Generating Semantically Valid Adversarial Questions for TableQA

Yi Zhu[♡], Yiwei Zhou[♢], Menglin Xia[♢]

[♡]Language Technology Lab, University of Cambridge, [♢]Amazon Alexa

AAAI-RSEML 2021

TableQA

Question answering on tabular data

- Input:
  - Natural language questions
  - Table/table schema
- Output:
  - Logical form
  - Final answer
- WikiSQL dataset [Zhong et al., 2017]
  - First large-scale dataset for TableQA (Text-to-SQL)
    - 24, 241 Wikipedia tables
    - 80, 654 pairs questions and SQL queries

Table:

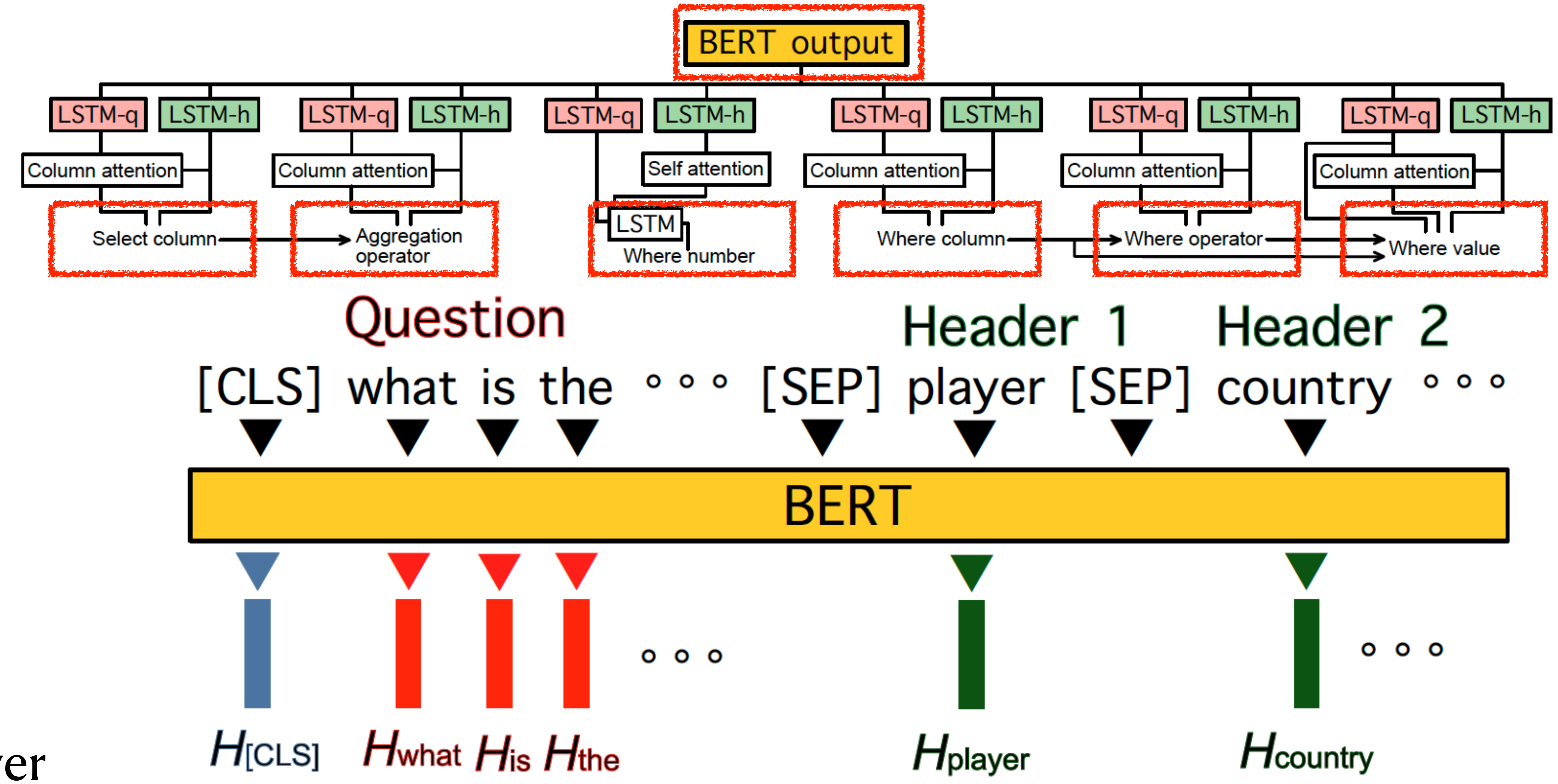
Rank	Nation	Gold	Silver	Bronze	Total
1	Russia	2	2	2	6
2	France	1	0	0	1
2	Hungary	1	0	0	1
4	Ukraine	0	1	1	2
5	Bulgaria	0	1	0	1
6	Poland	0	0	**1**	1

