AI Day 2022 Empower the Future

Recent Advances in Pre-trained Models for Vietnamese Language Processing

Dat Quoc Nguyen

Head of NLP, VinAI

https://datquocnguyen.github.io

VinAI

Introduction

- Transformer architecture for machine translation [1]

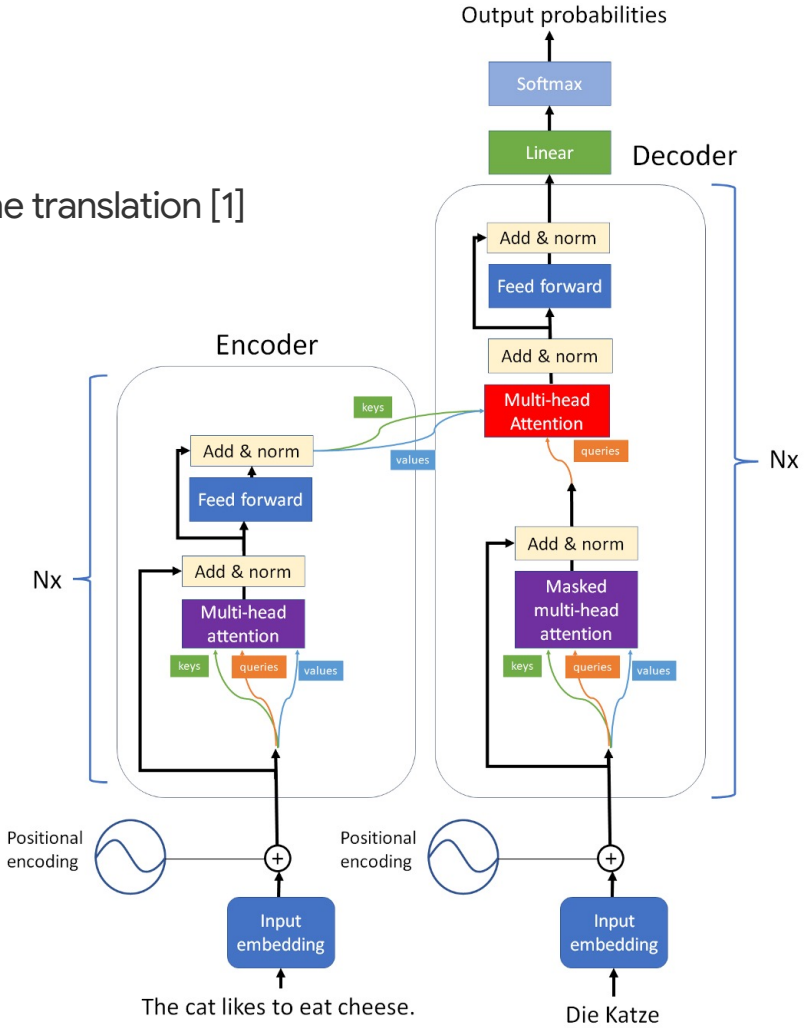


Image by Vincent Mueller: https://medium.com/towards-data-science/attention-please-85bd0abac41

Introduction

- Transformer encoder-based models, pre-trained on large-scale corpora with a masked language modeling objective, e.g. BERT [2], RoBERTa [3], ELECTRA [4]
- Pre-trained BERT-type models for Vietnamese:

Model	Type	Date	Data (vi)	
mBERT [2]	Multi.	11/2018	01GB	Syllable
XLM-R [5]	Multi.	11/2019	137GB	Syllable
PhoBERT [6]	Mono.	03/2020	20GB	Word
viBert [7]	Mono.	10/2020	10GB	Syllable
vELECTRA [7]	Mono.	10/2020	60GB	Syllable

"Data (vi)" denotes the size of Vietnamese text data used for pre-training

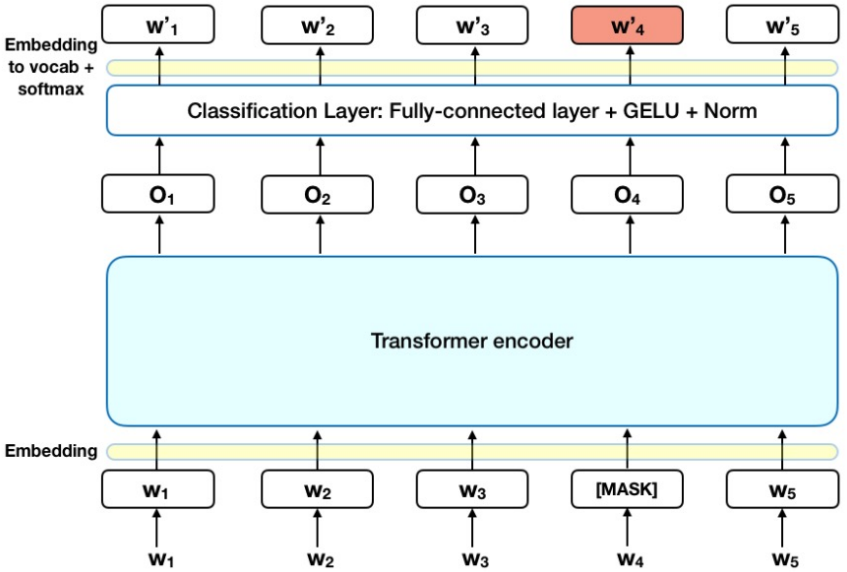


Image by Rani Horev: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Introduction

- Transformer decoder-based models, pre-trained on large-scale corpora with a standard language modeling objective, e.g. GPT-n [8-10], DialoGPT [11], LaMDA [12]
- Pre-trained GPT-type models for Vietnamese:
  - No study/research using these models for Vietnamese text generation ➔ This talk does not further cover these models in detail

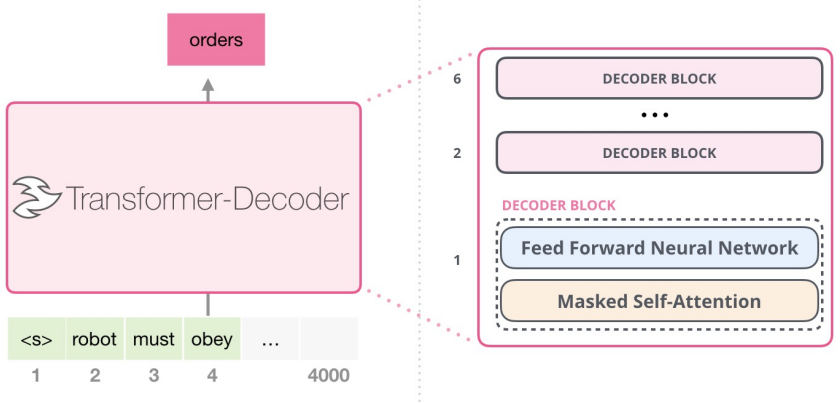


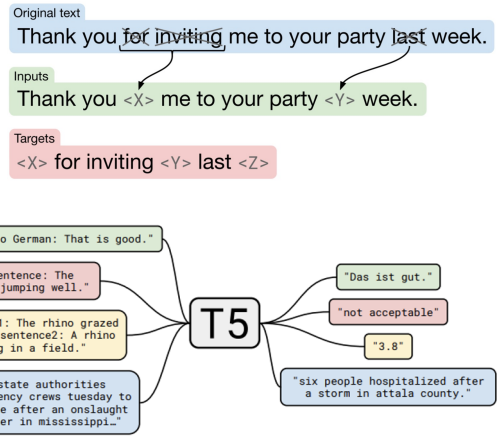
Image by Jay Alammar: https://jalammar.github.io/illustrated-gpt2/

Model	Type	Date	Data (vi)
XGLM [13]	Multi.	12/2021	50GB
BLOOM [14]	Multi.	07/2022	40GB
gpt-j-6B-vietnamese-news	Mono.	09/2021	65GB
gpt-neo-1.3B-vietnamese-news	Mono.	09/2021	65GB

