A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese

Anonymous submission

Abstract

Semantic parsing is an important NLP task. However, Vietnamese is a low-resource language in this research area. In this paper, we present the first public large-scale Text-to-SQL semantic parsing dataset for Vietnamese. We extend and evaluate two strong semantic parsing baselines EditSQL (Zhang et al., 2019) and IRNet (Guo et al., 2019) on our dataset. We compare two baselines with key configurations and find that: automatic Vietnamese word segmentation improves the parsing results of both baselines; the normalized pointwise mutual information (NPMI) score (Bouma, 2009) is useful for schema linking; latent syntactic features extracted from a dependency parser for Vietnamese also improve the results; and the monolingual language model PhoBERT for Vietnamese (Nguyen and Nguyen, 2020) helps produce higher performances than the recent best multilingual language model XLM-R (Conneau et al., 2020).

1 Introduction

Semantic parsing is the task of converting natural language sentences into meaning representations such as logical forms or standard SQL database queries (Mooney, 2007), which serves as an important component in many NLP systems such as Question Answering (QA) and Task-oriented Dialogue (Androustopoulos et al., 1995; Moldovan et al., 2003; Guo et al., 2018). The significant availability of the world’s knowledge stored in relational databases leads to the creation of large-scale Text-to-SQL datasets such as WikiSQL (Zhong et al., 2017) and Spider (Yu et al., 2018) which help boost the development of various state-of-the-art sequence-to-sequence (seq2seq) semantic parsers (Bogin et al., 2019; Zhang et al., 2019; Guo et al., 2019). Compared to WikiSQL, Spider presents challenges not only in handling complex questions but also in generalizing to unseen databases during evaluation.

Most SQL semantic parsing benchmarks such as WikiSQL and Spider are exclusively for English. Thus the development of semantic parsers has largely been limited to the English language. As SQL is a database interface and universal semantic representation, it is worth investigating the Text-to-SQL semantic parsing task for languages other than English. Especially, the difference in linguistic characteristics could add difficulties in applying state-of-the-art seq2seq semantic parsing models to the non-English languages (Min et al., 2019). For example, about 85% of word types in Vietnamese are composed of at least two syllables (Thang et al., 2008). Unlike English, in addition to marking word boundaries, white space is also used to separate syllables that constitute words in Vietnamese written texts; e.g. a 8-syllable written text “Có bao nhiêu quốc gia ở châu Âu” (How many countries in Europe) forms 5 words “Có bao nhiêu quốc gia ở châu Âu”. Thus it is interesting to study the influence of word segmentation in Vietnamese on its SQL parsing, i.e. syllable level vs. word level.

In terms of 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; Nguyen et al., 2012; Tung et al., 2015; Nguyen et al., 2017). Recently, Vuong et al. (2019) formulate the Vietnamese Text-to-SQL task as a sequence labeling-based slot filling problem and solve it using a conventional CRF model with handcrafted features, due to the simple structure of the input questions they deal with. Note that seq2seq-based semantic parsers have not yet been explored in any previous work w.r.t. Vietnamese.

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

In this paper, we introduce the first public large-scale Text-to-SQL dataset for the Vietnamese semantic parsing task. In particular, we create this dataset by manually translating the Spider dataset into Vietnamese. We empirically evaluate strong seq2seq baseline parsers EditSQL (Zhang et al., 2019) and IRNet (Guo et al., 2019) on our dataset. Extending the baselines, we extensively investigate key configurations and find that: (1) Our human-translated dataset is far more reliable than a dataset consisting of machine-translated questions, and the overall obtained result for Vietnamese is comparable to that for English. (2) Automatic Vietnamese word segmentation improves the performances of the baselines. (3) The NPMI score (Bouma, 2009) is useful for linking a cell value mentioned in a question to a column in the database schema. (4) Latent syntactic features dumped from a pre-trained dependency parser (Nguyen and Verspoor, 2018) for Vietnamese also helps improve the performances. (5) Highest improvements are accounted for the use of pre-trained language models, where PhoBERT (Nguyen and Nguyen, 2020) helps produce higher results than XLM-R (Conneau et al., 2020).

We hope that our dataset can serve as a starting point for future Vietnamese semantic parsing research and applications. We publicly release our dataset and code at: anonymous-url.