Question: What is the bronze value associated with ranks over 5?
SQL query: `SELECT` Bronze `WHERE` Rank > 5
Answer: 1

TableQA systems for WikiSQL

“Question”: “question_text”
“SQL” : {
“SELECT”: (agg , scol)
“WHERE”: [(wcol1 , = , value1) , (wcol2 , op2 , value2) , ...]
“TABLE”: “table_id”
}

- BERT-based encoder [Devlin et al. 2019]
- SQL generation -> multiple classification tasks
- SQLova as target system
  - 80.7 Q-Acc and 86.2 A-Acc
- Evaluation
  - Query Acc. $\text{Q-Acc} = \frac{\#\text{correct SQL}}{\#\text{test example}}$
  - Answer Acc. $\text{A-Acc} = \frac{\#\text{correct answer}}{\#\text{test example}}$



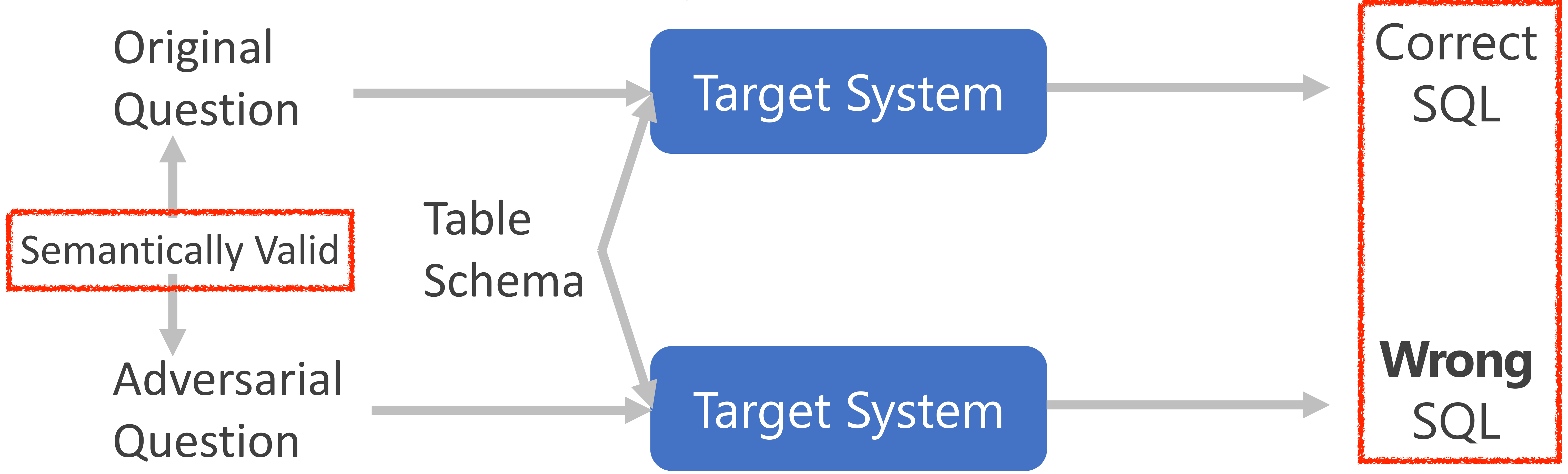
[Hwang et al. 2019]

Motivation

- **To what extent can the TableQA systems understand natural language questions and reason with given tables?**
- To answer it, we leverage white-box adversarial questions

What is the bronze value associated with ranks over 5?

