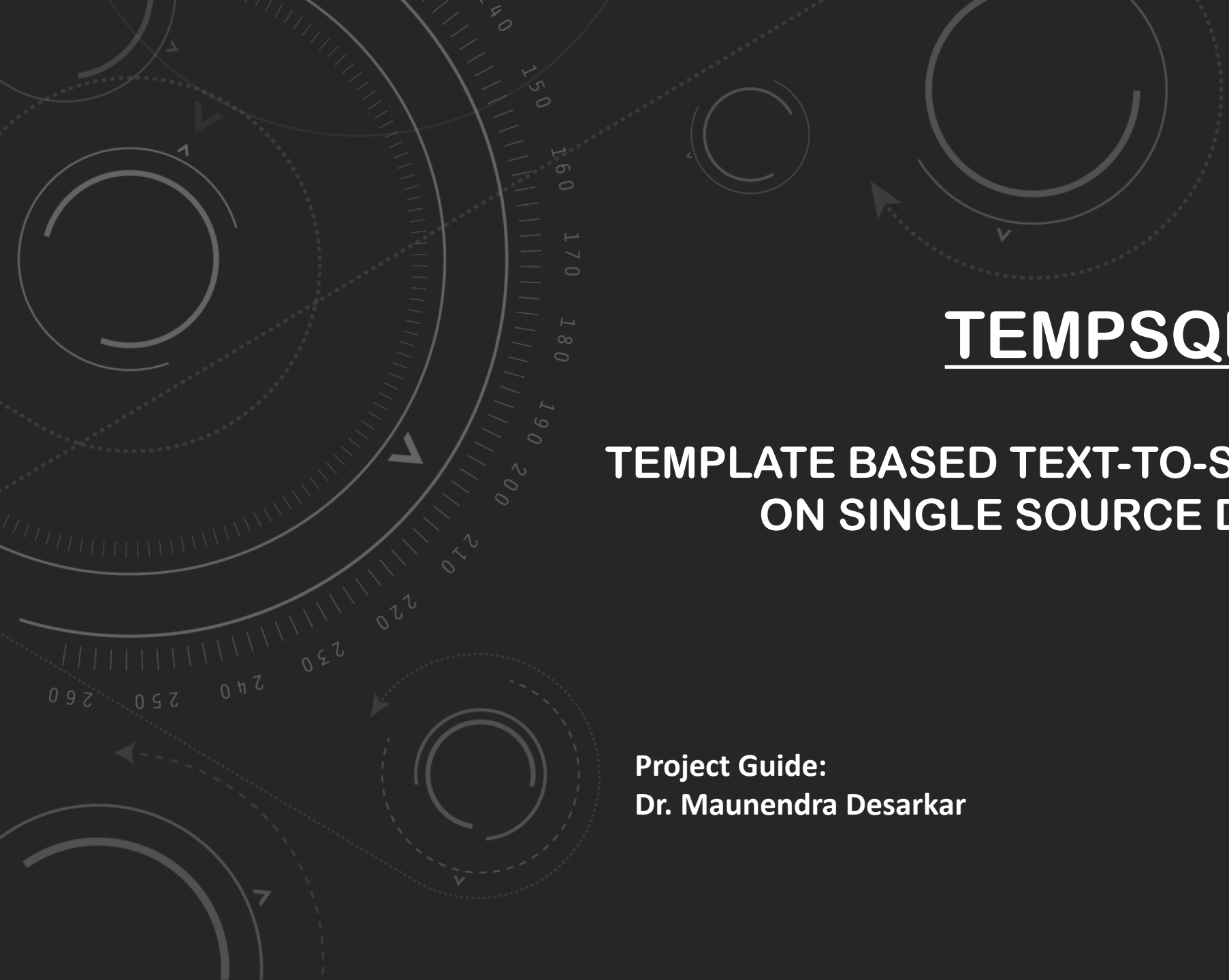


भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

TEMPSQL

TEMPLATE BASED TEXT-TO-SQL GENERATION ON SINGLE SOURCE DATABASE

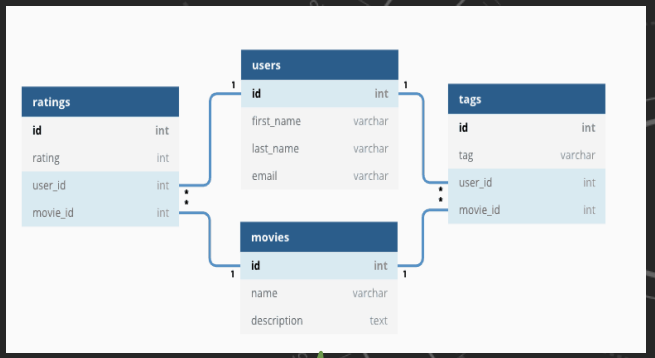
Project Guide:
Dr. Maunendra Desarkar

Submitted by:
Abhishek Bhatt
Anand Kumar Rai

SQL has always been too technical for non-technical users

Dilbert,1995

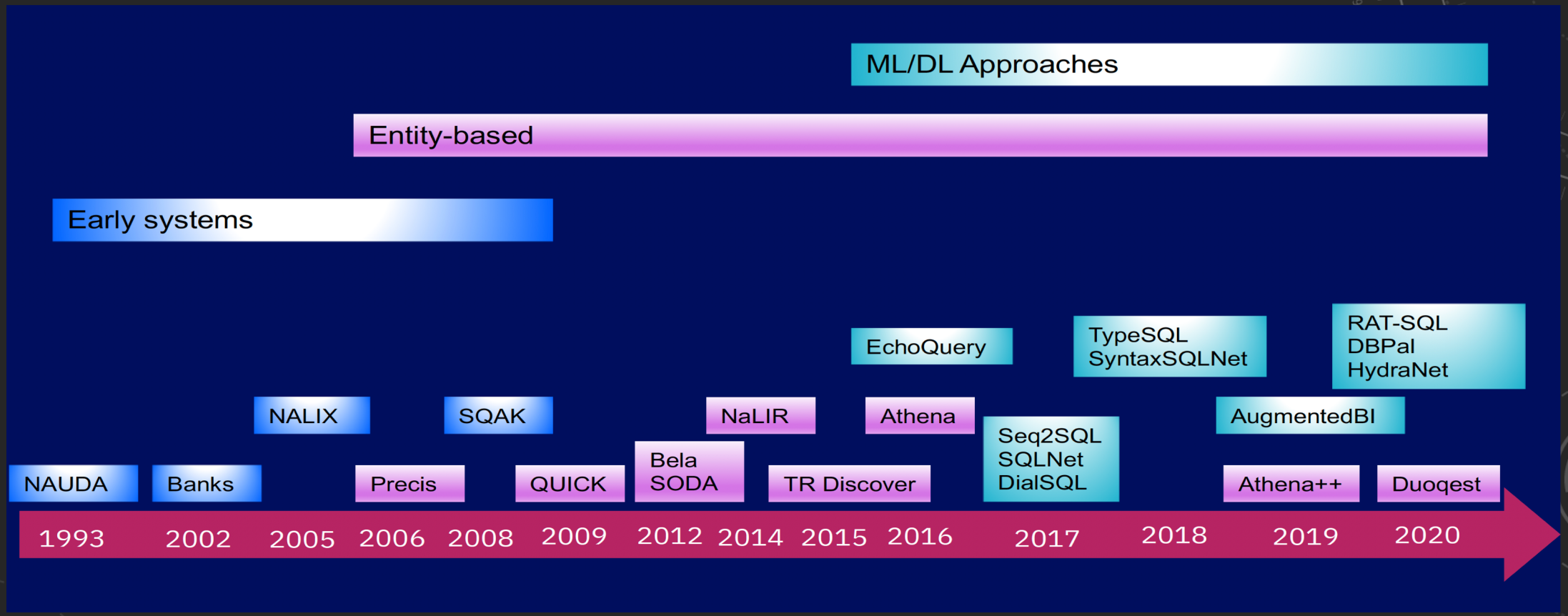
PROBLEM STATEMENT



- Input:
  - Natural Language Query
  - Database Schema

- Desired Result:
  - SQL equivalent of NL Query
  - Result corresponding to Query

HISTORICAL PERSPECTIVE

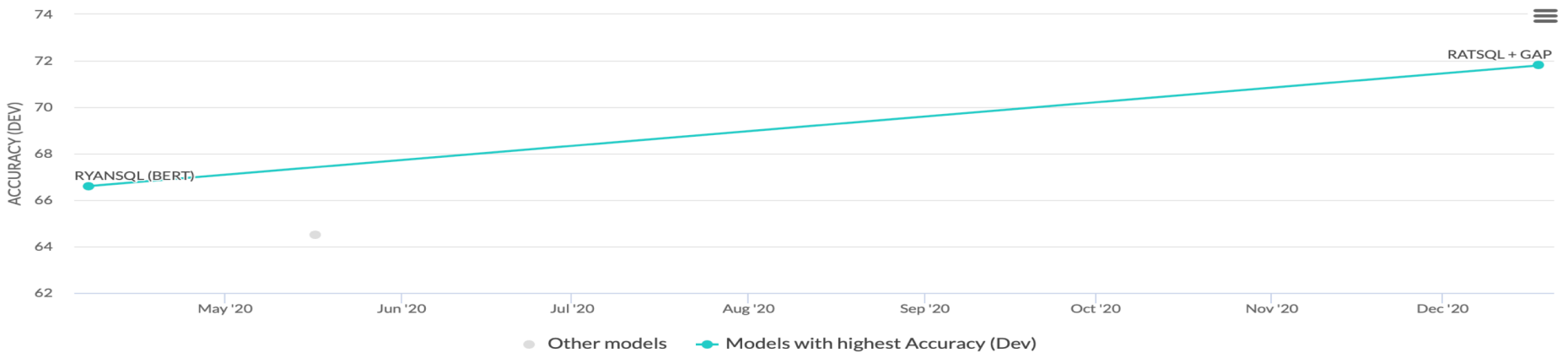


Source:*State of the Art and Open Challenges in NL Interfaces to Data, Ozcan et al , IBM, SIGMOD 2020*

CURRENT BENCHMARK

Text-To-Sql on spider

Leaderboard | Dataset



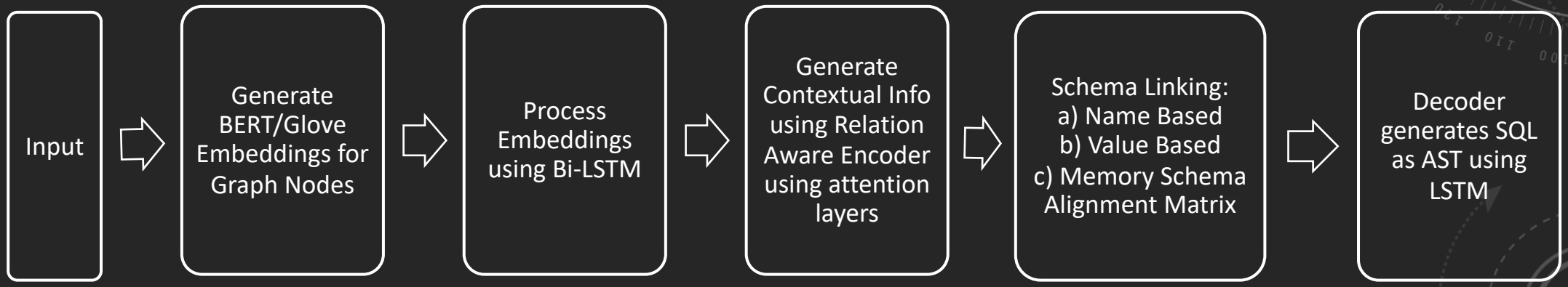
View Accuracy (Dev)

Edit Leaderboard

Rank	Model	Accuracy (Dev)	Accuracy (Test)	Paper	Code	Result	Year
1	**RATSQL + GAP**	71.8	69.7	Learning Contextual Representations for Semantic Parsing with Generation-Augmented Pre-Training			2020
2	**RYANSQL** (BERT)	66.6	58.2	RYANSQL: Recursively Applying Sketch-based Slot Fillings for Complex Text-to-SQL in Cross-Domain Databases			2020
3	**MAPO + TABERTLarge** (K = 3)	64.5		TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data			2020

LITERATURE SURVEY: RAT-SQL

- Approach
  - Input: Question Contextualised Database schema is represented as Graph G(V,E) ,
    V: {column names , table names, Question Words}, E: Relation between Question words & Schema Members



- Dataset Used:
  a) Spider Dataset (Yu et al., 2018b) GitHub - taoyds/spider: scripts and baselines for Spider: Yale complex and cross-domain semantic parsing and text-to-SQL challenge

  b) WIKISQL Dataset(Zhong et al., 2017). GitHub - salesforce/WikiSQL: A large annotated semantic parsing corpus for developing natural language interfaces.

