Structure-Guided Large Language Models for Text-to-SQL Generation

Qinggang Zhang · Hao Chen · Junnan Dong · Shengyuan Chen · Feiran Huang · Xiao Huang

The Hong Kong Polytechnic University 香港理工大學

Abstract

Recent advancements in large language models (LLMs) have shown promise in bridging the gap between natural language queries and database management systems, enabling users to interact with databases without the background of SQL. However, LLMs often struggle to fully exploit and comprehend the user intention and complex structures of databases. Decomposition-based methods have been proposed to enhance the performance of LLMs on complex tasks, but decomposing SQL generation into subtasks is non-trivial due to the declarative structure of SQL syntax and the intricate connections between query concepts and database elements. In this paper, we propose a novel Structure GUided text-to-SQL framework (SGU-SQL) that incorporates syntax-based prompting to enhance the SQL generation capabilities of LLMs. Specifically, SGU-SQL establishes structure-aware links between user queries and database schema and recursively decomposes the complex generation task using syntax-based prompting to guide LLMs in incrementally constructing target SQLs. Extensive experiments on two benchmark datasets demonstrate that SGU-SQL consistently outperforms state-of-the-art text-to-SQL baselines.

Introduction

Current LLM-based text-to-SQL methods face several challenges like ambiguous user intent, sophisticated database schema which often lacks proper documentations, and complex syntax structure of the SQL queries.

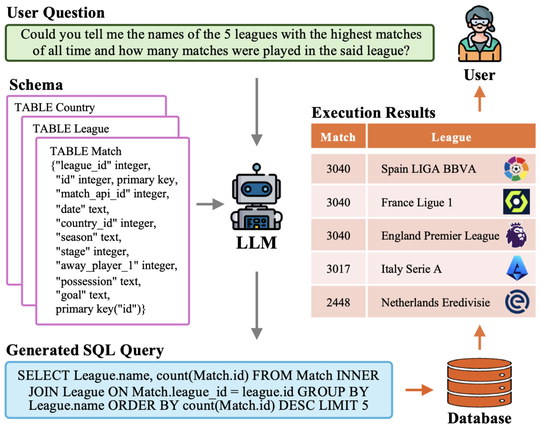

Figure 1. The pipeline of text-to-SQL models.

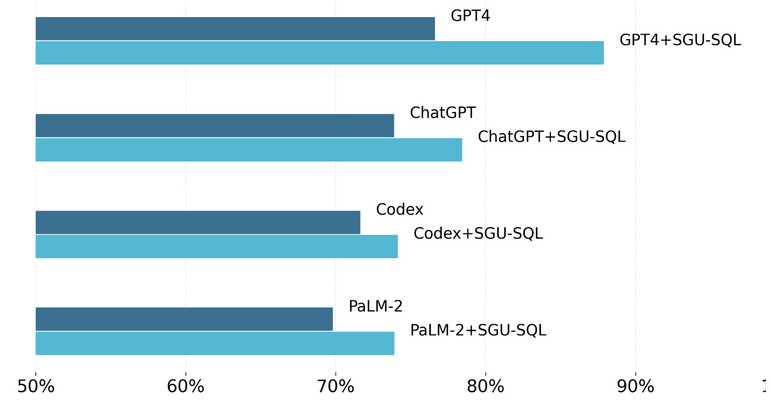

Figure 2. Performance of existing LLM-based text-to-SQL models.

Methods

1. How could we to map the natural language query to the relevant database elements?
2. How could we break down the complex generation task in a syntax-aware manner?

