TP-Link: Fine-grained Pre-Training for Text-to-SQL Parsing with Linking Information

Ziqiang Liu[1], Shujie Li[2], Zefeng Cai[2], Xiangyu Li[2], Yunshui Li[1], Lei Zhang[1], Chengming Li[3*], Xiping Hu[3], Ruifeng Xu[4], Min Yang[1*]
[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]University of Science and Technology of China, [3]Shenzhen MSU-BIT University
[4]Harbin Institute of Technology (Shenzhen)
{zq.liu4, ys.li, lei.zhang2, min.yang}@siat.ac.cn
{ustclsj, galileoc, xyli1}@mail.ustc.edu.cn

Abstract

In this paper, we introduce an innovative pre-training framework ***TP-Link***, which aims to improve context-dependent ***T***ext-to-SQL ***P***arsing by leveraging ***Link***ing information. This enhancement is achieved through better representation of both natural language utterances and the database schema, ultimately facilitating more effective text-to-SQL conversations. We present two novel pre-training objectives: (i) utterance linking prediction (ULP) task that models intricate syntactic relationships among natural language utterances in context-dependent text-to-SQL scenarios, and (ii) schema linking prediction (SLP) task that focuses on capturing fine-grained schema linking relationships between the utterances and the database schema. Extensive experiments demonstrate that our proposed TP-Link achieves state-of-the-art performance on two leading downstream benchmarks (*i.e.* , SParC and CoSQL).

Keywords: text-to-SQL, pre-training, semantic parsing

1. Introduction

Text-to-SQL semantic parsing task (Zhong et al., 2017; Yu et al., 2018) plays a crucial role in transforming natural language utterances into corresponding SQL queries, which can be executed on structured ontologies like databases or knowledge bases. Furthermore, it has evolved into conversational text-to-SQL semantic parsing (Yu et al., 2019a,b), which allows the conversion of user's real-time utterance into precise SQL query within multi-turn dialogues. The refined task not only extends the scope of the original text-to-SQL semantic parsing but also integrates historical utterances to generate accurate and comprehensive SQL queries, allowing users to engage in more intricate inquiries, streamlining knowledge retrieval and aligning effectively with practical requirements. Pre-trained language models (PLMs) have been demonstrated to be powerful in enhancing text-to-SQL parsing task, achieving impressive performance attributed to rich prior language knowledge in large-scale corpora. However, previous research (Yin et al., 2020; Yu et al., 2021a) has shown inherent differences in the distributions between tables and plain text, leading to sub-optimal performance of general pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020). Recent studies (Shi et al., 2021; Yu et al., 2021a; Liu et al., 2022; Deng et al., 2021) have mitigated

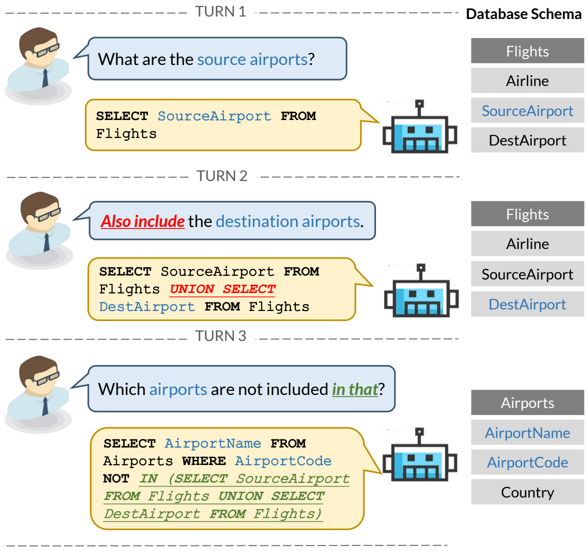

Figure 1: An example of cross-domain context-dependent text-to-SQL.

the aforementioned limitations by designing custom table-based language models for text-to-SQL parsing, which encode both natural language utterances and tables simultaneously. As the scale of pre-trained language models increases, large language models (LLMs) have emerged. Unlike models like BERT and T5 (Raffel et al., 2023) that require fine-tuning with a small amount of data, large language models like GPT-3 (Brown et al., 2020) require prompt engineering to generate tar-

* Corresponding authors.
Equal contribution.

LREC-COLING 2024, pages 16686–16697
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

get outputs, which exhibits remarkable zero-shot and few-shot capabilities. Recent study (Liu et al., 2023) indicates that even without using any training data, ChatGPT* still possesses strong text-to-SQL capabilities, although there remains some gap compared to current state-of-the-art models.

While PLMs and LLMs have shown promising results for text-to-SQL parsing tasks, multi-turn text-to-SQL task has consistently encountered two major challenges:

Firstly, leveraging contextual historical information from the dialogue is pivotal, particularly the syntactic coreference relationships within the context, to ensure the accurate generation of SQL statements. As shown in Figure 1, there are numerous instances of coreference (*e.g.*, **"in that"**) and ellipsis (*e.g.*, **"Also include"**), making this task very challenging. SCoRe (Yu et al., 2021b) focuses on identifying semantic switches in nearby utterances, neglecting to address long-range semantic dependencies. STaR (Cai et al., 2022) captures semantic dependencies at the sentence level through SQL similarity comparison, yet it does not effectively model contextual coreference relationships at finer-grained word/token level.

