Recent advances in Vietnamese language modeling and understanding

Dat Quoc Nguyen – VinAI Research, Vietnam

(12/11/2020 - Updated 19/04/2021)

VinAI RESEARCH

1st conference papers @ 1st KSE 2009

1. Dat Quoc Nguyen, Dai Quoc Nguyen, Son Bao Pham and The Duy Bui. **2009**. A Fast Template-based Approach to Automatically Identify Primary Text Content of a Web Page. In *Proceedings of KSE*, pages 232-236.
2. Dai Quoc Nguyen, Dat Quoc Nguyen and Son Bao Pham. **2009**. A Vietnamese Question Answering System. In *Proceedings of KSE*, pages 26-32.
   - **The first ontology-based question answering system for Vietnamese**

KSE 2009: Small-scale structured domain-specific knowledge
KSE 2020: Large-scale; unstructured; domain-general ➔ Pre-trained LMs

KSE 2009: Rule templates to parse input questions
KSE 2020: Neural semantic parsing

Outline

- PhoBERT: Pre-trained language models for Vietnamese
  - Dat Quoc Nguyen and Anh Tuan Nguyen. **2020**. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the ACL: EMNLP 2020*.

  External results taken from:

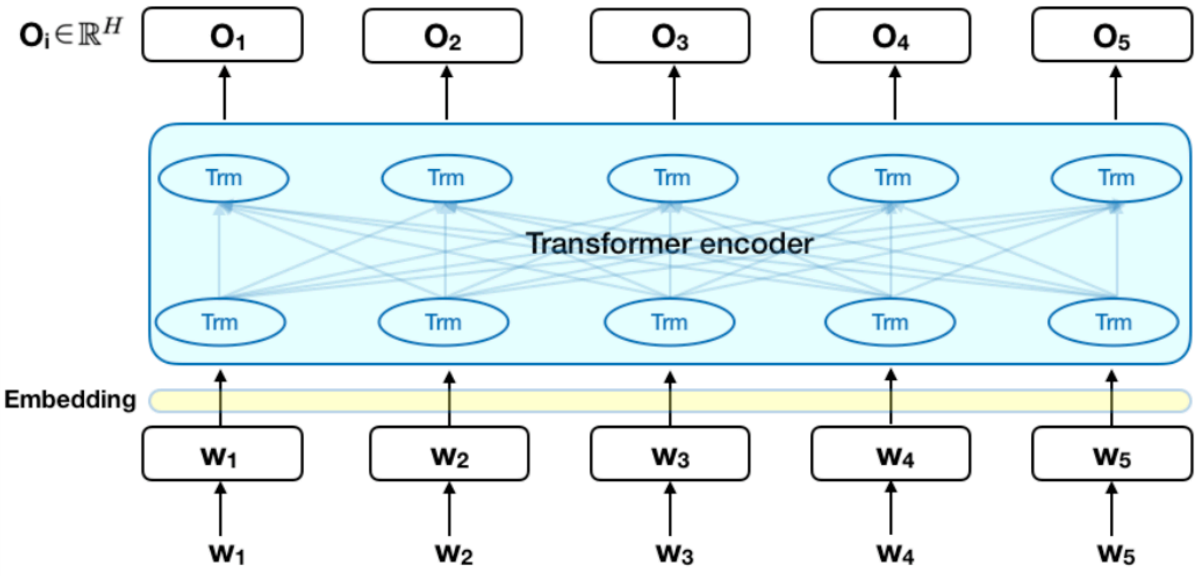
  - Thinh Hung Truong, Mai Hoang Dao and Dat Quoc Nguyen. **2021**. COVID-19 named entity recognition for Vietnamese. In *Proceedings of NAACL 2021*, to appear.
  - Dang Van Thin, Lac Si Le, Vu Xuan Hoang and Ngan Luu-Thuy Nguyen. **2021**. Investigating Monolingual and Multilingual BERT Models for Vietnamese Aspect Category Detection. *ArXiv Preprint*, arXiv:2103.09519.
- PhoNLP: A PhoBERT-based multi-task learning model for Vietnamese Part-of-Speech tagging, Named entity recognition and Dependency parsing
  - Linh The Nguyen and Dat Quoc Nguyen. **2021**. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. *In Proceedings of NAACL 2021 Demonstrations, to appear.*
- Text-to-SQL semantic parsing for Vietnamese
  - Anh Tuan Nguyen, Mai Hoang Dao and Dat Quoc Nguyen. **2020**. A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese. In *Findings of the ACL: EMNLP 2020*.

Outline

- **PhoBERT: Pre-trained language models for Vietnamese**
- PhoNLP: A PhoBERT-based multi-task learning model for Vietnamese Part-of-Speech tagging, Named entity recognition and Dependency parsing
- Text-to-SQL semantic parsing for Vietnamese

Motivation

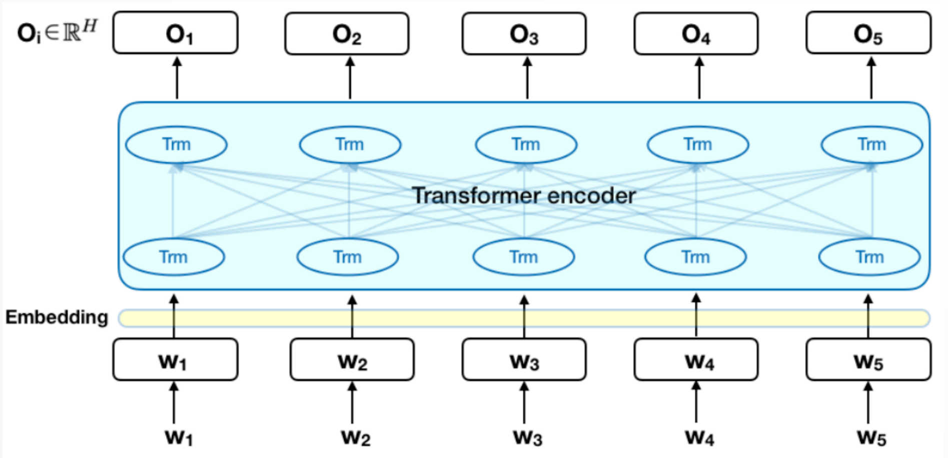
- Language model BERT—Bidirectional Encoder Representations from Transformers (Devlin et al., 2019)—is a recent breakthrough in NLP
  - BERT and its variants, pretrained on large-scale corpora, help improve the state-of-the-art performances of various NLP research & application tasks
  - Represent words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



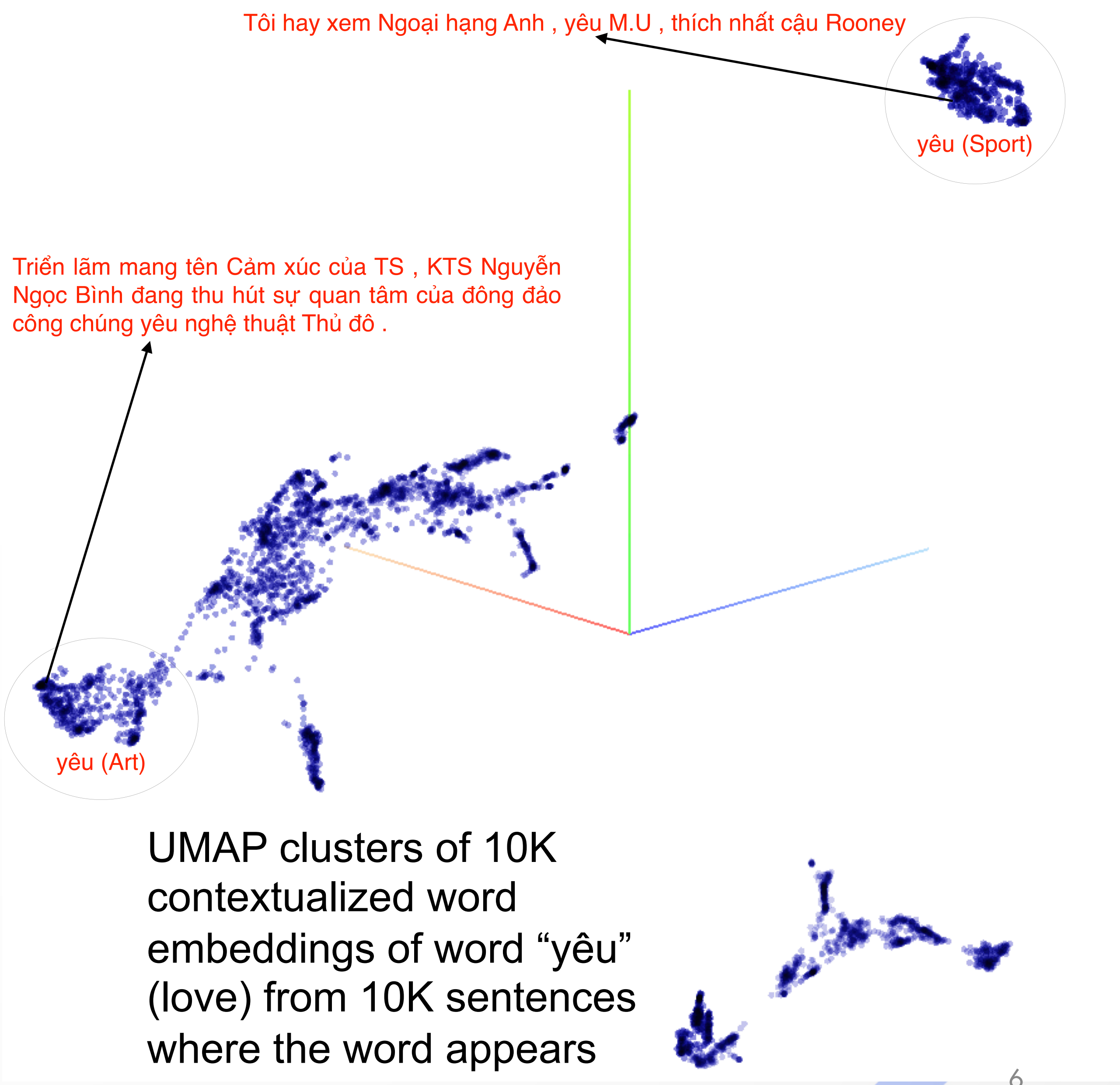
https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

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 English
  - 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*
- 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

- 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

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:
  - 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*

Downstream task evaluation

- **Aspect-based sentiment analysis:** To identify the aspect categories mentioned in user-generated reviews from a set of pre-defined categories (Thin et al., 2021)
  - Use a linear prediction layer on top of the PhoBERT output for the classification token [CLS]—the first token of the input sequence