LITERATURE SURVEY: RAT-SQL

- Results

Model	Dev	Test
IRNet (Guo et al., 2019)	53.2	46.7
Global-GNN (Bogin et al., 2019b)	52.7	47.4
IRNet V2 (Guo et al., 2019)	55.4	48.5
RAT-SQL (ours)	**62.7**	**57.2**
With BERT:		
EditSQL + BERT (Zhang et al., 2019)	57.6	53.4
GNN + Bertrand-DR (Kelkar et al., 2020)	57.9	54.6
IRNet V2 + BERT (Guo et al., 2019)	63.9	55.0
RYANSQL V2 + BERT (Choi et al., 2020)	**70.6**	60.6
RAT-SQL + BERT (ours)	69.7	**65.6**

Table 2: Accuracy on the Spider development and test sets, compared to the other approaches at the top of the dataset leaderboard as of May 1st, 2020. The test set results were scored using the Spider evaluation server.

Split	Easy	Medium	Hard	Extra Hard	All
RAT-SQL					
Dev	80.4	63.9	55.7	40.6	62.7
Test	74.8	60.7	53.6	31.5	57.2
RAT-SQL + BERT					
Dev	86.4	73.6	62.1	42.9	69.7
Test	83.0	71.3	58.3	38.4	65.6

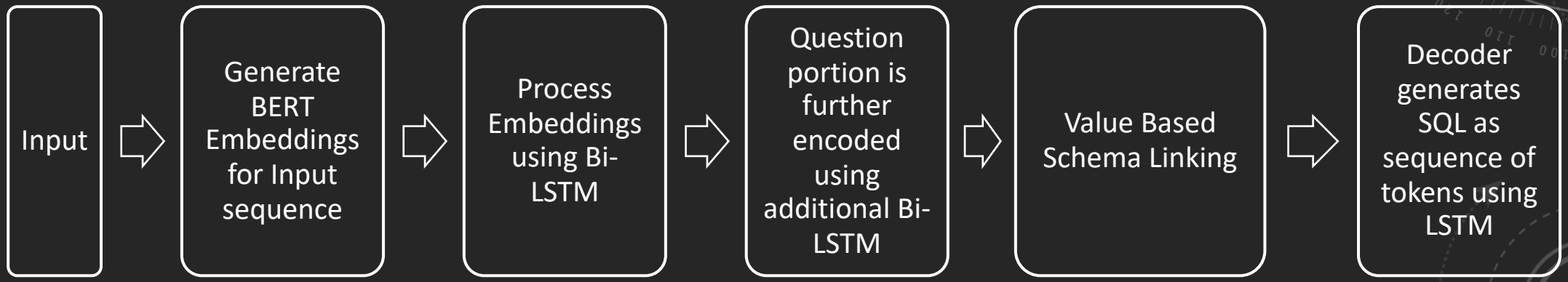
Table 3: Accuracy on the Spider development and test sets, by difficulty as defined by Yu et al. (2018b).

Model	Accuracy (%)
RAT-SQL + value-based linking	**60.54 ± 0.80**
RAT-SQL	55.13 ± 0.84
w/o schema linking relations	40.37 ± 2.32
w/o schema graph relations	35.59 ± 0.85

- Reference : Wang, Bailin, et al. “Rat-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers.” *arXiv preprint arXiv:1911.04942* (2019)

LITERATURE SURVEY: PHOTON

- Approach
  - Input: Combination of Question and DB schema as a tagged sequence



  - SQL correctness checking and the question corrector for ambiguous & untranslatable questions using user feedback are additional steps taken for improving performance
  - Dataset Used:
    a) Spider Dataset (Yu et al., 2018b) GitHub - taoyds/spider: scripts and baselines for Spider: Yale complex and cross-domain semantic parsing and text-to-SQL challenge

