Tracking Interaction States for Multi-Turn Text-to-SQL Semantic Parsing

Run-Ze Wang[1], **Zhen-Hua Ling**[*1], **Jing-Bo Zhou**[2], **Yu Hu**[3,4]
[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China
[2]Business Intelligence Lab, Baidu Research
[3]University of Science and Technology of China
[4]iFLYTEK Research
wrz94520@mail.ustc.edu.cn, zhling@ustc.edu.cn, zhoujingbo@baidu.com, yuhu@iflytek.com

Abstract

The task of multi-turn text-to-SQL semantic parsing aims to translate natural language utterances in an interaction into SQL queries in order to answer them using a database which normally contains multiple table schemas. Previous studies on this task usually utilized contextual information to enrich utterance representations and to further influence the decoding process. While they ignored to describe and track the interaction states which are determined by history SQL queries and are related with the intent of current utterance. In this paper, two kinds of interaction states are defined based on schema items and SQL keywords separately. A relational graph neural network and a non-linear layer are designed to update the representations of these two states respectively. The dynamic schema-state and SQL-state representations are then utilized to decode the SQL query corresponding to current utterance. Experimental results on the challenging CoSQL dataset demonstrate the effectiveness of our proposed method, which achieves better performance than other published methods on the task leaderboard.

Instruction

Querying relational databases and acquiring information from it have been studied extensively when designing natural language interfaces to databases(Zelle and Mooney 1996; Zhang et al. 2020). As such, a large body of research has focused on the task of translating natural language user utterances into SQL queries that existing database software can execute (Zhong, Xiong, and Socher 2017; Xu, Liu, and Song 2017; Yu et al. 2018a). While most of these works focus on precisely mapping stand-alone utterances to SQL queries (Krishnamurthy, Dasigi, and Gardner 2017; Dong and Lapata 2018), users often access information by multi-turn interactions in real-word applications. Therefor, designing a text-to-SQL system for multi-turn scenario has attracted more and more attention recently (Guo et al. 2019; Bogin, Gardner, and Berant 2019b; Yu et al. 2019c, 2018c).

The multi-turn text-to-SQl task is more complex and more challenging than traditional stand-alone text-to-SQL task. The main challenge is how to utilize the contextual information in history utterances to help system better parse current utterance. Figure 1 is an interaction example extracted from CoSQL dataset (Yu et al. 2019a). We can see from this figure that the phrase "those dorms" in U_2 refers to the phrase "all the dorms" in U_1. And the phrases "what dorms" and "those" in U_3 also refer to the same things as above. Besides, each predicted SQL query has obvious overlaps with history ones. Furthermore, each interaction corresponds to a database, called "schema", with several tables in it. The usage of table information in history queries also contributes to decode current query. For example, the SQL query Q_2 mentions the "dorm", "has_amenity" and "dorm_amenity" tables in the schema. Thus, the new clause "EXCEPT" in the SQL query Q_3 is more likely to select its contents from these tables.

U1: What are the names of all the dorms?
Q1: SELECT dorm_name FROM dorm

U2: Which of those dorms have a TV lounge?
Q2: SELECT T1.dorm_name FROM dorm AS T1 JOIN has_amenity AS T2 ON T1.dormid = T2.dormid JOIN dorm_amenity AS T3 ON T2.amenid = T3.amenid WHERE T3.amenity_name = 'TV Lounge'

U3: What dorms have no study rooms as amenities? | Do you mean among those with TV Lounges? | Yes.
Q3: SELECT T1.dorm_name FROM dorm AS T1 JOIN has_amenity AS T2 ON T1.dormid = T2.dormid JOIN dorm_amenity AS T3 ON T2.amenid = T3.amenid WHERE T3.amenity_name = 'TV Lounge' EXCEPT SELECT T1.dorm_name FROM dorm AS T1 JOIN has_amenity AS T2 ON T1.dormid = T2.dormid JOIN dorm_amenity AS T3 ON T2.amenid = T3.amenid WHERE T3.amenity_name = 'Study Room'
.....