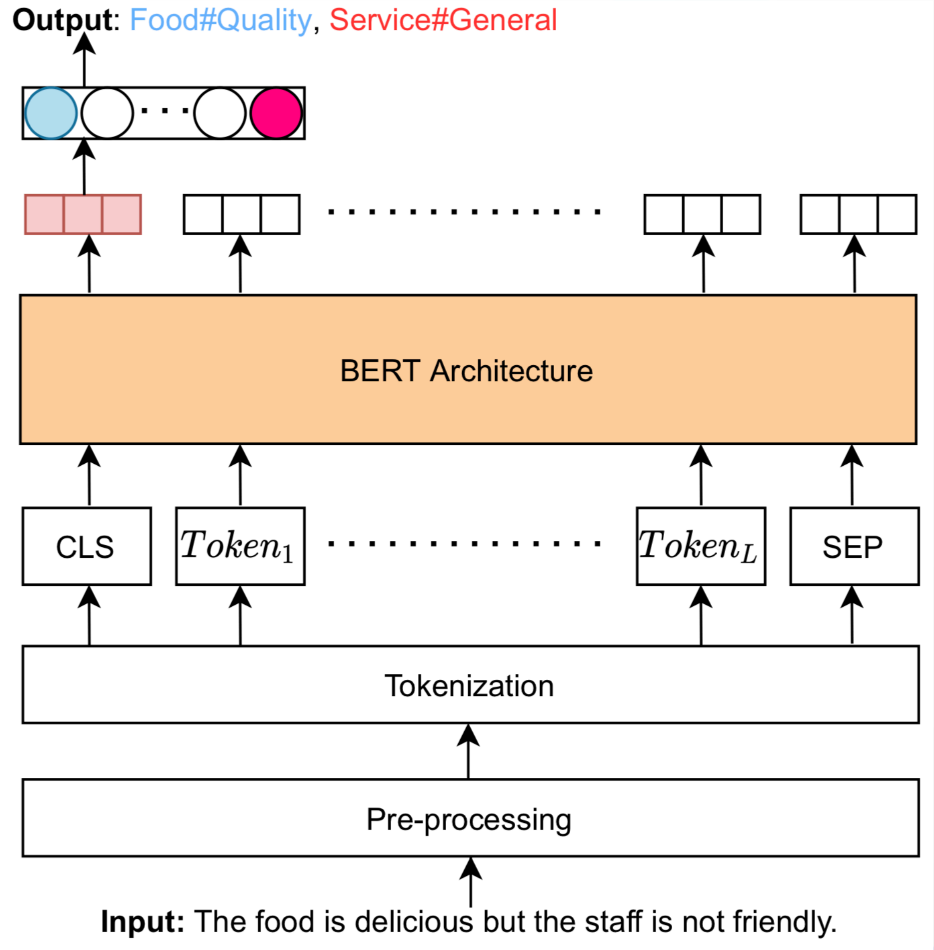


Figure taken from Thin et al. (2021)

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"

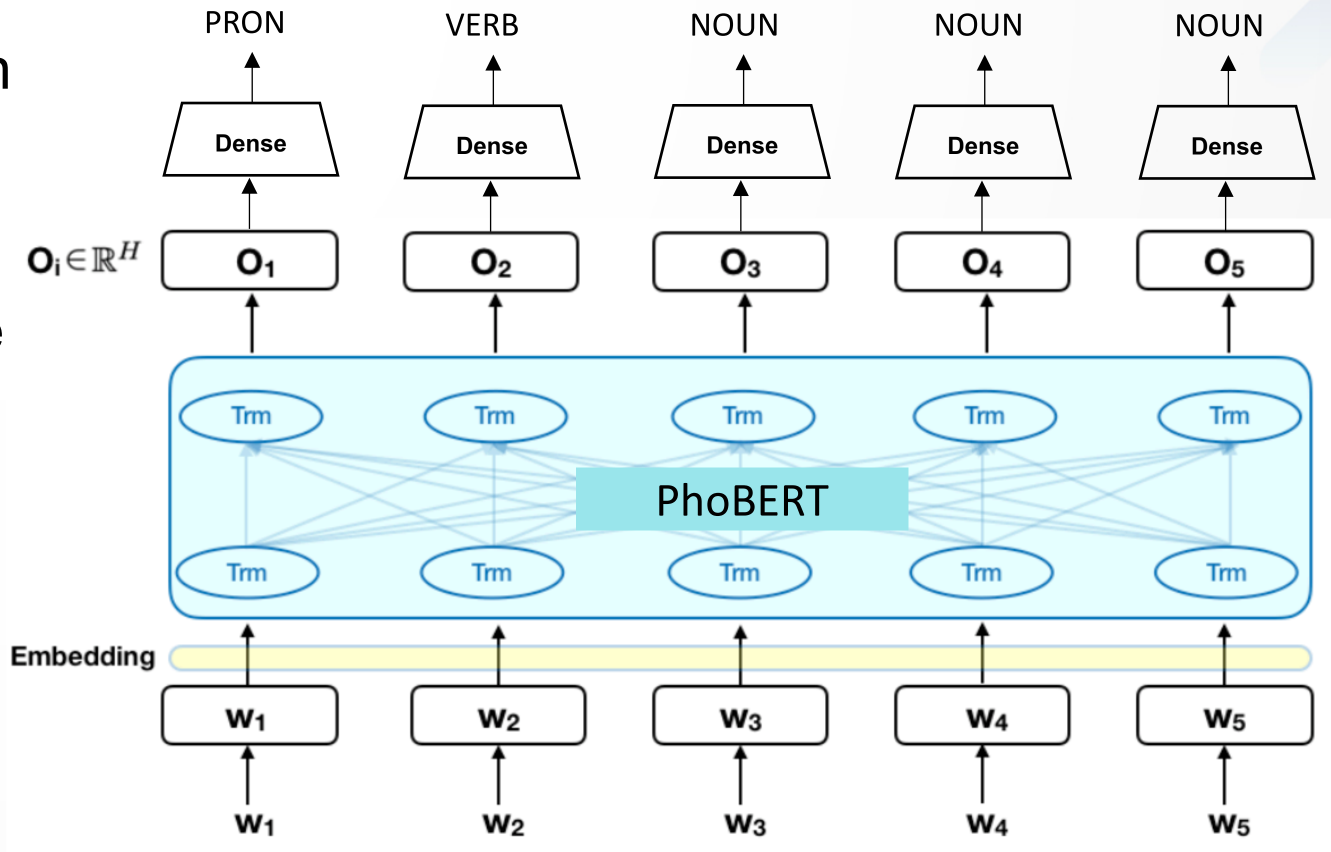
True (Entailment): "[CLS] Thông báo phản đối luật sư và tòa án hoặc cơ quan hành chính sẽ phải được gửi đi [SEP] [SEP] Ban cố vấn độc lập và toà án sẽ nhận được thông báo [SEP]"

(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

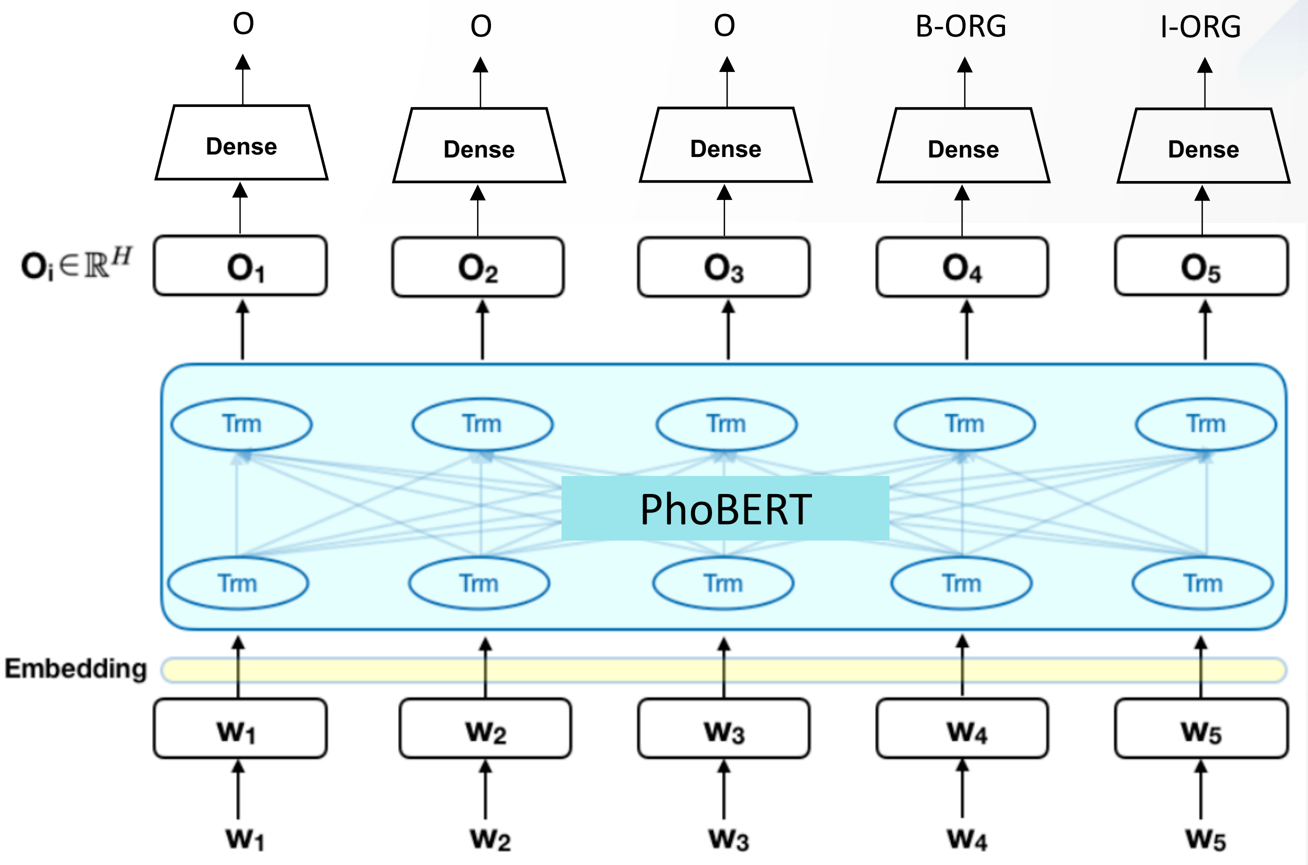


Drawn based on https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

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

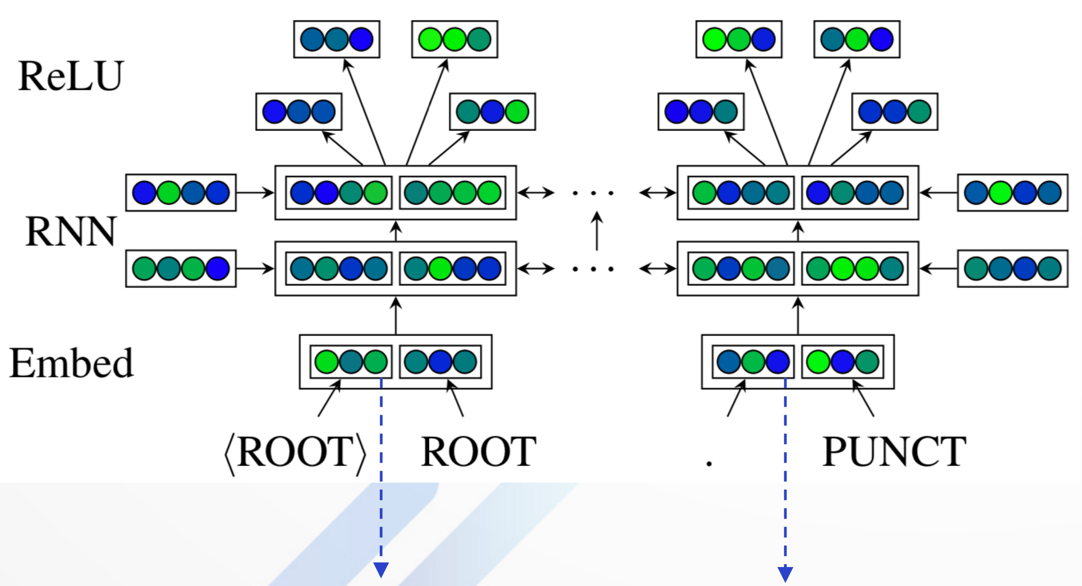


Drawn based on https://www.lyrn.ai/wp-content/uploads/2018/11/transformer.png

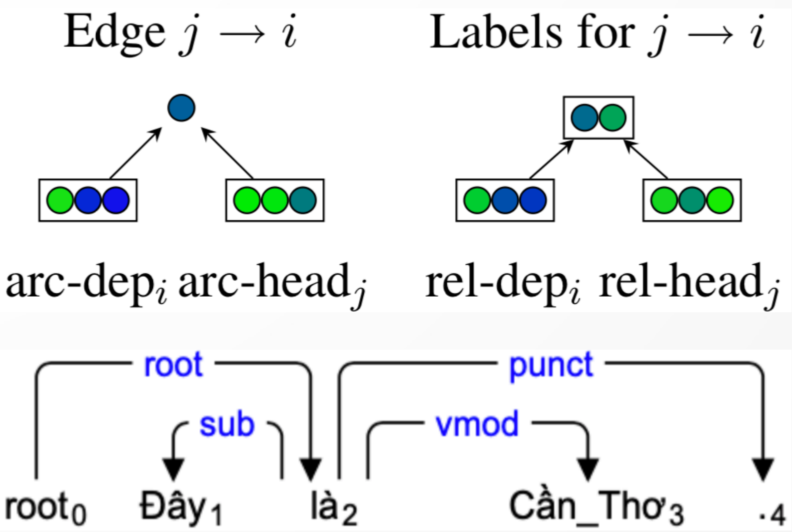
Downstream task evaluation

- **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 (Dozat and Manning, 2017) with the PhoBERT-based contextualized word embeddings as part of the input

