Text-To-SQL

Presented By:

Wasiim Ouro-sama

Jacob Austin

Task, Dataset and Experimental Setup

- Task - Given a query in natural language (English), the task is to translate it into the equivalent SQL query
- Dataset - Spider dataset, consisting of 7000 (NL, SQL) pairs spanning a wide variety of databases, tables, SQL queries, and foreign key relationships
- Experimental Setup
  - Metrics
    - Exact match(exact match)
    - Execution match (with values)
    - Execution match (without values)
    - In addition to an overall score, each metric is also grouped based on query difficulty of easy, medium and hard

Easy

What is the number of cars with more than 4 cylinders?

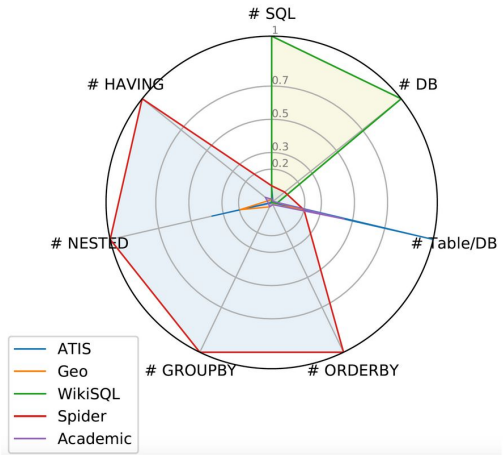
```
SELECT COUNT(*)
FROM cars_data
WHERE cylinders > 4
```

Meidum

For each stadium, how many concerts are there?

```
SELECT T2.name, COUNT(*)
FROM concert AS T1 JOIN stadium AS T2
ON T1.stadium_id = T2.stadium_id
GROUP BY T1.stadium_id
```

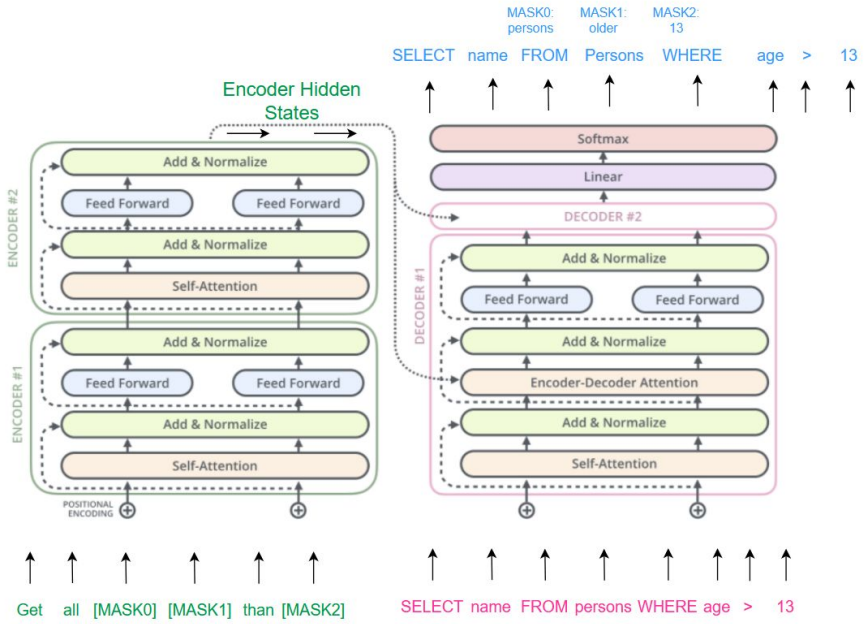
Why Spider?



CodeT5 Model

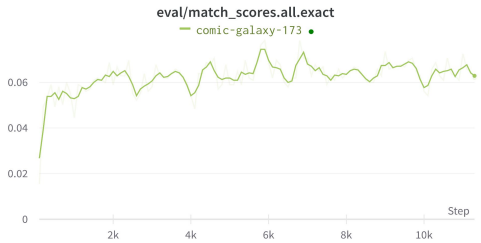
Training input: Get all persons older than 13

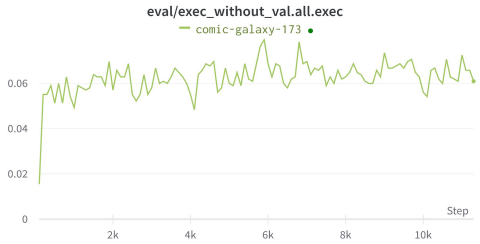
Training label: SELECT name FROM persons WHERE age > 13

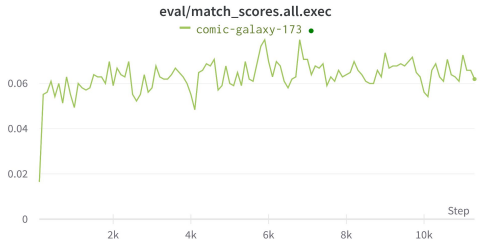


Results, Baselines, and Hyperparameters

CodeT5 Base Model Evaluation Metric







System	Development EM%	Development EX%	Test EM%	Test EX%
BRIDGE v2 + BERT (ensemble)† (Lin et al., 2020)	71.1	70.3	67.5	68.3
SmBoP + GraPPa† (Rubin and Berant, 2021)	74.7	75.0	69.5	71.1
RATSQL + GAP† (Shi et al., 2021)	71.8	-	69.7	-
DT-Fixup SQL-SP + RoBERTa† (Xu et al., 2021)	75.0	-	70.9	-
LGESQL + ELECTRA† (Cao et al., 2021)	75.1	-	**72.0**	-
T5-Base (Shaw et al., 2021)	57.1	-	-	-
T5-3B (Shaw et al., 2021)	70.0	-	-	-
T5-Base (ours)	57.2	57.9	-	-
T5-Base+PICARD	65.8	68.4	-	-
T5-Large	65.3	67.2	-	-
T5-Large+PICARD	69.1	72.9	-	-
T5-3B (ours)	69.9	71.4	-	-
T5-3B+PICARD	74.1	76.3	-	-
T5-3B†	71.5	74.4	68.0	70.1
T5-3B+PICARD†	**75.5**	**79.3**	**71.9**	**75.1**

CodeT5 base -	7.8%	7.9%
CodeT5 small no pretraining -	0%	0%
TypeSQL (Yale, 2018) -	8.2%	
Seq2Seq + attention (UEdinburgh, 2016) -	4.8%	

Hyperparameters:

batch_size = 16
num_train_epochs = 30
learning_rate = 1e-4
Weight_decay = .01
Warmup_steps = 200