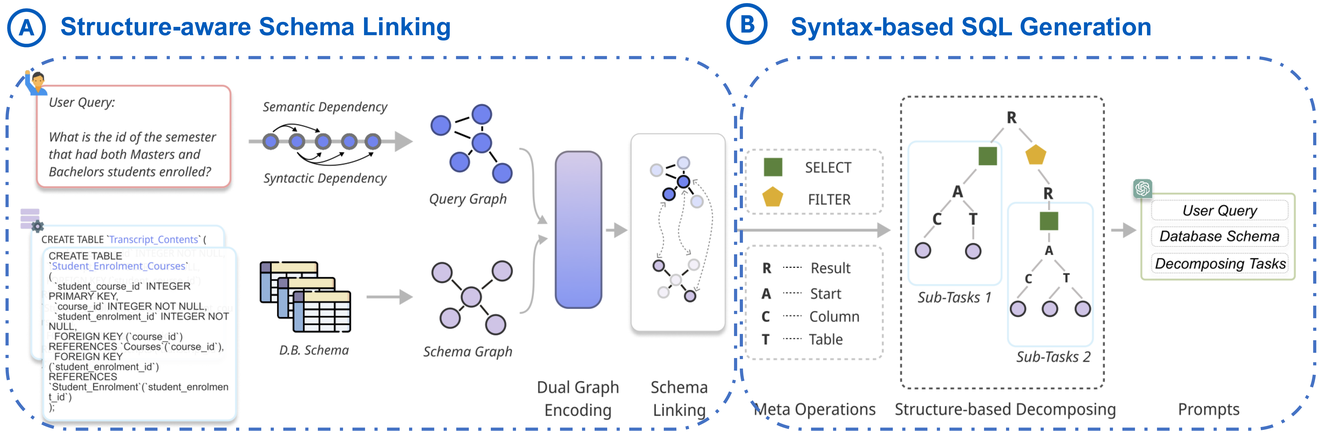

Figure 3. The overall framework of SGU-SQL. It first proposes graph-based structure construction to comprehend user query and database structure and then link query and database with dual-graph encoding, and then introduces tailored structure-decomposed generation strategies to decompose queries with syntax trees and then incrementally generate accurate SQL with LLM.

Results

Text-to-SQL Method	Backbone LM/LLM	Finetuning	Structure Information	Prompt Strategy	SPIDER Easy	Medium	Hard	Extra	Overall
Baichuan2	Baichuan2-7B	SFT	✘	✘	0.5775±0.0106	0.3521±0.0130	0.2010±0.0089	0.0667±0.0115	0.3353±0.0125
		LoRA	✘	✘	0.8714±0.0073	0.6305±0.0069	0.4489±0.0063	0.2958±0.0084	0.6035±0.0079
		QLoRA	✘	✘	0.8919±0.0057	0.6367±0.0071	0.4885±0.0053	0.3306±0.0079	0.6242±0.0061
	Baichuan2-13B	SFT	✘	✘	0.5805±0.0093	0.4133±0.0085	0.2644±0.0067	0.1875±0.0078	0.3927±0.0081
		LoRA	✘	✘	0.9024±0.0075	0.7015±0.0069	0.5688±0.0083	0.3915±0.0071	0.6776±0.0080
		QLoRA	✘	✘	0.8951±0.0103	0.6746±0.0123	0.5809±0.0115	0.3434±0.0109	0.6692±0.0114
LLaMA2	LLaMA2-7B	LoRA	✘	✘	0.8868±0.0016	0.6410 ±0.0041	0.4892±0.0030	0.3311±0.0017	0.6259±0.0022
		QLoRA	✘	✘	0.8472±0.0025	0.6234±0.0032	0.4658±0.0021	0.3309±0.0027	0.6083±0.0035
	LLaMA2-13B	LoRA	✘	✘	0.9066±0.0037	0.7292±0.0045	0.5517±0.0029	0.3430±0.0055	0.6809±0.0030
		QLoRA	✘	✘	0.9110±0.0043	0.7004±0.0059	0.5523±0.0032	0.3190±0.0061	0.6648±0.0045
	LLaMA2-70B	SFT	✘	✘	0.4110±0.0093	0.2293±0.0075	0.1906±0.0081	0.0725±0.0090	0.2414±0.0108
		LoRA	✘	✘	0.9151±0.0069	0.7323±0.0080	0.5575±0.0049	0.3921±0.0035	0.6869±0.0040
CodeLlama	CodeLlama-7B	SFT	✘	✘	0.2136±0.0150	0.1769±0.0161	0.0921±0.0169	0.0363±0.0144	0.1487±0.0163
		LoRA	✘	✘	0.9228±0.0105	0.7562±0.0134	0.5863±0.0096	0.3485±0.0126	0.7018±0.0108
		QLoRA	✘	✘	0.9115±0.0127	0.7506±0.0142	0.5982±0.0120	0.3310±0.0085	0.6961±0.0104
	CodeLlama-13B	SFT	✘	✘	0.6980±0.0115	0.6015±0.0121	0.4073±0.0109	0.2708±0.0145	0.5288±0.0140
		LoRA	✘	✘	0.9414±0.0086	0.7885±0.0073	0.6842±0.0081	0.4041±0.0069	0.7462±0.0092
		QLoRA	✘	✘	0.9402±0.0053	0.7445±0.0066	0.6263±0.0085	0.3915±0.0061	0.7270±0.0085
	CodeLlama-70B	SFT	✘	✘	0.7223±0.0143	0.6245±0.0120	0.4432±0.0131	0.3028±0.0147	0.5675±0.0144
		LoRA	✘	✘	**0.9621±0.0053**	0.8122±0.0069	0.7167±0.0055	0.4324±0.0069	0.7710±0.0061
Qwen	Qwen-7B	SFT	✘	✘	0.3956±0.0155	0.2561±0.0131	0.1384±0.0137	0.0427±0.0169	0.2356±0.0140
		LoRA	✘	✘	0.8546±0.0060	0.6876±0.0089	0.5743±0.0076	0.3340±0.0065	0.6519±0.0073
		QLoRA	✘	✘	0.9110±0.0045	0.6747±0.0081	0.5750±0.0076	0.3436±0.0055	0.6623±0.0069
	Qwen-14B	SFT	✘	✘	0.8713±0.0105	0.6323±0.0140	0.3686±0.0139	0.1810±0.0120	0.5735±0.0135
		LoRA	✘	✘	0.8946±0.0110	0.7021±0.0103	0.5517±0.0125	0.3669±0.0118	0.6625±0.0121
		QLoRA	✘	✘	0.9185±0.0075	0.7439±0.0060	0.5976±0.0081	0.4583±0.0083	0.7010±0.0090
	Qwen-72B	SFT	✘	✘	0.8313±0.0100	0.6345±0.0077	0.4886±0.0065	0.2772±0.0123	0.6033±0.0110
		LoRA	✘	✘	0.9269±0.0075	0.7563±0.0059	0.6215±0.0083	0.3673±0.0136	0.7127±0.0094
RAT-SQL	BERT-Large	✘	✔	✘	0.8044±0.0107	0.6395±0.0082	0.5573±0.0124	0.4036±0.0101	0.6271±0.0119
		SFT	✔	✘	0.8643±0.0119	0.7367±0.0145	0.6210±0.0093	0.4279±0.0116	0.6955±0.0124
LGESQL	BERT-Large	✘	✔	✘	0.8633±0.0097	0.6952±0.0065	0.6154±0.0093	0.4106±0.0118	0.6768±0.0109
		SFT	✔	✘	0.9150±0.0103	0.7647±0.0065	0.6673±0.0107	0.4888±0.0078	0.7421±0.0096
Graphix-T5	T5-Large	SFT	✔	✘	0.8993±0.0075	0.7874±0.0068	0.5980±0.0102	0.4401±0.0083	0.7263±0.097
	T5-3B	SFT	✔	✘	0.9193±0.0038	0.8164±0.0062	0.6157±0.0053	0.5006±0.0081	0.7562±0.0065
RESDSQL	T5-Base	SFT	✔	✘	0.9190±0.0047	0.8369±0.0051	0.6841±0.0070	0.5183±0.0065	0.7797±0.0073
	T5-Large	SFT	✔	✘	0.9355±0.0040	0.8543±0.0051	0.7241±0.0070	0.5361±0.0045	0.8008±0.0063
	T5-3B	SFT	✔	✔	0.9476±0.0081	0.8767±0.0104	0.7299±0.0120	0.5602±0.0094	0.8182±0.0100
DTS-SQL	DeepSeek-7B	SFT	✘	✔	0.9274±0.0091	0.9013±0.0075	0.7414±0.0090	0.5663±0.0103	0.8269±0.0094
CodeS	CodeLlama-13B	SFT	✘	✔	0.9274±0.0084	0.8789±0.0052	0.7069±0.0079	0.5904±0.0038	0.8150±0.0070
C^3-SQL	GPT-3.5	✘	✘	✔	0.9136±0.0068	0.8402±0.0094	0.7731±0.0064	0.6153±0.0080	0.8108±0.0095
DIN-SQL	GPT-4	✘	✘	✔	0.9234±0.0059	0.8744±0.0080	0.7644±0.0091	0.6265±0.0103	0.8279±0.0098
DAIL-SQL	GPT-4	✘	✘	✔	0.9153±0.0103	0.8924±0.0125	0.7701±0.0098	0.6024±0.0107	0.8308±0.0110
EPI-SQL	GPT-4	✘	✘	✔	0.9310±0.0121	0.9053±0.0085	0.8178±0.0108	0.6189±0.0097	0.8511±0.0114
SuperSQL	GPT-4	✘	✘	✔	0.9435±0.0074	0.9126±0.0050	0.8333±0.0062	0.6867±0.0055	0.8682±0.0068
PURPLE	GPT-4	✘	✘	✔	0.9404±0.0086	**0.9206±0.0041**	0.8268±0.0055	0.6715±0.0080	0.8670±0.0072
SGU-SQL	GPT-4	✘	✔	✔	0.9352±0.0061	0.9190±0.0043	**0.8437±0.0045**	**0.7213±0.0067**	**0.8795±0.0063**

	Dataset	Spider			BIRD		
	Metric	EX Acc	EM Acc	VES	EX Acc	EM Acc	VES
Finetuning-based	Baichuan2-7B	0.6035	0.5793	0.6082	0.1719	0.0547	0.2097
	Baichuan2-13B	0.6776	0.6078	0.6545	0.1766	0.0455	0.2126
	LlaMA2-7B	0.6083	0.5816	0.5795	0.1675	0.0469	0.1670
	LlaMA2-13B	0.6809	0.6400	0.6712	0.1993	0.0743	0.1739
	LlaMA2-70B	0.6869	0.6555	0.6779	0.2414	0.0778	0.1987
	CodeLlama-7B	0.7018	0.6431	0.7357	0.2370	0.1283	0.2504
	CodeLlama-13B	0.7462	0.7056	0.7391	0.2944	0.2551	0.3004
	CodeLlama-70B	0.7710	0.7139	0.7463	0.3287	0.2557	0.3428
	Qwen-7B	0.6519	0.6106	0.6625	0.1709	0.0439	0.1915
	Qwen-14B	0.6625	0.6238	0.6757	0.2286	0.0645	0.2396
	Qwen-72B	0.7127	0.6812	0.7082	0.2392	0.0894	0.2488
Structure Learning	RAT-SQL	0.6955	0.6597	0.6734	0.2639	0.2431	0.2431
	BRIDGE	0.6928	0.7053	0.6893	0.2459	0.2068	0.2574
	LGESQL	0.7421	0.7251	0.7067	0.2837	0.2493	0.2889
	S^{2}SQL	0.7643	0.7385	0.7539	0.2960	0.2649	0.3143
	RESDSQL	0.8182	0.7580	0.8226	0.3312	0.3174	0.3286
	Graphix-T5	0.7562	0.7463	0.7643	0.2984	0.2538	0.3062
	METASQL	0.7695	0.7288	0.7498	0.3180	0.3011	0.3225
In-Context Learning	GPT-3.5	0.7394	0.5327	0.7457	0.3562	0.3041	0.3415
	GPT-4	0.7665	0.5892	0.7390	0.4633	0.4255	0.4794
	PaLM-2	0.6985	0.4438	0.7148	0.2735	0.2543	0.3061
	CodeX	0.7167	0.4905	0.7011	0.3438	0.3019	0.3496
	C^3-GPT	0.8108	0.7036	0.8009	0.5020	0.4143	0.5077
	DIN-SQL	0.8279	0.7187	0.8173	0.5072	0.4398	0.5879
	DAIL-SQL	0.8308	0.7443	0.8317	0.5434	0.4581	0.5576
	DTS-SQL	0.8269	0.7260	0.8163	0.5581	0.4825	0.6038
	CodeS	0.8150	0.7069	0.8092	0.5714	0.4893	0.6120
	SuperSQL	0.8682	0.7589	0.8410	0.5860	0.4745	0.6067
	MAC-SQL	0.8635	0.7545	0.8541	0.5759	0.4906	0.5872
	SGU-SQL	**0.8795**	**0.7826**	**0.8652**	**0.6180**	**0.5144**	**0.6393**

