AI DAY 2020 RISING TO THE CHALLENGES

PhoBERT: Pre-trained language models for Vietnamese

SPEAKER:

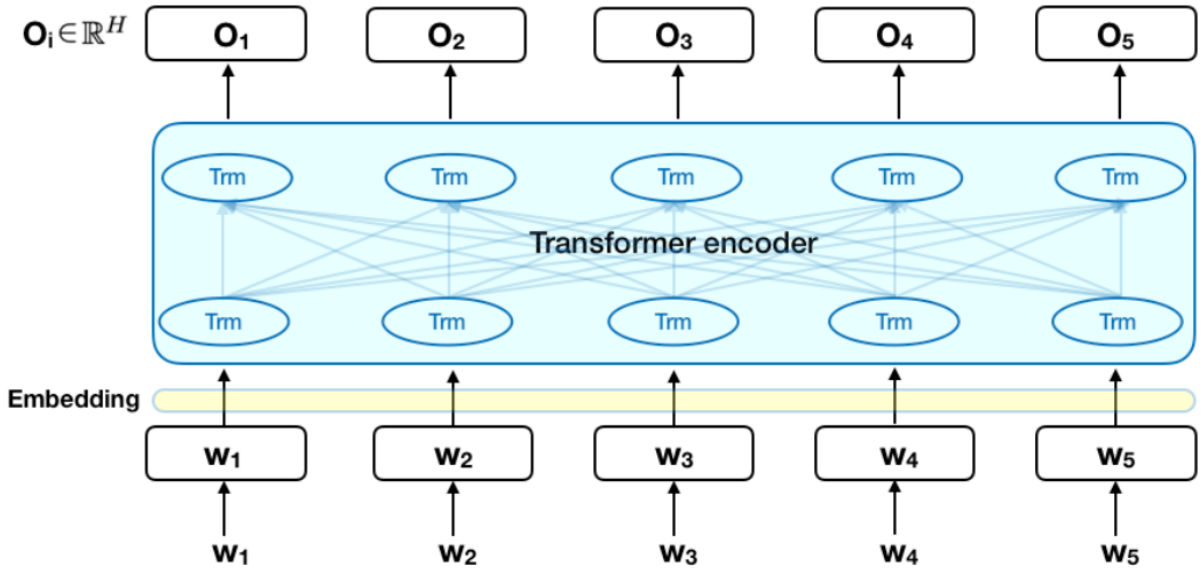
Dat Quoc Nguyen – VinAI Research, Vietnam

Joint work with **Anh Tuan Nguyen** – Former VinAI intern – Research engineer at NVIDIA, USA

VinAI RESEARCH

Motivation

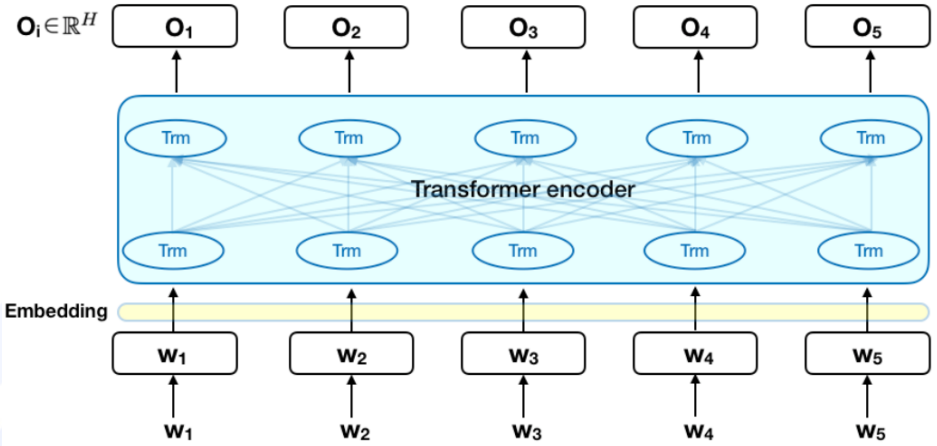
- Language model BERT—Bidirectional Encoder Representations from Transformers (Devlin et al., 2019)—is a recent breakthrough in NLP
  - BERT and its variants, pretrained on large-scale corpora, help improve the state-of-the-art performances of various NLP research & application tasks
  - Represent words by embedding vectors which encode the contexts where the words appear, i.e. contextualized word embeddings



