PhoBERT: Pre-trained language models for Vietnamese

Dat Quoc Nguyen[1] and **Anh Tuan Nguyen**[2,*]
[1]VinAI Research, Vietnam; [2]NVIDIA, USA
v.datnq9@vinai.io, tuananhn@nvidia.com

Abstract

We present **PhoBERT** with two versions—$\text{PhoBERT}_{\text{base}}$ and $\text{PhoBERT}_{\text{large}}$—the *first* public large-scale monolingual language models pre-trained for Vietnamese. Experimental results show that PhoBERT consistently outperforms the recent best pre-trained multilingual model XLM-R (Conneau et al., 2020) and improves the state-of-the-art in multiple Vietnamese-specific NLP tasks including Part-of-speech tagging, Dependency parsing, Named-entity recognition and Natural language inference. We release PhoBERT to facilitate future research and downstream applications for Vietnamese NLP. Our PhoBERT models are available at: https://github.com/VinAIResearch/PhoBERT.

1 Introduction

Pre-trained language models, especially BERT (Devlin et al., 2019)—the Bidirectional Encoder Representations from Transformers (Vaswani et al., 2017), 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 (Cui et al., 2019; de Vries et al., 2019; Vu et al., 2019; Martin et al., 2020) or employ existing pre-trained multilingual BERT-based models (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020).

In terms of Vietnamese language modeling, to the best of our knowledge, there are two main concerns as follows:

- The Vietnamese Wikipedia corpus is the only data used to train monolingual language models (Vu et al., 2019), and it also is the only Vietnamese dataset which is included in the pre-training data used by all multilingual language models except XLM-R. 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 pre-training data (Liu et al., 2019).
- All publicly released monolingual and multilingual BERT-based language models 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.[1] For example, a 6-syllable written text "Tôi là một nghiên cứu viên" (I am a researcher) forms 4 words "Tôi_{I} là_{am} một_{a} $\text{nghiên\_cứu\_viên}_{\text{researcher}}$".
  Without doing a pre-process step of Vietnamese word segmentation, those models directly apply Byte-Pair encoding (BPE) methods (Sennrich et al., 2016; Kudo and Richardson, 2018) to the syllable-level Vietnamese pre-training data.[2] Intuitively, for word-level Vietnamese NLP tasks, those models pre-trained on syllable-level data might not perform as good as language models pre-trained on word-level data.

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 four downstream Vietnamese NLP tasks: the common word-level ones of Part-of-speech (POS) tagging, Dependency parsing and Named-entity recogni-

*Work done during internship at VinAI Research.

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

[2]Although performing word segmentation before applying BPE on the Vietnamese Wikipedia corpus, ETNLP (Vu et al., 2019) in fact does not publicly release any pre-trained BERT-based language model (https://github.com/vietnlp/etnlp). In particular, Vu et al. (2019) release a set of 15K BERT-based word embeddings specialized only for the Vietnamese NER task.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1037–1042

tion (NER), and a language understanding task of Natural language inference (NLI) which can be formulated as either a syllable- or word-level task. Experimental results show that our models obtain state-of-the-art (SOTA) results on all these tasks. Our contributions are summarized as follows:

- We present the *first* large-scale monolingual language models pre-trained for Vietnamese.
- Our models help produce SOTA performances on four downstream tasks of POS tagging, Dependency parsing, NER and NLI, thus showing the effectiveness of large-scale BERT-based monolingual language models for Vietnamese.
- To the best of our knowledge, we also perform the *first* set of experiments to compare monolingual language models with the recent best multilingual model XLM-R in multiple (i.e. four) different language-specific tasks. The experiments show that our models outperform XLM-R on all these tasks, thus convincingly confirming that dedicated language-specific models still outperform multilingual ones.
- We publicly release our models under the name PhoBERT which can be used with `fairseq` (Ott et al., 2019) and `transformers` (Wolf et al., 2019). We hope 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 that we use for PhoBERT.

Architecture: Our PhoBERT has two versions, $\text{PhoBERT}_{\text{base}}$ and $\text{PhoBERT}_{\text{large}}$, using the same architectures of $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$, respectively. PhoBERT pre-training approach is based on RoBERTa (Liu et al., 2019) which optimizes the BERT pre-training procedure for more robust performance.

