Text-to-SQL Semantic Parsing

Group 4
Nguyen Tan Viet
Ngo Quoc Bao

March 2, 2022

Problem

Definition of problem

Given a natural language Q and the schema $S = \langle \mathcal{T}, \mathcal{C} \rangle$ for a relational database, the parser needs to generate the corresponding SQL query Y. The schema consists of tables $\mathcal{T} = \{t_1, .., t_N\}$ and fields $\mathcal{C} = \{c_{11}, ..., c_{1|T_1|}, ..., c_{n1}, ..., c_{N|T_N|}\}$.Each table t_i and each field c_{ij} has a textual name. Some fields are primary keys, used for uniquely indexing each data record, and some are foreign keys, used to reference a primary key in a different table. In addition, each field has a data type, $\tau \in \{number, text, time, boolean, etc\}$

Literal review

New point in research

1. Model has acess to value of each field called picklists (e.g *Property_type_code* can have one of the following values $\{"Apartment","Field","House","Shop","Other"\}$)
   $\rightarrow$ protect individual data record and sensitive fields such as User IDs or credit numbers.
2. Use LSTM-based pointer generator with multihead-attention as decoder

New point in research

Question-schema serialization and encoding

1. $X = [CLS], Q, [SEP], [T], t_1, [C], c_{11}, ..., c_{1|T_1|}$, hspace24pt$[T], t_2, [C], c_{21}, ..., c_{N|T_N|}, [SEP]$
2. X is encoded with Bert
3. Folowed by a directional LSTM $\mathbf{h}_X \in \mathbb{R}^{|X|xn}$
4. The question $\mathbf{h}_X$ is passed through another bi-LSTM $\mathbf{h}_Q \in \mathbb{R}^{|Q|xn}$
5. Dense lookup features to represent meta-data of schema $f_{pri} \in \mathbb{R}^{2xn}$, $f_{for} \in \mathbb{R}^{2xn}$, $f_{type} \in \mathbb{R}^{|\tau|xn}$

New point in research

Question-schema serialization and encoding

Meta-data features and base encoding are fused via a feed-forward layer $g(\mathbb{R}^{4n} \rightarrow \mathbb{R}^{n})$:

1. $\mathbf{h}_S^{t_i} = g([\mathbf{h}_X^p; \mathbf{0}; \mathbf{0}; \mathbf{0}])$
2. $\mathbf{h}_S^{c_{ij}} = g([\mathbf{h}_X^q; f_{pri}^u, f_{for}^v, f_{type}^w])$
   $= ReLU(\mathbf{W}_g[\mathbf{h}_X^q; f_{pri}^u, f_{for}^v, f_{type}^w] + \mathbf{b}_g)$
3. $\mathbf{h}_S = [\mathbf{h}^{t_1}, ..., \mathbf{h}^{t_{|\tau|}}, \mathbf{h}^{c_{11}}, ..., \mathbf{h}^{c_{N|T_N|}}] \in \mathbb{R}^{|S|xn}$

New point in research

Bridging

Anchor text, perform fuzzy matching between question Q and the picklists of each field, if found the matched values in DB, the mathched field values are inserted in X, succeeding corresponding field name and seperated by token $[V]$.

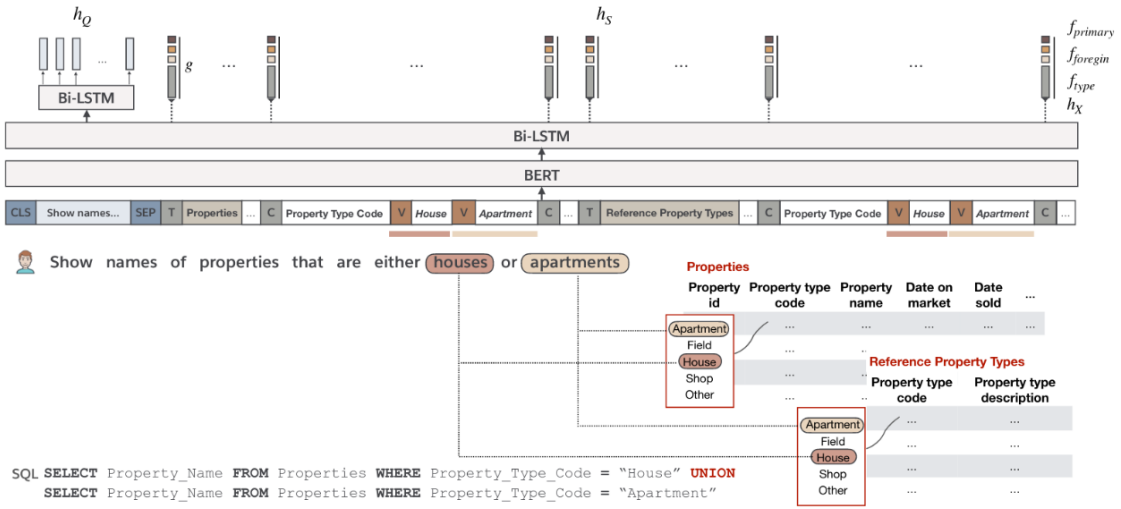

Figure 2: The BRIDGE encoder. The two phrases "houses" and "apartments" in the input question both matched to two DB fields. The matched values are appended to the corresponding field names in the hybrid sequence.

New point in research

Decoder

The decoder starts from the final state of the question encoder. At each step, the decoder performs one of the follow- ing actions: generating a token from the vocabulary V, copying a token from the question Q or copying a schema component from S. Mathematically, at each step t, given the decoder state s_t and the encoder representation $[h_Q; h_S] \in \mathbb{R}^{(|Q|+|S|)xn}$

$$e_{tj}^{(h)} = \frac{s_t W_U^{(h)} (h_j W_V^{(h)})^\top}{\sqrt{n/H}}; \alpha_{tj}^{(h)} = \underset{j}{softmax}\{e_{tj}^{(h)}\}$$

$$z_t^{(h)} = \sum_{j=1}^{|Q|+|S|} \alpha_{tj}^{(h)} (h_j W_V^{(h)}); z_t = [z_t^{(1)}, ..., z_t^{(H)}]$$

$$p_{gen}^t = sigmoid(s_t W_{gen}^s + z_t W_{gen}^z + b_{gen})$$

$$p_{out}^t = p_{gen}^t P_{\mathcal{V}}(y_t) + (1 - p_{gen}^t) \sum_{j:\tilde{x}_j = y_t} \alpha_{tj}^{(H)})$$

New point in research

Schema-Consistency Guided Decoding

a simple pruning strategy for sequence decoders, based on the fact that the DB fields ap- peared in each SQL clause must only come from the tables in the FROM clause.

Lemma 1 Let Y_{exec} be a SQL query with clauses arranged in execution order, then any table field in Y_{exec} must appear after the table.

As a result, we adopt a binary attention mask ξ

$\overset{\sim}{\alpha}_t^{(H)} = \alpha_t^{(H)}.\xi$

which initially has entries corresponding to all fields set to 0 Once a table t_i is decoded, all entries in ξ corresponding to that table to 1, allows the decoder to only search in the space specified by the condition in Lemma 1 with little overhead in decoding speed.

experiment result

ablation study

error analysis

Demo