Recent advances in Vietnamese language modeling and understanding

Dat Quoc Nguyen – VinAI Research, Vietnam
(12/11/2020 - Updated 19/04/2021)


- 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

External results taken from:

• 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
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

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.

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
theworld về trí tuệ nhân tạo

Words

VinAI công bố các kết quả nghiên cứu khoa học tại hội nghị hàng đầu
theworld 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 có văn Đàc lập và tòa 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

<table>
<thead>
<tr>
<th>ID</th>
<th>Form</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tôi 1</td>
<td>PRON</td>
</tr>
<tr>
<td>2</td>
<td>là ăn</td>
<td>VERB</td>
</tr>
<tr>
<td>3</td>
<td>sinh viên</td>
<td>NOUN</td>
</tr>
<tr>
<td>4</td>
<td>Đại học</td>
<td>NOUN</td>
</tr>
<tr>
<td>5</td>
<td>Công nghệ</td>
<td>NOUN</td>
</tr>
</tbody>
</table>

Downstream task evaluation

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Form</th>
<th>NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tôi</td>
<td>O</td>
</tr>
<tr>
<td>2</td>
<td>là_am</td>
<td>O</td>
</tr>
<tr>
<td>3</td>
<td>sinh_vién student</td>
<td>O</td>
</tr>
<tr>
<td>4</td>
<td>Đại_học university</td>
<td>B-ORG</td>
</tr>
<tr>
<td>5</td>
<td>Công_nghệ technology</td>
<td>I-ORG</td>
</tr>
</tbody>
</table>
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.
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)

<table>
<thead>
<tr>
<th>Information</th>
<th>PhoBERT</th>
<th>viBert4news</th>
<th>viBert_FPT</th>
<th>vELECTRA_FPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Domain</td>
<td>News+Wiki</td>
<td>News</td>
<td>News</td>
<td>News</td>
</tr>
<tr>
<td>Data Size</td>
<td>20G</td>
<td>20GB</td>
<td>10GB</td>
<td>60GB</td>
</tr>
<tr>
<td>Tokenization</td>
<td>Subword</td>
<td>Syllable</td>
<td>Subword</td>
<td>Subword</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>64K</td>
<td>62K</td>
<td>32K</td>
<td>32K</td>
</tr>
<tr>
<td>Word segmentation</td>
<td>True</td>
<td>False</td>
<td>False</td>
<td>False</td>
</tr>
</tbody>
</table>

### The experimental results of various mono-lingual and multi-lingual pre-trained BERT models on Vietnamese aspect category detection task for the restaurant domain.

<table>
<thead>
<tr>
<th>Types</th>
<th>Models</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-lingual</td>
<td>mBERT</td>
<td>81.39</td>
<td>76.34</td>
<td>78.78</td>
</tr>
<tr>
<td></td>
<td>mDistilBert</td>
<td>80.35</td>
<td>76.07</td>
<td>78.16</td>
</tr>
<tr>
<td></td>
<td>XLM-R</td>
<td>82.98</td>
<td>81.40</td>
<td>82.18</td>
</tr>
<tr>
<td>Mono-lingual</td>
<td>viBert4news</td>
<td>79.26</td>
<td>77.48</td>
<td>78.36</td>
</tr>
<tr>
<td></td>
<td>viBert_FPT</td>
<td>80.65</td>
<td>79.12</td>
<td>79.88</td>
</tr>
<tr>
<td></td>
<td>vELECTRA_FPT</td>
<td>83.08</td>
<td>79.54</td>
<td>81.27</td>
</tr>
<tr>
<td></td>
<td>PhoBERT</td>
<td><strong>85.60</strong></td>
<td><strong>87.49</strong></td>
<td><strong>86.53</strong></td>
</tr>
</tbody>
</table>

### The experimental results of various mono-lingual and multi-lingual pre-trained BERT models on Vietnamese aspect category detection task for the hotel domain.