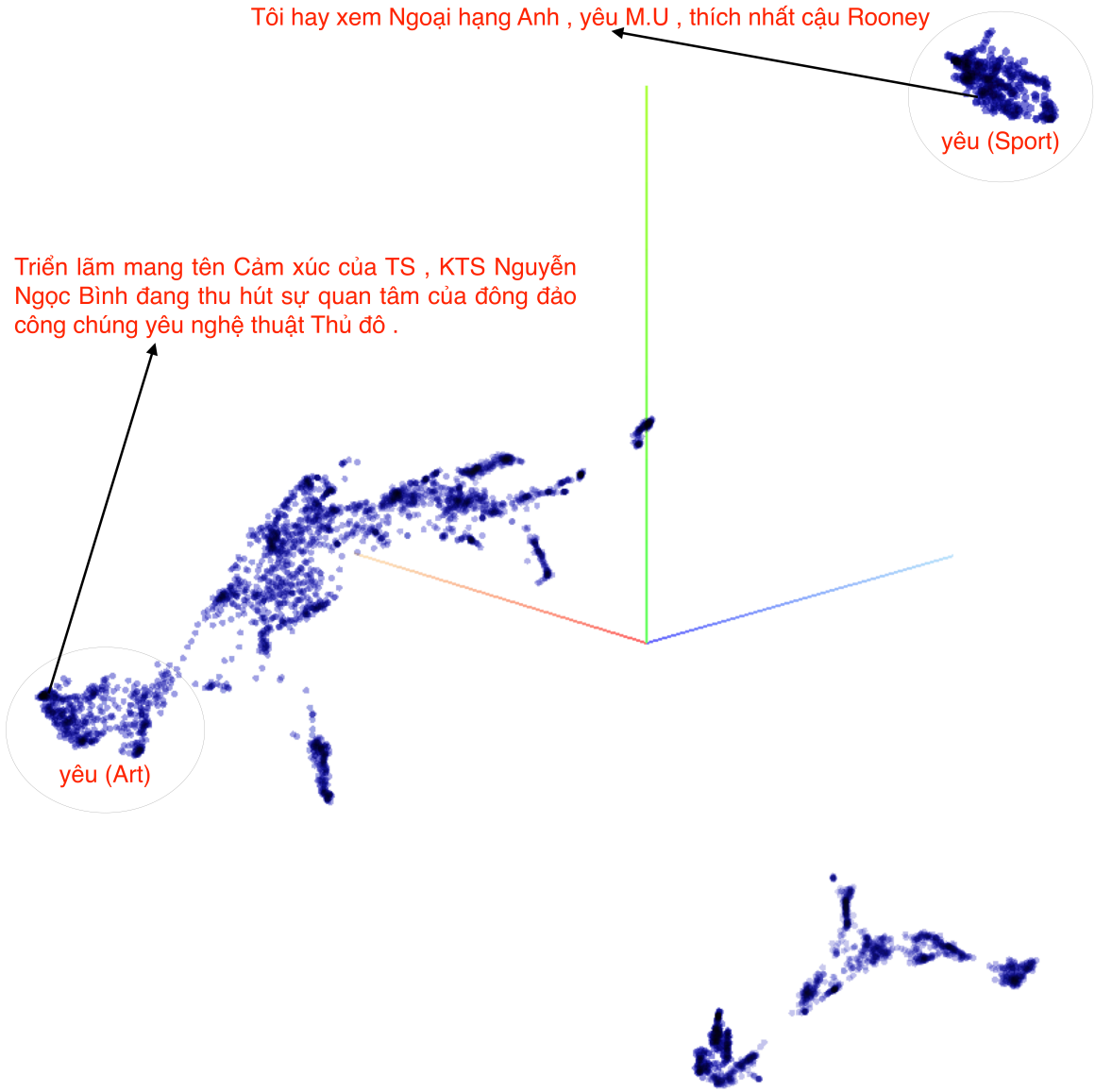
https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

Motivation

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



https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png



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

Motivation

- The success of BERT and its variants has largely been limited to the English language
  - Most pre-trained BERT-based models were learnt 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*
- Vietnamese language: 85% of Vietnamese word types are composed of at least 2 syllables (âm/tiếng)

Syllable-level *VinAI công bố các kết quả nghiên cứu khoa học tại hội nghị hàng đầu thế giới về trí tuệ nhân tạo*

Word-level *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

- Public pre-trained monolingual BERT-based language models for Vietnamese:
  - Used the Vietnamese Wikipedia corpus which is relatively small (**1GB**)
    (Note that pre-trained models can be significantly improved by using more data)
  - Trained at the syllable level, i.e. 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

Syllable-level *VinAI công bố các kết quả nghiên cứu khoa học tại hội nghị hàng đầu thế giới về trí tuệ nhân tạo*

Word-level *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)

