An optimized NL2SQL system for enterprise data mart

Kaiwen Dong, Kai Lu, Xin Xia, David Cieslak and Nitesh Chawla

aunalytics

Natural Language to SQL

Table: Transaction Summary

Table: Account

Table: Customer

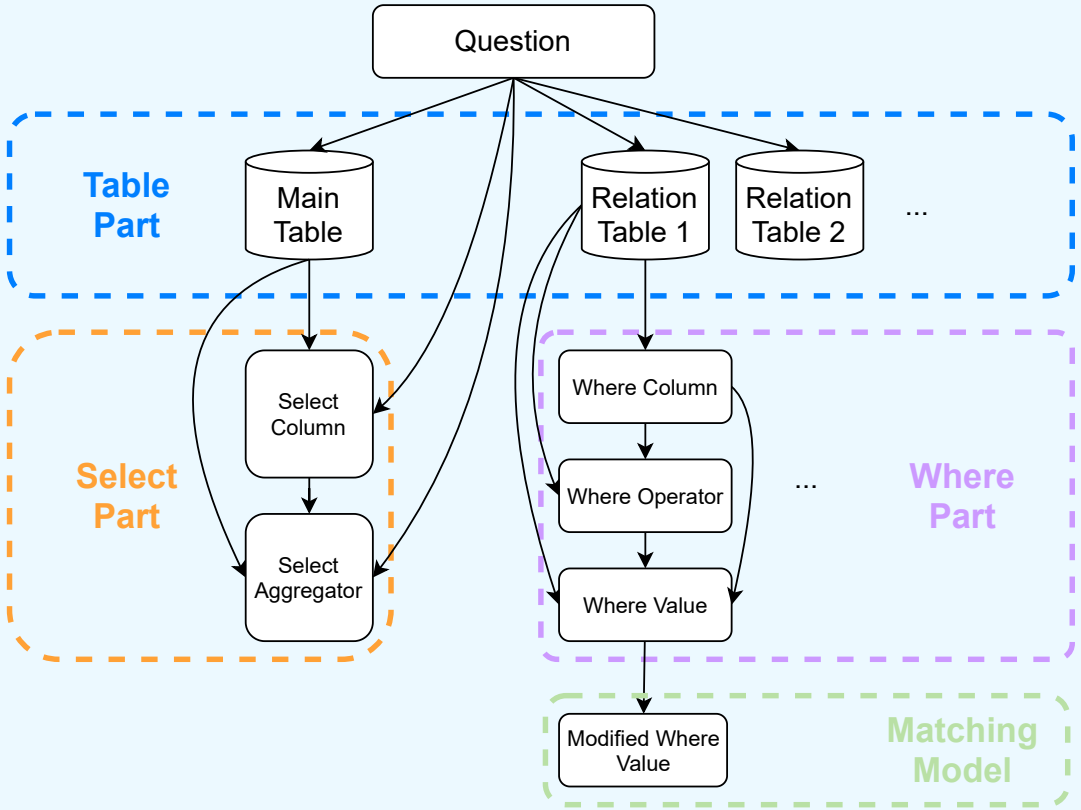
Customer_ID	Address	Customer_Age	City	CurrentCreditScore	DateJoined	DateofBirth
00FMJ	8725 WESTMORE RD B, San Diego, 92126	66	San Diego	444.0	2017-07-05	1953-11-11
1841	12720 CARMEL COUNTRY RD , San Diego, 92130	18	San Diego	745.0	2016-01-09	2002-03-28
04UVM	11134 CAMINITO INOCENTA , San Diego, 92126	14	San Diego	732.0	2016-05-21	2006-06-22
06S8H	4478 40TH ST , San Diego, 92116		San Diego	686.0	2018-07-10	2003-10-07
...	...	...	...	...	...	...