<table>
<thead>
<tr>
<th>Types</th>
<th>Models</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-lingual</td>
<td>mBERT</td>
<td>77.93</td>
<td>76.26</td>
<td>77.09</td>
</tr>
<tr>
<td></td>
<td>mDistilBert</td>
<td>78.59</td>
<td>74.97</td>
<td>76.73</td>
</tr>
<tr>
<td></td>
<td>XLM-R</td>
<td>78.86</td>
<td>76.56</td>
<td>77.70</td>
</tr>
<tr>
<td>Mono-lingual</td>
<td>viBert4news</td>
<td>79.39</td>
<td>74.83</td>
<td>77.04</td>
</tr>
<tr>
<td></td>
<td>viBert_FPT</td>
<td>81.14</td>
<td>74.54</td>
<td>77.70</td>
</tr>
<tr>
<td></td>
<td>vELECTRA_FPT</td>
<td>79.82</td>
<td>76.07</td>
<td>77.90</td>
</tr>
<tr>
<td></td>
<td>PhoBERT</td>
<td><strong>81.49</strong></td>
<td><strong>76.96</strong></td>
<td><strong>79.16</strong></td>
</tr>
</tbody>
</table>
### Downstream task evaluation

- **Vietnamese NLI results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiLSTM-max (Conneau et al., 2018)</td>
<td>66.4</td>
</tr>
<tr>
<td>mBiLSTM (Artetxe and Schwenk, 2019)</td>
<td>72.0</td>
</tr>
<tr>
<td>multilingual BERT (Devlin et al., 2019)</td>
<td>69.5</td>
</tr>
<tr>
<td>XLMMLM+TLM (Conneau and Lample, 2019)</td>
<td>76.6</td>
</tr>
<tr>
<td>XLM-R_{base} (Conneau et al., 2020)</td>
<td>75.4</td>
</tr>
<tr>
<td>XLM-R_{large} (Conneau et al., 2020)</td>
<td>79.7</td>
</tr>
<tr>
<td>PhoBERT_{base}</td>
<td>78.5</td>
</tr>
<tr>
<td>PhoBERT_{large}</td>
<td><strong>80.0</strong></td>
</tr>
</tbody>
</table>
Downstream task evaluation

- Vietnamese POS tagging results

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDRPOSTagger (Nguyen et al., 2014a) [♣]</td>
<td>95.1</td>
</tr>
<tr>
<td>BiLSTM-CNN-CRF (Ma and Hovy, 2016) [♣]</td>
<td>95.4</td>
</tr>
<tr>
<td>VnCoreNLP-POS (Nguyen et al., 2017) [♣]</td>
<td>95.9</td>
</tr>
<tr>
<td>jPTDP-v2 (Nguyen and Verspoor, 2018) [★]</td>
<td>95.7</td>
</tr>
<tr>
<td>jointWPD (Nguyen, 2019) [★]</td>
<td>96.0</td>
</tr>
<tr>
<td>XLM-R_{base} (our result)</td>
<td>96.2</td>
</tr>
<tr>
<td>XLM-R_{large} (our result)</td>
<td>96.3</td>
</tr>
<tr>
<td>PhoBERT_{base}</td>
<td>96.7</td>
</tr>
<tr>
<td>PhoBERT_{large}</td>
<td><strong>96.8</strong></td>
</tr>
</tbody>
</table>
# Downstream task evaluation

- Vietnamese NER results

**VLSP 2016 NER dataset**

<table>
<thead>
<tr>
<th>Model</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiLSTM-CNN-CRF [♦]</td>
<td>88.3</td>
</tr>
<tr>
<td>VnCoreNLP-NER (Vu et al., 2018) [♦]</td>
<td>88.6</td>
</tr>
<tr>
<td>VNER (Nguyen et al., 2019b)</td>
<td>89.6</td>
</tr>
<tr>
<td>BiLSTM-CNN-CRF + ETNLP [♣]</td>
<td>91.1</td>
</tr>
<tr>
<td>VnCoreNLP-NER + ETNLP [♣]</td>
<td>91.3</td>
</tr>
<tr>
<td>XLM-R_{base} (our result)</td>
<td>92.0</td>
</tr>
<tr>
<td>XLM-R_{large} (our result)</td>
<td>92.8</td>
</tr>
<tr>
<td>PhoBERT_{base}</td>
<td>93.6</td>
</tr>
<tr>
<td>PhoBERT_{large}</td>
<td>94.7</td>
</tr>
</tbody>
</table>

**PhoNER_COVID19 dataset**

<table>
<thead>
<tr>
<th>Model</th>
<th>Mic-F₁</th>
<th>Mac-F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiL-CRF</td>
<td>0.906</td>
<td>0.858</td>
</tr>
<tr>
<td>XLM-R_{base}</td>
<td>0.925</td>
<td>0.879</td>
</tr>
<tr>
<td>XLM-R_{large}</td>
<td>0.938</td>
<td>0.911</td>
</tr>
<tr>
<td>PhoBERT_{base}</td>
<td>0.910</td>
<td>0.875</td>
</tr>
<tr>
<td>PhoBERT_{large}</td>
<td>0.945</td>
<td>0.931</td>
</tr>
</tbody>
</table>

(See Truong et al. (2021) for details)
Downstream task evaluation

- Vietnamese dependency parsing results

<table>
<thead>
<tr>
<th>Model</th>
<th>LAS / UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>VnCoreNLP-DEP (Vu et al., 2018) [★]</td>
<td>71.38 / 77.35</td>
</tr>
<tr>
<td>jPTDP-v2 [★]</td>
<td>73.12 / 79.63</td>
</tr>
<tr>
<td>jointWPD [★]</td>
<td>73.90 / 80.12</td>
</tr>
<tr>
<td>Biaffine (Dozat and Manning, 2017) [★]</td>
<td>74.99 / 81.19</td>
</tr>
<tr>
<td>Biaffine w/ XLM-R_{base} (our result)</td>
<td>76.46 / 83.10</td>
</tr>
<tr>
<td>Biaffine w/ XLM-R_{large} (our result)</td>
<td>75.87 / 82.70</td>
</tr>
<tr>
<td>Biaffine w/ PhoBERT_{base}</td>
<td>78.77 / 85.22</td>
</tr>
<tr>
<td>Biaffine w/ PhoBERT_{large}</td>
<td>77.85 / 84.32</td>
</tr>
</tbody>
</table>
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\textsubscript{MLM+TLM} on NLI
  • PhoBERT does better than XLM-R on 5 downstream evaluation tasks
    • \textit{PhoBERT uses far fewer parameters than XLM-R}: 135M (PhoBERT-base) vs. 250M (XLM-R-base); 370M (PhoBERT-large) vs. 560M (XLM-R-large)
    • \textit{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
    • \textit{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

<table>
<thead>
<tr>
<th>ID</th>
<th>Form</th>
<th>POS</th>
<th>NER</th>
<th>Head</th>
<th>Rel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Đấy</td>
<td>PRON</td>
<td>O</td>
<td>2</td>
<td>sub</td>
</tr>
<tr>
<td>2</td>
<td>là</td>
<td>VERB</td>
<td>O</td>
<td>0</td>
<td>root</td>
</tr>
<tr>
<td>3</td>
<td>Cần_Thơ</td>
<td>NOUN</td>
<td>B-LOC</td>
<td>2</td>
<td>vmod</td>
</tr>
</tbody>
</table>
PhoNLP model

Dependency parsing

NER

POS Tagging

PhoBERT-based embeddings

BIAFFINE

Dense

CRF

Dense

B-LOC

Dense

PRON

softmax

Dense

VERB

softmax

Dense

NOUN

softmax

Day this

lai

Can Tho

Can Tho
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)
- Data leakage that has not been pointed out before: 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

<table>
<thead>
<tr>
<th>Task</th>
<th>Training</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS tagging (leakage)</td>
<td>27000</td>
<td>870</td>
<td>2120</td>
</tr>
<tr>
<td>POS tagging (Re-split)</td>
<td>23906</td>
<td>2009</td>
<td>3481</td>
</tr>
<tr>
<td>NER</td>
<td>14861</td>
<td>2000</td>
<td>2831</td>
</tr>
<tr>
<td>Dependency parsing</td>
<td>8977</td>
<td>200</td>
<td>1020</td>
</tr>
</tbody>
</table>
Evaluation

• Test results (employing the PhoBERT-base model)

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>POS</th>
<th>NER</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leak.</td>
<td>Single-task</td>
<td>96.7†</td>
<td>93.69</td>
<td>78.77†</td>
<td>85.22†</td>
</tr>
<tr>
<td></td>
<td>PhoNLP</td>
<td>96.76</td>
<td>94.41</td>
<td>79.11</td>
<td>85.47</td>
</tr>
<tr>
<td>Re-spl</td>
<td>Single-task</td>
<td>93.68</td>
<td>93.69</td>
<td>77.89</td>
<td>84.78</td>
</tr>
<tr>
<td></td>
<td>PhoNLP</td>
<td>93.88</td>
<td>94.51</td>
<td>78.17</td>
<td>84.95</td>
</tr>
</tbody>
</table>

Table 2: Performance results (in %) on the test sets for POS tagging (i.e. accuracy), NER (i.e. F1-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

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

What is the number of cars with more than 4 cylinders?

```sql
SELECT count(*) FROM CARS_DATA WHERE Cylinders > 4
```

Cho biết số lượng những chiếc xe có nhiều hơn 4 xi lanh

```sql
SELECT count(*) FROM [dữ liệu xe] WHERE [số lượng xi lanh] > 4
```
Semantic parsing

• Why do we want to do semantic parsing?

Sentence $\xrightarrow{\text{Semantic parser}}$ Meaning representation e.g. logical form $\xrightarrow{\text{Executor}}$ Response

Instructing a Robot

At the chair, turn left
Why do we want to do semantic parsing?

- 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

Example:

Which provinces border Ha Noi?

- Thai Nguyen, Vinh Phuc, Ha Nam, Hoa Binh, Bac Giang, Bac Ninh, Hung Yen, Phu Tho
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

![Text-to-SQL semantic parser](https://vinai.io)

Vietnamese examples:

```
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
```

```
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](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
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

<table>
<thead>
<tr>
<th></th>
<th>#Qu.</th>
<th>#SQL</th>
<th>#DB</th>
<th>#T/D</th>
<th>#Easy</th>
<th>#Med.</th>
<th>#Hard</th>
<th>#ExH</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>9691</td>
<td>5263</td>
<td>166</td>
<td>5.3</td>
<td>2233</td>
<td>3439</td>
<td>2095</td>
<td>1924</td>
</tr>
<tr>
<td>train</td>
<td>6831</td>
<td>3493</td>
<td>99</td>
<td>5.4</td>
<td>1559</td>
<td>2255</td>
<td>1502</td>
<td>1515</td>
</tr>
<tr>
<td>dev</td>
<td>954</td>
<td>589</td>
<td>25</td>
<td>4.2</td>
<td>249</td>
<td>405</td>
<td>191</td>
<td>109</td>
</tr>
<tr>
<td>test</td>
<td>1906</td>
<td>1193</td>
<td>42</td>
<td>5.7</td>
<td>425</td>
<td>779</td>
<td>402</td>
<td>300</td>
</tr>
</tbody>
</table>
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 làn.
SELECT count(*) FROM [dữ liệu xe] WHERE [số lượng xi làn] > 4

Original (Hard question—with a nested SQL query):
Which countries in europe have at least 3 car manufacturers?
WHERE T2.Continent = “europa” 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
Main results

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

<table>
<thead>
<tr>
<th>Approach</th>
<th>dev</th>
<th>test</th>
<th>Approach</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>EditSQL</td>
<td>28.6</td>
<td>24.1</td>
<td>IRNet</td>
<td>43.3</td>
<td>38.2</td>
</tr>
<tr>
<td>EditSQL XLM-R</td>
<td>55.2</td>
<td>51.3</td>
<td>IRNet XLM-R</td>
<td>58.6</td>
<td>52.8</td>
</tr>
<tr>
<td>EditSQL [MT]</td>
<td>22.8</td>
<td>17.4</td>
<td>IRNet [MT]</td>
<td>27.4</td>
<td>21.6</td>
</tr>
<tr>
<td>EditSQL</td>
<td>33.7</td>
<td>30.2</td>
<td>IRNet</td>
<td>49.7</td>
<td>43.6</td>
</tr>
<tr>
<td>EditSQL DeP</td>
<td>45.3</td>
<td>42.2</td>
<td>IRNet DeP</td>
<td>52.2</td>
<td>47.1</td>
</tr>
<tr>
<td>EditSQL PhoBERT</td>
<td>56.7</td>
<td>52.6</td>
<td>IRNet PhoBERT</td>
<td>60.2</td>
<td>53.2</td>
</tr>
<tr>
<td>En EditSQL [RoBERTa]</td>
<td>58.3</td>
<td>53.6</td>
<td>IRNet [RoBERTa]</td>
<td>63.8</td>
<td>55.3</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Approach</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>EditSQL [MT]</td>
<td>21.5</td>
<td>16.8</td>
</tr>
<tr>
<td>EditSQL</td>
<td>28.6</td>
<td>24.1</td>
</tr>
<tr>
<td>EditSQL$_{XLM-R}$</td>
<td>55.2</td>
<td>51.3</td>
</tr>
<tr>
<td>EditSQL [MT]</td>
<td>22.8</td>
<td>17.4</td>
</tr>
<tr>
<td>EditSQL$_{DeP}$</td>
<td>33.7</td>
<td>30.2</td>
</tr>
<tr>
<td>EditSQL$_{PhoBERT}$</td>
<td>45.3</td>
<td>42.2</td>
</tr>
<tr>
<td>EditSQL$_{RoBERTa}$</td>
<td>56.7</td>
<td>52.6</td>
</tr>
<tr>
<td><strong>En</strong> EditSQL$_{RoBERTa}$</td>
<td><strong>58.3</strong></td>
<td><strong>53.6</strong></td>
</tr>
<tr>
<td><strong>IRNet [MT]</strong></td>
<td><strong>25.4</strong></td>
<td><strong>20.3</strong></td>
</tr>
<tr>
<td>IRNet</td>
<td><strong>43.3</strong></td>
<td><strong>38.2</strong></td>
</tr>
<tr>
<td>IRNet$_{XLM-R}$</td>
<td><strong>58.6</strong></td>
<td><strong>52.8</strong></td>
</tr>
<tr>
<td>IRNet$_{DeP}$</td>
<td><strong>52.2</strong></td>
<td><strong>47.1</strong></td>
</tr>
<tr>
<td>IRNet$_{PhoBERT}$</td>
<td><strong>60.2</strong></td>
<td><strong>53.2</strong></td>
</tr>
<tr>
<td>IRNet$_{RoBERTa}$</td>
<td><strong>63.8</strong></td>
<td><strong>55.3</strong></td>
</tr>
</tbody>
</table>
Main results

- Exact matching accuracy categorized by 4 different hardness levels, and F1 scores of different SQL components on the test set

<table>
<thead>
<tr>
<th>Approach</th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
<th>ExH</th>
<th>SELECT</th>
<th>WHERE</th>
<th>ORDER BY</th>
<th>GROUP BY</th>
<th>KEYWORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>EditSQL_{DeP}</td>
<td>65.7</td>
<td>46.1</td>
<td>37.6</td>
<td>16.8</td>
<td>75.1</td>
<td>44.6</td>
<td>65.6</td>
<td>63.2</td>
<td>73.5</td>
</tr>
<tr>
<td>EditSQL_{XLM-R}</td>
<td>75.1</td>
<td>56.2</td>
<td>45.3</td>
<td>22.4</td>
<td>82.7</td>
<td>60.3</td>
<td>70.7</td>
<td>67.2</td>
<td>79.8</td>
</tr>
<tr>
<td>EditSQL_{PhoBERT}</td>
<td>75.6</td>
<td>58.0</td>
<td>47.4</td>
<td>22.7</td>
<td>83.3</td>
<td>61.8</td>
<td>72.5</td>
<td>67.9</td>
<td>80.6</td>
</tr>
<tr>
<td>IRNet_{DeP}</td>
<td>71.8</td>
<td>51.5</td>
<td>47.4</td>
<td>18.5</td>
<td>79.3</td>
<td>48.7</td>
<td>71.8</td>
<td>63.4</td>
<td>74.3</td>
</tr>
<tr>
<td>IRNet_{XLM-R}</td>
<td>76.2</td>
<td>57.8</td>
<td>46.8</td>
<td>23.5</td>
<td>83.5</td>
<td>59.1</td>
<td>74.4</td>
<td>68.3</td>
<td>80.5</td>
</tr>
<tr>
<td>IRNet_{PhoBERT}</td>
<td>76.8</td>
<td>57.5</td>
<td>47.2</td>
<td>24.8</td>
<td>84.5</td>
<td>59.3</td>
<td>76.6</td>
<td>68.2</td>
<td>80.3</td>
</tr>
</tbody>
</table>
Error analysis

• Causes of errors from 382 failed examples on the dev. set by IRNet\textsubscript{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 [sức chứa] 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 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!
Labels for $j \rightarrow i$

$\text{rel-dep}_i \text{ rel-head}_j$

$$s_{i,j} = \text{Biaffine}(h_i^{\text{head}}, h_j^{\text{dep}})$$

$\text{Biaffine}(y_1, y_2) = y_1^T U y_2 + W (y_1 \circ y_2) + b$

Bilinear  Linear