Secondly, context-dependent text-to-SQL parsing task is challenging due to potential SQL semantic inheritance in consecutive dialogue turns, which means a current SQL query might be modified by previous queries. However, when a context switch occurs, these information becomes redundant, affecting the final SQL generation. Therefore, it's essential to model a more nuanced relationship between the context and the schema. As shown in Figure 1, there exists considerable linking information between utterances and database schemas(*e.g.*, **"source airports"** in utterance and **"SourceAirport"** in database schema). In recent study, RASAT (Qi et al., 2022) adjusts its self-attention mechanism to relational self-attention, incorporating diverse relation information to boost encoding capabilities, but it does not consider information redundancy. CQR-SQL (Xiao et al., 2022) simplifies the schema linking information fused with downstream parsing models by rewriting multi-turn dialogues, but requires additional training and extensive complex annotation work. MIGA (Fu et al., 2022) integrates referential relationship and schema linking information through a multi-task approach but neglects redundancy. SCoRe (Yu et al., 2021b) predict SQL keywords using only the current turn's information, overlooking previous dialogues' schema linking information. STaR (Cai et al., 2022) tracks states at the schema level but does not consider fine-grained relationship between contextual utterances and the schema.

In this paper, we present an innovative pre-training framework TP-Link aimed at improving context-dependent text-to-SQL parsing task by leveraging linking information to address all the challenges mentioned above. This enhancement is achieved through better representation of both natural language utterances and the database schema, ultimately facilitating more effective text-to-SQL conversations. We present two novel pre-training objectives: utterance linking prediction (ULP) task, which models intricate syntactic relationships among natural language utterances in context-dependent text-to-SQL scenarios, and schema linking prediction (SLP) task, focused on fine-grained schema linking relationships between the utterance and the database schema.

We evaluated TP-Link on SParC (Yu et al., 2019b) and CoSQL (Yu et al., 2019a) datasets, and our main contributions of this work are summarized as follows:

- We introduce a utterance linking prediction (ULP) task to explicitly model word-level coreference relation within the context, effectively addressing complex coreference and ellipsis issues in multi-turn dialogues.

- We introduce a fine-grained schema linking prediction (SLP) task to ensure more precise schema linking, and enable the current utterance to focus on critical schema linking information from preceding utterances. Subsequent to the application of similarity filtering, the model allocates a greater degree of attention to pertinent schema linking information.

- Experimental results show that TP-Link achieves new state-of-the-art results on two context-dependent text-to-SQL datasets, SParC and CoSQL.

2. Methodology

2.1. Task Definition

In this section, we first introduce the task definition of conversational text-to-SQL parsing. The objective of the multi-turn text-to-SQL task is to generate a SQL query q_t corresponding to the current turn t given the user's current utterance u_t, the history of utterances $H_t = [u_1, u_2, ..., u_i, ..., u_{t-1}]$, and the database schema $S = [s_1, s_2, ..., s_j, ..., s_m]$ composed of m tables. Specifically, the i-th utterance is composed of n_i words and can be formally represented as $u_i = [w_{1_i}, w_{2_i}, ..., w_{n_i}]$. The j-th table consists of k_j columns and can be formally represented as $s_j = [t_j, c_{1_j}, c_{2_j}, ..., c_{k_j}]$, where t_j and c_j represent the table name and column names of the schema, respectively.

*https://chat.openai.com/

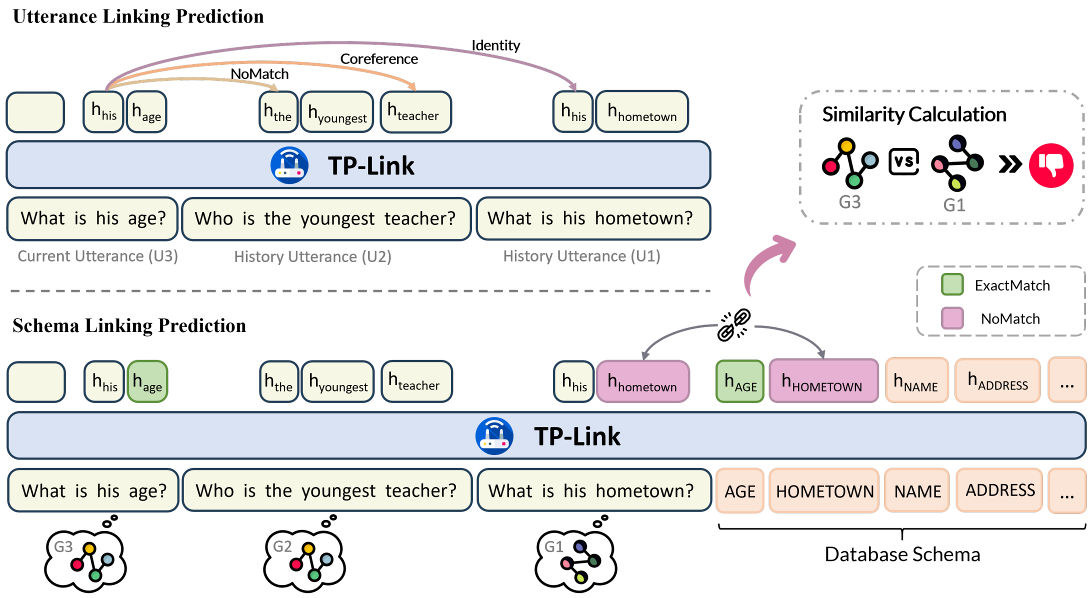

Figure 2: The overview of the proposed TP-LINK framework consisting of two novel pre-training objectives: Utterance Linking Prediction ($\mathcal{L}_{ULP}$) and Schema Linking Prediction ($\mathcal{L}_{SLP}$). The top part of the figure shows ULP and the bottom part shows SLP. For brevity, we do not show all the relationships and the masked language modeling objective ($\mathcal{L}_{MLM}$) here.

