Improving Text-to-SQL Evaluation Methodology

Catherine Finegan-Dollak,[1*] Jonathan K. Kummerfeld,[1*] Li Zhang,[1]
Karthik Ramanathan,[1] Sesh Sadasivam,[1] Rui Zhang,[2] and Dragomir Radev[2]
[1]University of Michigan, [2]Yale University

Evaluations should measure how well systems generalize to realistic unseen data. Yet standard train/test splits, which ensure that no *English question* is in both train and test, permit the same *SQL query* to appear in both. Using a simple classifier with a slot-filler as a basline, we show how the standard question-based split fails to evaluate a system's generalizability. In addition, *b*y analyzing properties of human-generated and automatically generated text-to-SQL datasets, we show the need to evaluate on more than one dataset to ensure systems perform well on realistic data. And we release improved resources to facilitate such evaluations.

Evaluate on more than one SQL dataset

Dataset size does not predict diversity.

- Especially small datasets from DB community had a higher ratio of unique queries to questions than the NLP community's datasets.
- Over half of the largest dataset follows one pattern.

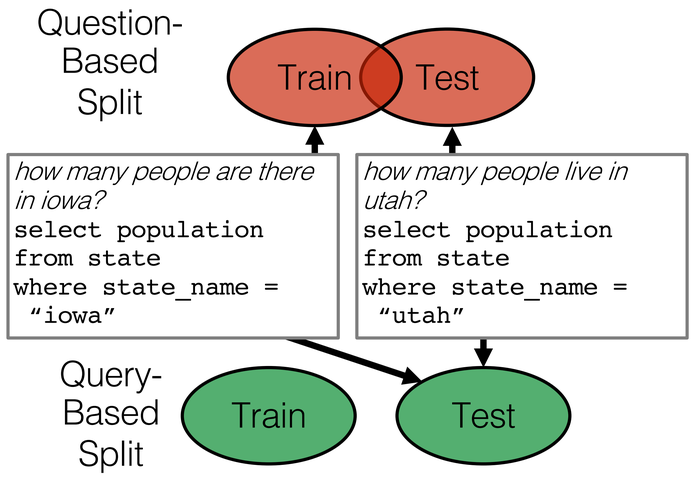
```
SELECT <field> AS <alias>
FROM <table>
WHERE <field> = <literal>;
```

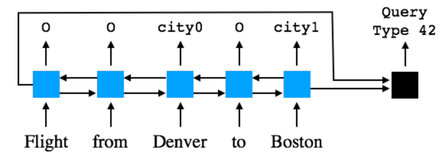
Human-generated data is more complex than auto-generated data.

- Questions humans ask require joins and nesting.
- Task-oriented datasets (Advising, ATIS) were particularly complex.

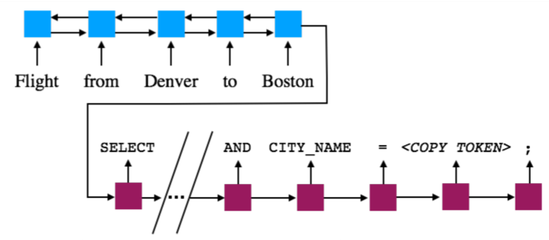
	Questions	Unique queries	Questions per pattern		Tables per query		Nesting depth	
			Mean	Max	Mean	Max	Mean	Max
Advising	4570	211	20.3	90	3.2	9	1.18	4
ATIS	5280	947	7.0	870	6.4	32	1.39	8
GeoQuery	877	246	8.9	327	1.4	5	2.03	7
Restaurants	378	23	22.2	81	2.6	5	1.17	2
Scholar	817	193	5.6	71	3.3	6	1.02	2
Academic	196	185	2.1	12	3.2	10	1.04	2
IMDB	131	89	2.5	21	1.9	5	1.01	2
Yelp	128	110	1.4	11	2.2	4	1	1
WikiSQL	80,654	77,840	165.3	42,816	1	1	1	1

Evaluate on *question split* and *query split*





Baseline



Attention-based copying seq2seq

(see paper for additional datasets and systems)	Advising		ATIS		GeoQuery		Scholar		Academic		IMDB	
	?	Q	?	Q	?	Q	?	Q	?	Q	?	Q
Baseline	**80**	0	46	0	57	0	52	0	0	0	0	0
s2s copying	70	0	**51**	**32**	**71**	20	**59**	5	**81**	**74**	**26**	**9**
Dong 2016	46	**2**	46	23	62	31	44	**6**	63	54	6	2
Iyer 2017	41	1	45	17	66	**40**	44	3	76	70	10	4
Baseline	89	0	56	0	56	0	66	0	0	0	7	0
Dong 2016	88	8	56	34	68	23	68	6	65	61	36	10
Iyer 2017	88	6	58	32	71	49	71	1	77	75	52	24
Baseline Oracle	100	0	69	0	78	0	84	0	11	0	47	0

Color Key: No oracle | Oracle entities | Full oracle

Abbreviation Key: ? = Question split | Q = Query split

- A baseline classifier + slot-filler performs comparably to state-of-the-art on question split.
- Query-based split tests generalizability to unseen queries.
- Variable anonymization noticeably decreases the difficulty of the task.

New resources to make these evaluations easier

- New Advising dataset, plus 7 existing-text-to-SQL datasets cleaned, variablized, and put into a single, standard format, with tools for easy use.
- Scan above or visit `https://github.com/jkkummerfeld/text2sql-data`

This material is based in part upon work supported by IBM.

* The first two authors contributed equally to this work.