LITERATURE SURVEY: PHOTON

- Results

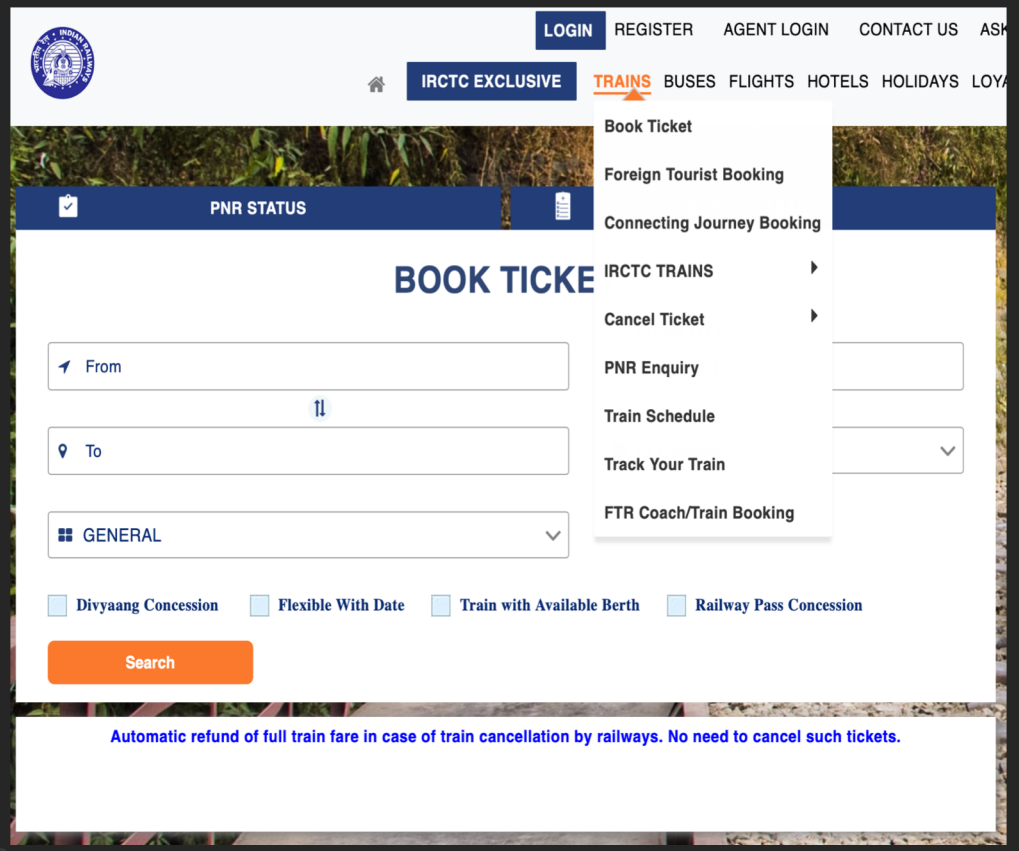
Model	EM Acc.
GNN (Bogin et al., 2019a)	40.7
Global-GNN (Bogin et al., 2019b)	52.7
EditSQL + BERT (Zhang et al., 2019)	57.6
GNN+Bertrand-DR† (Kelkar et al., 2020)	57.9
EditSQL+Bertrand-DR† (Kelkar et al., 2020)	58.5
IRNet + BERT (Guo et al., 2019)	61.9
RYANSQL + BERT † (Choi et al., 2020)	66.6
PHOTON	63.2

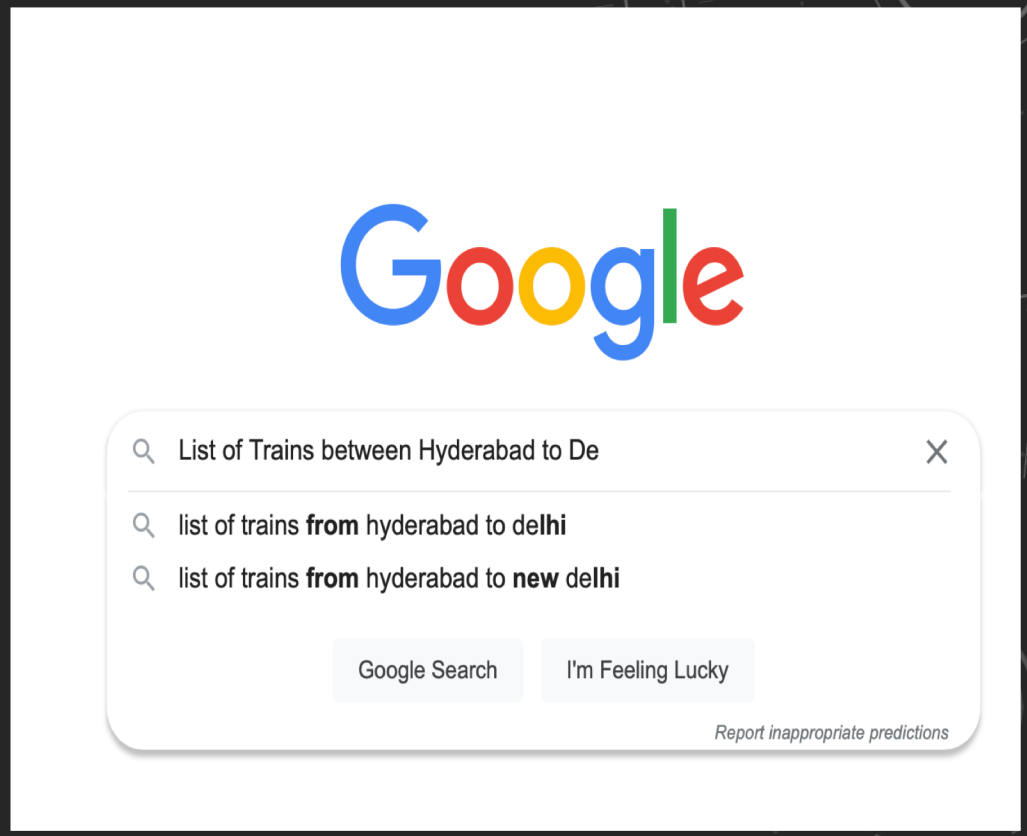
† denotes unpublished work on arXiv.

Table 3: Experimental results on the Spider Dev set (%). EM Acc. denotes the exact set match accuracy.

