

LREC-COLING 2024

The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

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

Ziqiang Liu, Shujie Li, Zefeng Cai, Xiangyu Li, Yunshui Li
Lei Zhang, Chengming Li, Xiping Hu, Ruifeng Xu, Min Yang



中国科学院大学
University of Chinese Academy of Sciences



中国科学院深圳先进技术研究院
SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

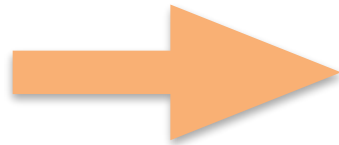
Conversational Text-to-SQL Semantic Parsing

- **Starting Point:** Build pre-trained **tabular** language model for downstream Text-to-SQL



Pre-trained **Language** Model
(PLM)

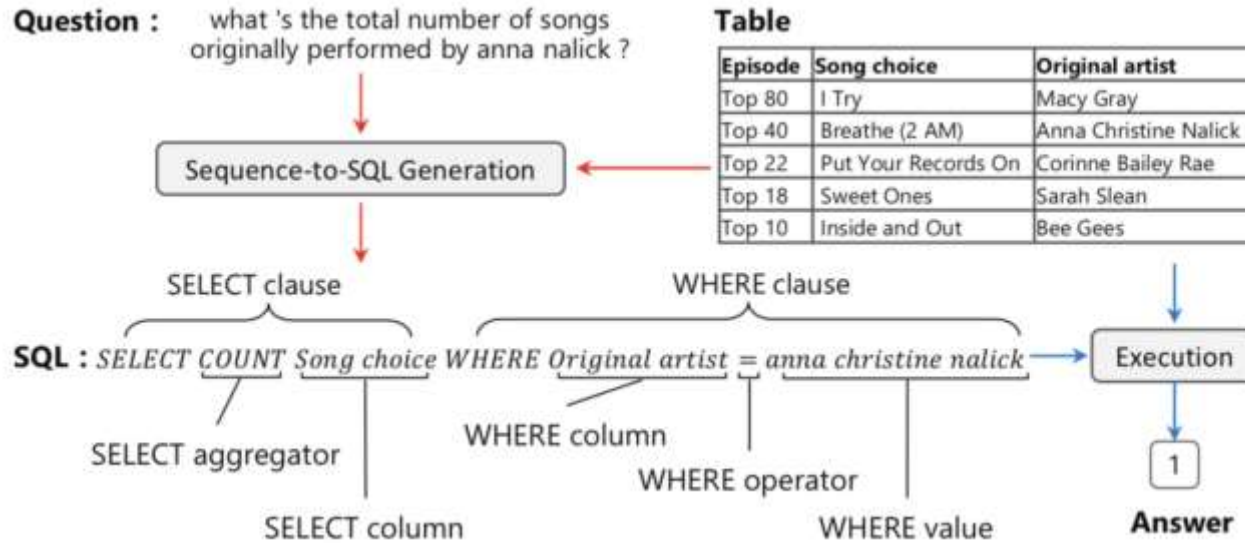
intrinsic discrepancy



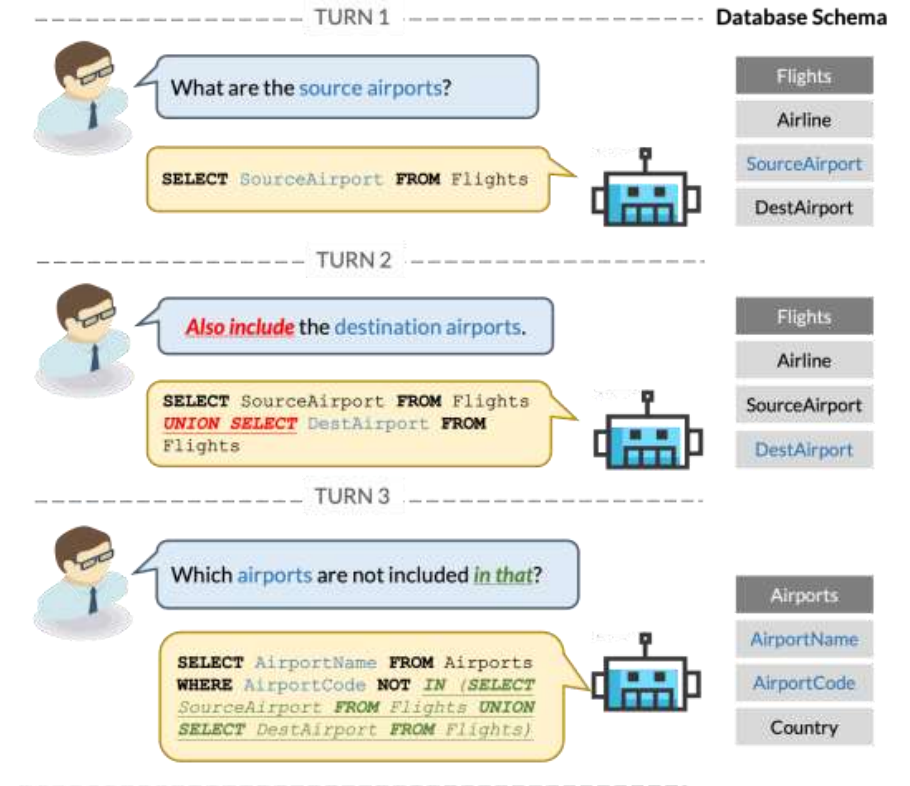
Pre-trained **Tabular** Language Model
(TaLM)

Conversational Text-to-SQL Semantic Parsing

- **Starting Point:** Build powerful multi-turn Text-to-SQL Semantic Parsing System