Introduction

- Standard encoder-decoder Transformer-based sequence-to-sequence models, pre-trained on large-scale corpora with a denoising objective, e.g. BART [15], T5 [16], ByT5 [17]
- Pre-trained sequence-to-sequence models for Vietnamese:

Model	Type	Date	Data (vi)
mBART [18]	Multi.	01/2020	137GB (25B syllables)
mT5 [19]	Multi.	10/2020	116B syllables
BARTpho [20]	Mono.	09/2021	20GB (4B syllables)
viT5 [21]	Mono.	07/2022	70GB



Figures taken from [16]

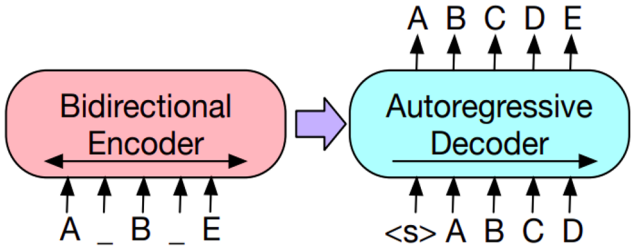


Figure taken from [15]

Introduction

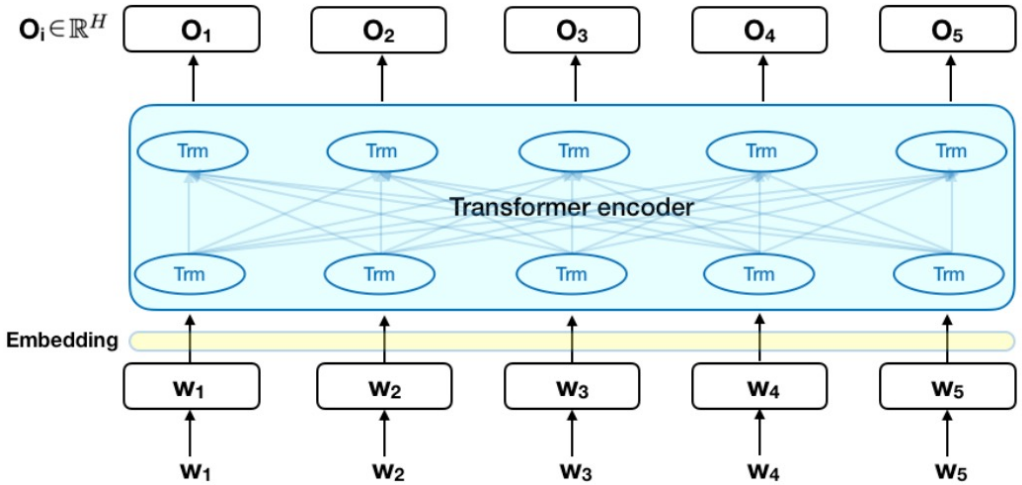
- Other pre-trained models for specific Vietnamese NLP tasks:
  - PhoNLP: A pre-trained model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing [22]
  - VinAI Translate: Pre-trained BART-type translation models for Vietnamese-to-English and English-to-Vietnamese [23]
- **Outline**
  - PhoBERT
  - BARTpho
  - VinAI Translate

Outline

- PhoBERT: Pre-trained language models for Vietnamese
- BARTpho
- VinAI Translate

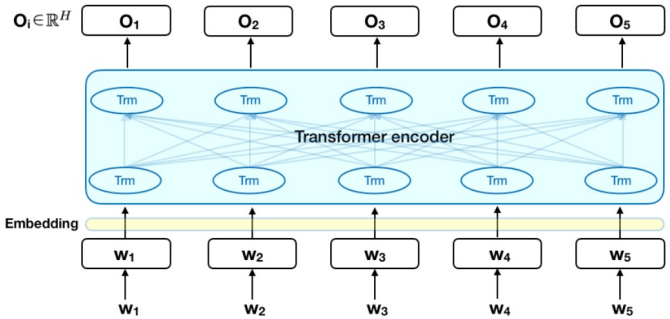
Motivation

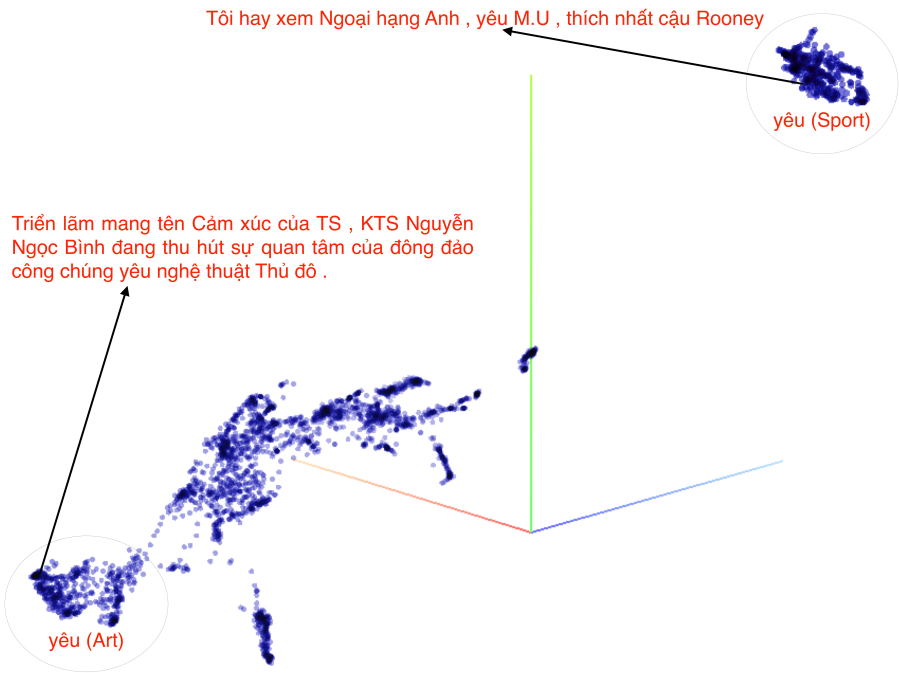
- Language model BERT—Bidirectional Encoder Representations from Transformers [2]—is a recent breakthrough in NLP
  - BERT and its variants, pretrained on large-scale corpora, help improve the state-of-the-art performances of various NLP research and application tasks
  - Represent words by embedding vectors which encode the contexts where the words appear, i.e. contextualized word embeddings



Motivation

- Illustration of how a BERT-based language model generates contextualized embeddings for the token “yêu” (love) depending on contextual sentences where “yêu” appears





UMAP clusters of 10K contextualized embeddings of the token “yêu” (love) from 10K sentences where the word appears

Motivation

- The success of BERT and its variants has largely been limited to English
  - Most pre-trained BERT-based models were learned using English corpus only, or data combined from different languages (i.e. pre-trained multilingual models)
- Multilingual BERT-based models are not aware of the **difference between Vietnamese syllables and word tokens**, thus using syllable-level pre-training Vietnamese texts
- 85% of Vietnamese word types are composed of at least 2 syllables (âm/tiếng)

Syllables: VinAI công bố các kết quả nghiên cứu khoa học tại hội nghị hàng đầu thế giới về trí tuệ nhân tạo

Words: VinAI công_bố các kết_quả nghiên_cứu khoa_học tại hội_nghị hàng_đầu thế_giới về trí_tuệ nhân_tạo

(VinAI publishes research outputs at world-leading conferences in Artificial Intelligence)

Motivation

- Previous monolingual BERT-based language models for Vietnamese:
  - Used the Vietnamese Wikipedia corpus which is relatively small (01GB)
    (Note that pre-trained models can be significantly improved by using more data)
  - Trained at the syllable level: without doing a pre-process step of Vietnamese word segmentation
- 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