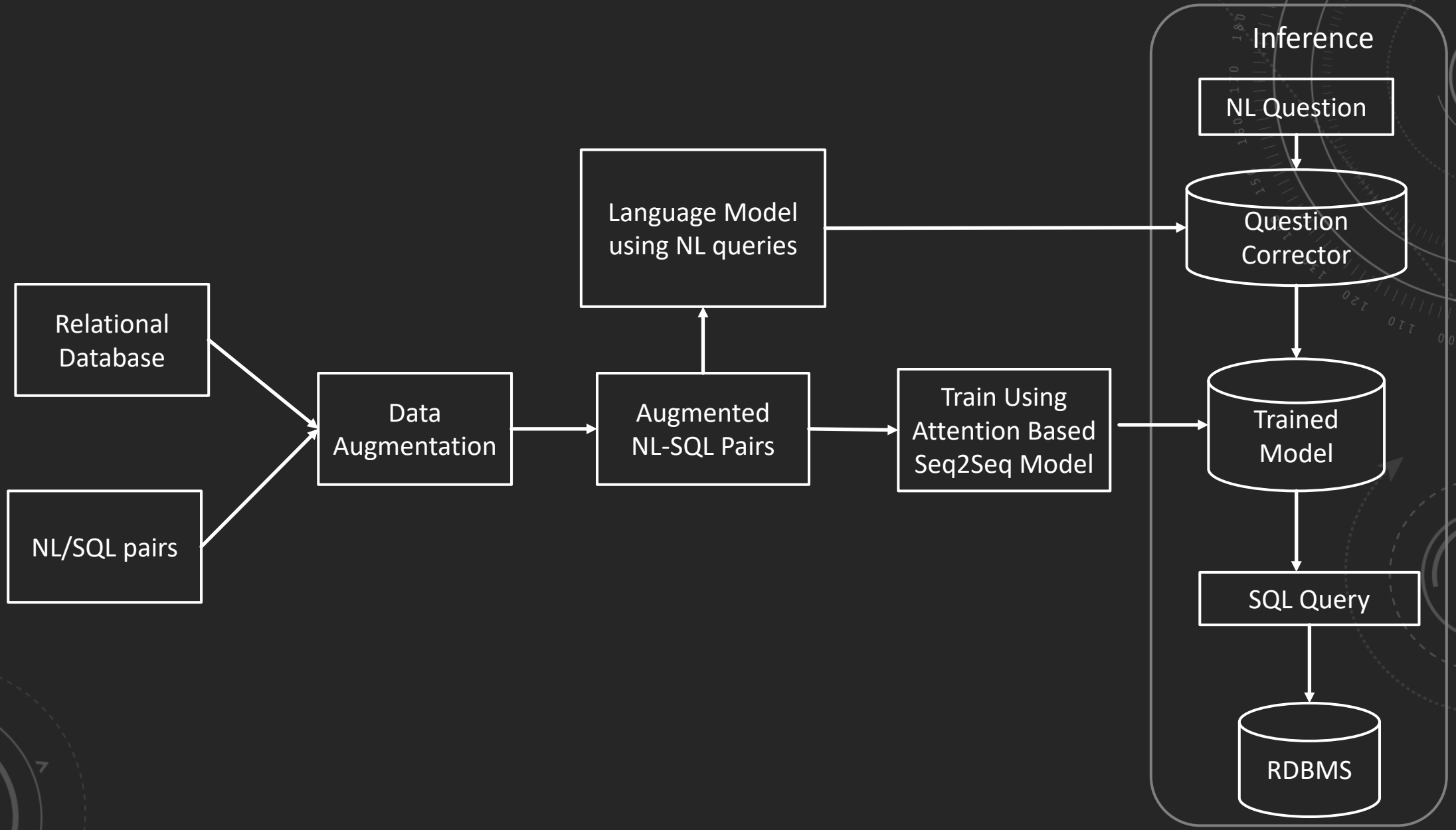
- Reference : Jichuan Zeng, Xi Victoria Lin, Caiming Xiong, Richard Socher, Michael R. Lyu, Irwin King, and Steven C. H. Hoi. Photon: A robust cross-domain text-to-sql system. CoRR, abs/2007.15280, 2020.
- Live Demo of the project is deployed at client (naturalsql.com)

MOTIVATION BEHIND OUR PROPOSED APPROACH





PROPOSED APPROACH: TEMPSQL SUMMARY



DATASET SELECTION

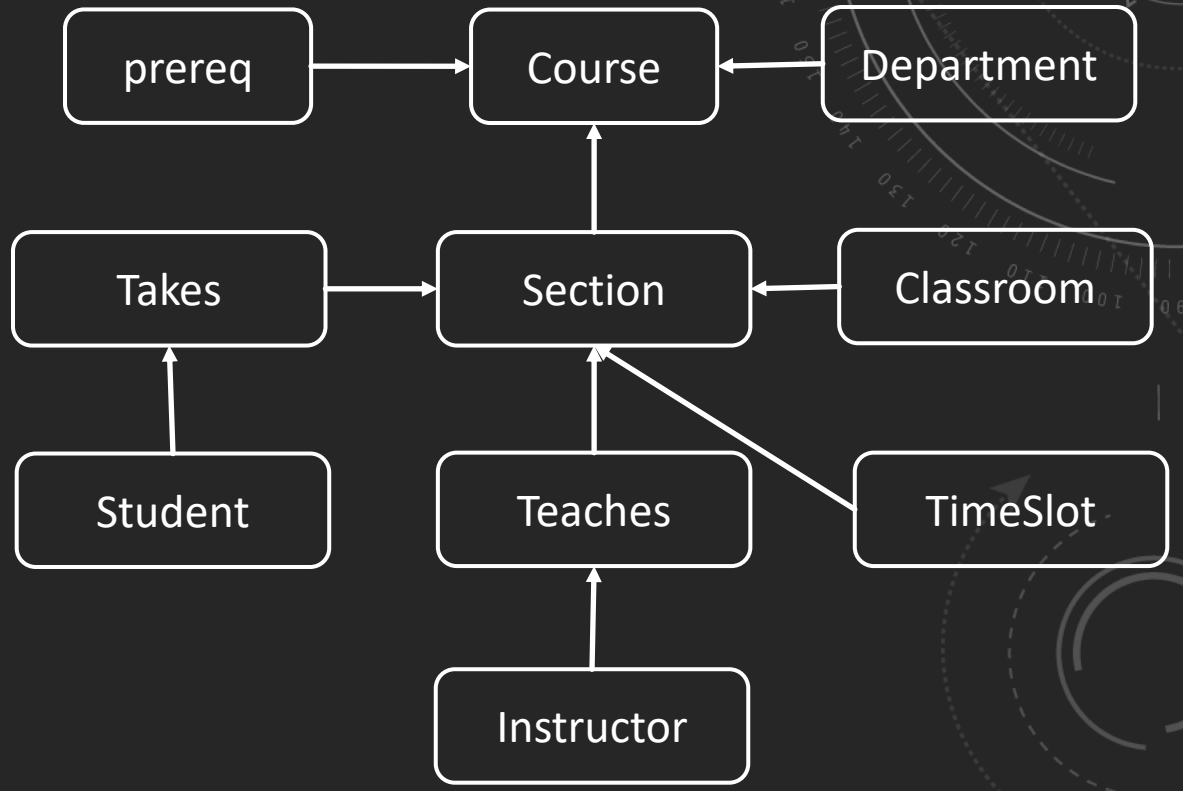
Dataset	# Q	# SQL	# DB	# Domain	# Table /DB	ORDER BY	GROUP BY	NESTED	HAVING
ATIS	5,280	947	1	1	32	0	5	315	0
GeoQuery	877	247	1	1	6	20	46	167	9
Scholar	817	193	1	1	7	75	100	7	20
Academic	196	185	1	1	15	23	40	7	18
IMDB	131	89	1	1	16	10	6	1	0
Yelp	128	110	1	1	7	18	21	0	4
Advising	3,898	208	1	1	10	15	9	22	0
Restaurants	378	378	1	1	3	0	0	4	0
WikiSQL	80,654	77,840	26,521	-	1	0	0	0	0
Spider	10,181	5,693	200	138	5.1	1335	1491	844	388