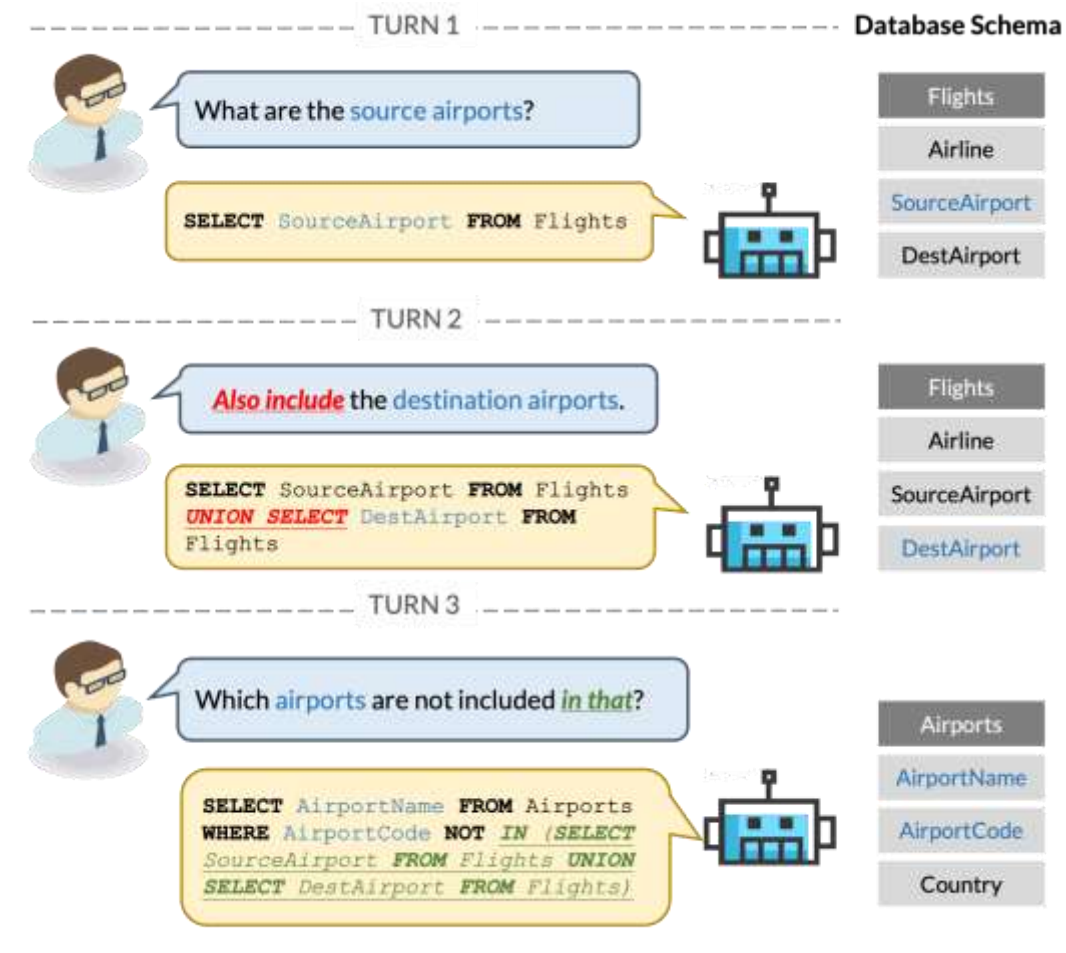
(a) Single-turn text-to-SQL



(b) Multi-turn text-to-SQL

Conversational Text-to-SQL Semantic Parsing

- **Motivation:** leveraging contextual **historical** information