2 Our Dataset

We 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. Note that the original Spider consists of 10181 questions with their corresponding 5693 SQL queries over 200 databases. However, only 9691 questions and their corresponding 5263 SQL queries over 166 databases for the training and development sets are publicly available. Thus we could only translate those available ones.

The 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; and finally the NLP researcher verifies the original and corrected versions and makes further revisions if needed. Note that in case we have literal translation for a question, we stick to the style of the original English question as much as possible. Otherwise, for complex questions, we will rephrase them based on the semantic meaning of the corresponding SQL queries to obtain the most natural language questions in Vietnamese.

Following Yu et al. (2018) and Min et al. (2019), we split our dataset into training, development and test sets such that no database overlaps between them as detailed in Table 1. Examples of question and SQL query pairs from our dataset are presented in the Appendix. Note that translated question and SQL query pairs in our dataset are written at the syllable level. To obtain a word-level dataset, we apply RDRSegmenter (Nguyen et al., 2018) from VnCoreNLP (Vu et al., 2018) to perform automatic Vietnamese word segmentation.

3 Baseline Models and Extensions

3.1 Baselines

Recent state-of-the-art results on the Spider dataset are reported for RYANSQL (Choi et al., 2020) and RAT-SQL (Wang et al., 2020), which are based on the seq2seq encoder-decoder architectures. However, their implementations are not published at the time of our empirical investigation (their codes are still not yet publicly available on 31/05/2020). Thus we select seq2seq based models EditSQL (Zhang et al., 2019) and IRNet (Guo et al., 2019) with publicly available implementations as our baselines, which produce near state-of-the-art results on Spider. Due to the space limit, we briefly describe EditSQL and IRNet in the Appendix. See Zhang et al. (2019) and Guo et al. (2019) for more details.

3.2 Our Extensions

NPMI for schema linking: IRNet essentially relies on the large-scale knowledge graph ConceptNet (Speer et al., 2017) to link a cell value mentioned in this question to a corresponding column in the database schema. For Vietnamese, we use the large-scale knowledge graph ConceptNet (Speer et al., 2017) to link a cell value mentioned in this question to a corresponding column in the database schema. However, the original ConceptNet is not fully accessible for Vietnamese. Thus we build a Vietnamese ConceptNet by manually translating the English-ConceptNet dictionary into Vietnamese. This Vietnamese ConceptNet is then used to link a cell value mentioned in a question to a column in the database schema. For example, in the question “Which table contains the column ‘name’?”, the cell value “name” is linked to the column “name” in the database schema. This linking process is crucial for the Vietnamese semantic parsing task, as it helps improve the performance of the baselines.

Table 1: Statistics of our human-translated dataset, “#Qu.”, “#SQL” and “#DB” denote the numbers of questions, SQL queries and databases, respectively. “#T/D” abbreviates the average number of tables per database. “#Easy”, “#Med.”, “#Hard” and “#ExH” denote the numbers of questions categorized by their SQL queries’ hardness levels of “easy”, “medium”, “hard” and “extra hard” (defined by Yu et al.), respectively.

<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>2333</td>
<td>3439</td>
<td>2095</td>
<td>1924</td>
</tr>
<tr>
<td>train</td>
<td>6831</td>
<td>3403</td>
<td>96</td>
<td>5.3</td>
<td>1599</td>
<td>2553</td>
<td>1302</td>
<td>715</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>

We hope that our dataset can serve as a starting point for future Vietnamese semantic parsing research and applications. We publicly release our dataset and code at: anonymous-url.
in a question to a column in the database schema based on two ConceptNet categories ‘is a type of’ and ‘related terms’. However, these two ConceptNet categories are not available for Vietnamese. Thus we propose a novel use of the NPMI collocation score (Bouma, 2009) for the schema linking in IRNet, which ranks the NPMI scores between the cell values and column names to match a cell value to its column.