SELECT Bronze WHERE Rank > 5



$$\arg\max_{\mathbf{L}\in\mathbb{L}} p(\mathbf{L}|\mathbf{y},\mathbf{t}) = \mathbf{L}_{\text{true}}(\mathbf{y})$$

$$\arg\max_{\mathbf{L}\in\mathbb{L}} p(\mathbf{L}|\hat{\mathbf{y}},\mathbf{t}) \neq \mathbf{L}_{\text{true}}(\mathbf{y}) \quad \text{s.t.} \quad \hat{\mathbf{y}} \overset{\text{semantic valid}}{=} \mathbf{y}$$

What is the bronze value whose rank is larger than 5?

SELECT **COUNT(Bronze)** WHERE Rank > 5

Semantically valid: **For humans**, adversarial questions yield the same correct SQL as original question

Problem definition and previous methods

- Adversarial loss for white-box adversarial questions

$$\mathcal{L}_{\text{adv}}(\hat{\mathbf{y}}, \mathbf{L}_{\text{true}}(\mathbf{y}), \mathbf{t}) = - \sum_{l \in \mathbf{L}_{\text{true}}(\mathbf{y})} \log(1 - p(l|\hat{\mathbf{y}}, \mathbf{t}))$$

- To produce *semantically valid* questions
  - Most previous methods constrained to **local** models [Abraham et al. 2018; Ren et al. 2019; Zhang et al. 2019, inter alia]
    - *Word/subword/character* level manipulation such as insertion/deletion/substitution
  - Few token swaps are less likely to lead to large semantic shift

$$\underset{1 \leq i \leq |\mathbf{y}|, \hat{\mathbf{y}}_i \in \mathcal{V}}{\arg\min} \; [\hat{\mathbf{y}}_i - \mathbf{y}_i]^T \nabla_{\mathbf{y}_i} \mathcal{L}_{\text{adv}}(\hat{\mathbf{y}}, \mathbf{L}_{\text{true}}(\mathbf{y}), \mathbf{t})$$