Figure 1: An interaction example from CoSQL dataset. U_i is the user utterance at the i-th turn while Q_i is its corresponding SQL query. The blue words are SQL keywords. Because this paper only focuses on the text-to-SQL task of CoSQL, we concatenate those utterances that cannot be translated into SQL queries with their responds offered in the original CoSQL dataset using "|", such as U_3.

*Zhen-Hua Ling is corresponding author.

Previous studies on multi-turn text-to-SQL usually adopted encoder-decoder architectures. The encoder aims to represent user utterances and table schemas of databases as vectors, while the decoder is designed to generate corresponding SQL queries based on the representations given by the encoder. Suhr, Iyer, and Artzi (2018) designed a context-dependent model for ATIS (Airline Travel Information System) (Hemphill, Godfrey, and Doddington 1990; Dahl et al. 1994), which is a single-domain multi-turn text-to-SQL task. They utilized an LSTM-based discourse state to integrate sentence-level representations of history user utterances at the encoding stage. Zhang et al. (2019) improved this method on SParC (Yu et al. 2019c) and CoSQL (Yu et al. 2019a) datasets. They concatenated all history utterance tokens behind current utterance tokens and fed them into a BERT-based encoder together with schema items in order to generate user utterance and table schema representations. At the decoding stage, both of these two methods employed a copy mechanism to make use of the tokens or segments in previous predicted queries.

This paper studies the multi-turn text-to-SQL task by comparing it to another popular NLP task, task-oriented dialogue system. Both tasks aims at accomplish specific goals, i.e., querying databases or solving problems. The user utterances in multi-turn text-to-SQL are similar to the user inputs in dialog systems, and SQL queries correspond to dialogue responses. Besides, both tasks strongly rely on history information when making predictions. An essential component in task-oriented dialogue systems is dialog state tracking (DST) (Wen et al. 2017; Budzianowski et al. 2018; Mrksic et al. 2017), which keeps track of user requests or intentions throughout a dialogue in the form of a set of slot-value pairs, i.e. dialogue states. However, for the multi-turn text-to-SQL task, how to define, track and utilize user intentions throughout an interaction has not yet been investigated in previous work.

Therefore, we propose a method of tracking interaction states for multi-turn text-to-SQL semantic parsing in this paper. Considering the goal of text-to-SQL is to generate SQL queries for executing on table-based databases, two types of interaction states (i.e., name-value pairs) are defined based on schema tables and SQL keywords respectively. For schema-states, their names are the column names of all tables in the schema. Their values come from SQL keywords and are extracted from the last predicted SQL query. For example, after generating Q_2 in Figure 1, the value of the schema-state "dorm.dorm_name" is set as {"SELECT"}. After generating Q_3, its value is updated into a set of keywords {"SELECT", "=", "EXCEPT", "SELECT"}. For SQL-states, their names are all the SQL keywords. Their values are column names and are also determined by the last predicted SQL query. For example, after generating Q_2 in Figure 1, the value of the SQL-state "SELECT" is "dorm.dorm_name". This value is not changed after generating Q_3 because the "SELECT" keyword in the nested clause is also followed with "dorm.dorm_name" in Q_3.

In order to encode and utilize these designed interaction states, a model of text-to-SQL parsing with interaction state tracking (IST-SQL) is designed. In this model, each state is updated by a state updater based on last predicted query. A relational graph neural network (RGNN) is employed to encode the schema-state representations. A schema-column-graph is built based on the foreign keys in the database for implementing the RGNN. Besides, an one-layer non-linear layer is adopted to encode SQL-states. The BERT model is used at the beginning of IST-SQL to generate the embedding vectors of current utterance and all schema column names. The utterance embedding is further fed into an utterance encoder to integrate the information of all history utterances. Finally, the utterance representations, schema-state representations and SQL-state representations are fed into a decoder with copy mechanism to predict the SQL query tokens in order.

We evaluate our model on the CoSQL dataset, which is the largest and the most difficult dataset for conversational and multi-turn text-to-SQL semantic parsing. Experimental results show that our model improves the question-matching accuracy (QM) of the previous best model (Zhang et al. 2019) from 40.8% to 41.8% and the interaction matching accuracy (IM) from 13.7% to 15.2%, respectively.