Syllables: VinAI công bố các kết quả nghiên cứu khoa học tại hội nghị hàng đầu thế giới về trí tuệ nhân tạo

Words: VinAI công_bố các kết_quả nghiên_cứu khoa_học tại hội_nghị hàng_đầu thế_giới về trí_tuệ nhân_tạo

(VinAI publishes research outputs at world-leading conferences in Artificial Intelligence)

Pre-training

- How VinAI trains PhoBERT to handle previous concerns:
  - Use a large-scale corpus of 20GB Vietnamese texts
  - Perform Vietnamese word segmentation before pre-training

    👉 Pre-training corpus of 145M word-segmented sentences (3B word tokens)
- PhoBERT pre-training procedure is based on RoBERTa [3] which optimizes BERT for more robust performance, e.g. removing the next-sentence pretraining objective
- Two versions: PhoBERT-base (150M parameters) and PhoBERT-large (350M parameters)
- Pre-train PhoBERT using 4 GPUs V100 16GB memory each in 8 weeks
- Publicly released: https://github.com/VinAIResearch/PhoBERT
- PhoBERT can be used with popular open-source libraries: **transformers** and **fairseq**

Downstream task evaluation

- **Aspect-based sentiment analysis**: To identify the aspect categories mentioned in user-generated reviews from a set of pre-defined categories [24]
  - Use a linear prediction layer on top of the PhoBERT output for the classification token [CLS]—the first token of the input sequence

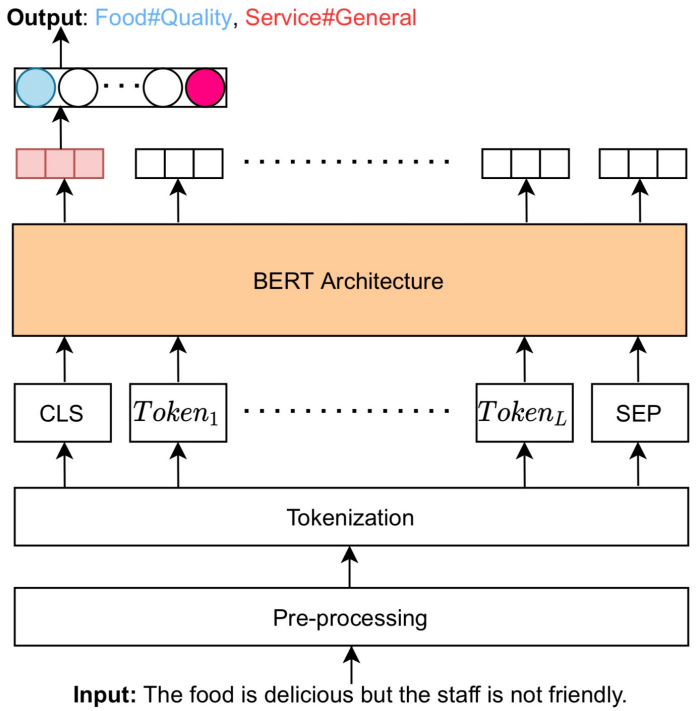


Figure taken from [24]

Downstream task evaluation

- **Natural language inference (NLI)**: To determine whether a “hypothesis” is true (entailment), false (contradiction), or undetermined (neutral) given a “premise” ➔ a sentence pair classification task
  - Use a linear prediction layer on top of the PhoBERT output for the [CLS] token—the first token of the input sequence when concatenating both “premise” and “hypothesis”

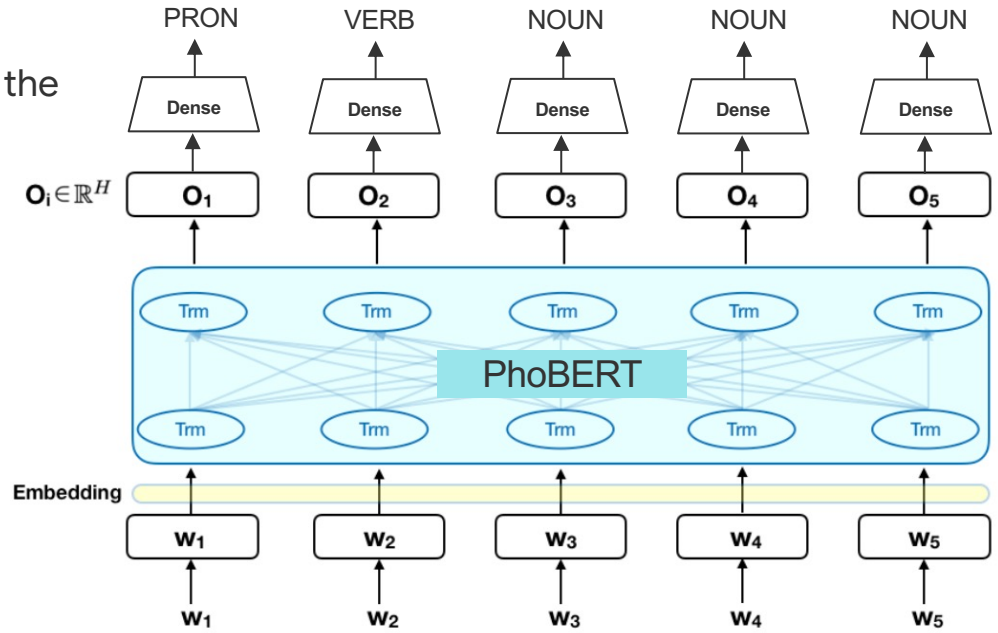
    True (Entailment): “[CLS] Thông báo phản đối luật sư và tòa án hoặc cơ quan hành chính sẽ phải được gửi đi [SEP] [SEP] Ban cố vấn độc lập và toà án sẽ nhận được thông báo [SEP]”

    (Dark red is the premise while dark blue is the hypothesis)

Downstream task evaluation

- **Part-of-Speech (POS) tagging**: To assign a lexical category tag to each word in a text
  - Use a linear prediction layer on top of the PhoBERT architecture

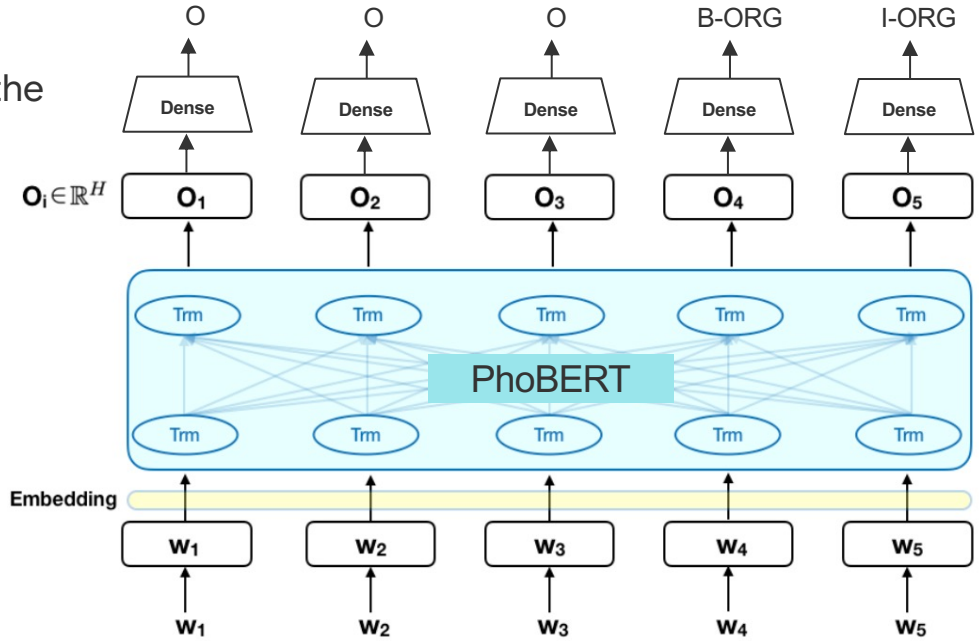
ID	Form	POS
1	Tôi I	PRON
2	là am	VERB
3	sinh_viên student	NOUN
4	Đại_học university	NOUN
5	Công_nghệ technology	NOUN