PhoBERT for Vietnamese

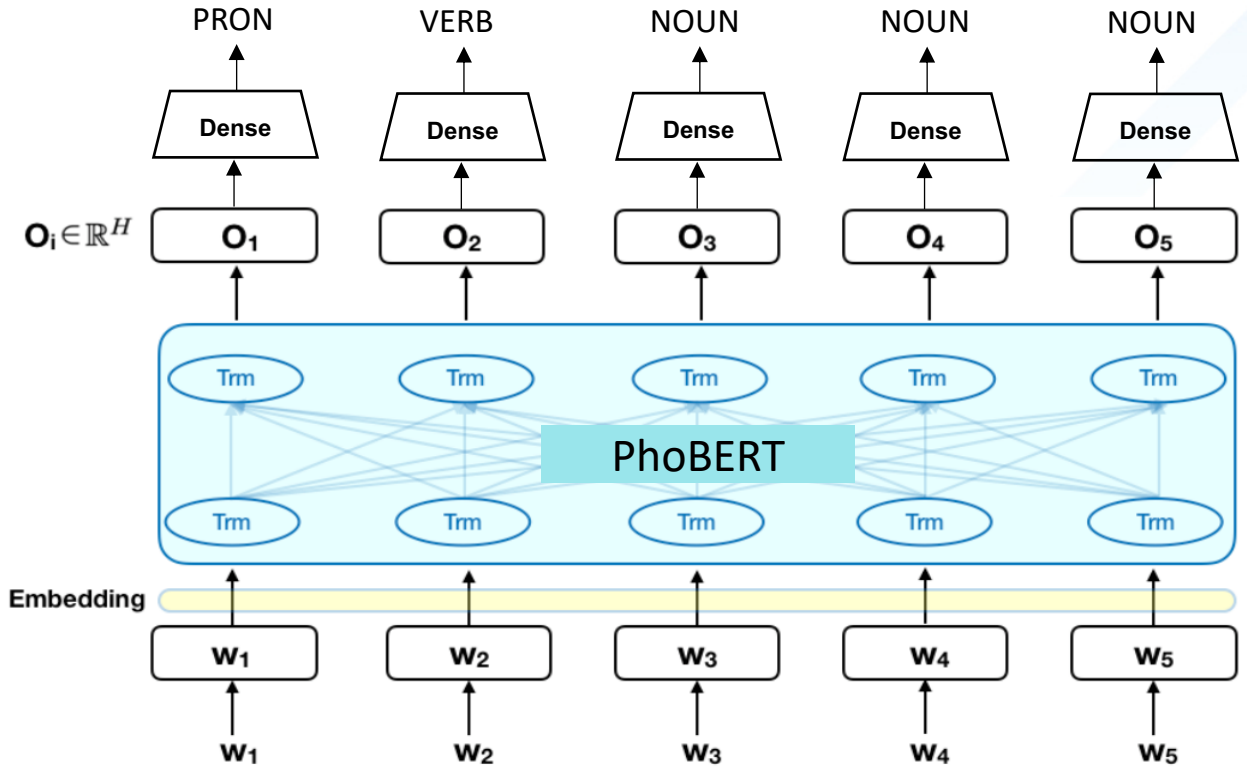
- How VinAI trains PhoBERT to handle previous concerns:
  - Used a large-scale corpus of 20GB Vietnamese texts
  - Performed Vietnamese word segmentation before pre-training

  👉 Pre-training corpus of 145M word-segmented sentences (3B word tokens)
- PhoBERT pre-training procedure is based on RoBERTa (Liu et. al., 2019) which optimizes BERT for more robust performance
- Two versions: PhoBERT-base (150M parameters) & PhoBERT-large (350M parameters)
- Pre-trained PhoBERT using 4 GPUs V100 16GB memory each in 8 weeks
- Publicly released under MIT license: https://github.com/VinAIResearch/PhoBERT
- PhoBERT can be used with popular open source libraries: transformers and fairseq

PhoBERT: Evaluation on 5 Vietnamese NLP tasks

- **Part-of-Speech (POS) tagging**: To assign a lexical category tag to each word in a text
  - Use the benchmark from the VLSP 2013 POS tagging task
  - 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