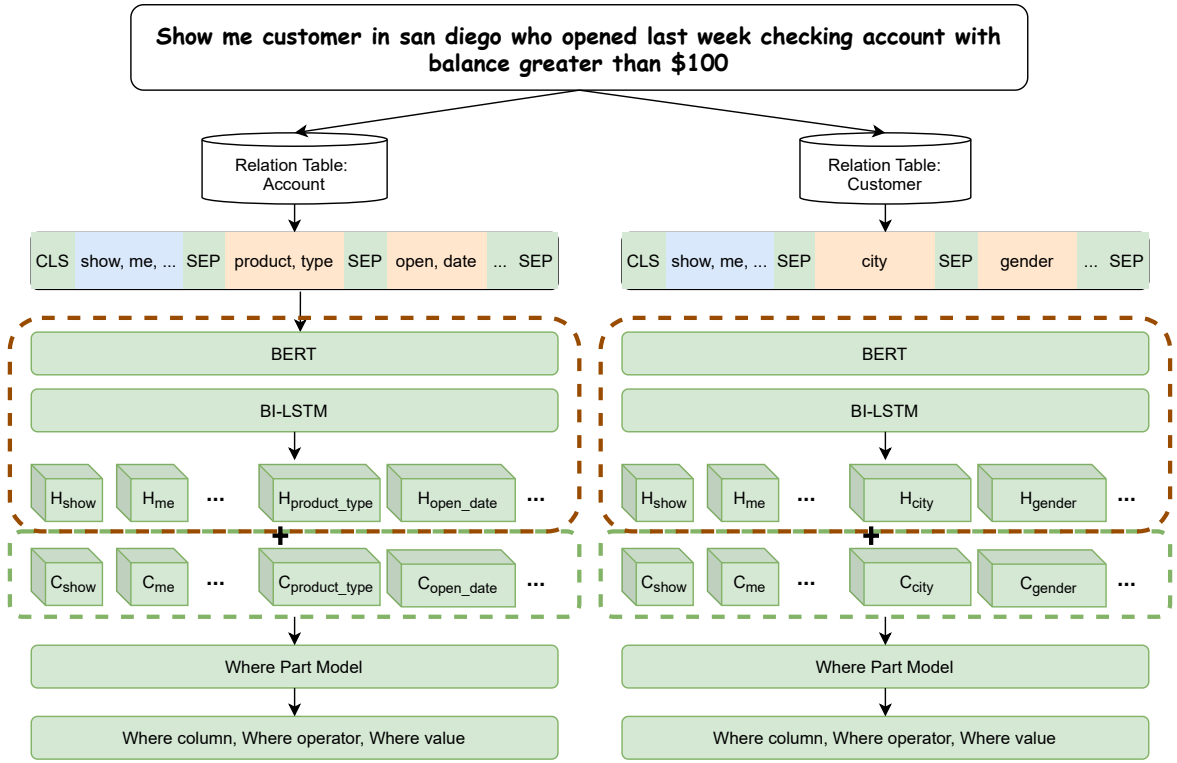
Question

Show me customers from San Diego with checking accounts

The inputs consist of a database (schema) and a question. The output is the corresponding SQL query to answer the question.

Table Expand





Breaking the database schema to table level allows the encoder to fit larger schema

SQL

The SQL defined in WikiSQL[1]:
SELECT $AGG $COLUMN FROM
(WHERE $COLUMN $OP $VALUE (AND $COLUMN $OP $VALUE)* FROM TABLE

Our version of SQL:
SELECT $AGG $COLUMN FROM
(WHERE $COLUMN $OP $VALUE (AND $COLUMN $OP $VALUE)* FROM $TABLE) JOIN ON …
(WHERE $COLUMN $OP $VALUE (AND $COLUMN $OP $VALUE)* FROM $TABLE) JOIN ON …
…

1. Zhong, V., Xiong, C., Socher, R.: Seq2sql: Generating structured queries from natural language using reinforcement learning. CoRR abs/1709.00103 (2017), http://arxiv.org/abs/1709.00103

Matching

Date Type	Problem	Question	Substring	Table Cell
Categorical	Case or form doesn't match	Show me **mortgages**	Account Type= "mortgages"	Account Type= "Mortgage"
	No cell value present in question	Which customer **doesn't have** mobile bank?	HasMobileBank= "doesn't have"	HasMobileBank= "No"
		Customers with **dda**	ProductCategory ="dda"	ProductCategory ="Demand Deposit"
Datetime	The date-time format doesn't match	Accounts opened since **2018**	OpenDate> "2018"	OpenDate> "2018-01-01T00:00:00.000Z"
	Can't parse relative time expression	Accounts opened **this year**	OpenDate="this year"	OpenDate≥ "2021-01-01T00:00:00.000Z And OpenDate< "2022-01-01T00:00:00.000Z"
Numeric	The data type doesn't match	Accounts with balance more than $**100**	CurrentBalance >"$100"	CurrentBalance> 100.0