2.2. Pre-training Objectives

In this section, we propose two innovative pre-training objectives **ULP** (**U**tterance **L**inking **P**rediction) and **SLP** (**S**chema **L**inking **P**rediction) with the aim of effectively capturing intricate syntax relations within utterances and establishing refined schema linking relations in context-dependent text-to-SQL scenarios. In the subsequent section, we will provide a comprehensive exposition of the details of the aforementioned pre-training objectives. The overall model architecture is depicted in Figure 2.

2.2.1. Utterance Linking Prediction

To address the challenge of coreference resolution and ellipsis in context-dependent text-to-SQL scenarios, we propose Utterance Linking Prediction (ULP) in a self-supervised manner to capture intricate syntax relations within utterances in context-dependent text-to-SQL scenarios and address coreference and ellipsis issues in multi-turn dialogues.

Longer Contextual Reference Modeling. Previous models (Yu et al., 2021b) either could not model long-distance contextual dependencies or could not accurately model coreference relationships within the long context. Theoretically, TP-LINK can model referential relationships between the current utterance and historical utterances in contexts of any length, thereby enhancing the model's ability to learn referential relationships in long contexts.

Label	ULP	SLP	Meaning
No-Match	✓		No syntactic linking relation.
Partial-Match	✓		Local-match relation.
Exact-Match	✓		Global-match relation.
Identity	✓	-	The same word.
Coreference	✓	-	Coreference relation.
Generic	-	✓	"$\star$" relation(SQL keyword).

Table 1: Supervised Labels and their meaning for ULP and SLP Tasks.

Finer-grained Coreference Resolution. Formally, we first utilize the coreference resolution tool *NeuralCoref*[†] to resolve the word-level syntactic relationships between the present utterance u_t and the entirety of utterances $H_t = \{u_1, ..., u_t\}$, resulting in the acquisition of supervised labels of ULP task. The proposed syntactic relationships are shown in Table 1. As shown in Figure 2, there is a coreference relationship between **"his"** and **"youngest teacher"**.

At t-th turn, the goal of ULP task is to predict the word-level syntactic relationships between U_t and U_t given all the utterances U_t and database schema S. That is, at the t-th turn, the input I_t of the ULP task is as:

$$I_t = [\{u_1, \dots, u_t\}; \{s_1, \dots, s_m\}] \quad (1)$$

which m denotes the total number of

[†]https://github.com/huggingface/neuralcoref

schema items across all tables. Then we can get the output representation $H_t = [h_t^1, ..., h_t^{|u_1|}, ..., h_t^{|u_1|+...+|u_t|}, ..., h_t^{|u_1|+...+|s_m|}]$, which $|\cdot|$ denotes the total number of tokens of utterances and schema items, and from this, we extract the token-level representation of all the utterances, denoted as $H_t^u = [h_t^1, ..., h_t^{|u_1|}, ..., h_t^{|u_1|+...+|u_t|}]$. Then, we aggregate subwords of each word to map tokens to words, obtaining word-level representations $H_t^{u_w} = [h_t^{w_1}, ..., h_t^{w_{n_1}}, ..., h_t^{w_{n_t}}]$. Next, we calculate the matrix multiplication of $H_t^{u_w}$ and $H_t^{u_w \top}$ as the heuristic representation and predict the word-level syntactic relationships, the functions are as follows:

$$P(u_t) = \text{softmax}(W_2(\tilde{H}_t^{u_w}) + b_2) \tag{2}$$

$$\tilde{H}_t^{u_w} = H_t^{u_w} \times H_t^{u_w \top} \tag{3}$$

$$H_t^{u_w} = \text{LayerNorm}(\text{GELU}((W_1(H_t^{u_w}) + b_1))) \tag{4}$$

$$H_t^{u_w} = \text{SubwordAggregation}(H_t^u) \tag{5}$$

where W_i and b_i are trainable parameters. We use attentive pooling function (Lin et al., 2017) for the implementation of subword aggregation.

Finally, the pre-training loss function of ULP task is defined as the cross-entropy between the heuristic representation $\tilde{H}_t^{u_w}$ and the gold word-level syntactic relationship labels Y_t^u :

$$\mathcal{L}_{\text{ULP}} = -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Y_i^j \log P(u_t)_i^j \tag{6}$$

which n denotes the total number of words of the whole utterances and i, j denotes the i-th and j-th word of the utterances, respectively.

2.2.2. Schema Linking Prediction

In context-dependent text-to-SQL scenarios, schema linking relationships are used to determine which table or column names in the database schema correspond to the entities or relationships mentioned in the natural language utterance. However, considering the redundancy issue, not all schema linking information is useful. We propose a refined Schema Linking Prediction (SLP) task to capture finer-grained schema linking relationships. We use SQL tree edit distance as structural similarity to filter schema linking relationships, which allows pre-trained language models to learn more accurate schema linking knowledge, further enhancing performance.

More accurate schema linking. Formally, our initial step involves the resolution of comprehensive schema linking relationships between utterances and database schema items at t-th turn. To address the issue of redundant schema linking relationships in context-dependent text-to-SQL scenarios, we propose a methodology to obtain refined schema linking relationships by measuring SQL structure similarity. Concretely, we employ the tree-based edited distance algorithm (Pawlik and Augsten, 2016) to compute the SQL structure similarity between the current utterance and historical utterances given the current SQL q_t and historical SQLs $\{q_1, ..., q_{t-1}\}$. Mathematically, we parse SQL q_t and $\{q_1, ..., q_{t-1}\}$ to tree-based structure G_t and $\{G_1, ..., G_{t-1}\}$, then compute the SQL structure similarity of G_t and $G_i, i \in [1, t-1]$ as:

$$f_{\text{similarity}}(G_t, G_i) = \text{APTED}(G_t, G_i) \tag{7}$$

where APTED denotes *All Path Tree Edit Distance*, we refer the readers to Pawlik and Augsten (2016) for the implementation details. And intuitively, when the SQL structure similarity is lower, it indicates a higher level of redundancy in the schema linking relationships. Therefore, we refined the entire schema linking relationships according to the SQL structure similarity with a similarity threshold α. This refinement process ultimately yields the final SLP task labels. The proposed schema linking relationships are shown in Table 1. As shown in Figure 2, although "hometown" in the utterance should ideally have an "ExactMatch" relationship with the column name "hometown" in the schema, it has been labeled as "NoMatch" due to the low similarity between G_3 and G_1 falling below the threshold α.