Latent syntactic features: Previous works have shown that syntactic features help improve semantic parsing (Monroe and Wang, 2014; Jie and Lu, 2018). Unlike these works which use handcrafted syntactic features extracted from dependency parse trees, and inspired by Zhang et al. (2017)’s relation extraction work, we investigate whether latent syntactic features extracted from the BiLSTM-based dependency parser jPTDP (Nguyen and Verspoor, 2018) pre-trained for Vietnamese would help improve Vietnamese Text-to-SQL parsing. In particular, our approach is that we 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. Pre-trained language models: Zhang et al. (2019) and Guo et al. (2019) make use of BERT (Devlin et al., 2019) to obtain near state-of-the-art on Spider. Thus we also extend EditSQL and IRNet with the use of pre-trained language models XLM-R-base (Conneau et al., 2020) and PhoBERT-base (Nguyen and Nguyen, 2020) for the syllable level and the word level, respectively. XLM-R is the recent best multi-lingual model, based on RoBERTa (Liu et al., 2019), pre-trained on a 2.5TB multilingual corpus which contains 137GB of syllable-level Vietnamese texts. PhoBERT is a monolingual variant of RoBERTa for Vietnamese, pre-trained on a 20GB of word-level Vietnamese texts.

4 Experiments

4.1 Experimental Setup

We conduct experiments to study a quantitative comparison between our human-translated dataset and a machine-translated dataset,\(^1\) the influence of Vietnamese word segmentation i.e. syllable level and word level, and the usefulness of the latent syntactic features, the pre-trained language models and the NPMI-based approach for schema linking.

\(^1\)We employ a well-known machine translation engine to translate the English questions into Vietnamese.

<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>21.5</td>
<td>16.8</td>
<td>IRNet</td>
<td>25.4</td>
<td>20.3</td>
</tr>
<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 ×LML-R</td>
<td>55.2</td>
<td>51.3</td>
<td>IRNet ×LML-R</td>
<td>58.6</td>
<td>52.8</td>
</tr>
<tr>
<td>EditSQL [MT]</td>
<td>22.8</td>
<td>17.5</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 ×DP</td>
<td>45.3</td>
<td>42.2</td>
<td>IRNet ×DP</td>
<td>52.2</td>
<td>47.1</td>
</tr>
<tr>
<td>EditSQL ×BERT</td>
<td>56.7</td>
<td>52.6</td>
<td>IRNet ×BERT</td>
<td>60.2</td>
<td>53.2</td>
</tr>
</tbody>
</table>

Table 2: Exact matching accuracies of EditSQL and IRNet. “Vi-Syllable” and “Vi-Word” denote the results w.r.t. the syllable level and the word level, respectively. [MT] denotes accuracy results with the machine-translated questions. The subscript “DeP” refers to the use of the latent syntactic features. Other subscripts denote the use of the pre-trained language models.

We run for 10 training epochs and evaluate the exact matching accuracy after each epoch on the development set, and then select the best model checkpoint to report the final result on the test set. Due to the space limit, we detail the experimental setup in the Appendix where readers can also find more details of evaluation metrics and hyper-parameter settings as well as training procedure.

4.2 Main Results

Table 2 shows the overall exact matching results of EditSQL and IRNet on the development and test sets. Clearly, IRNet does better than EditSQL, which is consistent with results obtained on the original Spider dataset.

We find that our human-translated dataset is far more reliable than a dataset consisting of machine-translated questions. In particular, at the word level, compared to the machine-translated dataset, our dataset obtains about 30.2-17.4 ≈ 13% and 43.6-21.6 = 22% absolute improvements in accuracies of EditSQL and IRNet, respectively (i.e. 75%–100% relative improvements). In addition, the word-based Text-to-SQL parsing obtains about 5+ absolute higher accuracies than the syllable-based Text-to-SQL parsing (EditSQL: 24.1%→30.2% ; IRNet: 38.2%→43.6%), i.e. automatic Vietnamese word segmentation improves the accuracy results.

Furthermore, latent syntactic features dumped from the pre-trained dependency parser jPTDP-v2 for Vietnamese helps improve the performances of the baselines (EditSQL: 30.2%→42.2% ; IRNet: 43.6%→47.1%). Also, biggest improvements are accounted for the use of pre-trained language mod-
els. In particular, PhoBERT helps produce higher results than XLM-R (EditSQL: 52.6% vs. 51.3%; IRNet: 53.2% vs. 52.8%).

We also retrain EditSQL and IRNet with the use of the strong pre-trained language model RoBERTa instead of BERT, but under our dataset split. We find that the overall results for Vietnamese are smaller but comparable to the English results. Thus Text-to-SQL semantic parsing for Vietnamese might not be significantly more challenging than that for English.