Downstream task evaluation

- **Named entity recognition (NER)**: To identify personal names, locations, organizations,...
  - Use a linear prediction layer on top of the PhoBERT architecture

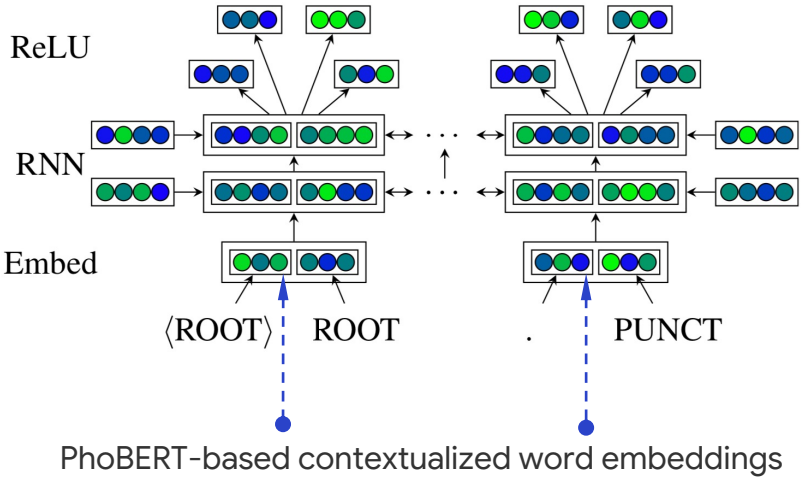
ID	Form	NER
1	Tôi I	O
2	là am	O
3	sinh_viên student	O
4	Đại_học university	B-ORG
5	Công_nghệ technology	I-ORG

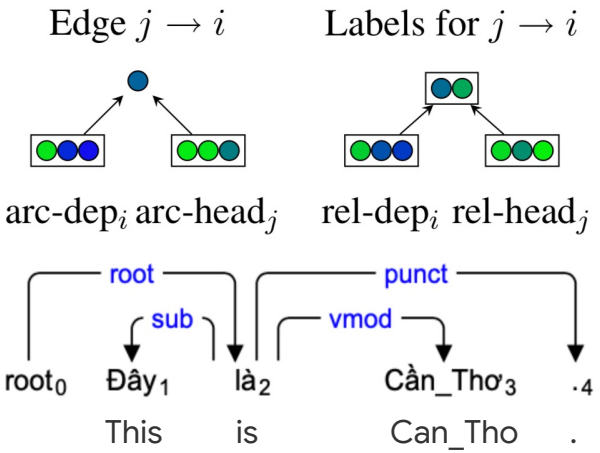


Downstream task evaluation

- **Dependency parsing**: To analyze the syntactic structure of a sentence by identifying grammatical relationships between "head" words and words which modify those heads
  - Extend the graph-based Biaffine parser [25] with the PhoBERT-based contextualized word embeddings as part of the input

Figures taken from [25]





Biaffine classifier

Downstream task evaluation

- Experimental Vietnamese benchmark datasets
  - *Aspect-based sentiment analysis*: Two large corpora for Vietnamese aspect-based sentiment analysis at sentence level [24]
  - *NLI*: The Vietnamese data from the cross-lingual NLI corpus v1.0 [26]
  - *POS tagging*: The VLSP 2013 POS tagging task
  - *NER*: The VLSP 2016 NER task's dataset [27] and PhoNER_COVID19 [28]
  - *Dependency parsing*: The Vietnamese dependency treebank VnDT [29]

- **Main baseline XLM-R** [5]—the pre-trained multilingual RoBERTa variant which uses 2.5 TB pre-training data, including 137 GB syllable-level Vietnamese text data

Downstream task evaluation

- Vietnamese aspect-based sentiment analysis (see [24] for details)

Model	Data (vi)	
mBERT [2]	01GB	Syllable
XLM-R [5]	137GB	Syllable
PhoBERT [6]	20GB	Word
viBert_FPT [7]	10GB	Syllable
vELECTRA_FPT [7]	60GB	Syllable
viBert4news	20GB	Syllable

THE EXPERIMENTAL RESULTS OF VARIOUS MONO-LINGUAL AND MULTI-LINGUAL PRE-TRAINED BERT MODELS ON VIETNAMESE ASPECT CATEGORY DETECTION TASK FOR THE RESTAURANT DOMAIN.

Types	Models	Precision	Recall	F1-score
Multi-lingual	mBERT	81.39	76.34	78.78
	mDistilBert	80.35	76.07	78.16
	XLM-R	82.98	81.40	82.18
Mono-lingual	viBert4news	79.26	77.48	78.36
	viBert_FPT	80.65	79.12	79.88
	vELECTRA_FPT	83.08	79.54	81.27
	PhoBERT	**85.60**	**87.49**	**86.53**

THE EXPERIMENTAL RESULTS OF VARIOUS MONO-LINGUAL AND MULTI-LINGUAL PRE-TRAINED BERT MODELS ON VIETNAMESE ASPECT CATEGORY DETECTION TASK FOR THE HOTEL DOMAIN.

Types	Models	Precision	Recall	F1-score
Multi-lingual	mBERT	77.93	76.26	77.09
	mDistilBert	78.59	74.97	76.73
	XLM-R	78.86	76.56	77.70
Mono-lingual	viBert4news	79.39	74.83	77.04
	viBert_FPT	81.14	74.54	77.70
	vELECTRA_FPT	79.82	76.07	77.90
	PhoBERT	**81.49**	**76.96**	**79.16**

Tables taken from [24]

Downstream task evaluation

- Vietnamese NLI results

NLI (syllable- or word-level)	
Model	Acc.
–	–
BiLSTM-max (Conneau et al., 2018)	66.4
mBiLSTM (Artetxe and Schwenk, 2019)	72.0
multilingual BERT (Devlin et al., 2019) [■]	69.5
$XLM_{MLM+TLM}$ (Conneau and Lample, 2019)	76.6
$XLM\text{-}R_{base}$ (Conneau et al., 2020)	75.4
$XLM\text{-}R_{large}$ (Conneau et al., 2020)	79.7
$PhoBERT_{base}$	78.5
$PhoBERT_{large}$	**80.0**

Downstream task evaluation

- Vietnamese POS tagging results

POS tagging (word-level)	
Model	Acc.
RDRPOSTagger (Nguyen et al., 2014a) [♣]	95.1
BiLSTM-CNN-CRF (Ma and Hovy, 2016) [♣]	95.4
VnCoreNLP-POS (Nguyen et al., 2017) [♣]	95.9
jPTDP-v2 (Nguyen and Verspoor, 2018) [★]	95.7
jointWPD (Nguyen, 2019) [★]	96.0
XLM-R_{base} (our result)	96.2
XLM-R_{large} (our result)	96.3
$PhoBERT_{base}$	<u>96.7</u>
$PhoBERT_{large}$	**96.8**

Downstream task evaluation

- Vietnamese NER results

VLSP 2016 NER dataset

NER (word-level)	
Model	F_1
BiLSTM-CNN-CRF [♦]	88.3
VnCoreNLP-NER (Vu et al., 2018) [♦]	88.6
VNER (Nguyen et al., 2019b)	89.6
BiLSTM-CNN-CRF + ETNLP [♠]	91.1
VnCoreNLP-NER + ETNLP [♠]	91.3
$XLM\text{-}R_{base}$ (our result)	92.0
$XLM\text{-}R_{large}$ (our result)	92.8
$PhoBERT_{base}$	93.6
$PhoBERT_{large}$	**94.7**