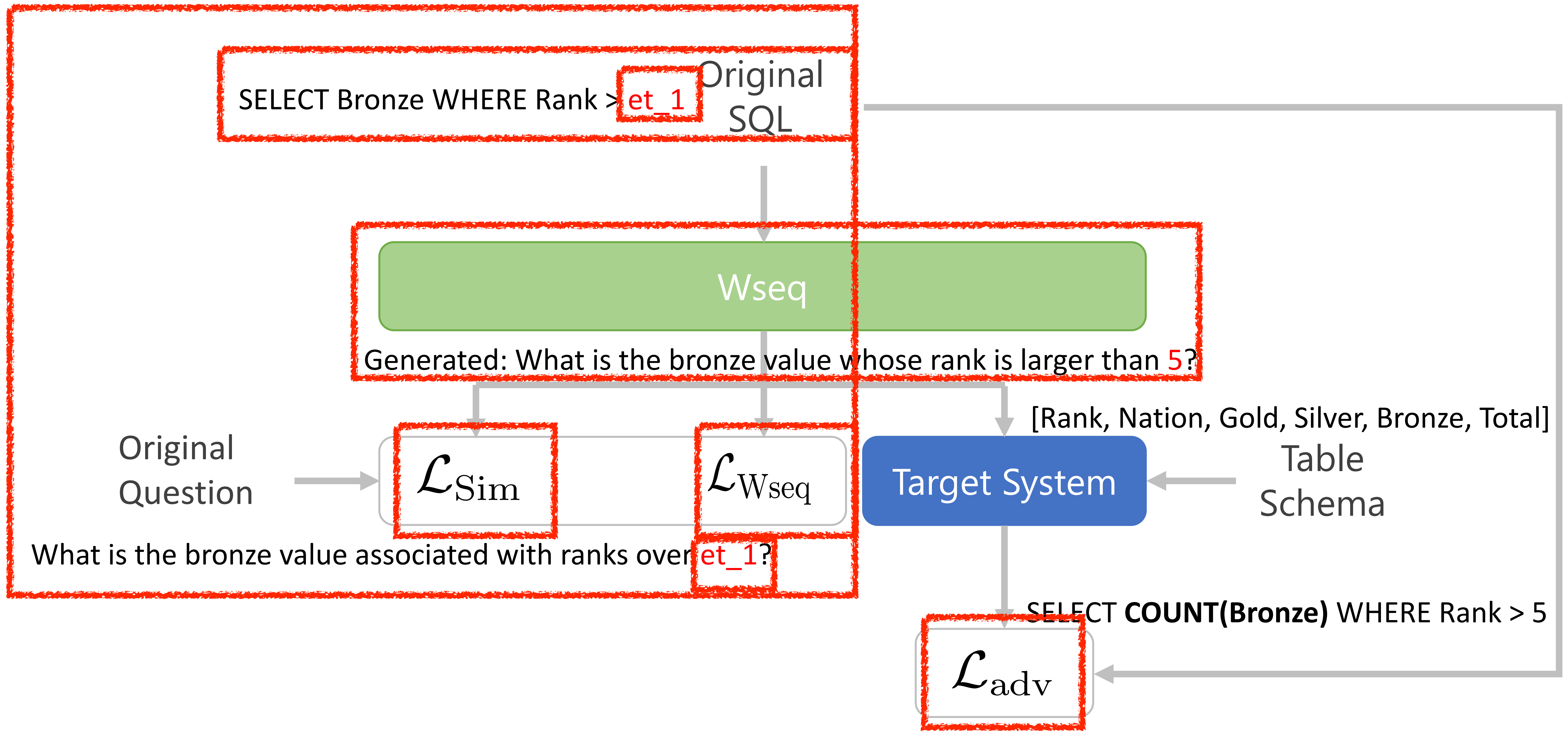
SAGE

- SAGE: **Semantically valid Adversarial GEnerator** for TableQA systems
- Generate adversarial questions at *sequence level*
  - Input: SQL query
  - Output: *Semantically valid* and *fluent* adversarial question that can fool TableQA systems

SAGE

Model architecture

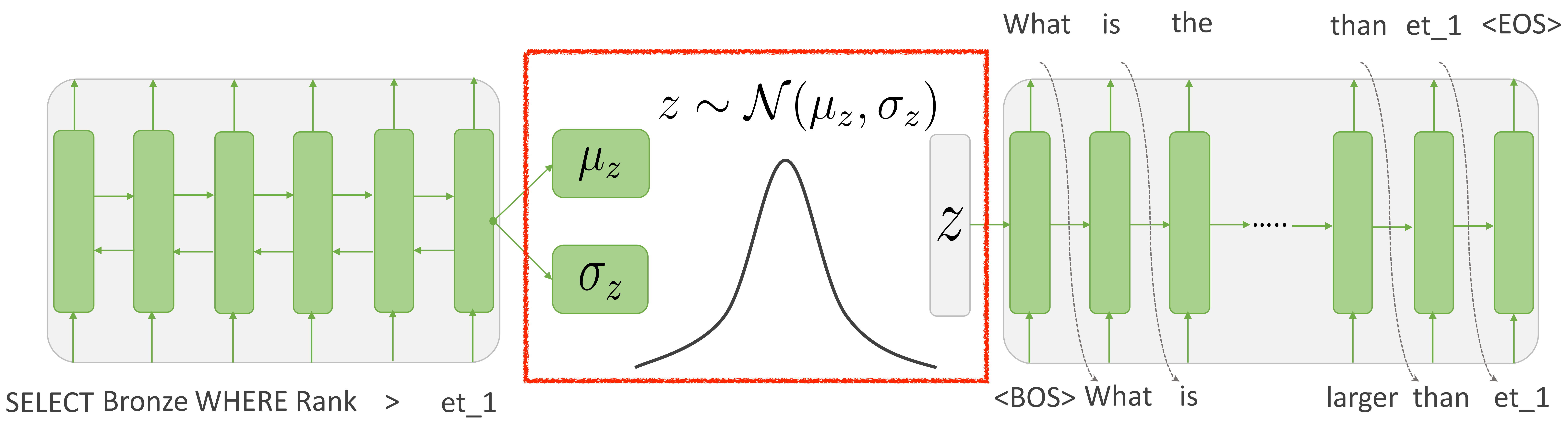
- SAGE has *three* components, each with a corresponding loss:
  1. Stochastic Wasserstein Seq2seq model for question generation (Wseq)
  2. Delexicalisation and minimum risk training with SIMILE to enhance semantic validity (Wseq-S)
  3. End-to-end training with adversarial loss using Gumbel-Softmax



$$\mathcal{L} = \mathcal{L}_{\text{Wseq}} + \sum_{(\mathbf{x},\mathbf{y},\mathbf{t})\in\mathcal{B}} \Big[\lambda_{\text{sim}}\mathcal{L}_{\text{sim}}(\mathbf{x},\mathbf{y}) + \sum_{\hat{\mathbf{y}}\in\mathcal{H}(\mathbf{x})} \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}(\hat{\mathbf{y}},\mathbf{x},\mathbf{t})\Big]$$