from the dialogue is pivotal

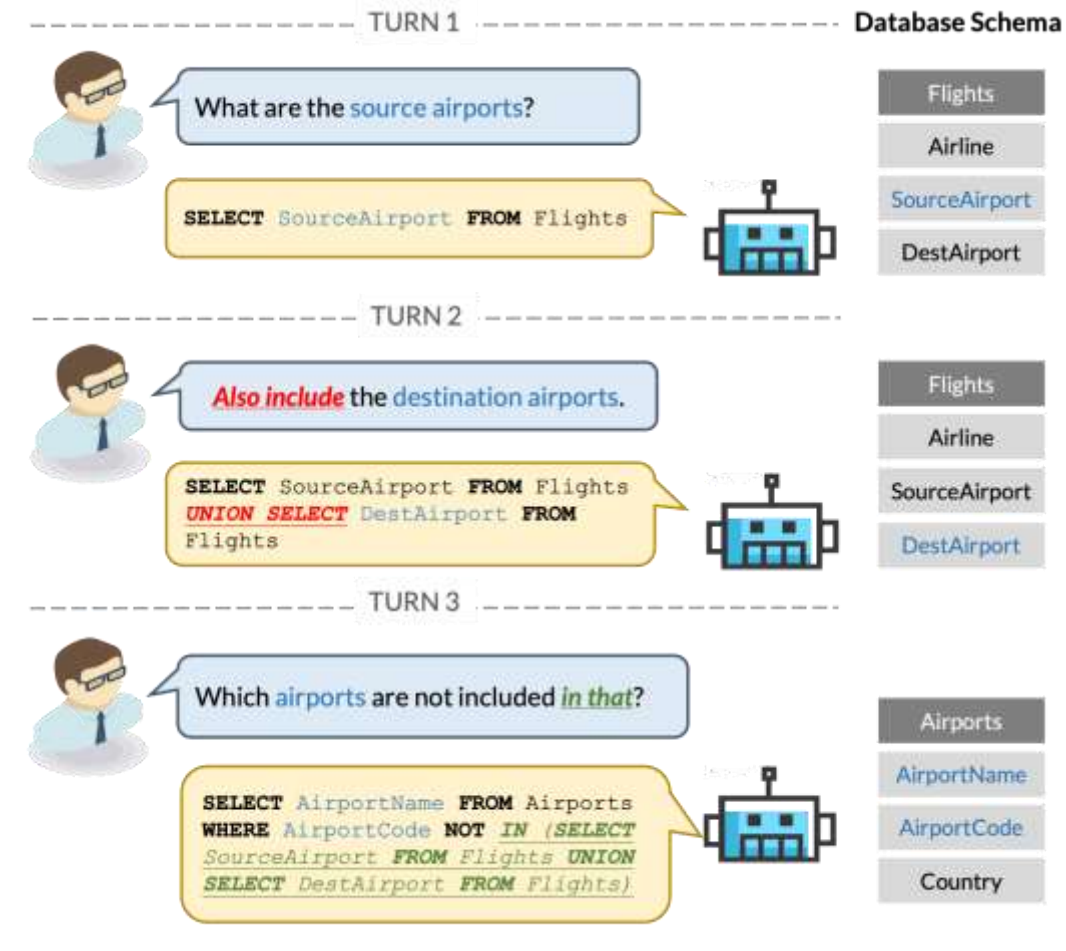


The **syntactic coreference** relationships within the context ensure the accurate generation of SQL statements

