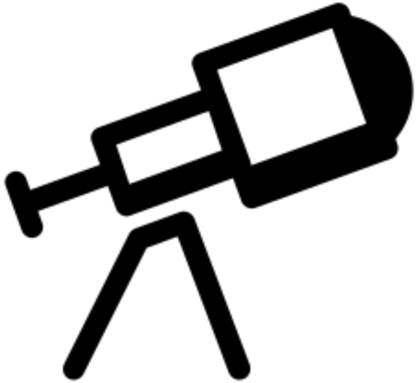
TypeSQL: From Natural Text to SQL Queries

DataBase Systems 2020

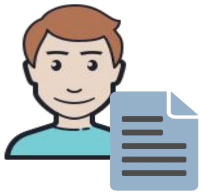
Apostolos Papatheodorou

Slides Structure

- *TypeSQL introduction*
- *Implementation*
- *Evaluation & Plan*



Presentation: Instead of Introduction 1/2



Question for a relation

Corresponding Sql query

Type Sql

a) ***Slot filling approach***
b) ***Knowledge based &***
Type aware

i.e. Assign each query word a type(Column, Number, KB entity)

c) ***Search DB rows for better accuracy (+9% improvement)***

Presenatation: Basic idea (2/2)

What are Slots ???

SELECT *$AGGR* *$SEL_COL*

WHERE *$COND_COL $OP $COND_VAL*

(AND *$COND_COL $OP $COND_VAL***)***

Query:

Was J. Biden v. president of US in 2019 ?

(none) (person) (column) (country) (year)

Three steps Formula

1) *Preprocessing: type recognition*
2) *Bi-directional LSTM for*
   a) *Type-word encoding*
   b) *Column's name encoding*
3) *Slot Prediction (more LSTMs)*

Implementation: Encoding (1/3)

Two Bi-LSTMs for
a) Quest: word-type → *H-QT*
b) Column schema → *H-COL*

Output

Slot Predictions

SELECT $AGGR $SEL_COL
WHERE $COND_COL $OP $COND_VAL
ND $COND_COL $OP $COND_VAL)*

Implementation of

Three LSTMs for prediction

- $AGG
- $OP, $COND_VAL,
- $SEL_COL, $COND_COL, $COND#

(5 LSMTs all together)

$$\alpha_{\mathrm{QT/COL}} = \mathbf{softmax}(\mathrm{H}_{\mathrm{COL}} \mathrm{W}_{ct} \mathrm{H}_{\mathrm{QT}}^{\top})$$
$$\mathrm{H}_{\mathrm{QT/COL}} = \alpha_{\mathrm{QT/COL}} \mathrm{H}_{\mathrm{QT}}$$

Implementation: Preprocessing (2/3)

Query:

Was J. Biden v. president of US in 2019 ?

(none) (person) (column) (country) (year)

1) **Tokenization k-grams, k=[2-6] for type assignment.**

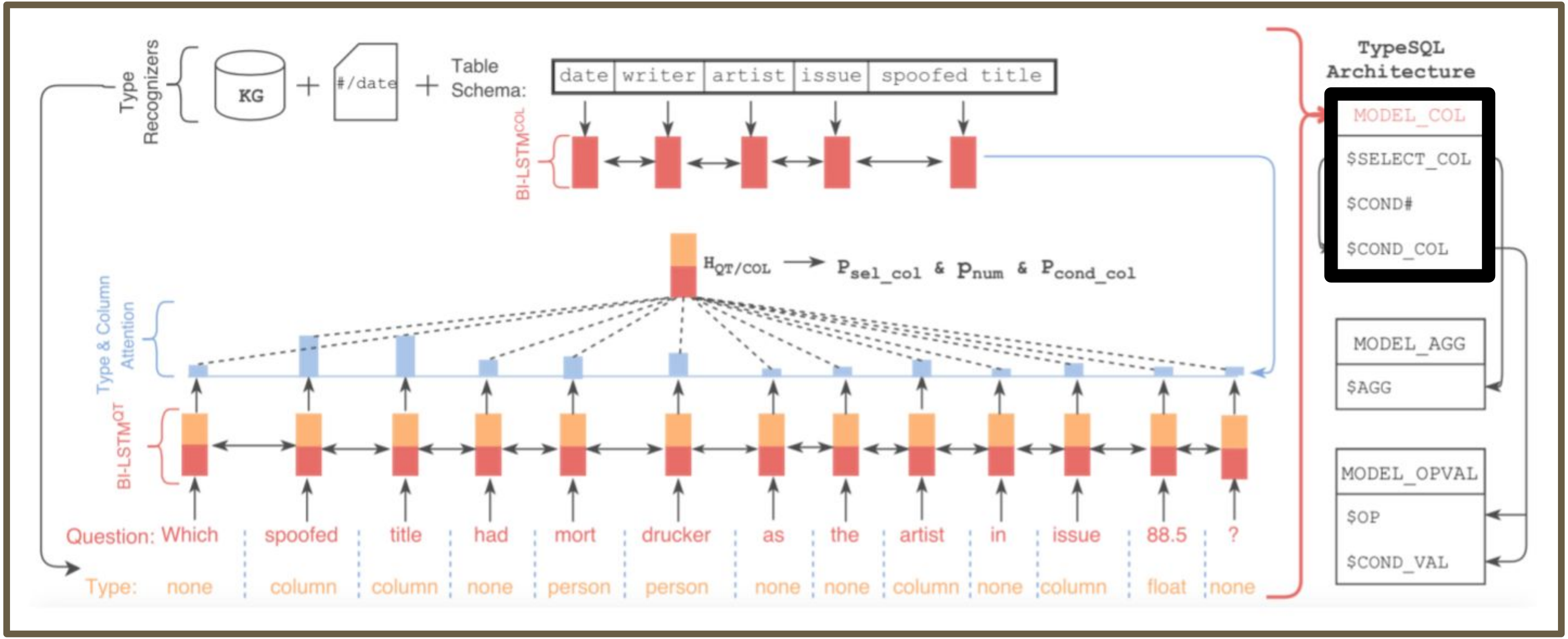
MyQuest: Random choice of k value?