Figures from Dozat et.al. (2017)



PhoBERT-based contextualized word embeddings



This is Can_Tho .

Biaffine classifier

Downstream task evaluation

- Experimental datasets
  - *Aspect-based sentiment analysis*: Two large corpora for Vietnamese aspect-based sentiment analysis at sentence level (Thin et al., 2021)
  - *NLI*: The Vietnamese data from the cross-lingual NLI corpus v1.0 (Conneau et al., 2018, Williams et al., 2018)
  - *POS tagging*: The VLSP 2013 POS tagging task
  - *NER*: The VLSP 2016 NER task's dataset (Nguyen et al., 2019); PhoNER_COVID19 (Truong et al., 2021)
  - *Dependency parsing*: The VnDT treebank (Nguyen et al., 2014)

- **Main 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

Downstream task evaluation

- Vietnamese aspect-based sentiment analysis (**See Thin et al. (2021) for details**)

Information	PhoBERT	viBert4news	viBert_FPT	vELECTRA_FPT
Data Domain	News+Wiki	News	News	News
Data Size	20G	20GB	10GB	60GB
Tokenization	Subword	Syllable	Subword	Subword
Vocabulary size	64K	62K	32K	32K
Word segmentation	True	False	False	False

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**

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
$\text{XLM}_{\text{MLM+TLM}}$ (Conneau and Lample, 2019)	76.6
$\text{XLM-R}_{\text{base}}$ (Conneau et al., 2020)	75.4
$\text{XLM-R}_{\text{large}}$ (Conneau et al., 2020)	<u>79.7</u>
$\text{PhoBERT}_{\text{base}}$	78.5
$\text{PhoBERT}_{\text{large}}$	**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$_{\text{base}}$ (our result)	96.2
XLM-R$_{\text{large}}$ (our result)	96.3
PhoBERT$_{\text{base}}$	<u>96.7</u>
PhoBERT$_{\text{large}}$	**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-R_{base} (our result)	92.0
XLM-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-R_{base}	0.925	0.879
	XLM-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 Truong et al. (2021) 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$_{\text{base}}$ (our result)	76.46 / 83.10
Biaffine w/ XLM-R$_{\text{large}}$ (our result)	75.87 / 82.70
Biaffine w/ PhoBERT$_{\text{base}}$	**78.77 / 85.22**
Biaffine w/ PhoBERT$_{\text{large}}$	<u>77.85</u> / <u>84.32</u>

Downstream task evaluation

- 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 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

Key takeaways

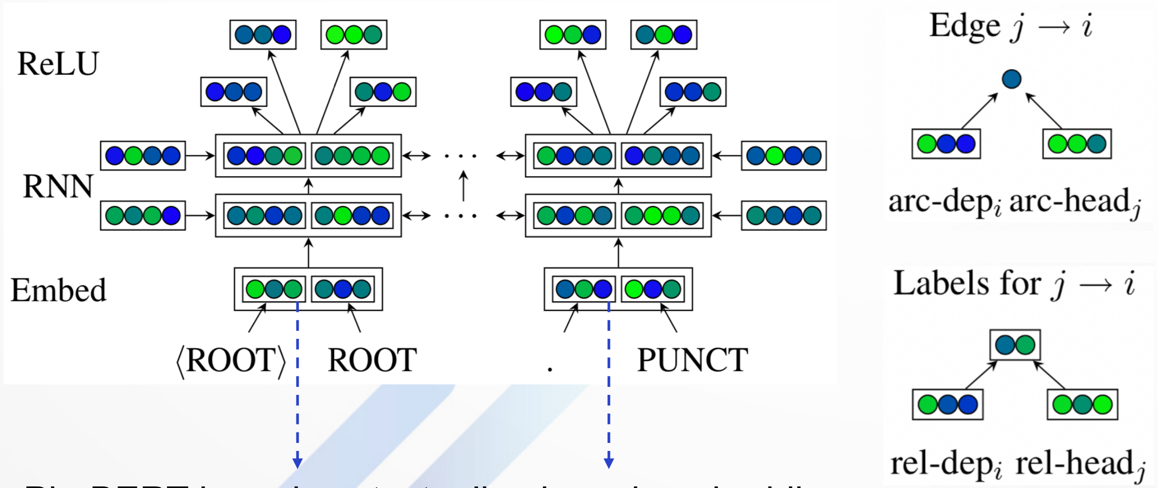
- PhoBERT with two versions PhoBERT-base and PhoBERT-large are the first public large-scale monolingual language models pre-trained for Vietnamese
- PhoBERT helps produce state-of-the-art performances on 5 downstream tasks
  - Aspect-based sentiment analysis, NLI, POS tagging, NER and Dependency parsing
  - 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

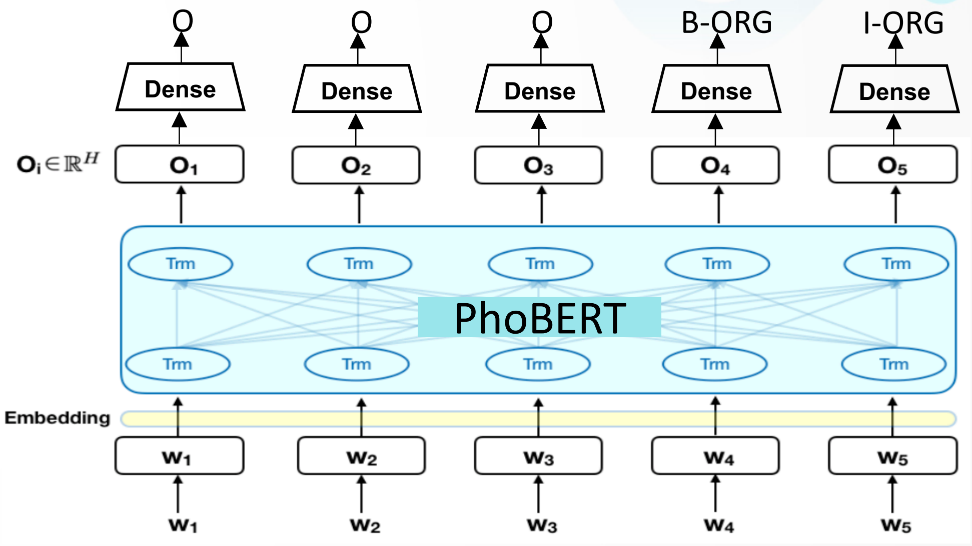
- PhoBERT: Pre-trained language models for Vietnamese
- **PhoNLP: A PhoBERT-based multi-task learning model for Vietnamese Part-of-Speech tagging, Named entity recognition and Dependency parsing**
- Text-to-SQL semantic parsing for Vietnamese

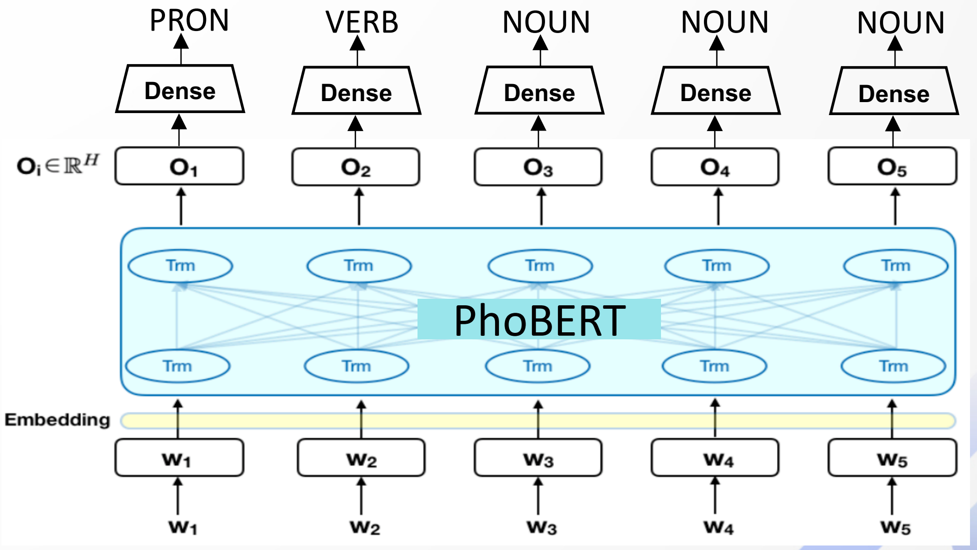
Motivation

- POS tagging, NER and dependency parsing
  - POS tags are used for dependency parsing (and might be used for NER)
  - Error propagation