- Coreference
in **that**
- Ellipsis
Also include

Conversational Text-to-SQL Semantic Parsing

- **Motivation:** leveraging contextual **historical** information from the dialogue is pivotal

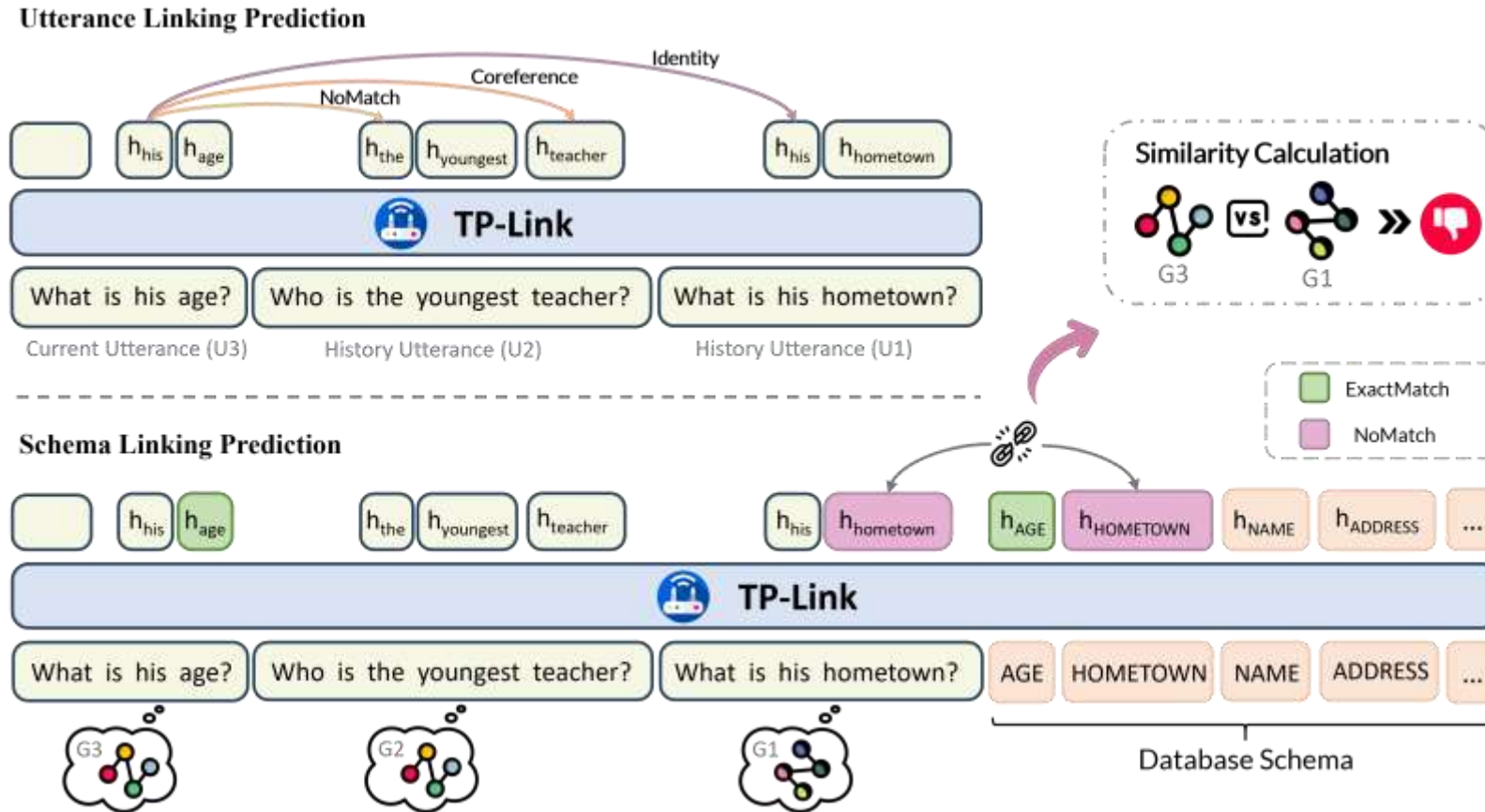


There exists considerable **linking information** between utterances and database schemas