Structured Query

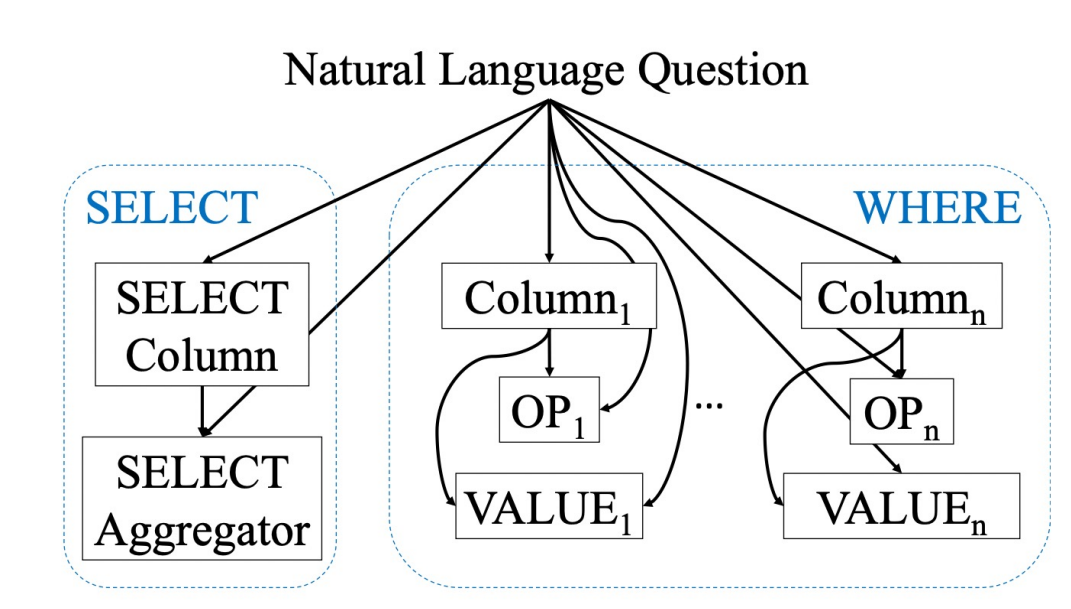
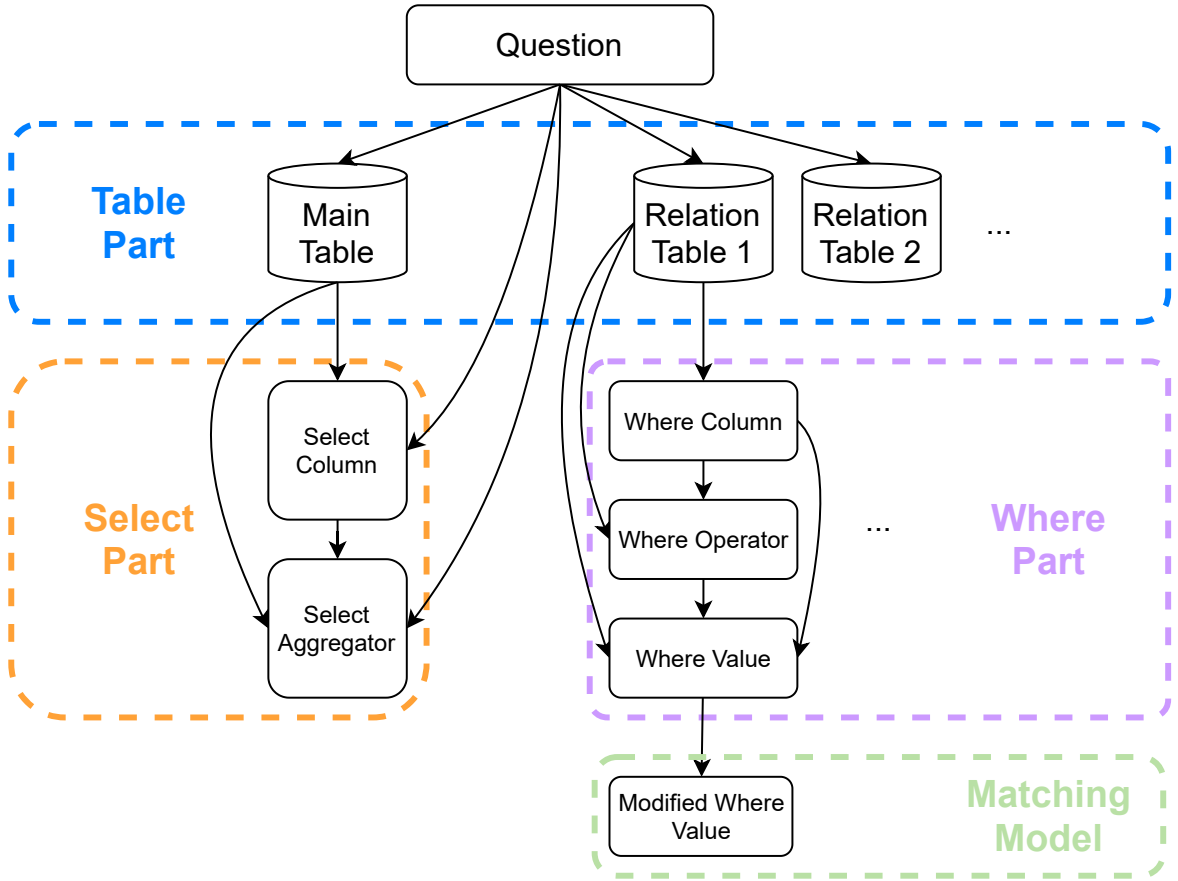


Figure from Xu[2]



2. Xu, X., Liu, C., Song, D.: Sqlnet: Generating structured queries from natural language without reinforcement learning. CoRR abs/1711.04436 (2017), http: //arxiv.org/abs/1711.04436