PhoBERT-based contextualized word embeddings

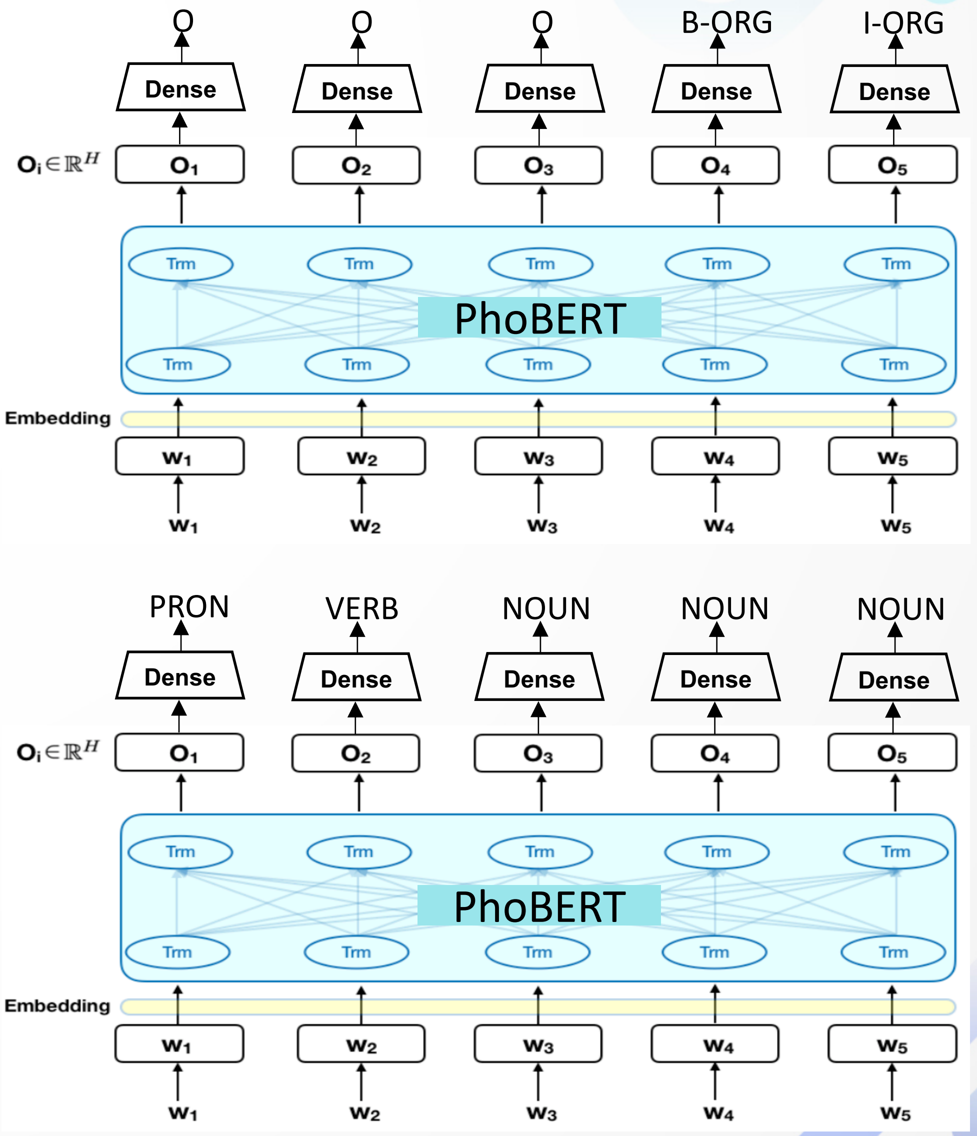


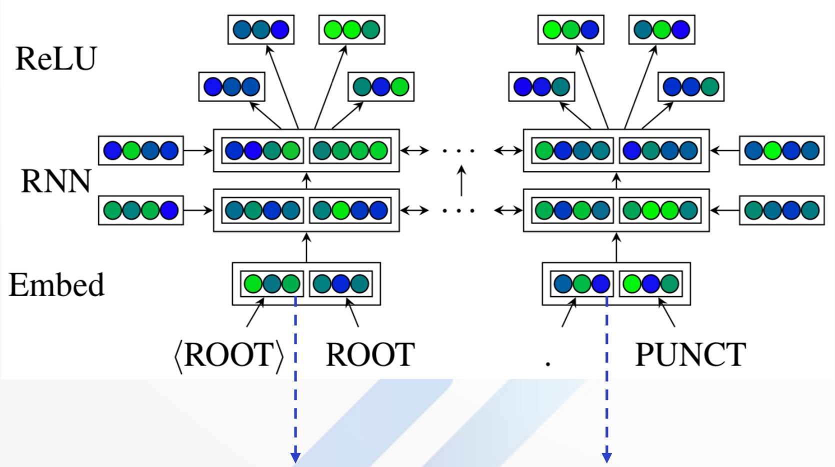


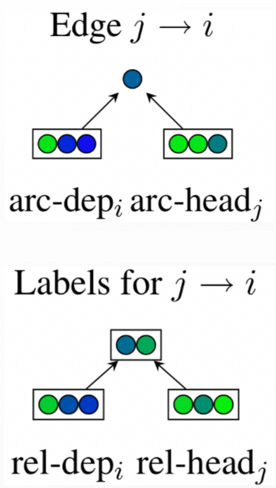
Motivation

- POS tagging, NER and dependency parsing
  - PhoBERT-base based fine-tuned model for each task (350MB)

  ➔ 1.0+GB for 3 task models



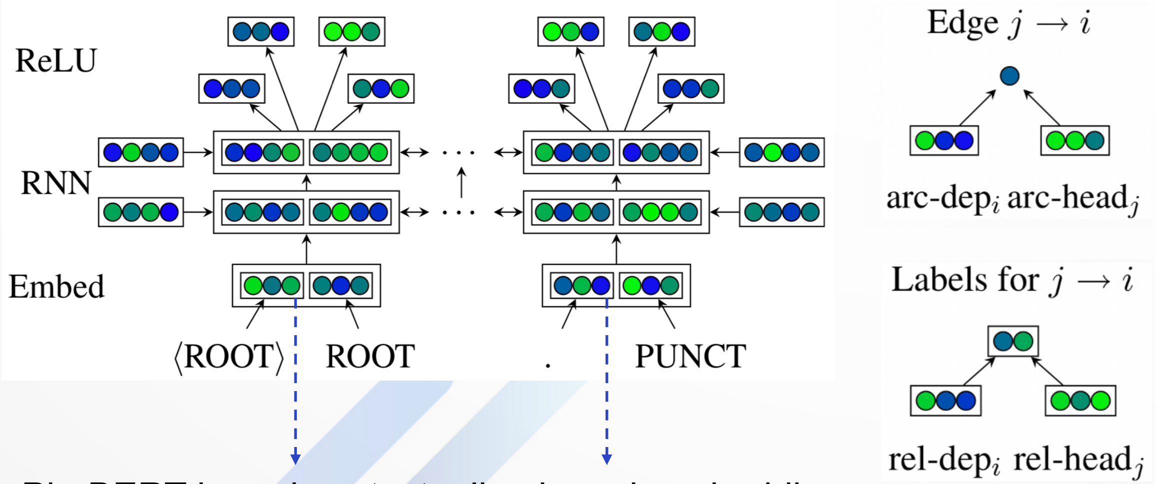




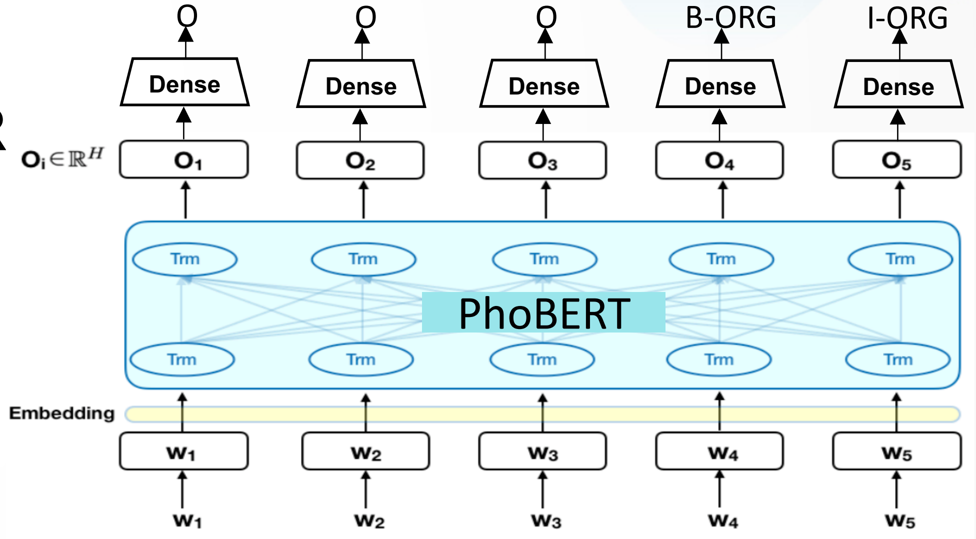
PhoBERT-based contextualized word embeddings

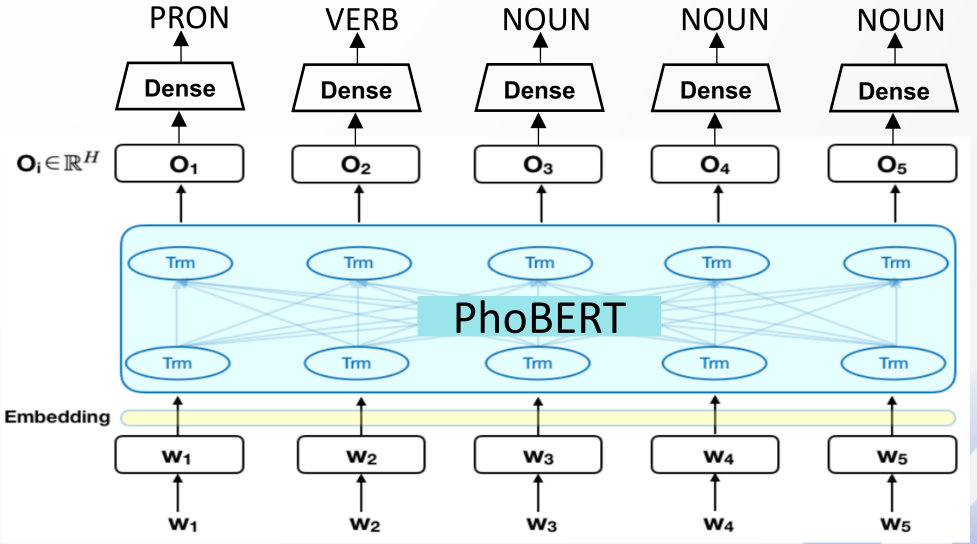
Motivation

- Joint multi-task learning for POS tagging, NER and dependency parsing
  - Might improve performance
  - Storage advantage