Table 3 shows the exact matching accuracies of EditSQL and IRNet w.r.t. different hardness levels of SQL queries and the F\textsubscript{1} scores w.r.t. different SQL components on the test set. Clearly, in most cases, the pre-trained language models help produce substantially higher improvements to the baselines compared to the latent syntactic features, especially for the WHERE component.

NPMI-based schema linking: We also investigate the contribution of our NPMI-based extension approach for schema linking in applying IRNet for Vietnamese. Without using NPMI for schema linking,\textsuperscript{2} we observe 6+% absolute decrease in the exact matching accuracies of IRNet on both development and test sets, thus showing the usefulness of our NPMI-based approach for schema linking.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
<th>ExH</th>
<th>SELECT WHERE ORDER BY GROUP BY KEYWORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>EditSQL\textsuperscript{DeP}</td>
<td>65.7</td>
<td>46.1</td>
<td>37.6</td>
<td>16.8</td>
<td>75.1</td>
</tr>
<tr>
<td>EditSQL\textsuperscript{XLM-R}</td>
<td>75.1</td>
<td>56.2</td>
<td>45.3</td>
<td>22.4</td>
<td>82.7</td>
</tr>
<tr>
<td>EditSQL\textsuperscript{PhoBERT}</td>
<td>75.6</td>
<td>58.0</td>
<td>47.4</td>
<td>22.7</td>
<td>83.3</td>
</tr>
<tr>
<td>IRNet\textsuperscript{DeP}</td>
<td>71.8</td>
<td>51.5</td>
<td>47.4</td>
<td>18.5</td>
<td>79.3</td>
</tr>
<tr>
<td>IRNet\textsuperscript{XLM-R}</td>
<td>76.2</td>
<td>57.8</td>
<td>46.8</td>
<td>23.5</td>
<td>83.5</td>
</tr>
<tr>
<td>IRNet\textsuperscript{PhoBERT}</td>
<td>76.8</td>
<td>57.5</td>
<td>47.2</td>
<td>24.8</td>
<td>84.5</td>
</tr>
</tbody>
</table>

Table 3: Exact matching accuracy categorized by different hardness levels and F\textsubscript{1} scores of different SQL components on the test set. “ExH” abbreviates Extra Hard.

4.3 Error Analysis

To understand the source of errors, we perform an error analysis on the development set which consists of 954 questions. Using IRNet\textsuperscript{PhoBERT} which produces the best result, we identify several causes of errors from 382/954 failed examples.

For 121/382 cases (32%), IRNet\textsuperscript{PhoBERT} makes incorrect predictions on the column names which are not mentioned or only partially mentioned in the questions. For example, given the question “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),\textsuperscript{3} the model produces an incorrect column name prediction of “tên” (name) instead of the correct one “tên bài hát” (song name). Errors related to column name predictions can either be missing the entire column names or inserting random column names into the WHERE component of the predicted SQL queries.

About 12% of failed examples (47/382) in fact have an equivalent implementation of their intent with a different SQL syntax. For example, the model produces a ‘failed’ SQL output “SELECT MAX [sức chứa] FROM [sân vận động]” which is equivalent to the gold SQL query of “SELECT [sức chứa] FROM [sân vận động] ORDER BY [sức chứa] DESC LIMIT 1”, i.e. the SQL output would be valid if we measure an execution accuracy.

About 22% of failed examples (84/382) are caused by nested and complex SQL queries which mostly belong to the Extra Hard category. With 18% of failed examples (70/382), incorrectly predicting operators is another common type of errors. For example, given the phrases “già nhất” (oldest) and “trẻ nhất” (youngest) in the question, the model fails to predict the correct operators max and min, respectively. The remaining 60/382 cases (16%) are accounted for an incorrect prediction of table names in a FROM clause.

5 Conclusion

We have created the first public large-scale dataset for Vietnamese Text-to-SQL semantic parsing. We also extensively experiment with key research configurations using two strong English models on our dataset and find that: the input representations, the NPMI-based approach for schema linking, the latent syntactic features and the pre-trained language models all have the influence on this Vietnamese-specific task. We hope that our dataset can serve as the starting point for further research and applications in Vietnamese QA and dialogue systems.

\textsuperscript{2}Word segmentation is not shown for simplification.

\textsuperscript{3}With our schema linking, IRNet assigns a ‘NONE’ type for column names.
References