Finer-grained schema linking. Similar to ULP task, we predict the refined schema linking relationships according to the heuristic representation $\tilde{H}_t^s$ between the utterances representation H_t^u and schema representation $H_t^s = [h_t^{|u_1|+...+|u_t|+1}, ..., h_t^{|u_1|+...+|s_m|}]$, the functions are as follows:

$$P(u_t, s) = \text{softmax}(W_4(\tilde{H}_t^{s_w}) + b_4) \tag{8}$$

$$\tilde{H}_t^{s_w} = H_t^{u_w} \times H_t^{s_w \top} \tag{9}$$

$$H_t^{s_w} = \text{LayerNorm}(\text{GELU}((W_3(H_t^{s_w}) + b_3))) \tag{10}$$

$$H_t^{s_w} = \text{SubwordAggregation}(H_t^s) \tag{11}$$

where W_i and b_i are trainable parameters.

Finally, the pre-training loss function of SLP task is defined the cross-entropy between the heuristic representation $\tilde{H}_t^{s_w}$ and the gold schema linking labels Y_t^s :

$$\mathcal{L}_{\text{SLP}} = -\frac{1}{n \cdot k}\sum_{i=1}^{n}\sum_{j=1}^{k} Y_i^j \log P(u_t, s)_i^j \tag{12}$$

where n denotes the total number of words in the whole utterances, k denotes the number of columns in the schema, i denotes the i-th word of utterances and j denotes the j-th column of schema.

2.2.3. Masked Language Modeling

Masked Language Modeling (MLM) is a pretraining task in BERT (Devlin et al., 2019), aiming to learn the contextual modeling ability of natural language text. To enhance the generalization ability of pre-trained language models, we retain the MLM task during the pretraining phase. Specifically, given the t-th turn of dialogue input I_t, the MLM task randomly selects a token and replaces it with a `[MASK]` token. Finally, it predicts the original token of `[MASK]` token based on context. We apply the original 15% probability of masking operations from BERT. The loss for the MLM task is denoted as $\mathcal{L}_{\mathrm{MLM}}$, and this loss function primarily minimizes the cross-entropy of the `[MASK]` tokens.

2.2.4. Joint Pre-training Objective

Following STaR (Cai et al., 2022), when integrating the loss functions of the three tasks, we do not directly use a weighted linear sum of these three loss functions. Instead, it takes into account the homoscedastic uncertainty of each loss function (Kendall et al., 2018), avoiding the need to spend a significant amount of time fine-tuning the weights of the loss functions. The summation formula for the loss functions based on homoscedastic uncertainty is as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{joint}} &= \frac{1}{2\delta_1^2}\mathcal{L}_{\mathrm{ULP}} + \frac{1}{2\delta_2^2}\mathcal{L}_{\mathrm{SLP}} + \frac{1}{2\delta_3^2}\mathcal{L}_{\mathrm{MLM}} \\ &+ log(1+\delta_1) + log(1+\delta_2) + log(1+\delta_3)\end{aligned} \tag{13}$$

where $\delta_1, \delta_2, \delta_3$ represent the observation noise parameters of the model.

3. Experimental Setup

3.1. Downstream Datasets

We conduct experiments on two prominent context-dependent semantic parsing benchmarks: (1) SParC (Yu et al., 2019b) is a cross-domain multi-turn text-to-SQL dataset, which encompasses $4,298$ question turns, featuring a substantial corpus of approximately $12k+$ natural language questions, each meticulously annotated with its corresponding SQL expression in the form of question-SQL pairs. (2) CoSQL (Yu et al., 2019a) is a conversational text-to-SQL corpus, which comprises a comprehensive collection of $30k+$ turns plus $10k+$ annotated SQL queries. Both CoSQL and SParC contain 200 complex databases spanning 138 distinct domains. Notably, CoSQL is particularly a more challenging benchmark due to its alignment with real-world application scenarios compared to SParC.

3.2. Baseline Models

We present a comparative analysis of various models to demonstrate the effectiveness of TP-Link. The models compared are primarily classified into two categories: semantic parsing methods and tabular knowledge pre-trained language models.

We first compare TP-Link against previous context-dependent parsing systems: **EditSQL** (Zhang et al., 2019), **GAZP** (Zhong et al., 2020), **IGSQL** (Cai and Wan, 2020), **RichContext** (Liu et al., 2020), **IST-SQL** (Wang et al., 2021), **R^{2}SQL** (Hui et al., 2021), **DELTA** (Chen et al., 2021), **RAT-SQL** (Wang et al., 2020),**PICARD** (Scholak et al., 2021),**UNIFIEDSKG** (Xie et al., 2022), **RASAT** (Qi et al., 2022), **HIE-SQL** (Zheng et al., 2022), **CQR-SQL** (Xiao et al., 2022), **MIGA** (Fu et al., 2022).

Since TP-Link mainly focus on pre-training improvements for context-dependent parsing models, we compare the performance of different pre-training model variants with the same downstream model. Specifically, the pre-trained models for comparison are as follows: **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019), **GraPPa** (Yu et al., 2021a), **SCoRe** (Yu et al., 2021b), **STaR** (Cai et al., 2022).

Furthermore, we also compare the performance of TP-Link with **ChatGPT** (Liu et al., 2023), one of the most powerful zero-shot models in context-dependent text-to-SQL scenarios.

3.3. Implementation Details

In the pretraining phase, we initialize our pretrained language model using the ELECTRA (Clark et al., 2020). We retained the *Replaced Token Detection* (RTD) task from ELECTRA as part of the masked language modeling task to further enhance the model's performance. The objective of this task is to improve the language understanding ability of the pretrained language model and prevent misleading predictions in downstream tasks. Next, we continually pre-train the ELECTRA on a synthetic text-to-SQL corpus consisting of $480k$ examples, following the methodology introduced by Cai et al. (2022). As a result, we obtain our proposed TP-Link.

Phrase	Params	Value
Pre-training	Model	ELECTRA
	Mask Rate	0.15
	Max Squence Length	256
	Optimizer	Adam
	Learning Rate	$1e-6$
	Gradient Clipping	1
	Batch Size	30
	Structure Similarity α	0.5
Fine-tuning	Model	LGESQL
	Batch Size	100

Table 2: Some of the parameters and values used during pre-training and fine-tuning.

In the downstream phase, we choose the LGESQL (Cao et al., 2021) as the downstream

MODEL	SPARC		COSQL	
	QM(%)	IM(%)	QM(%)	IM(%)
Previous Parsing Systems.				
EDITSQL + BERT (Zhang et al., 2019)	47.2	29.5	39.9	12.3
GAZP + BERT (Zhong et al., 2020)	48.9	29.7	42.0	12.3
IGSQL + BERT (Cai and Wan, 2020)	50.7	32.5	44.1	15.8
RICHCONTEXT + BERT (Liu et al., 2020)	52.6	29.9	41.0	14.0
IST-SQL + BERT (Wang et al., 2021)	47.6	29.9	44.4	14.7
R^2SQL + BERT (Hui et al., 2021)	54.1	35.2	45.7	19.5
DELTA + BART (Chen et al., 2021)	58.6	35.6	51.7	21.5
RAT-SQL + SCORE (Wang et al., 2020)	62.2	42.5	52.1	22.0
T5-3B + PICARD (Scholak et al., 2021)	-	-	56.9	24.2
UNIFIEDSKG (Xie et al., 2022)	61.5	41.9	54.1	22.8
RASAT + PICARD (Qi et al., 2022)	66.7	47.2	58.8	27.0
HIE-SQL + GRAPPA (Zheng et al., 2022)	64.7	45.0	56.4	28.7
CQR-SQL + ELECTRA (Xiao et al., 2022)	**67.8**	48.1	58.4	29.4
MIGA (Fu et al., 2022)	67.3	**48.9**	**59.0**	**29.8**
Zero-shot Models.				
ChatGPT (Liu et al., 2023)	37.6	20.1	37.9	13.0
Pre-trained Models.				
LGESQL	52.4	31.3	41.2	15.0
w. BERT (Devlin et al., 2019)	59.8	40.5	50.7	20.8
w. ROBERTA (Liu et al., 2019)	61.6	41.2	51.9	20.8
w. GRAPPA (Yu et al., 2021a)	62.5	42.4	52.6	21.5
w. SCORE (Yu et al., 2021b)	62.3	43.6	52.3	22.5
w. STAR (Cai et al., 2022)	**66.9**	**46.9**	**59.7**	**30.0**
w. **TP-LINK**	**68.0 (↑ 0.2 / 1.1)**	**50.0 (↑ 1.1 / 3.1)**	**60.7 (↑ 1.7 / 1.0)**	**31.7 (↑ 1.9 / 1.7)**

Table 3: Experimental results of various methods in terms of question match (QM) accuracy and interaction match (IM) accuracy on both SPARC and COSQL dev datasets. "-" means that the results are not accessible. The indicators in the parentheses represent the improvements of our model in comparison to the best results from the previous parsing system and pre-trained models, respectively. All percentage changes in this paper are reported as absolute values.

inference model which performs well in single-turn text-to-SQL semantic parsing tasks. Following that, we replace the original ELECTRA in LGESQL with our pre-trained TP-LINK, followed by fine-tuning on the downstream datasets. In this paper, the parameters of the downstream LGESQL are kept mostly consistent with the original model, except for the direct concatenation of the current utterance and historical utterances as part of the input. Furthermore, to thoroughly validate the effectiveness of our model, we conduct additional combined experiments with other pre-trained language models based on LGESQL as the foundational downstream model. Table 2 lists some of the parameters and values used during model pre-training and fine-tuning phrase.

3.4. Evaluation Metrics

We use two evaluation metrics to intuitively demonstrate the performance of our model. One is the *Question Match Accuracy* (QM), which indicates whether the SQL query generated by the model matches the actual SQL query exactly. The other metric is the *Interaction Match Accuracy* (IM), which accounts for the QM score of each question in a multi-turn dialogue interaction.

4. Experiment

4.1. Main Result

The final results are shown in Table 3. TP-LINK represents the model proposed in this paper. Some experimental results are referenced from STAR (Cai et al., 2022) and MIGA (Fu et al., 2022). For the results of ChatGPT, we directly utilize the inference results from Liu et al. (2023) to recalculate the QM and IM metrics.