- User utterance
source airports
- Database schema
SourceAirport

Conversational Text-to-SQL Semantic Parsing

- **Method:** propose **ULP** and **SLP** for Text-to-SQL Semantic Parsing



Conversational Text-to-SQL Semantic Parsing

- **Method:** Utterance Linking Prediction

1. Finer-grained Coreference Resolution:

Utilize the coreference resolution tool [NeuralCoref](#) to resolve the word-level syntactic relationships between the present utterance and the entirety of utterances.

2. At t -th turn, the goal of ULP task is to predict the word-level syntactic relationships within U_t given all the utterances U_t and database schema S .

3. The pre-training loss function of ULP task is defined as the cross-entropy between the heuristic representation and the gold word-level syntactic relationship labels:

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

Conversational Text-to-SQL Semantic Parsing

- **Method:** Schema Linking Prediction

1. More accurate schema linking:

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.

2. Similar to ULP task, we predict the refined schema linking relationships according to the heuristic representation between the utterances representation and schema representation

3. The pre-training loss function of SLP task is defined the cross-entropy between the heuristic representation and the gold schema linking labels:

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

Conversational Text-to-SQL Semantic Parsing

- **Method:** Pretraining and finetuning

1. We continually pre-train the **ELECTRA** on a synthetic text-to-SQL corpus consisting of **480k** examples, following the methodology introduced by **STaR**. As a result, we obtain our proposed **TP-Link** model.
2. We choose the **LGESQL** as the downstream 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.

- **Metrics:** QM and IM

1. Question Match Accuracy (**QM**) indicates whether the SQL query generated by the model matches the actual SQL query exactly.
2. Interaction Match Accuracy (**IM**) accounts for the QM score of each question in a multi-turn dialogue interaction.

Conversational Text-to-SQL Semantic Parsing

- Overall results: **New SOTA** on CoSQL and SParC

MODEL	SPARC		CoSQL	
	QM(%)	IM(%)	QM(%)	IM(%)
<i>Previous Parsing Systems.</i>				
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 ² SQL + 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
<i>Zero-shot Models.</i>				
ChatGPT (Liu et al., 2023)	37.6	20.1	37.9	13.0
<i>Pre-trained Models.</i>				
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)

Conversational Text-to-SQL Semantic Parsing

- Ablation study

Model	SPARC		CoSQL	
	QM(%)	IM(%)	QM(%)	IM(%)
TP-LINK	68.0	50.0	60.7	31.7
<i>w/o</i> ULP	66.3 (↓1.7)	46.9 (↓3.1)	59.4 (↓1.3)	30.4 (↓1.3)
<i>w/o</i> SLP	66.0 (↓2.0)	46.7 (↓3.3)	58.8 (↓1.9)	29.4 (↓2.3)
<i>w/o</i> 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
<i>w.</i> SLP	66.3 (↓1.7)	46.9 (↓3.1)	59.4 (↓1.3)	30.4 (↓1.3)
<i>w.</i> 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.

Conversational Text-to-SQL Semantic Parsing

- Case study

(b) a hard Case of CoSQL

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

Conversational Text-to-SQL Semantic Parsing

- **Summary**

1. 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.
2. 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.
3. TP-Link achieves new state-of-the-art (**SOTA**) results on two context-dependent text-to-SQL datasets, SParC and CoSQL.



LREC-COLING 2024

The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Thank You!



中国科学院大学
University of Chinese Academy of Sciences



中国科学院深圳先进技术研究院
SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES