arXiv:2003.00744v1 [cs.CL] 2 Mar 2020

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 PhoBERT$_{\text{base}}$ and PhoBERT$_{\text{large}}$, using the same configuration as BERT$_{\text{base}}$ and 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 ($\sim$1GB), and (ii) the second corpus ($\sim$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 $\sim$145M word-segmented sentences ($\sim$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.