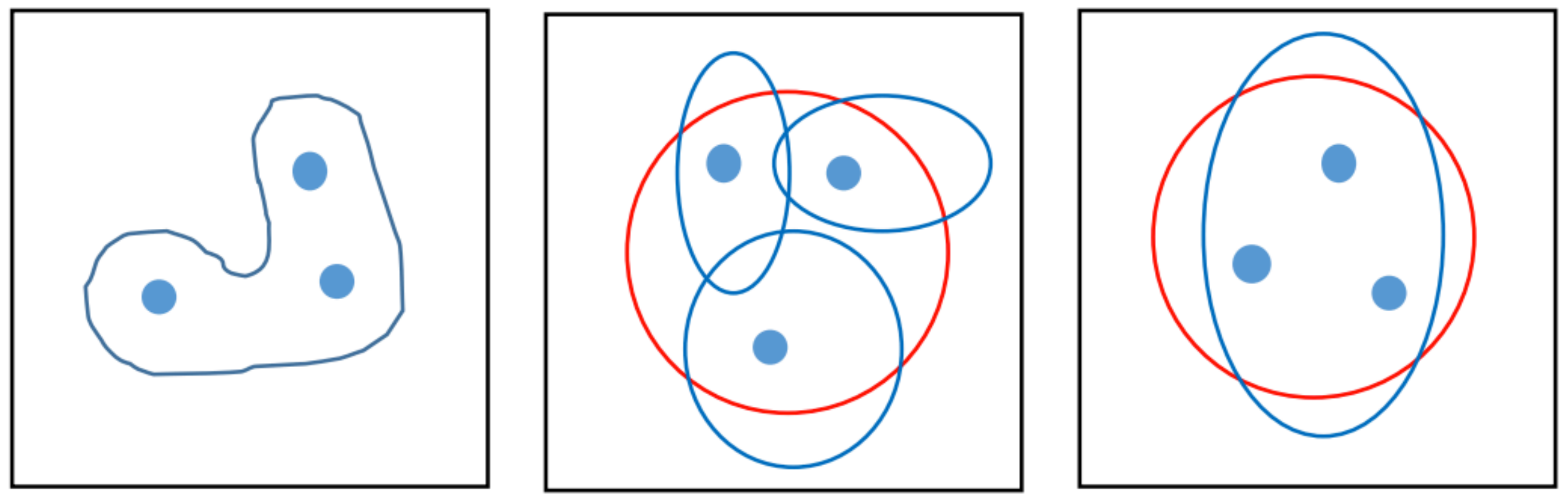
SAGE

Stochastic Wasserstein Seq2seq Model (Wseq)



$$\mathcal{L}_{\text{Wseq}} = -\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{B}} \sum_{i}^{|\mathbf{y}|} \log p(y_i|\mathbf{z}, \mathbf{y}_{<i}, \mathbf{x}) + \lambda_{\text{Wseq}} \boxed{\hat{D}_{\mathbf{z}}(q(\mathbf{z}), p(\mathbf{z}))}$$

$$\hat{D}_{\mathbf{z}}(q(\mathbf{z}), p(\mathbf{z})) = \sum_{i\neq j} \frac{k(\mathbf{z}_i, \mathbf{z}_j) + k(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j)}{n(n-1)} - 2\sum_{i,j} \frac{k(\mathbf{z}_i, \tilde{\mathbf{z}_j})}{n^2}$$



(a) DAE (b) VAE (c) WAE

[Bahuleyan et al. 2019]

SAGE

Delexicalisation

- Delexicalisation
  - Adversarial questions need contain all the entities in order to maintain semantic validity
  - *i-th entity* (WHERE values) -> *et_i*
    - Reduce the length of the entity tokens
    - Improve entity coverage

Table:

Rank	Nation	Gold	Silver	Bronze	Total
1	Russia	2	2	2	6
2	France	1	0	0	1
2	Hungary	1	0	0	1
4	Ukraine	0	1	1	2
5	Bulgaria	0	1	0	1
6	Poland	0	0	**1**	1

Question: What is the bronze value associated with ranks over 5?
SQL query: `SELECT` Bronze `WHERE` Rank > 5
Answer: 1

Delexicalised SQL query

SELECT Bronze WHERE Rank > et_1?