PhoNER_COVID19 dataset

	Model	Mic-F_1	Mac-F_1
Syllable	BiL-CRF	0.906	0.858
	$XLM\text{-}R_{base}$	0.925	0.879
	$XLM\text{-}R_{large}$	0.938	0.911
Word	BiL-CRF	0.910	0.875
	$PhoBERT_{base}$	0.942	0.920
	$PhoBERT_{large}$	**0.945**	**0.931**

(See [28] for details)

Downstream task evaluation

- Vietnamese dependency parsing results

Dependency parsing (word-level)	
Model	LAS / UAS
–	–
VnCoreNLP-DEP (Vu et al., 2018) [★]	71.38 / 77.35
jPTDP-v2 [★]	73.12 / 79.63
jointWPD [★]	73.90 / 80.12
Biaffine (Dozat and Manning, 2017) [★]	74.99 / 81.19
Biaffine w/ XLM-R_{base} (our result)	76.46 / 83.10
Biaffine w/ XLM-R_{large} (our result)	75.87 / 82.70
Biaffine w/ $PhoBERT_{base}$	**78.77 / 85.22**
Biaffine w/ $PhoBERT_{large}$	77.85 / 84.32

Downstream task evaluation

- Using more pre-training data can significantly improve the quality of the pre-trained language models [3]
- PhoBERT does better than XLM-R on 5 downstream evaluation tasks
  - PhoBERT uses far fewer parameters than XLM-R: 135M (PhoBERT-base) vs. 250M (XLM-R-base); 370M (PhoBERT-large) vs. 560M (XLM-R-large)
  - XLM-R uses a 2.5TB multilingual pre-training corpus which contains 137GB of Vietnamese texts, i.e. 137 / 20 ~ 7 times bigger than the PhoBERT's monolingual pre-training corpus
  - XLM-R uses syllable-level Vietnamese texts # PhoBERT uses word-level texts

    👉 Dedicated language-specific models outperform multilingual ones

Takeaways

- PhoBERT-base and PhoBERT-large are the first public large-scale monolingual language models pre-trained for Vietnamese
- PhoBERT helps produce state-of-the-art performances on 5 downstream tasks
  - Aspect-based sentiment analysis, NLI, POS tagging, NER and Dependency parsing
  - PhoBERT outperforms XLM-R on all these tasks
- PhoBERT can serve as a strong baseline for future Vietnamese NLP research and applications: https://github.com/VinAIResearch/PhoBERT

Outline

- PhoBERT: Pre-trained language models for Vietnamese
- BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese
- VinAI Translate

Motivation

- Pre-trained masked language models (BERT [2] and its variants) have achieved state-of-the-art results in various natural language understanding tasks
- Due to a bidirectionality nature, it “might” be difficult to directly apply those pre-trained language models to natural language generation tasks, e.g. text summarization

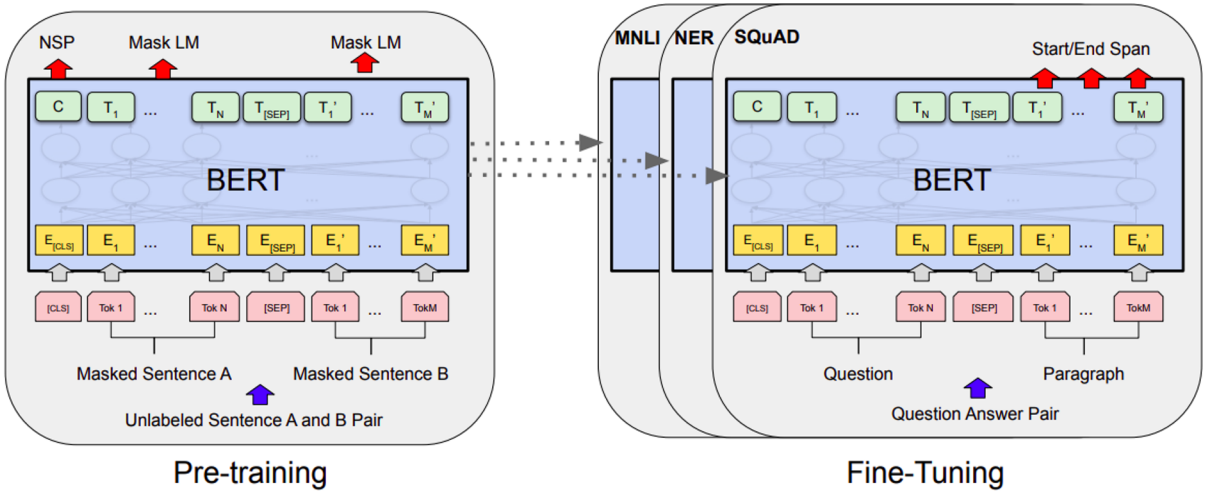


Figure taken from [2]

Motivation

- Pre-trained sequence-to-sequence models (e.g. BART [15], T5 [16]) have been proposed to obtain state-of-the-art performances for generative NLP tasks

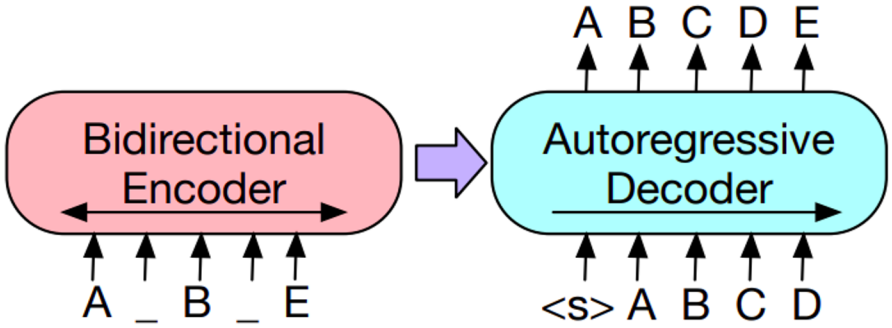
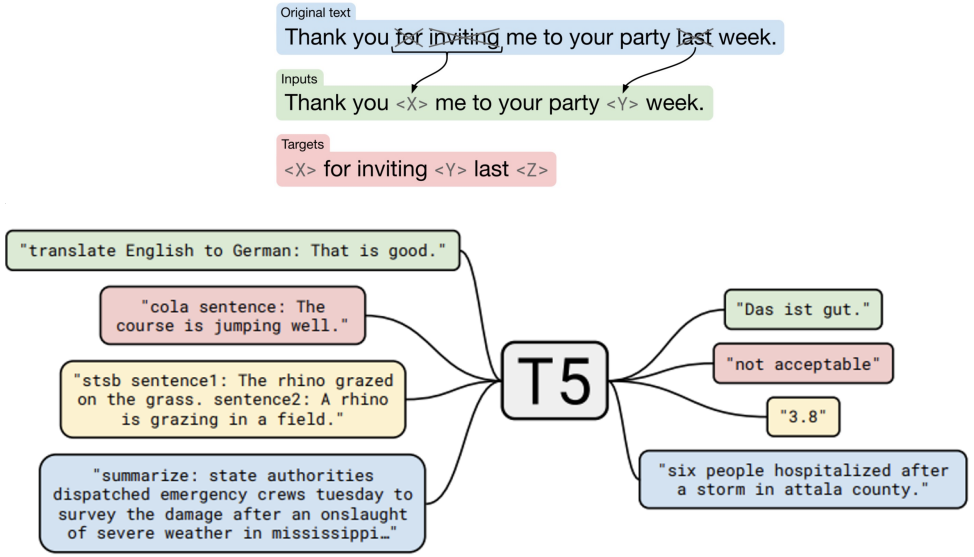


Figure taken from [15]



Figures taken from [16]

Motivation

- Public pre-trained sequence-to-sequence models used for Vietnamese:
  - Multilingual models: mBART [18], mT5 [19]
  - There is not a previous existing public monolingual model for Vietnamese
- Monolingual models are preferable as dedicated language-specific models still outperform multilingual ones

