays a crucial role for the Vietnamese language understanding task of NLI, i.e. word segmentation is necessary to improve the NLI performance. This reconfirms that dedicated language-specific models still outperform multilingual ones [Martin *et al.*, 2019].

Experiments also show that using a straightforward fine-tuning manner as we do can lead to SOTA results. Note that we might boost our downstream task performances even further by doing a more careful hyper-parameter fine-tuning.

4 Conclusion

In this paper, we have presented the first public large-scale PhoBERT language models for Vietnamese. We demonstrate the usefulness of PhoBERT by producing new state-of-the-art performances for three Vietnamese NLP tasks of POS tagging, NER and NLI. By publicly releasing PhoBERT, we hope that it can foster future research and applications in Vietnamse NLP. Our PhoBERT and its usage are available at: `https://github.com/VinAIResearch/PhoBERT`.

References

[Artetxe and Schwenk, 2019] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, 7:597–610, 2019.

[Conneau and Lample, 2019] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, pages 7059–7069, 2019.

[Conneau et al., 2018] Alexis Conneau, Ruty Rinott, Guillaume Lample, Holger Schwenk, Ves Stoyanov, Adina Williams, and Samuel R. Bowman. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485, 2018.

[Conneau et al., 2019] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint*, arXiv:1911.02116, 2019.

[de Vries et al., 2019] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. BERTje: A Dutch BERT Model. *arXiv preprint*, arXiv:1912.09582, 2019.

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

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint*, arXiv:1412.6980, 2014.

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

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of ICLR*, 2019.

[Ma and Hovy, 2016] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of ACL*, pages 1064–1074, 2016.

[Martin et al., 2019] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a Tasty French Language Model. *arXiv preprint*, arXiv:1911.03894, 2019.

[Nguyen and Verspoor, 2018] Dat Quoc Nguyen and Karin Verspoor. An improved neural network model for joint POS tagging and dependency parsing. In *Proceedings of the CoNLL 2018 Shared Task*, pages 81–91, 2018.

[Nguyen et al., 2014] Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham. RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger. In *Proceedings of the Demonstrations at EACL*, pages 17–20, 2014.

[Nguyen et al., 2017] Dat Quoc Nguyen, Thanh Vu, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. From word segmentation to POS tagging for Vietnamese. In *Proceedings of ALTA*, pages 108–113, 2017.

[Nguyen et al., 2018] Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. A Fast and Accurate Vietnamese Word Segmenter. In *Proceedings of LREC*, pages 2582–2587, 2018.

[Nguyen et al., 2019a] Huyen Nguyen, Quyen Ngo, Luong Vu, Vu Tran, and Hien Nguyen. VLSP Shared Task: Named Entity Recognition. *Journal of Computer Science and Cybernetics*, 34(4):283–294, 2019.

[Nguyen et al., 2019b] Kim Anh Nguyen, Ngan Dong, and Cam-Tu Nguyen. Attentive neural network for named entity recognition in vietnamese. In *Proceedings of RIVF*, 2019.

[Nguyen, 2019] Dat Quoc Nguyen. A neural joint model for Vietnamese word segmentation, POS tagging and dependency parsing. In *Proceedings of ALTA*, pages 28–34, 2019.

[Ott et al., 2019] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[Sennrich et al., 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725, 2016.

[Vu et al., 2018] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In *Proceedings of NAACL: Demonstrations*, pages 56–60, 2018.

[Vu et al., 2019] Xuan-Son Vu, Thanh Vu, Son Tran, and Lili Jiang. ETNLP: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. In *Proceedings of RANLP*, pages 1285–1294, 2019.

[Wu and Dredze, 2019] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP*, pages 833–844, 2019.