Delexicalised question

What is the bronze value associated with ranks over et_1?

SAGE

Minimum risk training with SIMILE (Wseq-S)

- SIMILE + minimum risk training [Wieting et al. 2019]
  - SIMILE
    - A pretrained neural network model calculating cosine similarity between embeddings of two sentences
  - Why SIMILE over other string matching based metrics like BLEU?
    - Our generated questions
      - Different in lexical/syntactic realizations
      - High semantic similarity
    - Correlate better with human judgement
  - Minimum risk training [Shen et al. 2016]

$$\mathcal{L}_{\text{sim}}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{p(\hat{\mathbf{y}}|\mathbf{x})}[1 - \textsc{SimiLe}(\mathbf{y}, \hat{\mathbf{y}})]$$

$$\triangleq \sum_{\hat{\mathbf{y}} \in \mathcal{H}(\mathbf{x})} (1 - \textsc{SimiLe}(\mathbf{y}, \hat{\mathbf{y}})) \frac{p(\hat{\mathbf{y}}|\mathbf{x})}{\sum_{\hat{\mathbf{y}}' \in \mathcal{H}(\mathbf{x})} p(\hat{\mathbf{y}}'|\mathbf{x})}$$

SAGE

End-to-end training with adversarial loss using Gumbel-Softmax

- To enable end-to-end training, we adopt the Gumbel-Softmax [Jang et al., 2017]

$$p(y_i) = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_j^{|\mathcal{V}|} \exp((\log(\pi_j) + g_j)/\tau)}$$

Experiments

Baselines

- Local $\underset{1 \le i \le |\mathbf{y}|, \hat{\mathbf{y}}_i \in \mathcal{V}}{\arg\min} \; [\hat{\mathbf{y}}_i - \mathbf{y}_i]^T \nabla_{\mathbf{y}_i} \mathcal{L}_{\text{adv}}(\hat{\mathbf{y}}, \mathbf{L}_{\text{true}}(\mathbf{y}), \mathbf{t})$
  - Unconstrained: Search within the whole embedding space
  - kNN: search within 10 nearest neighbors of the original token embedding
  - CharSwap: Swap or add a character to the original token to change it to <unk>
- Seq2seq-based
  - Seq2seq without delexicalisation
  - Seq2seq
  - Wseq (ours)
  - Wseq-S (ours)

Results

Automatic evaluation

		Semantic validity				Flip rate		Fluency
		BLEU	METEOR	SIMLE	Ecr (%)	Qfr (%)	Afr (%)	Perplexity
	Original Questions	-	-	-	-	-	-	816
Local	Unconstrained	79.26	51.93	87.35	100	49.46	41.23	1596
Local	kNN	80.39	56.03	93.30	100	23.80	18.23	1106
Local	CharSwap	80.76	53.91	90.51	100	26.10	22.09	2658
Seq2seq-based	Seq2seq w/o delex	32.69	35.77	80.09	68.97	12.62	11.25	515
Seq2seq-based	Seq2seq	34.91	37.58	82.79	99.38	8.98	6.69	561
Seq2seq-based	**Wseq (ours)**	33.72	37.70	82.18	98.91	8.37	6.91	**474**
Seq2seq-based	**Wseq-S (ours)**	**36.05**	**37.94**	**84.32**	**99.46**	7.76	6.14	610
Seq2seq-based	**SAGE (ours)**	33.54	36.35	82.38	99.11	**17.61**	**14.46**	710

- Entity coverage rate $\mathrm{Ecr} = \frac{v}{m}$; v = |generated questions with all required entities|, m = |all generated questions|
- Query flip rate $\mathrm{Qfr} = \frac{q}{m}$; Answer flip rate $\mathrm{Afr} = \frac{a}{m}$