PhoBERT-based contextualized word embeddings

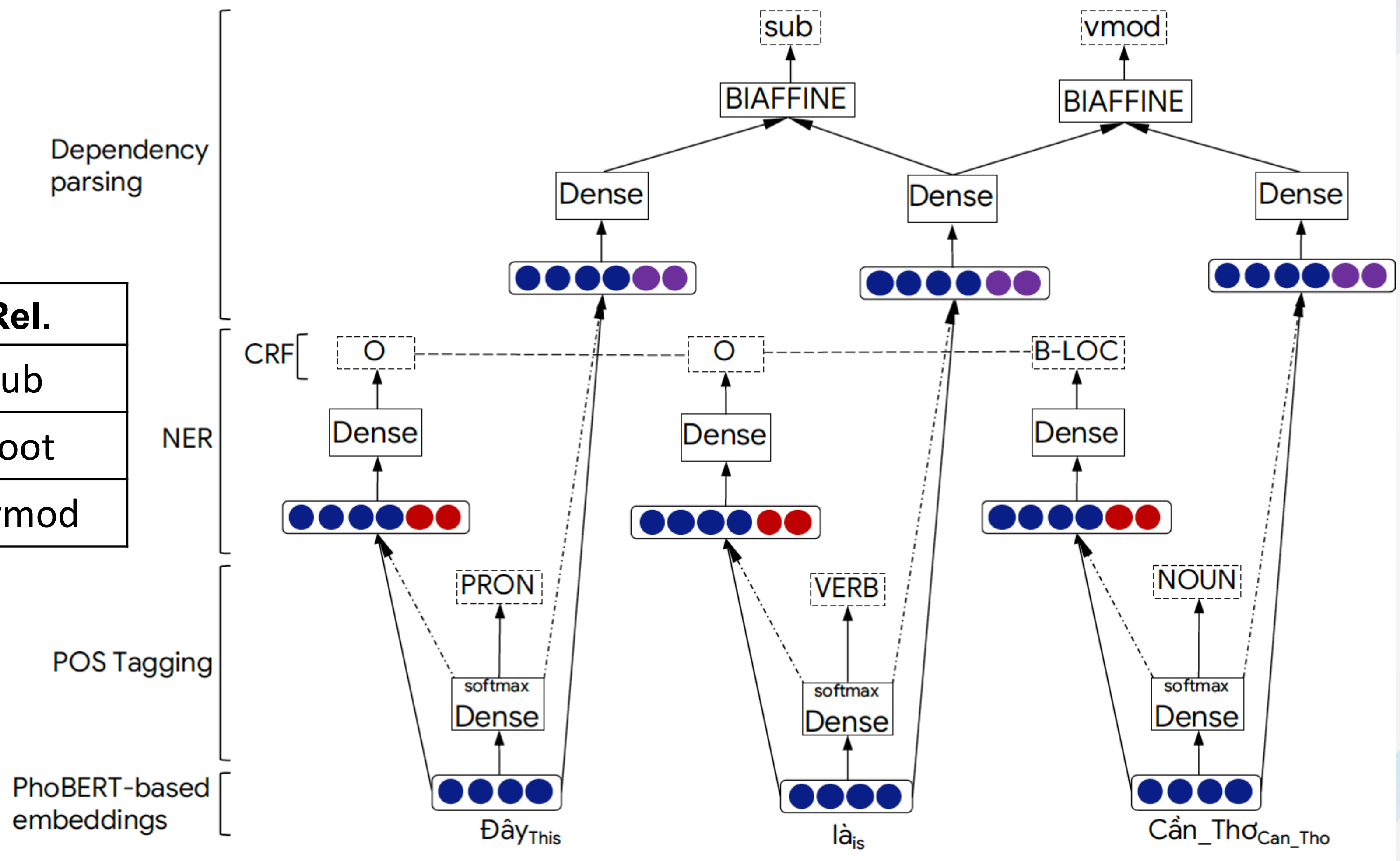




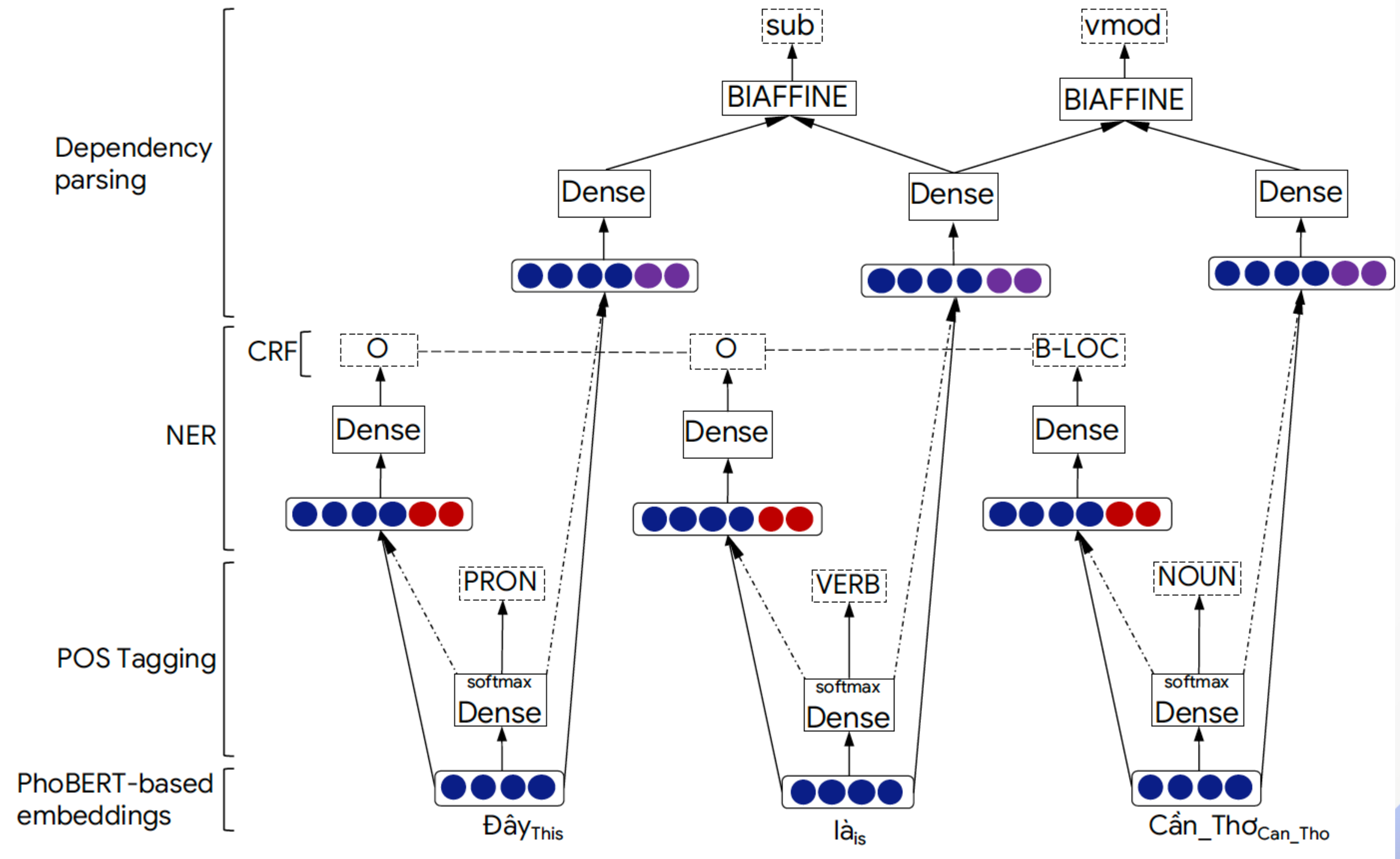
PhoNLP model

- Joint multi-task learning for POS tagging, NER and dependency parsing

ID	Form	POS	NER	Head	Rel.
1	Đây	PRON	O	2	sub
2	là	VERB	O	0	root
3	Cần_Thơ	NOUN	B-LOC	2	vmod

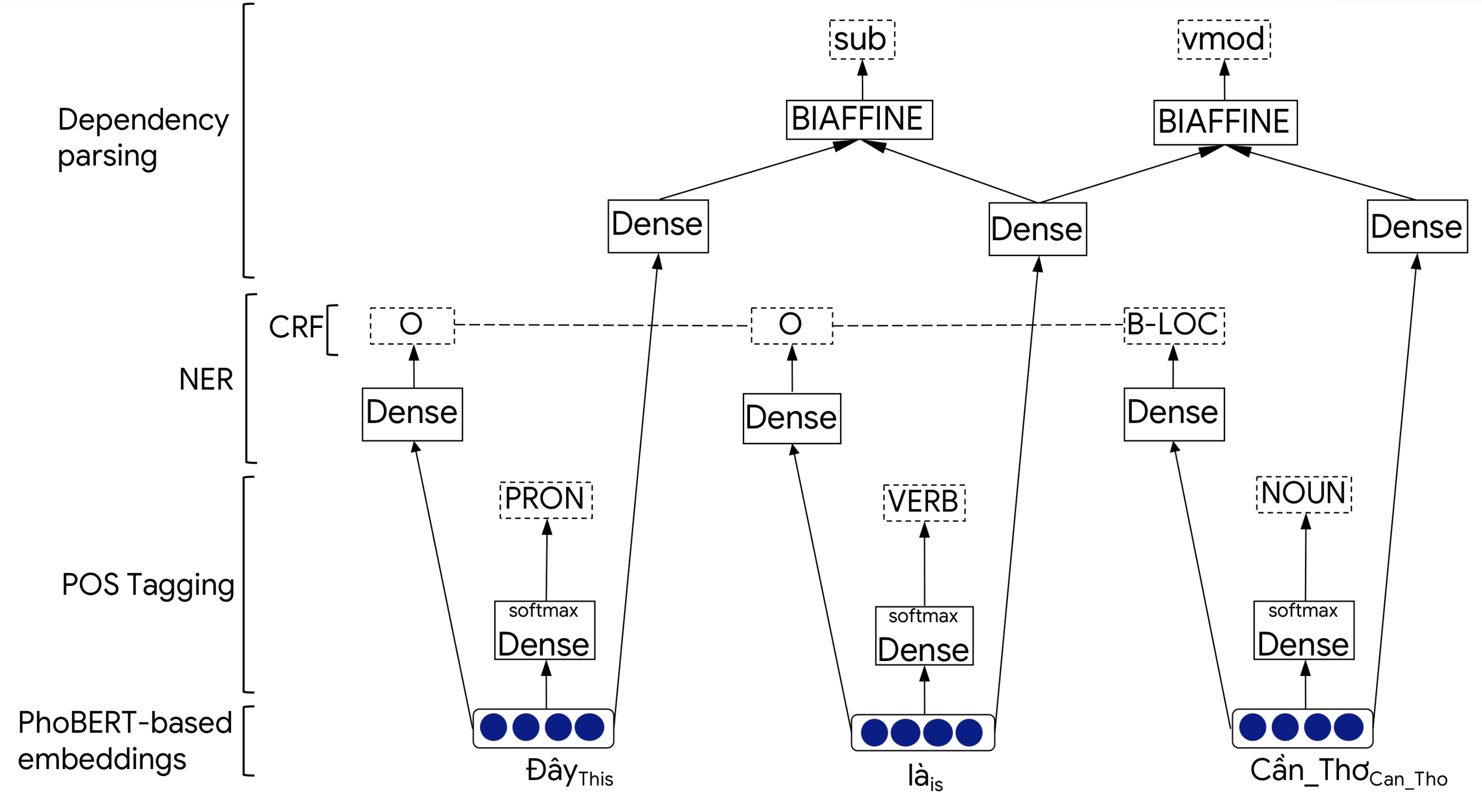


PhoNLP model



PhoNLP model

- In preliminary experiments, a non-hierarchical multi-task learning manner produced lower scores than PhoNLP



Evaluation

- VLSP 2013 POS tagging dataset, VLSP 2016 NER dataset & VnDT v.1.1 dependency treebank that was converted from Vietnamese constituent treebank (Nguyen et.al, 2009)
- <u>Data leakage that has not been pointed out before</u>: *All sentences from the VLSP 2016 and VnDT datasets are included in the VLSP 2013 dataset*
  - 90+% of sentences from validation and test sets for NER and dependency parsing are included in the POS tagging training set

➔ Re-split the VLSP 2013 POS tagging dataset to avoid the data leakage

Task	Training	Validation	Test
~~POS tagging (leakage)~~	~~27000~~	~~870~~	~~2120~~
POS tagging (Re-split)	23906	2009	3481
NER	14861	2000	2831
Dependency parsing	8977	200	1020