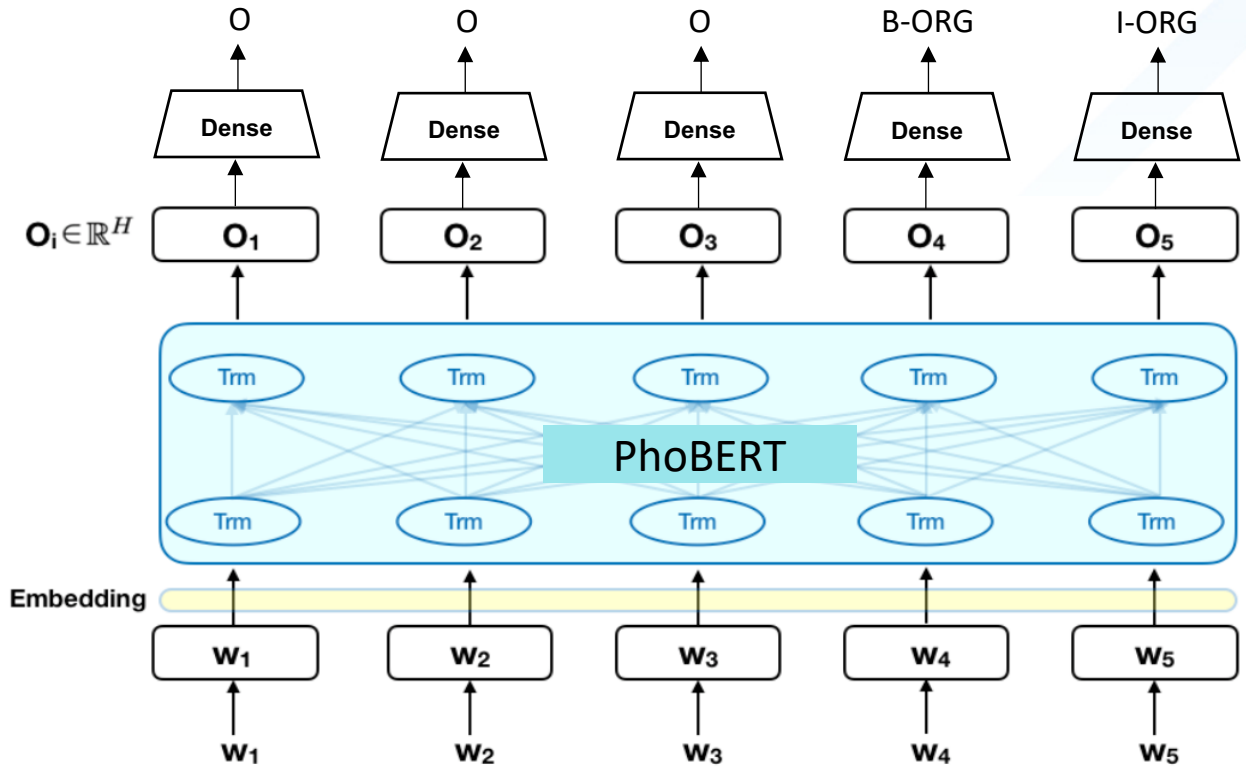


Drawn based on https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

PhoBERT: Evaluation on 5 Vietnamese NLP tasks

- **Named entity recognition (NER)**: To identify personal names, locations, organizations and the like
  - Use the benchmark from the VLSP 2016 NER task (Nguyen et al., 2019)
  - 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



Drawn based on https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

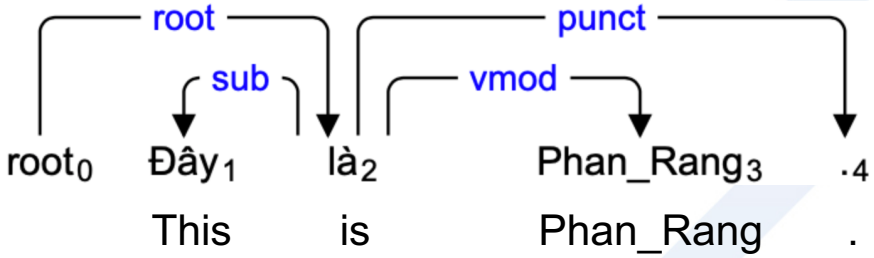
PhoBERT: Evaluation on 5 Vietnamese NLP tasks

- **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 the Vietnamese data from the cross-lingual NLI corpus v1.0 (Conneau et al., 2018, Williams et al.,2018)
  - 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]”

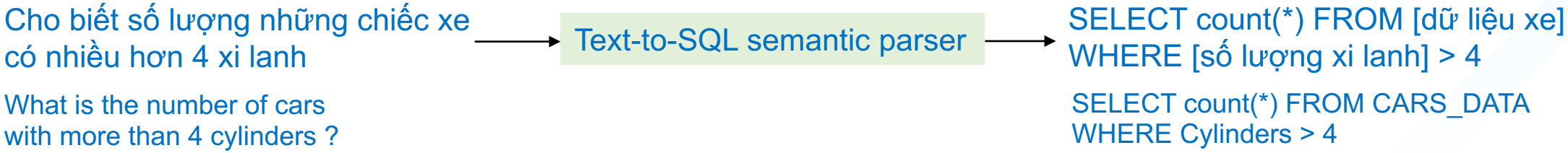
(Dark red is the premise while dark blue is the hypothesis)