Specifically, compared to methods in previous advanced parsing systems, TP-LINK has achieved significant and consistent improvements in both QM and IM. On the SPARC dataset, QM and IM have improved by at least 0.2% and 1.1% respectively, and on the COSQL dataset, QM and IM have improved by at least 1.7% and 1.9%. When considering a unified downstream model, TP-LINK has also demonstrated notable improvements compared to various pre-trained models. On the SPARC dataset, QM and IM have improved by at least 1.1% and 3.1%, and on the COSQL dataset, there have been improvements of at least 1.0% and 1.7% respectively. The SOTA results highlight TP-LINK's strong performance in multi-turn text-to-SQL tasks.

When compared to zero-shot ChatGPT, TP-

Turn 1	: What is China's population?
GOLD	`SELECT population FROM country WHERE name = 'China'`
STAR & TP-Link	`SELECT country.Population FROM country WHERE country.Name = "China"` ✅
Turn 2	: How many Asian countries have a population greater than *that of* Nigeria?
GOLD	`SELECT count ( Name ) FROM country WHERE Continent = "Asia" AND population > ( SELECT population FROM country WHERE name = 'Nigeria' )`
STAR	`SELECT COUNT(*) FROM country WHERE country.Continent = "Asia" AND country.Population > "Nigeria"`
TP-Link	`SELECT COUNT(*) FROM country WHERE country.Continent = "Asia" AND country.Population > ( SELECT country.Population FROM country WHERE country.Name = "Nigeria" )` ✅
Turn 3	: Can you list *those countries*?
GOLD	`SELECT Name FROM country WHERE Continent = "Asia" AND population > ( SELECT population FROM country WHERE name = 'Nigeria' )`
STAR	`SELECT country.Name FROM country WHERE country.Population > ( SELECT MAX(country.Population) FROM country WHERE country.Continent = "Nigeria" )` Missing "Asia" Information ❌
TP-Link	`SELECT country.Name FROM country WHERE country.Continent = "Asia" AND country.Population > ( SELECT country.Population FROM country WHERE country.Name = "Nigeria" )` ✅

Figure 3: A hard case on CoSQL dataset. TP-LINK gives the correct predictions while STAR fails.

Model	SPARC		CoSQL	
	QM(%)	IM(%)	QM(%)	IM(%)
TP-LINK	**68.0**	**50.0**	**60.7**	**31.7**
w/o ULP	66.3 (↓1.7)	46.9 (↓3.1)	59.4 (↓1.3)	30.4 (↓1.3)
w/o SLP	66.0 (↓2.0)	46.7 (↓3.3)	58.8 (↓1.9)	29.4 (↓2.3)
w/o ULP & SLP	65.3 (↓2.7)	45.6 (↓4.4)	57.0 (↓3.7)	27.3 (↓4.4)

Table 4: Ablation study of pretraining objectives in terms of QM and IM on the dev sets of both SPARC and CoSQL.

Model	SPARC		CoSQL	
	QM(%)	IM(%)	QM(%)	IM(%)
TP-LINK	**68.0**	**50.0**	**60.7**	**31.7**
w. SLP	66.3 (↓1.7)	46.9 (↓3.1)	59.4 (↓1.3)	30.4 (↓1.3)
w. SLP(full)	65.5 (↓2.5)	45.3 (↓4.7)	58.5 (↓2.2)	28.7 (↓3.0)

Table 5: Ablation study about refined schema linking information of TP-LINK in terms of QM and IM on the dev sets of both SPARC and CoSQL.

LINK's QM and IM on SPARC improve by 30.4% and 29.9% respectively, and on CoSQL, QM and IM improve by 22.8% and 18.7% respectively. The results demonstrate the difficulty of context-dependent text-to-SQL task, as well as the gap in zero-shot performance compared to fine-tuning, thereby indicating the superiority of TP-LINK.

In summary, TP-LINK consistently outperforms other models in comparisons, demonstrating its effectiveness and generalizability.

4.2. Ablation Study

4.2.1. Effectiveness of Pretraining Objectives

In order to independently validate the effectiveness of the two pretraining objectives proposed in this paper, we conducted ablation experiments on the pretraining objectives, and the results are shown in Table 4. We performed ablation experiments for Utterance Linking Prediction (ULP, *i.e.* , *w/o* SLP experiment), and Schema Linking Prediction (SLP, *i.e.* , *w/o* ULP experiment) based on SQL tree edit distance. Results thoroughly validate that each pretraining objective has a standalone improvement effect, and the best experimental results are achieved when all pretraining objectives are used simultaneously.

4.2.2. Effectiveness of Similarity Filtering

To thoroughly validate the effectiveness of SQL structure similarity filtering, we conduct relevant ablation experiments. The results are shown in Table 5, where `full` indicates the usage of complete schema linking information, *i.e.* , no schema linking filtering based on similarity thresholds α. The results demonstrate a significant improvement when applying SQL similarity filtering compared to not using it, indicating the presence of considerable redundancy in schema linking during multi-turn dialogues, which has a substantial impact on performance. This experiment validates the necessity and effectiveness of the similarity filtering method.

4.3. Case Study

Figure 3 illustrates a hard case on CoSQL. In the second turn, STAR did not understand what **"that"** referred to and treated *"Nigeria"* as the comparison object. However, *"Nigeria"* itself is not comparable, the comparable entity is the "*population*". TP-LINK successfully understood that **"that"** referred to the *"population"* and consequently arrived at the correct answer. In the third turn, STAR had lost the information about *"Asia"* and mistakenly considered *"Nigeria"* as a continent. On the other hand, TP-LINK could comprehend that **"those countries"** referred to the results from the previous turn, thereby

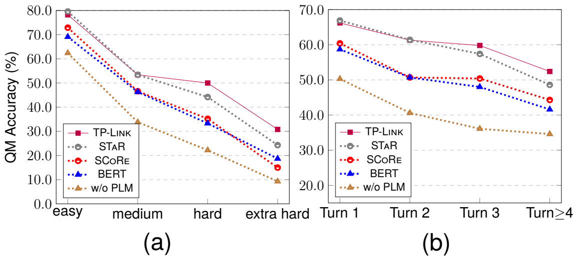

Figure 4: The results of TP-LINK and baselines on CoSQL dev sets (a) by varying the difficulty levels of the data and (b) by varying the conversation turns.

inheriting the SQL statement from the previous turn and correctly generating the SQL for the current turn, accurately identifying *"Nigeria"* as a country.