Ref: https://www.aclweb.org/anthology/D18-1425.pdf

- Summary of SPIDER Dataset
  - 166 databases
  - 138 domains
  - Average 20-50 NL – SQL pairs per database
  - Large SQL pattern coverage
    SELECT, WHERE, GROUP BY, HAVING, ORDER BY, LIMIT, JOIN, INTERSECT, EXCEPT, UNION, NOT IN, OR, AND, EXISTS, LIKE

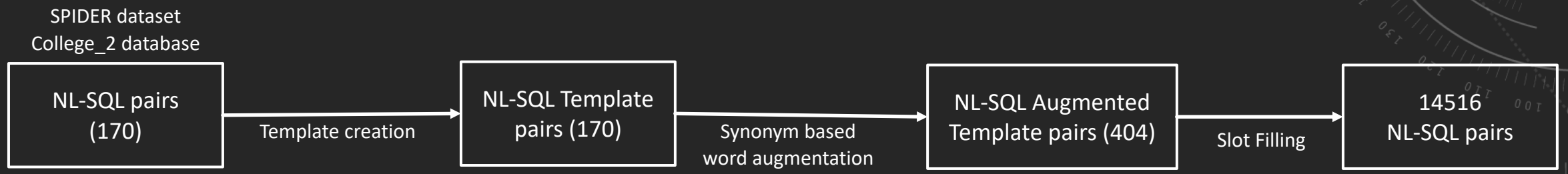
DATABASE SELECTION

- “College_2” was selected out of the 166 databases:
  - Highest number of annotated NL - SQL pair (170)
  - Queries involve JOIN, INTERSECT, EXCEPT, UNION, IN & nested queries



College_2 database schema

DATA AUGMENTATION



DATA AUGMENTATION

Slot Filling

NL Query

List the instructors from Physics Department

List the instructors from Cybernetics Department
List the instructors from Chemistry Department
List the instructors from Mech Department
…

SQL Query

SELECT name FROM instructor WHERE dept_name = 'Physics'

SELECT name FROM instructor WHERE dept_name = 'Cybernetics'
SELECT name FROM instructor WHERE dept_name = 'Chemistry'
SELECT name FROM instructor WHERE dept_name = 'Mech'
…

Table : Instructor
Column : dept_name
Physics Cybernetics Chemistry Mech …

DATA AUGMENTATION

Slot Filling

NL Query

List the instructors from Physics Department

NL Template

List the instructors from {instructor.dept_name} department?

Generated NL queries

List the instructors from Physics Department
List the instructors from Cybernetics Department
List the instructors from Chemistry Department
List the instructors from Mech Department
...

SQL Query

SELECT name FROM instructor WHERE dept_name = 'Physics'

SQL Template

SELECT name FROM instructor WHERE dept_name = '{instructor.dept_name}'

Generated SQL queries

SELECT name FROM instructor WHERE dept_name = 'Physics'
SELECT name FROM instructor WHERE dept_name = 'Cybernetics'
SELECT name FROM instructor WHERE dept_name = 'Chemistry'
SELECT name FROM instructor WHERE dept_name = 'Mech'
...

DATA AUGMENTATION

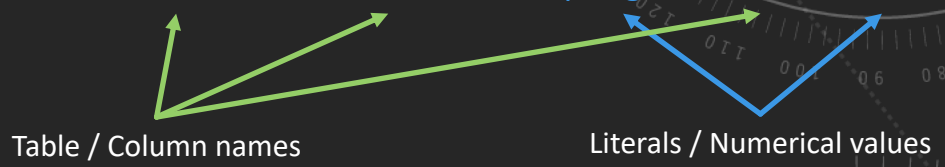
Automation of Template Creation

NL Query from Spider dataset

How many instructors teach a course in the Spring of 2010?

SQL Query from Spider dataset

SELECT COUNT (DISTINCT ID) FROM teaches WHERE semester = 'Spring' AND YEAR = 2010



Auto generated NL Template

How many instructors teach a course in the {teaches.semester} of {teaches.YEAR}?

Auto generated SQL Template

SELECT COUNT (DISTINCT ID) FROM teaches WHERE semester = {teaches.semester} AND YEAR = {teaches.YEAR}

Template Generation for College_2 database		
	Template	**NL-SQL pair**
Manual	22 (28%)	5431 (37%)
Automatic	57 (72%)	9175 (63%)
Total	79	14516