Results

Human evaluation

- Randomly sample 100 questions from the WikiSQL test set
- Three native expert annotators to annotate the generated adversarial questions
  - Semantic validity
    - Whether they use the *same* columns & rows in table for the *same* answer
  - Fluency
    - Rank questions including the original one in terms of *fluency* and *naturalness*

	Validity (%)↑	**Fluency (rank)↓**
Original Questions	-	2.2
Unconstrained	20.3	4.39
kNN	64.0	3.39
Seq2seq w/o delex	78.7	2.99
Seq2seq	89.3	2.56
Wseq (ours)	88.7	**2.42**
Wseq-S (ours)	**90.3**	2.61
SAGE (ours)	78.7†	2.71‡

†: Significant compared to kNN ($p < 0.01$).
‡: Significant compared to kNN ($p < 0.01$) and Seq2seq w/o delex ($p < 0.05$).

Results

Qualitative analysis

	Question	SQL	H
Semantic Validity	What is the sum of wins after 1999 ? (**Original**)	`SELECT` SUM(Wins) `WHERE` Year $>$ 1999	-
	What is the sum of wins **downs** 1999 ? (**Unconstrained**)	✓	N
	What is the sum of wins after 1999 **is** (**kNN**)	`SELECT` Wins `WHERE` Year $>$ 1999	N
	How many wins in the years after 1999 ? (**Seq2seq**)	✓	Y
	What is the total wins for the year after 1999 ? (**Wseq**)	✓	Y
	What is the sum of wins in the year later than 1999 ? (**Wseq-S**)	`SELECT` COUNT(Wins) `WHERE` YEAR $>$ 1999	Y
	How many wins have a year later than 1999 ? (**SAGE**)	`SELECT` COUNT(Wins) `WHERE` YEAR $>$ 1999	Y
Fluency	What was the date when the opponent was at South Carolina ? (**Original**)	`SELECT` Date `WHERE` Opponent $=$ at South Carolina	3.0
	What was the date when the **jord** was at South Carolina ? (**Unconstrained**)	✓	5.3
	What was the date when the opponent was at South Carolina **,** (**kNN**)	✓	4.0
	What date was the opponent at South carolina ? (**Seq2seq**)	✓	1.3
	What is the date of the game against at South Carolina ? (**Wseq**)	✓	4.7
	What is the date of the opponent at South Carolina ? (**Wseq-S**)	✓	4.0
	On what date was the opponent at South Carolina ? (**SAGE**)	✓	1.7

Adversarial Training with SAGE

Test performance

- Target systems
  - SQLova-B with BERT Base encoder
  - SQLova-L with BERT Large encoder
1. Train **SAGE-B** and **SAGE-L** for both systems
2. Generate adversarial questions on WikiSQL training set **AdvData-B** and **AdvData-L**
3. Retrain two SQLova-B models with
   - WikiSQL training set + AdvData-B
   - WikiSQL training set + AdvData-L
4. Evaluate the two SQLova-B models on WikiSQL test set

	AdvData-B		**AdvData-L**	
	Q-Acc	A-Acc	Q-Acc	A-Acc
Before Aug.	79.0	84.5	79.0	84.5
+30k	**79.5**	85.2	79.3	85.0
+56k	79.4	**85.5**	**79.6**	**85.3**

Adversarial Training with SAGE

Robustness

- We attack the retrained two SQLova-B models with different methods

	Before Aug.		AdvData-B		AdvData-L	
Attack model	Qfr	Afr	Qfr	Afr	Qfr	Afr
Unconstrained	53.97	46.07	53.46	45.15	**51.01**	**43.26**
kNN	27.36	21.85	**25.29**	**19.83**	25.57	20.51
SAGE	16.55	12.31	**10.30**	**8.09**	14.21	12.19

Conclusion

- We proposed SAGE, the first sequence-level model for white-box adversarial attack on TableQA systems
  - Wasserstein Seq2seq model
  - Delexicalization and semantic similarity regularization
  - Adversarial loss with Gumbel-Softmax
- SAGE is effective in consolidating semantic validity and fluency while maintaining high flip rate of generated adversarial questions
- Generated adversarial questions have been demonstrated to improve TableQA systems' performance and robustness

Thank you! Questions?

Contact: yz568@cam.ac.uk