4.4. Discussion

To further validate the effectiveness of the proposed TP-LINK, we compare the performance of TP-LINK with some baseline methods under different interaction turns and SQL difficulties.

4.4.1. Comparison of Different SQL Query Difficulty Levels

We also conducted comparative experiments on SQL query statements of different difficulty levels, as shown in Figure 4(a). The experimental results indicate that in the case of `[hard]` and `[extra hard]` difficulty levels of SQL generation, TP-LINK demonstrates best performance, which suggests that TP-LINK can effectively handle the generation of more challenging SQL query statements, highlighting the effectiveness of schema linking modeling.

4.4.2. Comparison of Different Dialogue Turns

We conduct comparative experiments, as depicted in Figure 4(b), to compare the performance across various dialogue turns. The experimental results indicate that TP-LINK outperforms other models in longer dialogue turns (*e.g.*`turn=3` or `turn`$\geq$`4`), which suggests that our model is more effective in generating SQL queries in longer dialogue turns, demonstrating the effectiveness of fine-grained utterance linking and schema linking.

5. Related Work

Context-free Text-to-SQL Context-free text-to-SQL refers to the process of taking a natural language utterance and a database schema as input and generating a SQL query as output. Currently, the mainstream dataset for context-free text-to-SQL tasks is SPIDER (Yu et al., 2018), of which each question corresponds to a SQL statement. Generally, there are two mainstream approaches for context-free text-to-SQL tasks. One is graph-based parsers, *e.g.*RAT-SQL (Wang et al., 2020), LGESQL (Cao et al., 2021), S^2SQL (Hui et al., 2022). The other is T5-based parsers, *e.g.*PICARD (Scholak et al., 2021), T5-SR (Li et al., 2023b), which have achieved impressive performance on SPIDER. Recently, BINDER (Cheng et al., 2023) utilizes Codex (Ouyang et al., 2022) for transforming natural language into SQL/Python and other programming languages, which requires only a small amount of annotation to adapt to various programming languages. Zhao et al. (2022) propose a data synthesis framework aiming to improve the quality of the generated natural language question to enhance performance. Recent study (Liu et al., 2023) indicates that even without using any training data, ChatGPT still possesses strong text-to-SQL capabilities. DIN-SQL (Pourreza and Rafiei, 2023) achieves SOTA performance on the SPIDER using GPT-4 through in-context learning (Brown et al., 2020; Min et al., 2022).

Context-dependent Text-to-SQL Nevertheless, context-free text-to-SQL struggles to handle complex queries in a single statement. Users often prefer interactive dialogues, where they can gradually achieve their goals using context. Context-dependent text-to-SQL semantic parsing (Yu et al., 2019a,b) enables users to achieve their goals through multi-turn conversations, continuously refining their questions based on query results during the dialogue. Following that, a series of methods emerged to address this task. Based on a copying mechanism, EditSQL (Zhang et al., 2019) considers the information from the SQL query of the previous turn when predicting the SQL query for the current turn of dialogue. R^2SQL (Hui et al., 2021) introduces a memory decay mechanism to simulate the changes in the database schema within the dialogue flow. In addition, some methods draw inspiration from dialogue system's dialogue state tracking(DST) modules. ISTSQL (Wang et al., 2021) treats the database schema as the dialogue state, which enhances the effectiveness by tracking the database schema state and SQL keyword state. CQR-SQL (Xiao et al., 2022) simplifies the schema linking information fused with downstream parsing models by rewriting multi-turn dialogues. MIGA (Fu et al., 2022) integrates referential relationship information and schema linking information through a multi-task approach. BIRD (Li et al., 2023a) proposes a more challenging benchmark for large-scale cross-domain text-to-SQL tasks.

6. Conclusion

In this paper, we propose TP-Link, a novel tabular pretraining framework for multi-turn text-to-SQL, incorporating two novel pretraining objectives. Utterance linking prediction models syntactic relationships in multi-turn dialogues, enabling the pretrained language model to learn syntactic knowledge in advance, which addresses issues of coreference and ellipsis that exist in multi-turn dialogues. Schema linking prediction filters accurate schema linking relationships based on tree-based edit distance and SQL structural similarity, allowing pretrained language models to learn precise schema linking knowledge and address redundancy issues in the schema linking relations. Extensive experiments demonstrate that our model achieves new state-of-the-art results on downstream datasets SParC and CoSQL.

Limitations

Since the emergence of large language models (LLMs) like GPT-3(Brown et al., 2020), there has been a trend towards using LLMs to accomplish various natural language processing tasks with in-context learning. The proposed TP-Link relies on supervised data and is challenging to directly extend to LLMs. Recent study (Liu et al., 2023) also demonstrates the potential of LLMs in handling multi-turn text-to-SQL task under zero-shot scenarios. In future work, we plan to expand our method to larger-scale models, and further into zero-shot scenarios.

Acknowledgments

Min Yang was supported by National Key Research and Development Program of China (2022YFF0902100), National Natural Science Foundation of China (62376262), Shenzhen Science and Technology Innovation Program (KQTD20190929172835662), Shenzhen Basic Research Foundation (JCYJ20210324115614039 and JCYJ20200109113441941).