Based on k-grams Search

a) ***Columns from schema***
b) ***Entities in Freebase KB***
c) ***numbers, dates, years***

D) IF RELATION AVAILABLE USE IT (+9% improvement)

Paper's paradigm



Implementation: Final Models(3/3)

MODEL_COL: (3 Slots)

$SELECT_COL
$COND#
$COND_Col

MODEL_COL- $SELECT_COL

$$s = \mathbf{V}^{sel}\mathbf{tanh}(\mathbf{W}_c^{sel}\mathbf{H}_{\mathrm{COL}}^\top + \mathbf{W}_{qt}^{sel}\mathbf{H}_{\mathrm{QT/COL}}^\top)$$
$$P_{sel\_col} = \mathbf{softmax}(s)$$

MODEL_COL- $SELECT_COND#

$$c = \mathbf{V}^{col}\mathbf{tanh}(\mathbf{W}_c^{col}\mathbf{H}_{\mathrm{COL}}^\top + \mathbf{W}_{qt}^{col}\mathbf{H}_{\mathrm{QT/COL}}^\top + \mathbf{W}_{qt}^{scol}\mathbf{H}_{\mathrm{QT/SCOL}}^\top)$$
$$P_{cond\_col} = \mathbf{softmax}(c)$$

Evaluation & Plan

	Dev			Test		
	Acc_{agg}	Acc_{sel}	Acc_{where}	Acc_{agg}	Acc_{sel}	Acc_{where}
Seq2SQL (Zhong et al., 2017)	90.0%	89.6%	62.1%	90.1%	88.9%	60.2%
SQLNet (Xu et al., 2017)	90.1%	91.5%	74.1%	90.3%	90.9%	71.9%
TypeSQL (ours)	90.3%	**93.1%**	**78.5%**	90.5%	**92.2%**	**77.8%**
TypeSQL+TC (ours)	90.3%	**93.5%**	**92.8%**	90.5%	**92.1%**	**87.9%**

Recap & Action Plan

For accomplishing the ultimate goal: **NLP text** to **SQL query**

Three LSTMs for prediction

Evaluation & Plan

	Dev			Test		
	Acc_{agg}	Acc_{sel}	Acc_{where}	Acc_{agg}	Acc_{sel}	Acc_{where}
Seq2SQL (Zhong et al., 2017)	90.0%	89.6%	62.1%	90.1%	88.9%	60.2%
SQLNet (Xu et al., 2017)	90.1%	91.5%	74.1%	90.3%	90.9%	71.9%
TypeSQL (ours)	90.3%	**93.1%**	**78.5%**	90.5%	**92.2%**	**77.8%**
TypeSQL+TC (ours)	90.3%	**93.5%**	**92.8%**	90.5%	**92.1%**	**87.9%**

Recap & Action Plan

For accomplishing the ultimate goal: **NLP text** to **SQL query**

Two Basic Competitors:

Seq2SQL & **SQLNet**

Three LSTMs for prediction

- ***$AGG***
- ***$OP, $COND_VAL,***
- ***$SEL_COL, $COND_COL, $COND#***

+2 LSMTs During preprocessing

Evaluation & Plan

	Dev			Test		
	Acc_{agg}	Acc_{sel}	Acc_{where}	Acc_{agg}	Acc_{sel}	Acc_{where}
Seq2SQL (Zhong et al., 2017)	90.0%	89.6%	62.1%	90.1%	88.9%	60.2%
SQLNet (Xu et al., 2017)	90.1%	91.5%	74.1%	90.3%	90.9%	71.9%
TypeSQL (ours)	90.3%	**93.1%**	**78.5%**	90.5%	**92.2%**	**77.8%**
TypeSQL+TC (ours)	90.3%	**93.5%**	**92.8%**	90.5%	**92.1%**	**87.9%**

Recap & Action Plan

For accomplishing the ultimate goal: **NLP text** to **SQL query**

Two Basic Competitors:

Seq2SQL & **SQLNet**

Implementation details

- WikiSQL Dataset
- **Pennington et al** & ***Wieting GImbel***
- Tuning: **dimens**, **dropout rate**, etc.
- **Adam** optimizer
- *My experiments on optimizers & tuning*

Thanks for watching