Pre-training data: To handle the first concern mentioned in Section 1, we use a 20GB pre-training dataset of uncompressed texts. This dataset is a concatenation of two corpora: (i) the first one is the Vietnamese Wikipedia corpus (∼1GB), and (ii) the second corpus (∼19GB) is generated by removing similar articles and duplication from a 50GB Vietnamese news corpus.[3] To solve the second concern, 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 of 64K subword types. On average there are 24.4 subword tokens per sentence.

Task	#training	#valid	#test
POS tagging†	27,000	870	2,120
Dep. parsing†	8,977	200	1,020
NER†	14,861	2,000	2,831
NLI‡	392,702	2,490	5,010

Table 1: Statistics of the downstream task datasets. "#training", "#valid" and "#test" denote the size of the training, validation and test sets, respectively. † and ‡ refer to the dataset size as the numbers of sentences and sentence pairs, respectively.

Optimization: We employ the RoBERTa implementation in `fairseq` (Ott et al., 2019). We set a maximum length at 256 subword tokens, thus generating 145M × 24.4 / 256 ≈ 13.8M sentence blocks. Following Liu et al. (2019), we optimize the models using Adam (Kingma and Ba, 2014). We use a batch size of 1024 across 4 V100 GPUs (16GB each) and a peak learning rate of 0.0004 for $\text{PhoBERT}_{\text{base}}$, and a batch size of 512 and a peak learning rate of 0.0002 for $\text{PhoBERT}_{\text{large}}$. We run for 40 epochs (here, the learning rate is warmed up for 2 epochs), thus resulting in 13.8M × 40 / 1024 ≈ 540K training steps for $\text{PhoBERT}_{\text{base}}$ and 1.08M training steps for $\text{PhoBERT}_{\text{large}}$. We pre-train $\text{PhoBERT}_{\text{base}}$ during 3 weeks, and then $\text{PhoBERT}_{\text{large}}$ during 5 weeks.

3 Experimental setup

We evaluate the performance of PhoBERT on four downstream Vietnamese NLP tasks: POS tagging, Dependency parsing, NER and NLI.

Downstream task datasets

Table 1 presents the statistics of the experimental datasets that we employ for downstream task evaluation. For POS tagging, Dependency parsing and NER, we follow the VnCoreNLP setup (Vu et al., 2018), using standard benchmarks of the VLSP 2013 POS tagging dataset,[4] the VnDT dependency

[3] `https://github.com/binhvq/news-corpus`, crawled from a wide range of news websites and topics.

[4] `https://vlsp.org.vn/vlsp2013/eval`

POS tagging (word-level)		**Dependency parsing** (word-level)	
Model	Acc.	Model	LAS / UAS
RDRPOSTagger (Nguyen et al., 2014a) [♣]	95.1	_	_
BiLSTM-CNN-CRF (Ma and Hovy, 2016) [♣]	95.4	VnCoreNLP-DEP (Vu et al., 2018) [★]	71.38 / 77.35
VnCoreNLP-POS (Nguyen et al., 2017) [♣]	95.9	jPTDP-v2 [★]	73.12 / 79.63
jPTDP-v2 (Nguyen and Verspoor, 2018) [★]	95.7	jointWPD [★]	73.90 / 80.12
jointWPD (Nguyen, 2019) [★]	96.0	Biaffine (Dozat and Manning, 2017) [★]	74.99 / 81.19
XLM-R_{base} (our result)	96.2	Biaffine w/ XLM-R_{base} (our result)	76.46 / 83.10
XLM-R_{large} (our result)	96.3	Biaffine w/ XLM-R_{large} (our result)	75.87 / 82.70
$PhoBERT_{base}$	96.7	Biaffine w/ $PhoBERT_{base}$	**78.77 / 85.22**
$PhoBERT_{large}$	**96.8**	Biaffine w/ $PhoBERT_{large}$	77.85 / 84.32

Table 2: Performance scores (in %) on the POS tagging and Dependency parsing test sets. "Acc.", "LAS" and "UAS" abbreviate the Accuracy, the Labeled Attachment Score and the Unlabeled Attachment Score, respectively (here, all these evaluation metrics are computed on all word tokens, including punctuation). [♣] and [★] denote results reported by Nguyen et al. (2017) and Nguyen (2019), respectively.

treebank v1.1 (Nguyen et al., 2014b) with POS tags predicted by VnCoreNLP and the VLSP 2016 NER dataset (Nguyen et al., 2019a).