Bibliographical References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yitao Cai and Xiaojun Wan. 2020. IGSQL: Database schema interaction graph based neural model for context-dependent text-to-SQL generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6903–6912, Online. Association for Computational Linguistics.

Zefeng Cai, Xiangyu Li, Binyuan Hui, Min Yang, Bowen Li, Binhua Li, Zheng Cao, Weijie Li, Fei Huang, Luo Si, and Yongbin Li. 2022. Star: Sql guided pre-training for context-dependent text-to-sql parsing.

Ruisheng Cao, Lu Chen, Zhi Chen, Yanbin Zhao, Su Zhu, and Kai Yu. 2021. LGESQL: Line graph enhanced text-to-SQL model with mixed local and non-local relations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2541–2555, Online. Association for Computational Linguistics.

Zhi Chen, Lu Chen, Hanqi Li, Ruisheng Cao, Da Ma, Mengyue Wu, and Kai Yu. 2021. Decoupled dialogue modeling and semantic parsing for multi-turn text-to-SQL. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3063–3074, Online. Association for Computational Linguistics.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding language models in symbolic languages.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiang Deng, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. 2021. Structure-grounded pretraining for text-to-SQL. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1337–1350, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yingwen Fu, Wenjie Ou, Zhou Yu, and Yue Lin. 2022. Miga: A unified multi-task generation framework for conversational text-to-sql.

Binyuan Hui, Ruiying Geng, Qiyu Ren, Binhua Li, Yongbin Li, Jian Sun, Fei Huang, Luo Si, Pengfei Zhu, and Xiaodan Zhu. 2021. Dynamic hybrid relation exploration network for cross-domain context-dependent semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13116–13124.

Binyuan Hui, Ruiying Geng, Lihan Wang, Bowen Qin, Bowen Li, Jian Sun, and Yongbin Li. 2022. S^2SQL: Injecting syntax to question-schema interaction graph encoder for text-to-sql parsers. In *ACL*.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics.

Jinyang Li, Binyuan Hui, Ge Qu, Binhua Li, Jiaxi Yang, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023a. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls.

Yuntao Li, Zhenpeng Su, Yutian Li, Hanchu Zhang, Sirui Wang, Wei Wu, and Yan Zhang. 2023b. T5-sr: A unified seq-to-seq decoding strategy for semantic parsing.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding.

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S. Yu. 2023. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability.

Qian Liu, Bei Chen, Jiaqi Guo, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020. How far are we from effective context modeling? an exploratory study on semantic parsing in context.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. Tapex: Table pre-training via learning a neural sql executor.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Mateusz Pawlik and Nikolaus Augsten. 2016. Tree edit distance: Robust and memory-efficient. *Information Systems*, 56:157–173.

Mohammadreza Pourreza and Davood Rafiei. 2023. Din-sql: Decomposed in-context learning of text-to-sql with self-correction.

Jiexing Qi, Jingyao Tang, Ziwei He, Xiangpeng Wan, Yu Cheng, Chenghu Zhou, Xinbing Wang, Quanshi Zhang, and Zhouhan Lin. 2022. Rasat: Integrating relational structures into pretrained seq2seq model for text-to-sql.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peng Shi, Patrick Ng, Zhi guo Wang, Henghui Zhu, Alexander Hanbo Li, J. Wang, C. D. Santos, and Bing Xiang. 2021. Learning contextual representations for semantic parsing with generation-augmented pre-training. In *AAAI*.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7567–7578, Online. Association for Computational Linguistics.

Run-Ze Wang, Zhen-Hua Ling, Jingbo Zhou, and Yu Hu. 2021. Tracking interaction states for multi-turn text-to-sql semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13979–13987.

Dongling Xiao, Linzheng Chai, Qian-Wen Zhang, Zhao Yan, Zhoujun Li, and Yunbo Cao. 2022. Cqr-sql: Conversational question reformulation enhanced context-dependent text-to-sql parsers.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir R. Radev, Richard Socher, and Caiming Xiong. 2021a. Grappa: Grammar-augmented pre-training for table semantic parsing. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter Lasecki, and Dragomir Radev. 2019a. CoSQL: A conversational text-to-SQL challenge towards cross-domain natural language interfaces to databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1962–1979, Hong Kong, China. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Alex Polozov, Christopher Meek, and Ahmed Hassan Awadallah. 2021b. Score: Pre-training for context representation in conversational semantic parsing. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019b. SParC: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523, Florence, Italy. Association for Computational Linguistics.

Rui Zhang, Tao Yu, Heyang Er, Sungrok Shim, Eric Xue, Xi Victoria Lin, Tianze Shi, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019. Editing-based SQL query generation for cross-domain context-dependent questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5338–5349, Hong Kong, China. Association for Computational Linguistics.

Yiyun Zhao, Jiarong Jiang, Yiqun Hu, Wuwei Lan, Henry Zhu, Anuj Chauhan, Alexander Li, Lin Pan, Jun Wang, Chung-Wei Hang, Sheng Zhang, Marvin Dong, Joe Lilien, Patrick Ng, Zhiguo Wang, Vittorio Castelli, and Bing Xiang. 2022. Importance of synthesizing high-quality data for text-to-sql parsing.

Yanzhao Zheng, Haibin Wang, Baohua Dong, Xingjun Wang, and Changshan Li. 2022. Hie-sql: History information enhanced network for context-dependent text-to-sql semantic parsing. *ArXiv preprint*, abs/2203.07376.

Victor Zhong, Mike Lewis, Sida I. Wang, and Luke Zettlemoyer. 2020. Grounded adaptation for zero-shot executable semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6869–6882, Online. Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.