Evaluation

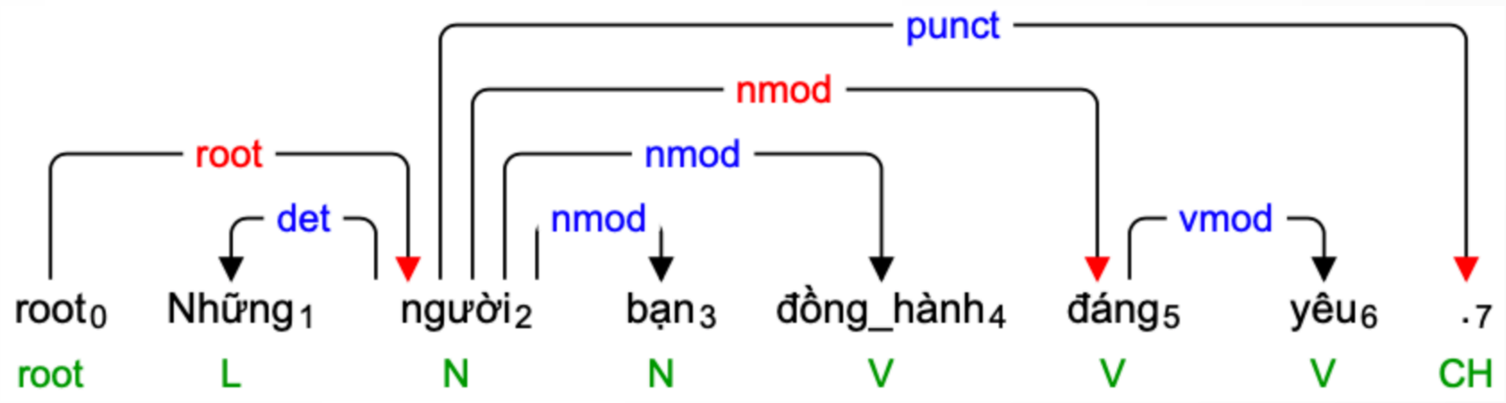
- Test results (employing the PhoBERT-base model)

	Model	POS	NER	LAS	UAS
Leak.	Single-task	96.7†	93.69	78.77†	85.22†
	PhoNLP	**96.76**	**94.41**	**79.11**	**85.47**
Re-spl	Single-task	93.68	93.69	77.89	84.78
	PhoNLP	**93.88**	**94.51**	**78.17**	**84.95**

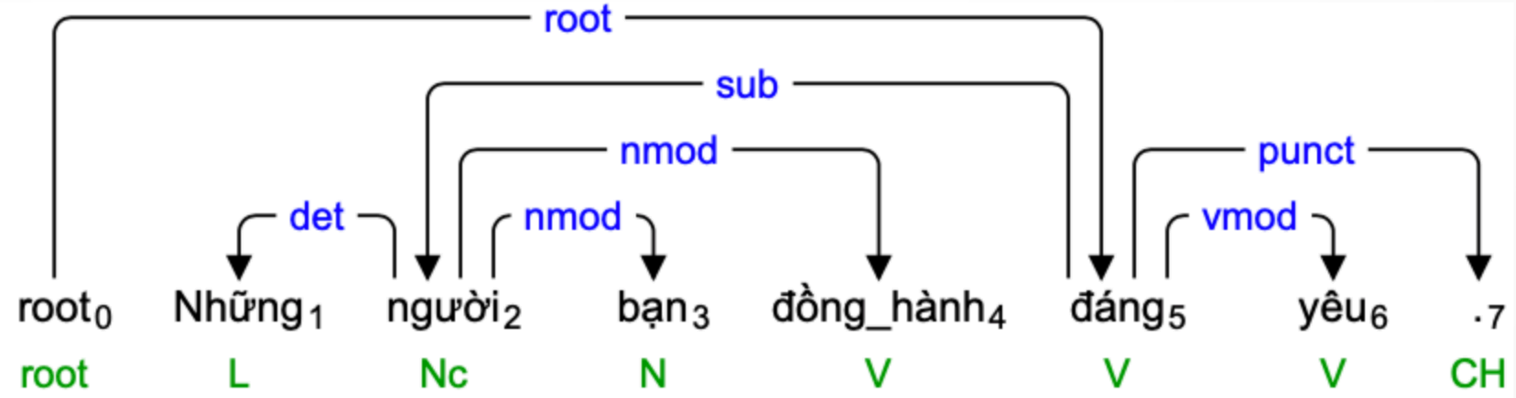
Table 2: Performance results (in %) on the test sets for POS tagging (i.e. accuracy), NER (i.e. F_1-score) and dependency parsing (i.e. LAS and UAS scores). “Leak.” abbreviates “leakage”, denoting the results obtained w.r.t. the data leakage issue. “Re-spl” denotes the results obtained w.r.t. the data re-split and duplication removal for POS tagging to avoid the data leakage issue. “Single-task” refers to as the single-task training approach. † denotes scores taken from the PhoBERT paper (Nguyen and Nguyen, 2020). Note that “Single-task” NER is not affected by the data leakage issue.

Evaluation

Single task learning



PhoNLP



Key takeaways

- A new multi-task learning model PhoNLP for jointly training 3 NLP tasks of POS tagging, NER and dependency parsing
  - Results on Vietnamese show that multi-task learning does better than single-task learning & helps produce state-of-the-art performances
- Data leakage
  - Re-split VLSP 2013 POS tagging data to handle this issue
- Future work is to adapt PhoNLP for other languages

Outline

- PhoBERT: Pre-trained language models for Vietnamese
- PhoNLP: A PhoBERT-based multi-task learning model for Vietnamese Part-of-Speech tagging, Named entity recognition and Dependency parsing
- **Text-to-SQL semantic parsing for Vietnamese**

Semantic parsing

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

Cho biết số lượng những chiếc xe có nhiều hơn 4 xi lanh → Text-to-SQL semantic parser → SELECT count(*) FROM [dữ liệu xe] WHERE [số lượng xi lanh] > 4

What is the number of cars with more than 4 cylinders ? → SELECT count(*) FROM CARS_DATA WHERE Cylinders > 4

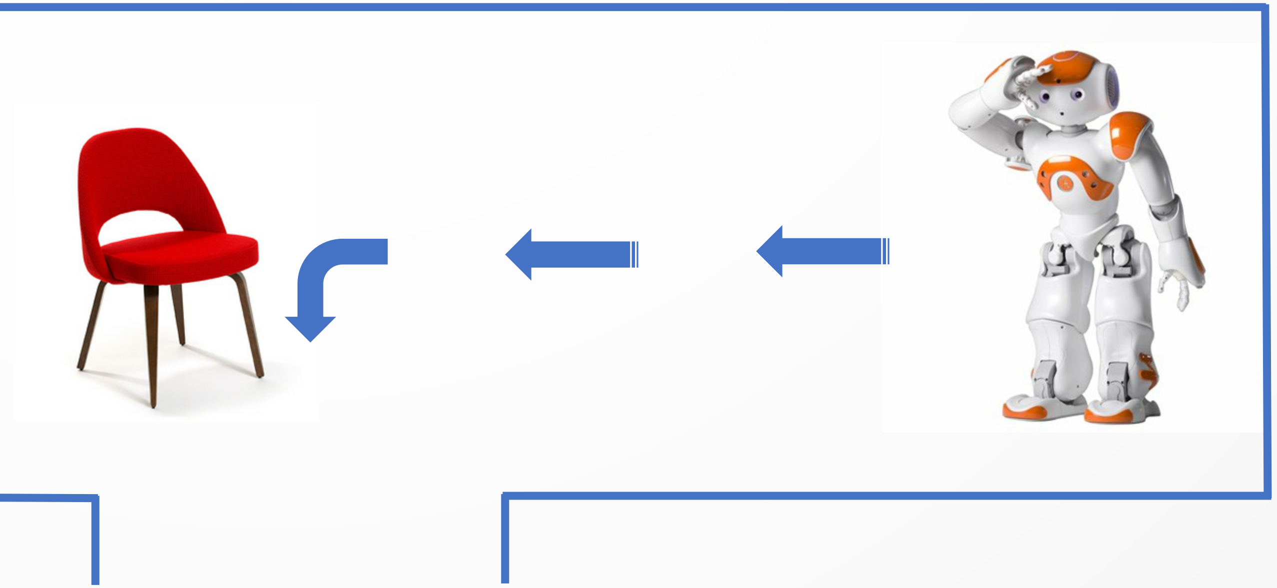
Semantic parsing

- Why do we want to do semantic parsing?

Sentence → Semantic parser → Meaning representation e.g. logical form → Executor → Response

Instructing a Robot

At the chair,
turn left



Semantic parsing

- Why do we want to do semantic parsing?

Sentence → Semantic parser → Meaning representation e.g. SQL query → Executor → Response

Which provinces border Ha Noi? →  → Thai Nguyen, Vinh Phuc, Ha Nam, Hoa Binh, Bac Giang, Bac Ninh, Hung Yen, Phu Tho

- Serve as an important component in many NLP systems such as Question answering and Task-oriented dialogue
- Access information stored in databases via natural language statements
- Users do not need to understand SQL query syntax as well as database schemas

Text-to-SQL semantic parsing

- The significant availability of the world’s knowledge stored in relational databases leads to the creation of large-scale Text-to-SQL benchmarks, e.g. WikiSQL (Zhong et al., 2017) and Spider (Yu et al., 2018)
  - Help boost the development of various state-of-the-art (SOTA) sequence-to-sequence (seq2seq) semantic parsers

- Most benchmarks are exclusively for English

Text-to-SQL semantic parsing

- SQL is a database interface and universal semantic representation
  - Worth investigating the Text-to-SQL parsing task for languages other than English
- The difference in linguistic characteristics could add difficulties in applying seq2seq semantic parsing models to the non-English languages (Min et al., 2019)
  - Study the influence of *word segmentation* in Vietnamese on its SQL parsing, i.e. syllable level vs. word level

Cho biết số lượng những chiếc xe có nhiều hơn 4 xi lanh → Text-to-SQL semantic parser → SELECT count(*) FROM [dữ liệu xe] WHERE [số lượng xi lanh] > 4

Cho biết số_lượng những chiếc xe có nhiều hơn 4 xi_lanh

SELECT count(*) FROM [dữ_liệu xe] WHERE [số_lượng xi_lanh] > 4

Vietnamese semantic parsing

- Previous approaches:
  - Construct rule templates to convert single database-driven questions into meaning representations (Nguyen and Le, 2008; Nguyen et al., 2009, 2012; Tung et al., 2015; Nguyen et al., 2017)
  - Vuong et al. (2019) formulate the Text-to-SQL semantic parsing task for Vietnamese as a sequence labeling-based slot filling problem
    - Use a conventional CRF model with handcrafted features
  - Seq2seq-based semantic parsers have not yet been explored in any previous work for Vietnamese

Vietnamese semantic parsing

- Semantic parsing datasets for Vietnamese:
  - A corpus of 5460 sentences for assigning semantic roles (Phuong et al., 2017)
  - A small Text-to-SQL dataset of 1258 simple structured questions over 3 databases (Vuong et al., 2019)
  - These two datasets are not publicly available for research community

- **Contributions of our work**
  - Introduce the first public large-scale Vietnamese dataset for Text-to-SQL semantic parsing: https://github.com/VinAIResearch/ViText2SQL
  - Extend strong seq2seq semantic parsers and compare them under various configurations on our dataset

Text-to-SQL semantic parsing dataset for Vietnamese

- Strategy to construct such a dataset
  - Manually translate an existing English dataset into Vietnamese
- WikiSQL and Spider are well-known large-scale Text-to-SQL benchmarks for English
  - Spider presents challenges not only in handling complex questions but also in generalizing to unseen databases during evaluation

👉 Manually translate Spider into Vietnamese

Text-to-SQL semantic parsing dataset for Vietnamese

- Manually translate all English questions and the database schema (i.e. table and column names as well as values in SQL queries) in Spider into Vietnamese
  - The original Spider dataset consists of 10181 questions with their corresponding 5693 SQL queries over 200 databases
  - Only 9691 questions and their corresponding 5263 SQL queries over 166 databases, which are used for training and validation, are publicly available