- **Dependency parsing**: To analyze the syntactic structure of a sentence by identifying grammatical relationships between "head" words and words which modify those heads
  - Use the benchmark VnDT treebank (Nguyen et al., 2014)
  - Extend the Biaffine parser (Dozat and Manning, 2017) with the input PhoBERT-based contextualized word embeddings



PhoBERT: Evaluation on 5 Vietnamese NLP tasks

- **Text-to-SQL semantic parsing**: To convert natural language statements into meaning representations of standard SQL database queries



  - Access information stored in databases via natural language statements
  - Users do not need to understand SQL query syntax as well as database schemas
  - *We created a large-scale Text-to-SQL dataset for the Vietnamese semantic parsing task: ~10K question and SQL query pairs over ~200 databases*
  - Extend strong sequence-to-sequence baseline parsers EditSQL (Zhang et al.,2019) and IRNet (Guo et al., 2019) with the input PhoBERT-based contextualized word embeddings
- **Baseline XLM-R** (Conneau et al., 2020)—the recent best multilingual pre-trained model which uses 2.5 TB pre-training data including 137GB syllable-level Vietnamese text data

Main evaluation results

- Vietnamese POS tagging and NER 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**

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-R_{base} (our result)	92.0
XLM-R_{large} (our result)	92.8
$PhoBERT_{base}$	<u>93.6</u>
$PhoBERT_{large}$	**94.7**

Main evaluation results

- Vietnamese NLI and Dependency parsing 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**

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/ $XLM\text{-}R_{base}$ (our result)	76.46 / 83.10
Biaffine w/ $XLM\text{-}R_{large}$ (our result)	75.87 / 82.70
Biaffine w/ $PhoBERT_{base}$	**78.77 / 85.22**
Biaffine w/ $PhoBERT_{large}$	77.85 / 84.32

Main evaluation results

- Exact matching accuracies for Vietnamese Text-to-SQL semantic parsing
  - Automatic Vietnamese word segmentation improves the performances of the baselines

	Approach	dev	test	Approach	dev	test
Syllable	w/ EditSQL	28.6	24.1	w/ IRNet	43.3	38.2
	w/ EditSQL w/ XLM-R_{base}	55.2	51.3	w/ IRNet w/ XLM-R_{base}	58.6	52.8
Word	w/ EditSQL	33.7	30.2	w/ IRNet	49.7	43.6
	w/ EditSQL w/ PhoBERT$_{base}$	56.7	52.6	w/ IRNet w/ PhoBERT$_{base}$	60.2	53.2

Main evaluation results

- Using more pre-training data can significantly improve the quality of the pre-trained language models (Liu et al., 2019):
  - Not surprising that PhoBERT helps produce better performance than ETNLP on NER, and the multilingual BERT and $XLM_{MLM+TLM}$ on NLI
- PhoBERT does better than XLM-R on all five 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 Vietnamese texts

  👉Dedicated language-specific models (PhoBERT) still outperform multilingual ones (XLM-R)

Conclusion

- PhoBERT with two versions 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 five downstream tasks: POS tagging, NER, NLI, Dependency parsing and Text-to-SQL semantic parsing
- The first set of experiments to compare monolingual language models with the recent best multilingual model XLM-R in five different language-specific tasks
  - 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

Thanks for your attention!

We are recruiting:

https://www.vinai.io/careers

Text-to-SQL semantic parsing: Exact matching accuracy categorized by different hardness levels and F1 scores of different SQL components on the test set

Approach	Easy	Medium	Hard	ExH	SELECT	WHERE	ORDER BY	GROUP BY	KEYWORDS
EditSQL w/ XLM-R_{base}	75.1	56.2	45.3	22.4	82.7	60.3	70.7	67.2	79.8
EditSQL w/ PhoBERT$_{base}$	75.6	58.0	47.4	22.7	83.3	61.8	72.5	67.9	80.6
IRNet w/ XLM-R_{base}	76.2	57.8	46.8	23.5	83.5	59.1	74.4	68.3	80.5
IRNet w/ PhoBERT$_{base}$	76.8	57.5	47.2	24.8	84.5	59.3	76.6	68.2	80.3