"moz_sql_parser" library for parsing SQL queries

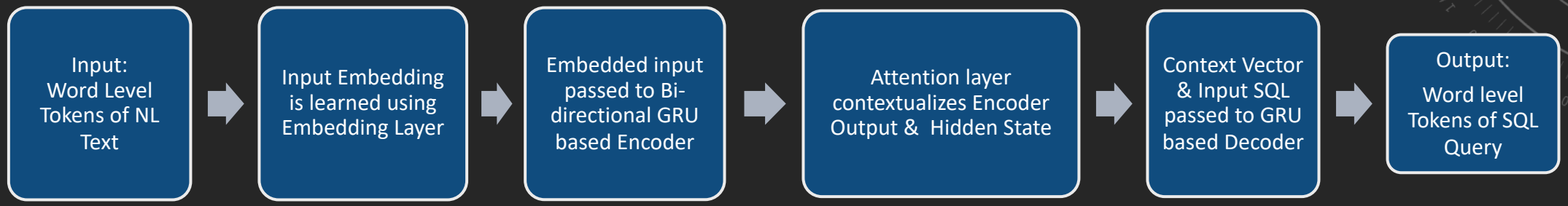
DATA AUGMENTATION

Synonym based word augmentation

NL Query	SQL Query
Find the name of the courses that do not have any prerequisite?	SELECT title FROM course WHERE course_id NOT IN (SELECT course_id FROM prereq)
Get the name of the courses that do not have any prerequisite?	SELECT title FROM course WHERE course_id NOT IN (SELECT course_id FROM prereq)
Retrieve the name of the courses that do not have any prerequisite?	SELECT title FROM course WHERE course_id NOT IN (SELECT course_id FROM prereq)
Obtain the name of the courses that do not have any prerequisite?	SELECT title FROM course WHERE course_id NOT IN (SELECT course_id FROM prereq)

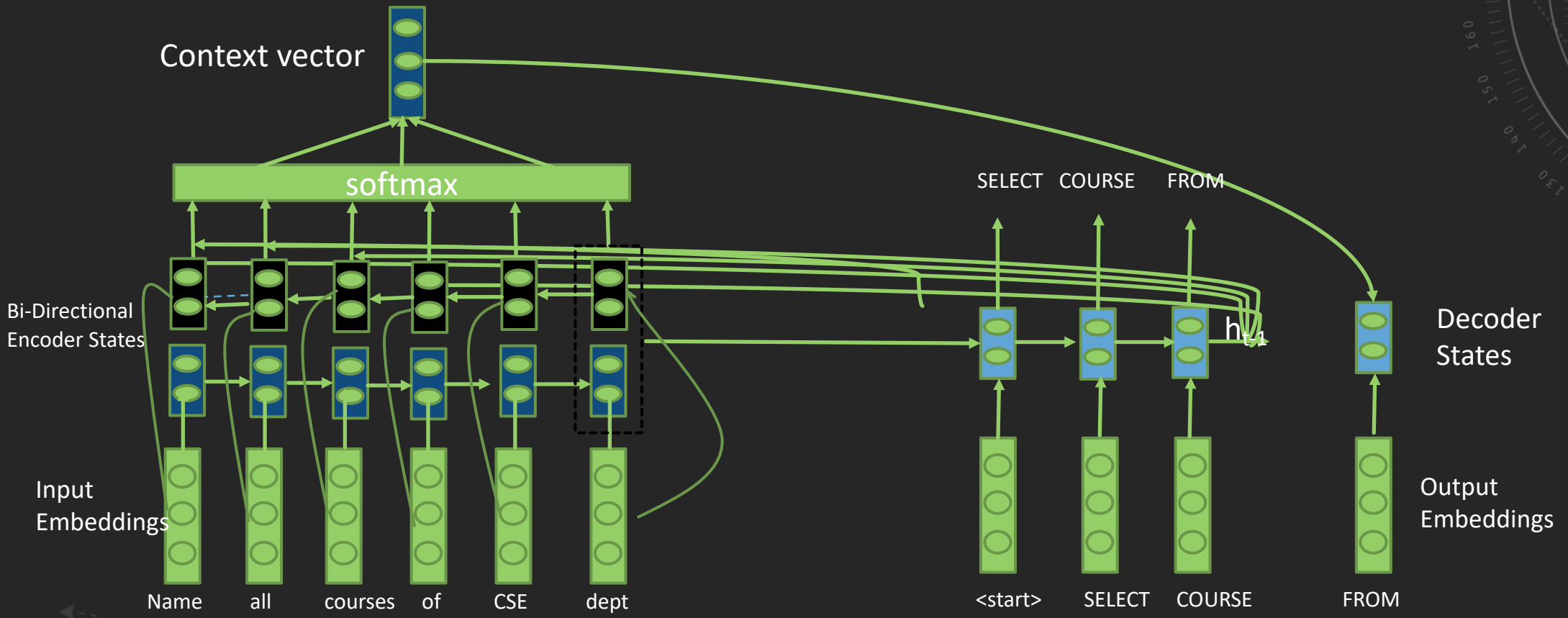
Synonyms from NLTK & PiDictionary
Find
Get
Retrieve
Obtain

TEMPSQL: MODEL DESCRIPTION




```
X= Embedding(X)                                          [NL]
EO, EH = Encoder(X)
attention_score      =FC(tanh(FC(EO) + FC(EH)))          [Badhnau Attention]
context_vector    = sum( softmax(attention_score ) * EO)
Y = Embedding(Y)                                         [SQL]
output     = Decoder(concat(Y, context vector))
```

TEMPSQL: MODEL DESCRIPTION



TEMPSQL: MODEL PERFORMANCE

Dataset: College_2	No. of Samples	Accuracy
Train	11612	72.91
Test	2904	71.56

Types of Data	Samples (%)	Test Accuracy (%)
Easy (Simple No Rel Operator)	4%	87.9
Medium (Rel Op Present)	38%	99.7
Hard (Nested Queries)	19%	95.8
Difficult (Multiple Joins, Intersect, Except)	39%	29.7