Text-to-SQL semantic parsing dataset for Vietnamese

- Translation work is performed by 1 NLP researcher and 2 computer science students (IELTS 7.0+)
  - Every question and SQL query pair from the same database is first translated by one student and then cross-checked and corrected by the second student
  - The NLP researcher verifies the original and corrected versions and makes further revisions if needed

Text-to-SQL semantic parsing dataset for Vietnamese

- In case of literal translation for a question: Stick to the style of the original English question
- Rephrase complex questions based on the semantic meaning of the corresponding SQL queries to obtain the most natural language questions in Vietnamese
- Split our dataset into training, development and test sets such that no database overlaps between them

	#Qu.	#SQL	#DB	#T/D	#Easy	#Med.	#Hard	#ExH
all	9691	5263	166	5.3	2233	3439	2095	1924
train	6831	3493	99	5.4	1559	2255	1502	1515
dev	954	589	25	4.2	249	405	191	109
test	1906	1193	42	5.7	425	779	402	300

Text-to-SQL semantic parsing dataset for Vietnamese

- Translated question and SQL query pairs in our dataset are written at the syllable level
- Apply RDRSegmenter from VnCoreNLP (Vu et al., 2018) to perform automatic Vietnamese word segmentation

Original (Easy question–involving one table in one database):
What is the number of cars with more than 4 cylinders?
SELECT count(*) FROM CARS_DATA WHERE Cylinders > 4

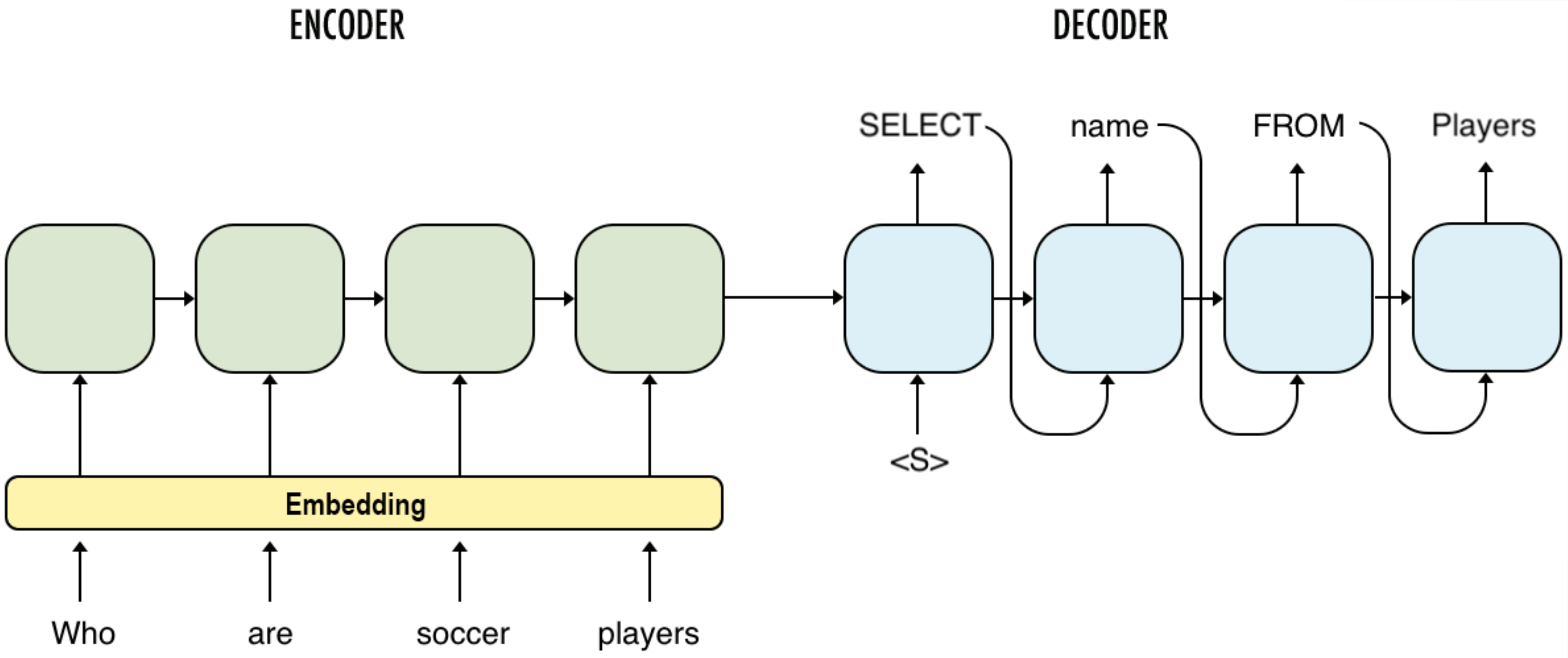
Translated:
Cho biết số lượng những chiếc xe có nhiều hơn 4 xi lanh.
SELECT count(*) FROM [dữ liệu xe] WHERE [số lượng xi lanh] > 4

Original (Hard question–with a nested SQL query):
Which countries in europe have at least 3 car manufacturers?
SELECT T1.CountryName FROM COUNTRIES AS T1 JOIN CONTINENTS AS T2 ON T1.Continent = T2.ContId JOIN CAR_MAKERS AS T3 ON T1.CountryId = T3.Country WHERE T2.Continent = "europe" GROUP BY T1.CountryName HAVING count(*) >= 3

Translated:
Những quốc gia nào ở châu Âu có ít nhất 3 nhà sản xuất xe hơi?
SELECT T1.[tên quốc gia] FROM [quốc gia] AS T1 JOIN [lục địa] AS T2 ON T1.[lục địa] = T2.[id lục địa] JOIN [nhà sản xuất xe hơi] AS T3 ON T1.[id quốc gia] = T3.[quốc gia] WHERE T2.[lục địa] = "châu Âu" GROUP BY T1.[tên quốc gia] HAVING count(*) >= 3

Baseline models

- Formulate the text-to-SQL semantic parsing task as a seq2seq problem
  - Employ seq2seq encoder-decoder architectures



Baseline models

- Select seq2seq based models EditSQL (Zhang et al., 2019) and IRNet (Guo et al., 2019) with publicly available implementations as our baselines, obtaining near SOTA scores on Spider
- EditSQL:
  - A BiLSTM-based question-table encoder to encode the question and table schema
  - A BiLSTM-based interaction encoder with attention to incorporate the recent question history
  - An LSTM-based table-aware decoder with attention, taking into account the outputs of both encoders to generate a SQL query

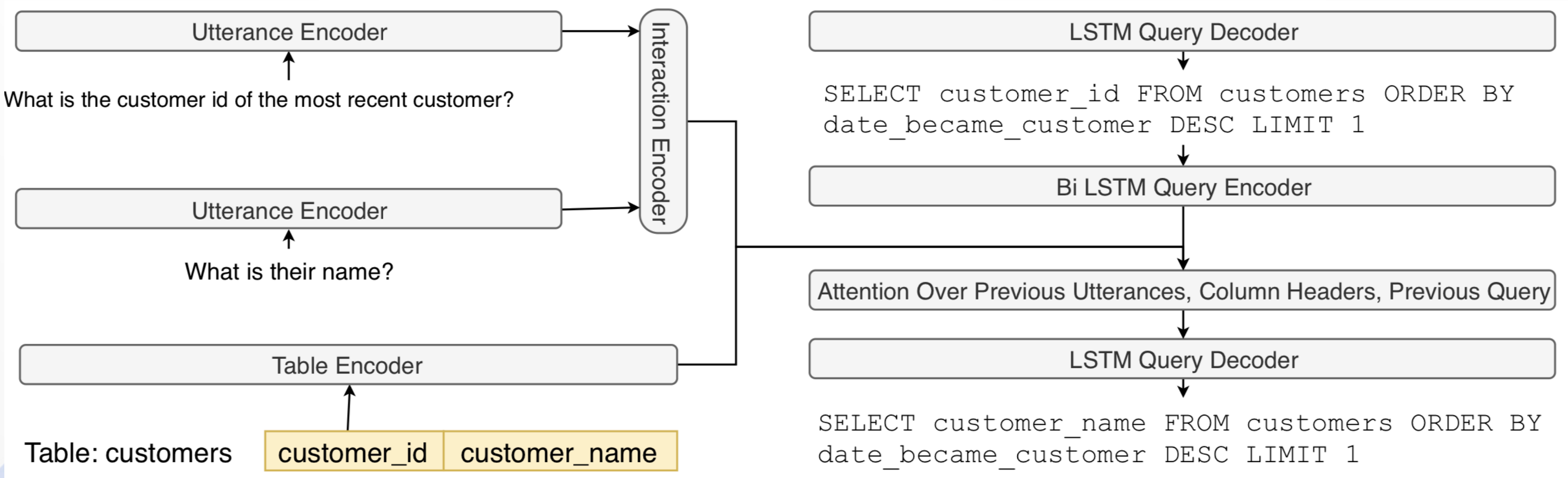


Figure taken from Zhang et al. (2019)

Baseline models

- IRNet:
  - N-gram matching-based schema linking to identify the columns and the tables in a question
  - Take the question, a database schema and the schema linking results as input to synthesize a tree-structured SemQL query
    - Performed by using a BiLSTM-based question encoder and an attention-based schema encoder together with a grammar-based LSTM decoder (Yin and Neubig, 2017)
  - Deterministically uses the SemQL query to infer a SQL query with domain knowledge

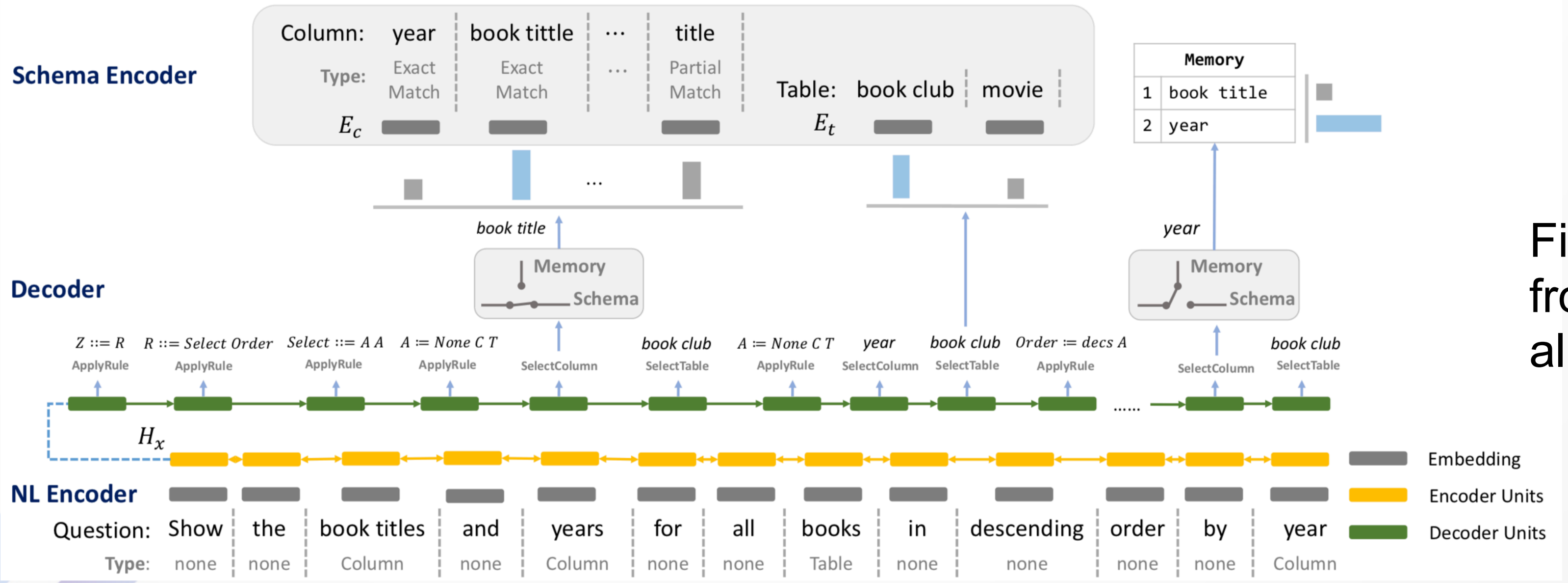


Figure taken from Guo et al. (2019)