- **In our work:**
  - Introduce BARTpho—the first large-scale monolingual sequence-to-sequence model pre-trained for Vietnamese
  - Show the effectiveness of BARTpho in a comparison with mBART on Vietnamese downstream tasks: Text summarization, Capitalization and Punctuation restoration

BARTpho pre-training

- Follow BART's architecture and pre-training procedure:
  - Standard sequence-to-sequence Transformer architecture, with an additional layer-normalization layer and GeLU activation
  - Two steps of pre-training:
    - Corrupt input text by noising function

    - Learn sequence-to-sequence model to reconstruct the original input text
- Why BART?
  - Produce the strongest performances on downstream tasks in comparison to other pre-trained sequence-to-sequence models under a comparable setting in terms of the relatively equal numbers of model parameters and pre-training data sizes

BARTpho pre-training

- Employ a “large” architecture: 12-layer encoder and 12-layer decoder
- Pre-training data: 20GB Vietnamese texts
- BARTpho-word (420M parameters)
  - Word-level model, employing PhobertTokenizer for subword BPE segmentation
- BARTpho-syllable (396M parameters)
  - Syllable-level model, employing the pre-trained SentencePiece model used in mBART [18] for BPE segmentation

BARTpho pre-training

- Implementation details:
  - Initialize parameter weights of the syllable-level model by those from mBART
  - Train with a batch size of 512 sequence blocks across 8 A100 GPUs (40GB each) and a peak learning rate of 0.0001
  - Run for 15 training epochs (the learning rate is warmed up for 1.5 epochs)
- Publicly released: https://github.com/VinAIResearch/BARTpho
- BARTpho can be used with popular open-source libraries: **transformers** and **fairseq**

Text summarization evaluation

- Given a single document, create a summary that represents the most important or relevant information within the original content
- Experimental dataset: the single-document summarization dataset VNDS [30]
  - Original training / validation / test: 105418 / 22642 / 22644
  - After filtering duplication: 99134 / 22184 / 22498
- Metrics: ROUGE scores and Human evaluation
- **Main baseline mBART** [18]—the pre-trained multilingual BART variant which uses 1.3+TB pre-training data, including 137GB syllable-level Vietnamese text data

Text summarization evaluation

- Vietnamese text summarization task results w.r.t. data duplicate filtering
  - Both BARTpho versions perform better than mBART in both automatic and human evaluations

Model	#params	Validation set			Test set			
		R-1	R-2	R-L	R-1	R-2	R-L	Human
mBART	680M	60.06	28.69	38.85	60.03	28.51	38.74	21/100
BARTpho-syllable	396M	60.29	29.07	39.02	60.41	29.20	39.22	37/100
BARTpho-word	420M	**60.55**	**29.89**	**39.73**	**60.51**	**29.65**	**39.75**	**42/100**

Text summarization evaluation

- BARTpho's results are higher than previously published ones on the original test set
  - All models are fine-tuned on the original training set

Model	Data (vi)	R-1	R-2	R-L
fastAbs [*]	N/A	54.52	23.01	37.64
PhoBERT2PhoBERT [**]	20GB (4B syllables)	60.37	29.12	39.44
mT5 [**]	116B syllables	58.05	26.76	37.38
mBART	137GB (25B syllables)	60.35	29.13	39.21
BARTpho-syllable	20GB (4B syllables)	60.88	29.90	39.64
BARTpho-word	20GB (4B syllables)	**61.14**	**30.31**	**40.15**

[*]: Taken from [30]
[**]: Taken from [31]

Capitalization and punctuation restoration evaluation

- Capitalization and punctuation restoration are important steps in ASR transcript post-processing
  - Reconstruct a well-formatted text from a transcript text without information about capitalization and punctuation
  - Formulate the tasks as a sequence-to-sequence problem, taking lowercase, unpunctuated texts as input and producing true case, punctuated texts as output
- Generate an experimental dataset automatically by leveraging the Vietnamese data of the PhoST dataset [32]
  - Training / validation / test: 327370 / 1933 / 1976

Capitalization and punctuation restoration evaluation

- Capitalization and punctuation restoration F1 scores (in %) on the test set
  - BARTpho performs better than mBART in both capitalization and punctuation restoration tasks

Model	Capitalization	Punctuation restoration			
		Comma	Period	Question	Overall
mBART	91.28	67.26	**92.19**	85.71	78.71
BARTpho-syllable	91.98	67.95	91.79	**88.15**	79.09
BARTpho-word	**92.41**	**68.39**	92.05	87.82	**79.29**

Takeaways

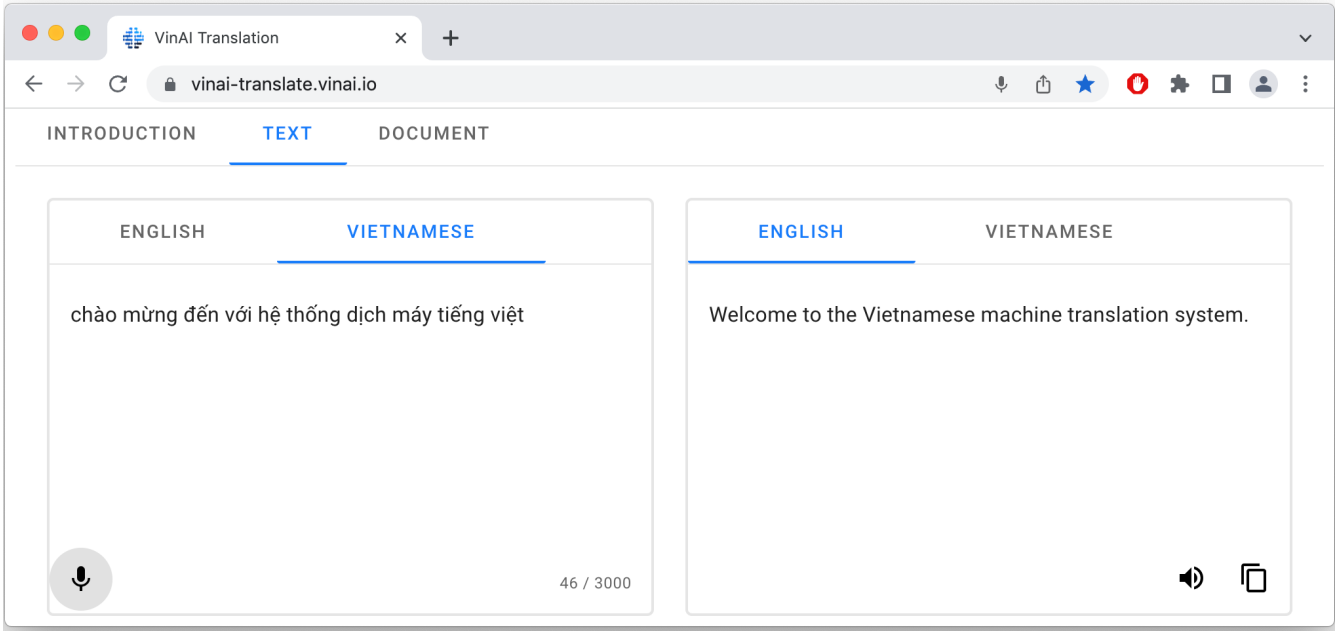
- BARTpho-word and BARTpho-syllable are the first public large-scale monolingual sequence-to-sequence models pre-trained for Vietnamese
- BARTpho performs better than its competitor mBART on 3 downstream tasks
  - Vietnamese text summarization, capitalization and punctuation restoration
  - Produce state-of-the-art performances
- BARTpho can serve as a strong baseline for future research and applications of generative Vietnamese NLP tasks: https://github.com/VinAIResearch/BARTpho
  - 09/2021: Release BARTpho-word and BARTpho-syllable
  - 08/2022: Release “base” architecture variants of BARTpho-word and BARTpho-syllable