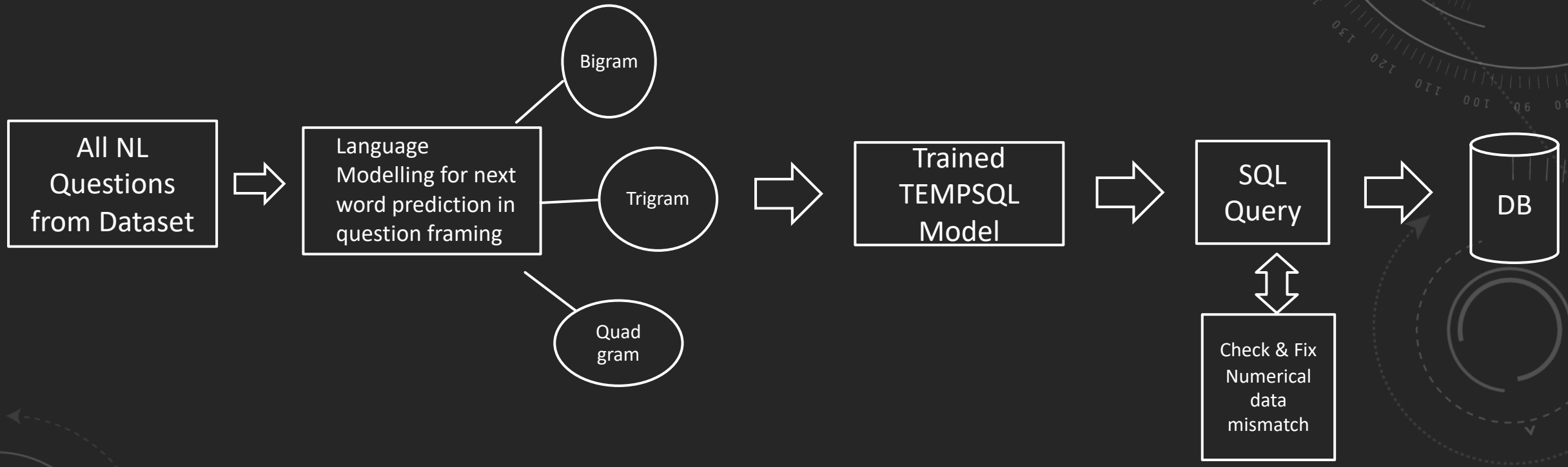
TEMPSQL: MODEL PERFORMANCE ON DIFFERENT DATASET

SPIDER dataset
Store_1 database

Dataset: store_1	No. of Samples	Accuracy
Train	28537	81.38 %
Test	7135	79.41 %

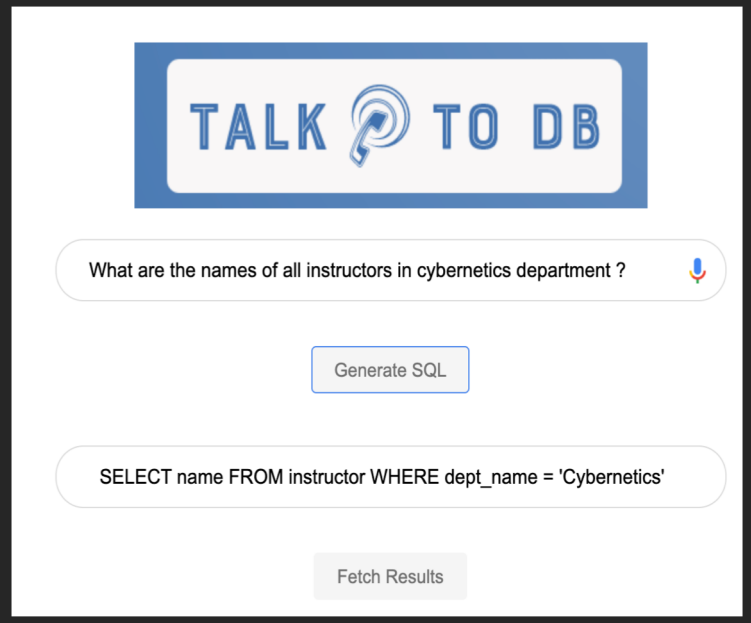
Types of Data	Samples (%)	Test Accuracy(%)
Easy (Simple No Rel Operator)	47.06%	98.27%
Medium (Rel Op Present)	8.35%	90.43
Hard (Multiple Joins Present)	0 %	NA
Difficult (Nested Queries)	44.45%	57.59 %

TALK2DB: INTERACTIVE WEB BASED UI FOR INFERENCE

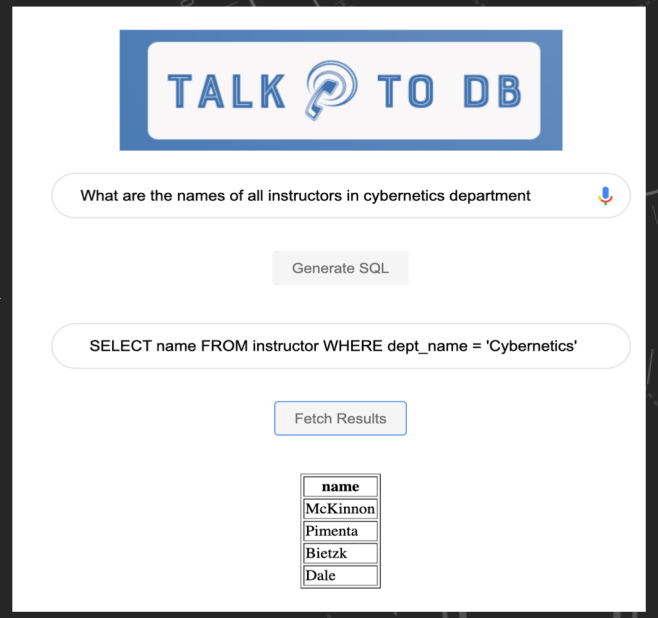


TALK2DB: INTERACTIVE WEB BASED UI FOR INFERENCE

Language Model for AutoComplete



SQL Generation from NL using trained TempSQL Model



Fetch Results from DB

BENEFITS OF PROPOSED APPROACH

- No dependency on external dataset
- Easily transition for existing industry reporting software
- Simplistic and Specialized model with no unnecessary learned embeddings
- Continual improvement in performance possible by adding templates
- Assistance to New users with no domain knowledge with auto complete

SCOPE FOR FURTHER WORK

- Template generation from DB Schema
- Upgradation to Inference UI to conversational mode
- Improvement of accuracy on complex queries

THANK YOU


IT'S CALLED A DASHBOARD. IT SHOWS THE CURRENT STATUS OF ALL OUR PROJECTS.
Dilbert.com @ScottAdamsSays
WITH A TOOL LIKE THIS, YOU NEVER NEED TO ASK US FOR STATUS UPDATES.
7-28-16 © 2016 Scott Adams, Inc. /Dist. by Universal Uclick
HOW'D THE FAKE DASHBOARD GAMBIT WORK OUT?
GREAT! HE HASN'T TALKED TO ME IN WEEKS.