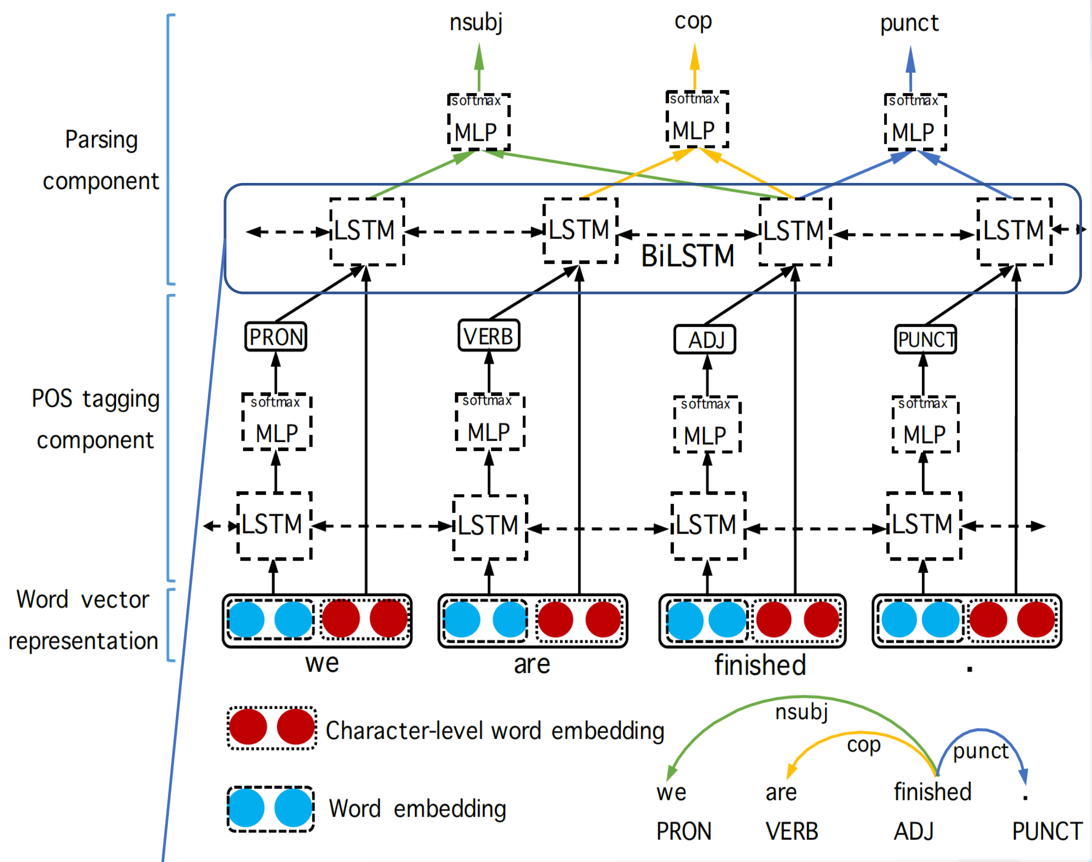
Extensions

- *Normalized pointwise mutual information (NPMI) for schema linking*
    - IRNet relies on the large-scale KG ConceptNet (Speer et al., 2017) to link a cell value mentioned in a question to a column in the database schema based on two ConceptNet categories “is a type of” and “related terms”
        - These two ConceptNet categories are not available for Vietnamese
    - Propose a novel use of the NPMI collocation score (Bouma, 2009) for the schema linking in IRNet

- *Incorporating contextualized word embeddings as part of input embeddings*
    - Extend baselines with the use of pre-trained language models XLM-R-base and PhoBERT-base for the syllable- and word-level settings, respectively

Extensions

- *Incorporating latent syntactic features:*
  - Hand-crafted syntactic features help improve semantic parsing (Monroe and Wang, 2014; Jie and Lu, 2018)
  - Investigate whether latent syntactic features would help improve Vietnamese Text-to-SQL parsing?
  - Dump latent feature representations from jPTDP's BiLSTM encoder given our word-level inputs, and directly use them as part of input embeddings of EditSQL and IRNet



jPTDP's BiLSTM encoder

Main results

- Our human-translated dataset vs. a machine-translated dataset
- The influence of Vietnamese word segmentation, i.e. syllable level vs. word level

	Approach	dev	test	Approach	dev	test
Vi-Syllable	**EditSQL** [MT]	21.5	16.8	**IRNet** [MT]	25.4	20.3
	EditSQL	28.6	24.1	**IRNet**	43.3	38.2
	EditSQL$_{\text{XLM-R}}$	55.2	51.3	**IRNet**$_{\text{XLM-R}}$	58.6	52.8
Vi-Word	**EditSQL** [MT]	22.8	17.4	**IRNet** [MT]	27.4	21.6
	EditSQL	33.7	30.2	**IRNet**	49.7	43.6
	EditSQL$_{\text{DeP}}$	45.3	42.2	**IRNet**$_{\text{DeP}}$	52.2	47.1
	EditSQL$_{\text{PhoBERT}}$	56.7	52.6	**IRNet**$_{\text{PhoBERT}}$	60.2	53.2
En	**EditSQL**$_{\text{RoBERTa}}$	58.3	53.6	**IRNet**$_{\text{RoBERTa}}$	63.8	55.3

Exact matching accuracy

Main results

- The usefulness of the latent syntactic features
- The usefulness of the pre-trained language models
- Without using NPMI for schema linking in IRNet ➔ 6+% absolute decrease

	Approach	dev	test	Approach	dev	test
Vi-Syllable	**EditSQL** [MT]	21.5	16.8	**IRNet** [MT]	25.4	20.3
	EditSQL	28.6	24.1	**IRNet**	43.3	38.2
	EditSQL$_{\text{XLM-R}}$	55.2	51.3	**IRNet**$_{\text{XLM-R}}$	58.6	52.8
Vi-Word	**EditSQL** [MT]	22.8	17.4	**IRNet** [MT]	27.4	21.6
	EditSQL	33.7	30.2	**IRNet**	49.7	43.6
	EditSQL$_{\text{DeP}}$	45.3	42.2	**IRNet**$_{\text{DeP}}$	52.2	47.1
	EditSQL$_{\text{PhoBERT}}$	56.7	52.6	**IRNet**$_{\text{PhoBERT}}$	60.2	53.2
En	**EditSQL**$_{\text{RoBERTa}}$	58.3	53.6	**IRNet**$_{\text{RoBERTa}}$	63.8	55.3

Exact matching accuracy

Main results

- Exact matching accuracy categorized by 4 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
$\text{EditSQL}_{\text{DeP}}$	65.7	46.1	37.6	16.8	75.1	44.6	65.6	63.2	73.5
$\text{EditSQL}_{\text{XLM-R}}$	75.1	56.2	45.3	22.4	82.7	60.3	70.7	67.2	79.8
$\text{EditSQL}_{\text{PhoBERT}}$	75.6	58.0	47.4	22.7	83.3	61.8	72.5	67.9	80.6
$\text{IRNet}_{\text{DeP}}$	71.8	51.5	47.4	18.5	79.3	48.7	71.8	63.4	74.3
$\text{IRNet}_{\text{XLM-R}}$	76.2	57.8	46.8	23.5	83.5	59.1	74.4	68.3	80.5
$\text{IRNet}_{\text{PhoBERT}}$	76.8	57.5	47.2	24.8	84.5	59.3	76.6	68.2	80.3

Error analysis

- Causes of errors from 382 failed examples on the dev. set by $\text{IRNet}_{\text{PhoBERT}}$
  - 121/382 cases: Incorrect predictions on the column names which are not mentioned or only partially mentioned in the questions
    - *Hiển thị tên và năm phát hành của những bài hát thuộc về ca sĩ trẻ tuổi nhất* (Show the name and the release year of the song by the youngest singer)
      The model produces an incorrect column name prediction of "tên" (name) instead of the correct one "tên bài hát" (song name)
  - 47/382 cases: having an equivalent implementation of their intent with a different SQL syntax
    - A 'failed' SQL output "*SELECT MAX [sức chứa] FROM [sân vận động]*" is equivalent to the gold SQL query of "*SELECT [sức chứa] FROM [sân vận động] ORDER BY [sùc chùa] DESC LIMIT 1*"
    - The SQL output would be valid if we measure an execution accuracy

Word segmentation is not shown for simplification

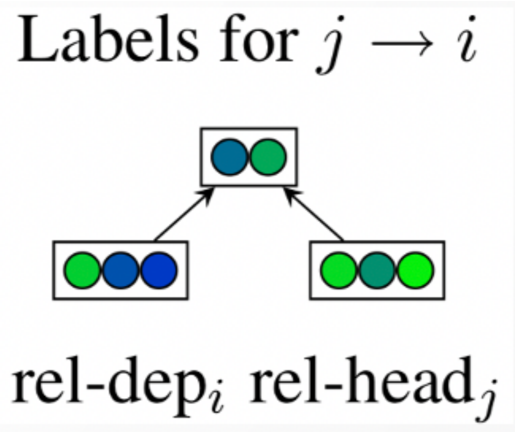
Error analysis

- Causes of errors from 382 failed examples on the dev. set by $\text{IRNet}_{\text{PhoBERT}}$
  - 84/382 cases are caused by nested and complex SQL queries which mostly belong to the Extra Hard category
  - 70/382 cases: Incorrectly predicting operators is another common type of errors, e.g. operators “*max*” and “*min*”
  - 60/382 cases are accounted for an incorrect prediction of table names in a FROM clause

Key takeaways

- The first public large-scale dataset for Vietnamese Text-to-SQL semantic parsing https://github.com/VinAIResearch/ViText2SQL
- Extensively experiment with key research configurations using two strong baseline models on our dataset and find that:
    1. Our human-translated dataset is far more reliable than a dataset consisting of machine-translated questions
    2. Automatic Vietnamese word segmentation improves the performances of the baselines
    3. The NPMI score is useful for linking a cell value mentioned in a question to a column in the database schema
    4. Latent syntactic features also help improve the performances
    5. Highest improvements are accounted for the use of pre-trained language models, where PhoBERT helps produce higher results than XLM-R

Thanks for your attention!



$$\mathbf{s}_{i,j} = \mathrm{Biaffine}\left(\mathbf{h}_i^{(\mathrm{head})}, \mathbf{h}_j^{(\mathrm{dep})}\right)$$

$$\mathrm{Biaffine}(\mathbf{y}_1, \mathbf{y}_2) = \underbrace{\mathbf{y}_1^{\top}\mathbf{U}\mathbf{y}_2}_{\text{Bilinear}} + \underbrace{\mathbf{W}(\mathbf{y}_1 \circ \mathbf{y}_2) + \mathbf{b}}_{\text{Linear}}$$