Outline

- PhoBERT: Pre-trained language models for Vietnamese
- BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese
- VinAI Translate: Pre-trained translation models for Vietnamese-to-English and English-to-Vietnamese

Motivation

- The demand for high-quality Vietnamese-English machine translation has rapidly increased
- Strategy:
  - *Train state-of-the-art machine translation models on a high-quality and large-scale parallel dataset*



Motivation

- Issues:
  - High-quality Vietnamese-English parallel corpora are either not publicly available or small-scale
  - Larger Vietnamese-English parallel corpora are noisy
- Approach:
  - Construct a high-quality and large-scale parallel dataset
    - PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation [33]
  - Fine-tune a strong pre-trained sequence-to-sequence model on this dataset
    - mBART: Multilingual Denoising Pre-training for Neural Machine Translation [18]

Dataset construction

- Construct PhoMT—a high-quality and large-scale Vietnamese-English parallel dataset
  1. Collecting parallel document pairs
  2. Pre-processing
  3. Aligning parallel sentence pairs
  4. Post-processing

Dataset construction

- PhoMT: Collecting parallel document pairs

Dataset construction

- PhoMT: Pre-processing
  - Manually inspect and remove low-quality document pairs from OpenSubtitles domain
  - Filter English paragraphs inside Vietnamese documents (and vice versa)
  - Perform sentence segmentation using VnCoreNLP [34] and Stanford CoreNLP [35]
- PhoMT: Align parallel sentence pairs
  - Translate English source sentences into Vietnamese using Google Translate
  - Align between translated source sentences and target sentences using 3 toolkits: Hunalign [36], Gargantua [37], Bleualign [38]
  - Select pairs that are aligned by at least 2/3 toolkits

Dataset construction

- PhoMT: Post-processing
  - Split the dataset into train/validation/test sets
  - Manually inspect validation and test sets and remove misaligned and low-quality sentence pairs (0.8%)
- PhoMT: A high-quality and large-scale Vietnamese-English parallel dataset consisting of 3.02M pairs

Domain	Total		Training			Validation			Test		
	#doc	#pair	#pair	#en/s	#vi/s	#pair	#en/s	#vi/s	#pair	#en/s	#vi/s
News	2559	41504	40990	24.4	32.0	257	22.3	30.3	257	26.8	34.5
Blogspot	1071	93956	92545	25.0	34.6	597	26.4	37.8	814	23.7	31.5
TED-Talks	3123	320802	316808	19.8	23.8	1994	20.0	24.6	2000	22.0	27.9
MediaWiki	38969	496799	490505	26.0	32.8	3024	25.3	32.3	3270	27.0	33.7
WikiHow	6616	513837	507379	18.9	22.4	3212	17.9	21.5	3246	17.5	21.5
OpenSub	3312	1548971	1529772	9.7	11.1	9635	9.5	10.7	9564	10.0	11.4
All	55650	3015869	2977999	15.7	19.0	18719	15.3	18.7	19151	16.2	19.8

https://github.com/VinAIResearch/PhoMT

Dataset construction

- Fine-tune mBART on the PhoMT training set of ~3M pairs for English-to-Vietnamese
- From each English-Vietnamese sentence pair in “noisy” datasets CCAligned [39] and WikiMatrix [40]
  - Employ the fine-tuned model to translate the English sentence into Vietnamese
  - Select pairs with a BLEU score between the Vietnamese-translated variant and the Vietnamese target sentence ranging from 0.15 to 0.95, resulting in 6M pairs
- A collection of 3M + 6M = 9M “high-quality” sentence pairs
- Simulate the ASR output: Lowercase and remove punctuations from the source sentences while keeping the target sentences intact, obtaining 9M pairs for each translation direction
- For each translation direction: 9M + 9M = 18M sentence pairs

Pre-trained VinAI Translate models

- Fine-tune mBART for each translation direction using 18M sentence pairs
  - Reduce mBART vocabulary from 250K tokens to 90K tokens belonging to English and Vietnamese
- Publicly released: https://github.com/VinAIResearch/VinAI_Translate

Model	#params	Max length
vinai/vinai-translate-vi2en	448M	1024
vinai/vinai-translate-en2vi	448M	1024

- Pre-trained VinAI Translate models can be used with the popular open-source library **transformers**
- These pre-trained models are currently used in the translation component of the VinAI Translate system [41]: https://vinai-translate.vinai.io
- Users can also try these models at: https://huggingface.co/spaces/vinai/VinAI_Translate

PhoMT evaluation results

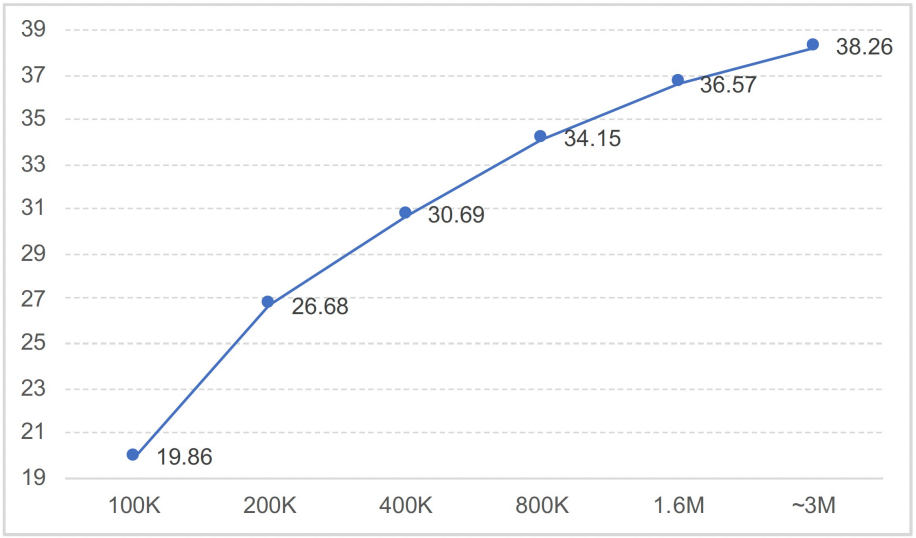
- Experimental results on the PhoMT validation and test sets while using the PhoMT training set of 2.97M pairs for training

Model	Validation set				Test set					
	En-to-Vi		Vi-to-En		En-to-Vi			Vi-to-En		
	TER↓	BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	Human↑	TER↓	BLEU↑	Human↑
Google Translate	45.86	40.10	44.69	36.89	46.52	39.86	23/100	45.86	35.76	10/100
Bing Translator	45.36	40.82	45.32	36.61	46.04	40.37	14/100	46.09	35.74	15/100
Transformer-base	42.77	43.01	43.42	38.26	43.79	42.12	13/100	44.28	37.19	13/100
Transformer-big	42.13	43.75	43.08	39.04	43.04	42.94	18/100	44.06	37.83	28/100
mBART	**41.56**	**44.32**	**41.44**	**40.88**	**42.57**	**43.46**	**32/100**	**42.54**	**39.78**	**34/100**

- mBART achieves the best performances, in both translation directions and on all metrics
- Neural MT baselines outperform automatic translation engines

PhoMT evaluation results

- BLEU scores of Transformer-base on the Vi- to-En validation set when varying training sizes on PhoMT



- Sample a set of 1.55M non-duplicate Vietnamese-English sentence pairs from OPUS's OpenSubtitles, which has the same size as the PhoMT's OpenSubtitles training subset:
  - OPUS's OpenSubtitles: 29.72 BLEU
  - PhoMT's OpenSubtitles: 31.11 BLEU

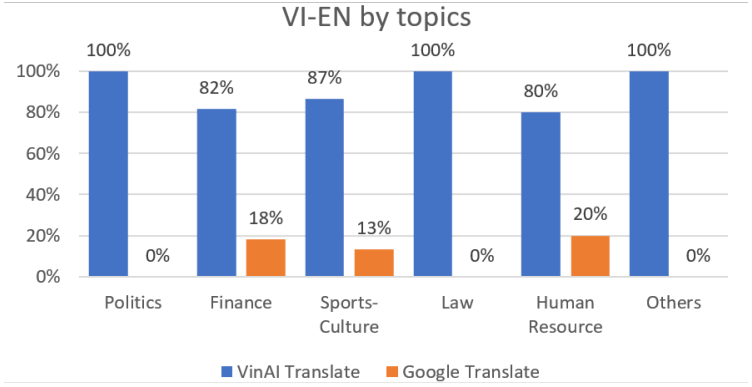
Our curation effort paid off!