For NLI, we use the manually-constructed Vietnamese validation and test sets from the cross-lingual NLI (XNLI) corpus v1.0 (Conneau et al., 2018) where the Vietnamese training set is released as a machine-translated version of the corresponding English training set (Williams et al., 2018). Unlike the POS tagging, Dependency parsing and NER datasets which provide the gold word segmentation, for NLI, we employ RDRSegmenter to segment the text into words before applying BPE to produce subwords from word tokens.

Fine-tuning

Following Devlin et al. (2019), for POS tagging and NER, we append a linear prediction layer on top of the PhoBERT architecture (i.e. to the last Transformer layer of PhoBERT) w.r.t. the first subword of each word token.[5] For dependency parsing, following Nguyen (2019), we employ a reimplementation of the state-of-the-art Biaffine dependency parser (Dozat and Manning, 2017) from Ma et al. (2018) with default optimal hyper-parameters. We then extend this parser by replacing the pre-trained word embedding of each word in an input sentence by the corresponding contextualized embedding (from the last layer) computed for the first subword token of the word.

For POS tagging, NER and NLI, we employ `transformers` (Wolf et al., 2019) to fine-tune PhoBERT for each task and each dataset independently. We use AdamW (Loshchilov and Hutter, 2019) with a fixed learning rate of 1.e-5 and a batch size of 32 (Liu et al., 2019). 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 checkpoint to report the final result on the test set (note that each of our scores is an average over 5 runs with different random seeds).

[5] In our preliminary experiments, using the average of contextualized embeddings of subword tokens of each word to represent the word produces slightly lower performance than using the contextualized embedding of the first subword.

4 Experimental results

Main results

Tables 2 and 3 compare PhoBERT scores with the previous highest reported results, using the same experimental setup. It is clear that our PhoBERT helps produce new SOTA performance results for all four downstream tasks.

For POS tagging, the neural model jointWPD for joint POS tagging and dependency parsing (Nguyen, 2019) and the feature-based model VnCoreNLP-POS (Nguyen et al., 2017) are the two previous SOTA models, obtaining accuracies at about 96.0%. PhoBERT obtains 0.8% absolute higher accuracy than these two models.

For Dependency parsing, the previous highest parsing scores LAS and UAS are obtained by the Biaffine parser at 75.0% and 81.2%, respectively. PhoBERT helps boost the Biaffine parser with about 4% absolute improvement, achieving a LAS at 78.8% and a UAS at 85.2%.

For NER, $PhoBERT_{large}$ produces 1.1 points higher F_1 than $PhoBERT_{base}$. In addition, $PhoBERT_{base}$ obtains 2+ points higher than the previous SOTA feature- and neural network-based models VnCoreNLP-NER (Vu et al., 2018) and BiLSTM-CNN-CRF (Ma and Hovy, 2016) which

NER (word-level)		**NLI** (syllable- or word-level)	
Model	F_1	Model	Acc.
BiLSTM-CNN-CRF [♦]	88.3	_	_
VnCoreNLP-NER (Vu et al., 2018) [♦]	88.6	BiLSTM-max (Conneau et al., 2018)	66.4
VNER (Nguyen et al., 2019b)	89.6	mBiLSTM (Artetxe and Schwenk, 2019)	72.0
BiLSTM-CNN-CRF + ETNLP [♠]	91.1	multilingual BERT (Devlin et al., 2019) [■]	69.5
VnCoreNLP-NER + ETNLP [♠]	91.3	$\text{XLM}_{\text{MLM+TLM}}$ (Conneau and Lample, 2019)	76.6
$\text{XLM-R}_{\text{base}}$ (our result)	92.0	$\text{XLM-R}_{\text{base}}$ (Conneau et al., 2020)	75.4
$\text{XLM-R}_{\text{large}}$ (our result)	92.8	$\text{XLM-R}_{\text{large}}$ (Conneau et al., 2020)	79.7
$\text{PhoBERT}_{\text{base}}$	93.6	$\text{PhoBERT}_{\text{base}}$	78.5
$\text{PhoBERT}_{\text{large}}$	**94.7**	$\text{PhoBERT}_{\text{large}}$	**80.0**

Table 3: Performance scores (in %) on the NER and NLI test sets. [♦], [♠] and [■] denote results reported by Vu et al. (2018), Vu et al. (2019) and Wu and Dredze (2019), respectively. Note that there are higher Vietnamese NLI results reported for XLM-R when fine-tuning on the concatenation of all 15 training datasets from the XNLI corpus (i.e. TRANSLATE-TRAIN-ALL: 79.5% for $\text{XLM-R}_{\text{base}}$ and 83.4% $\text{XLM-R}_{\text{large}}$). However, those results might not be comparable as we only use the monolingual Vietnamese training data for fine-tuning.

are trained with the set of 15K BERT-based ETNLP word embeddings (Vu et al., 2019).