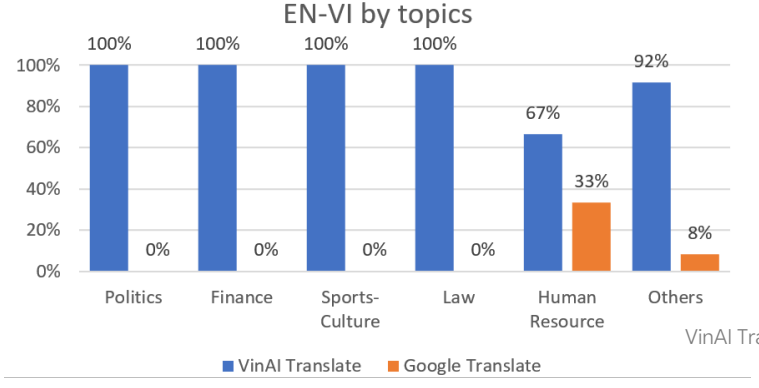
VinAI Translate evaluation results

- Automatic evaluation results

Model	Validation set		Test set	
	EN-VI	VI-EN	EN-VI	VI-EN
Google Translate	40.10	36.89	39.86	35.76
PhoMT	44.32	40.88	43.46	39.78
VinAI Translate	**45.31**	**41.41**	**44.29**	**40.42**

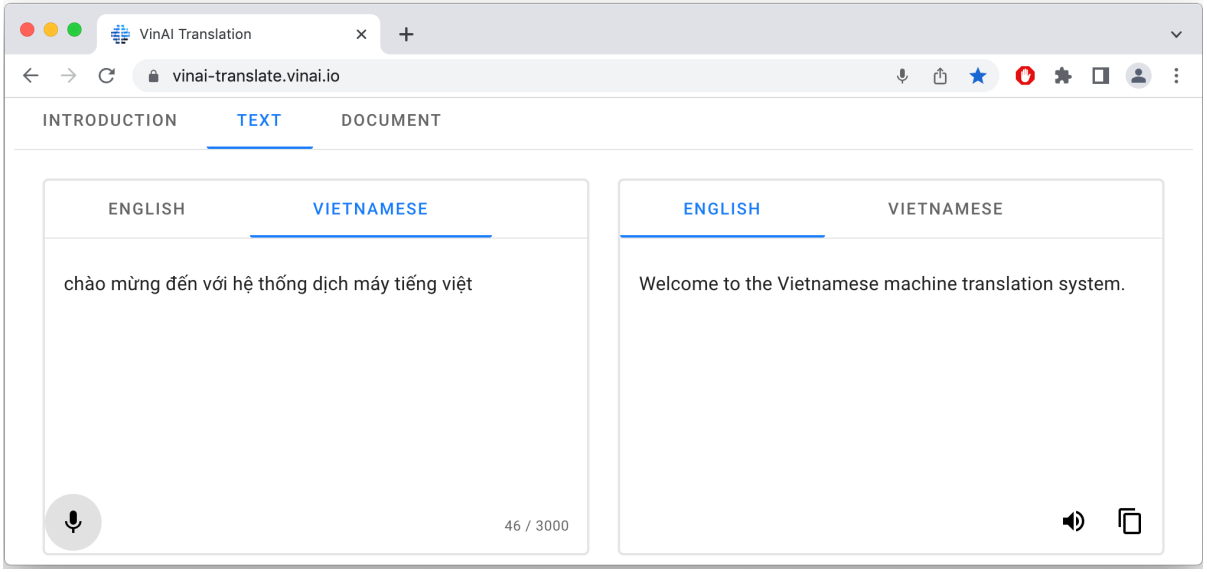
- Human evaluation results





Takeaways

- PhoMT—A high-quality and large-scale Vietnamese-English parallel dataset: https://github.com/VinAIResearch/PhoMT
- State-of-the-art translation models pre-trained for Vietnamese-to-English and English-to-Vietnamese: https://github.com/VinAIResearch/VinAI_Translate



AI Day 2022 Empower the Future

Thank you!

Public resources for Vietnamese NLP from VinAI

- PhoST (INTERSPEECH 2022): A high-quality and large-scale dataset for English-Vietnamese speech translation.
- VinAI_Translate (INTERSPEECH 2022): Pre-trained text translation models for Vietnamese-to-English and English-to-Vietnamese.
- BARTpho (INTERSPEECH 2022): Pre-trained sequence-to-sequence models for Vietnamese.
- QA-CarManual (IUI 2022): Demo video of a Vietnamese speech-based question answering over car manuals.
- PhoMT (EMNLP 2021): A high-quality and large-scale benchmark dataset for Vietnamese-English machine translation.
- PhoATIS (INTERSPEECH 2021): An intent detection and slot filling dataset for Vietnamese.
- PhoNLP (NAACL 2021): A BERT-based multi-task learning toolkit for Vietnamese POS tagging, named entity recognition and dependency parsing.
- PhoNER_COVID19 (NAACL 2021): A dataset for Vietnamese named entity recognition.
- ViText2SQL (EMNLP 2020 Findings): A dataset for Vietnamese Text2SQL semantic parsing.
- PhoBERT (EMNLP 2020 Findings): Pre-trained language models for Vietnamese.
- PhoW2V (2020): Pre-trained Word2Vec syllable- and word-level embeddings for Vietnamese.

VinAI

References

1. Attention Is All You Need
2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
3. RoBERTa: A Robustly Optimized BERT Pretraining Approach
4. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators
5. Unsupervised Cross-lingual Representation Learning at Scale
6. PhoBERT: Pre-trained language models for Vietnamese
7. Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models
8. Improving Language Understanding by Generative Pre-Training
9. Language Models are Unsupervised Multitask Learners
10. Language Models are Few-Shot Learners
11. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation
12. LaMDA: Language Models for Dialog Applications
13. Few-shot Learning with Multilingual Language Models
14. BigScience Large Open-science Open-access Multilingual Language Model
15. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

References

16. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
17. ByT5: Towards a token-free future with pre-trained byte-to-byte models
18. Multilingual Denoising Pre-training for Neural Machine Translation
19. mT5: A massively multilingual pre-trained text-to-text transformer
20. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese
21. ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation
22. A Vietnamese-English Neural Machine Translation System
23. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing
24. Investigating Monolingual and Multilingual BERTModels for Vietnamese Aspect Category Detection
25. Deep Biaffine Attention for Neural Dependency Parsing
26. XNLI: Evaluating Cross-lingual Sentence Representations
27. VLSP Shared Task: Named Entity Recognition
28. COVID-19 Named Entity Recognition for Vietnamese
29. From Treebank Conversion to Automatic Dependency Parsing for Vietnamese
30. VNDS: A Vietnamese Dataset for Summarization

References

31. VieSum: How Robust Are Transformer-based Models on Vietnamese Summarization?
32. A High-Quality and Large-Scale Dataset for English-Vietnamese Speech Translation
33. PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation
34. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit
35. The Stanford CoreNLP Natural Language Processing Toolkit
36. Parallel corpora for medium density languages
37. Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora
38. Iterative, MT-based Sentence Alignment of Parallel Texts
39. CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs
40. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia
41. A Vietnamese-English Neural Machine Translation System