For NLI, PhoBERT outperforms the multilingual BERT (Devlin et al., 2019) and the BERT-based cross-lingual model with a new translation language modeling objective $\text{XLM}_{\text{MLM+TLM}}$ (Conneau and Lample, 2019) by large margins. PhoBERT also performs better than the recent best pre-trained multilingual model XLM-R but using far fewer parameters than XLM-R: 135M ($\text{PhoBERT}_{\text{base}}$) vs. 250M ($\text{XLM-R}_{\text{base}}$); 370M ($\text{PhoBERT}_{\text{large}}$) vs. 560M ($\text{XLM-R}_{\text{large}}$).

Discussion

We find that $\text{PhoBERT}_{\text{large}}$ achieves 0.9% lower dependency parsing scores than $\text{PhoBERT}_{\text{base}}$. One possible reason is that the last Transformer layer in the BERT architecture might not be the optimal one which encodes the richest information of syntactic structures (Hewitt and Manning, 2019; Jawahar et al., 2019). Future work will study which PhoBERT's Transformer layer contains richer syntactic information by evaluating the Vietnamese parsing performance from each layer.

Using more pre-training data can 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 $\text{XLM}_{\text{MLM+TLM}}$ on NLI (here, PhoBERT uses 20GB of Vietnamese texts while those models employ the 1GB Vietnamese Wikipedia corpus).

Following the fine-tuning approach that we use for PhoBERT, we carefully fine-tune XLM-R for the remaining Vietnamese POS tagging, Dependency parsing and NER tasks (here, it is applied to the first sub-syllable token of the first syllable of each word).[6] Tables 2 and 3 show that our PhoBERT also does better than XLM-R on these three word-level tasks. It is worth noting that XLM-R uses a 2.5TB pre-training corpus which contains 137GB of Vietnamese texts (i.e. about 137 / 20 $\approx$ 7 times bigger than our pre-training corpus). Recall that PhoBERT performs Vietnamese word segmentation to segment syllable-level sentences into word tokens before applying BPE to segment the word-segmented sentences into subword units, while XLM-R directly applies BPE to the syllable-level Vietnamese pre-training sentences. This reconfirms that the dedicated language-specific models still outperform the multilingual ones (Martin et al., 2020).[7]

5 Conclusion

In this paper, we have presented the first large-scale monolingual PhoBERT language models pre-trained for Vietnamese. We demonstrate the usefulness of PhoBERT by showing that PhoBERT performs better than the recent best multilingual model XLM-R and helps produce the SOTA performances for four downstream Vietnamese NLP tasks of POS tagging, Dependency parsing, NER and NLI. By publicly releasing PhoBERT models, we hope that they can foster future research and applications in Vietnamese NLP.

[6] For fine-tuning XLM-R, we use a grid search on the validation set to select the AdamW learning rate from {5e-6, 1e-5, 2e-5, 4e-5} and the batch size from {16, 32}.

[7] Note that Martin et al. (2020) only compare their model CamemBERT with XLM-R on the French NLI task.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *TACL*, 7:597–610.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Proceedings of NeurIPS*, pages 7059–7069.

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

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv preprint*, arXiv:1906.08101.

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.

Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *Proceedings of ICLR*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*, pages 4129–4138.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of ACL*, pages 3651–3657.

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

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.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of ICLR*.

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

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-Pointer Networks for Dependency Parsing. In *Proceedings of ACL*, pages 1403–1414.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of ACL*, pages 7203–7219.

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

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

Dat Quoc Nguyen, Dai Quoc Nguyen, Son Bao Pham, Phuong-Thai Nguyen, and Minh Le Nguyen. 2014b. From Treebank Conversion to Automatic Dependency Parsing for Vietnamese. In *Proceedings of NLDB*, pages 196–207.

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

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

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

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

Kim Anh Nguyen, Ngan Dong, and Cam-Tu Nguyen. 2019b. Attentive Neural Network for Named Entity Recognition in Vietnamese. In *Proceedings of RIVF*.

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.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

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

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

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of NAACL*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint*, arXiv:1910.03771.

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