The main contributions of this paper are twofolds. First, we propose to track interaction states for multi-turn text-to-SQL. Two types of interaction states are defined and are updated, encoded and utilized in our proposed IST-SQL model. Second, our proposed model performs better than the advanced benchmarks on the CoSQL dataset.

Related Work

Text-to-SQL Datasets and Methods

Text-to-SQL generation task is one of the most important tasks in semantic parsing, which aims to map natural language utterances into formal representations, such as logical forms (Zelle and Mooney 1996; Clarke et al. 2010), lambda calculus (Zettlemoyer and Collins 2005; Artzi and Zettlemoyer 2011) and executable programming languages (Miller et al. 1996; Yin and Neubig 2017).

At the earliest time, researchers focused on the converting single utterances while their corresponding SQL queries were simple and always produced in one domain, such as GeoQuery (Zelle and Mooney 1996) and Overnight (Wang, Berant, and Liang 2015). Recently, some large-scale datasets with open domains database have been released to attract attentions to the unseen-domain problem (Bogin, Gardner, and Berant 2019a; Dong and Lapata 2016, 2018; Bogin, Gardner, and Berant 2019b; Finegan-Dollak et al. 2018). WikiSQL (Zhong, Xiong, and Socher 2017) is a popular open domain text-to-SQL task but the SQL queries in it are still simple, just contains "SELECT", "WHERE" and "FROM" clauses. To study the complex SQL queries, Yu et al. (2018c) released a complex cross-domain text-to-SQL dataset names Spider, which contains most of the SQL clauses.

There are also some dataset that can support the studies on multi-turn text-to-SQL semantic parsing, such as ATIS (Hemphill, Godfrey, and Doddington 1990), SequentialQA (Iyyer, Yih, and Chang 2017), CoSQL (Yu et al. 2019a) and SParC (Yu et al. 2019b). SParC (Yu et al. 2019b) was

built based on Spider by dividing its single utterance into multiple ones and writing their corresponding SQL queries by annotators. CoSQL (Yu et al. 2019a) was released in 2019, which is the main evaluation dataset of this paper. Its user utterances were all collected from Wizard-of-Oz (WOZ) conversations and the corresponding SQL queries were written by annotators.

Previous methods on text-to-SQL semantic parsing always adopted encoder-decoder architectures. LSTM and BERT were popularly employed to obtain the representations of user utterances and database items (Xu, Liu, and Song 2017; Dong and Lapata 2018; Guo et al. 2019). Classification-based (Yu et al. 2018a; Xu, Liu, and Song 2017; Yu et al. 2018b), sequence-based (Zhang et al. 2019), and abstract-syntax-tree-based (Dong and Lapata 2018; Guo et al. 2019; Bogin, Gardner, and Berant 2019b) decoders have been developed for this task. In order to deal with the multi-turn scenario, previous studies usually integrated history utterance information into the representation of current utterance and introduced copy mechanism into the decoder for utilizing previous predicted SQL queries (Hemphill, Godfrey, and Doddington 1990; Zhang et al. 2019). In contrast, we propose to track interaction states for multi-turn text-to-SQL and define two type of interaction states to record history information in this paper.

Dialog State Tracking Task

Our proposed method is inspired by the dialog state tracking (DST) component in task-oriented dialogue systems(Henderson, Thomson, and Williams 2014; Gu et al. 2019; Gu, Ling, and Liu 2019). The goal of task-oriented dialogue systems is to help users accomplish a specific task such as hotel reservation, flight booking or travel information searching. (Wen et al. 2017; Budzianowski et al. 2018; Mrksic et al. 2017). Dialog state tracking is to records the dialog process and the user intention at each turn (Wang et al. 2020) in task-oriented dialogue systems. Here, dialogue states, i.e., a set of slot-value pairs, are usually pre-defined manually. Various methods have been proposed for dialog state tracking. For example, Ouyang et al. (2020) used a connection model to connect current states with previous states and copy previous values. Hu et al. (2020) designed an slot attention module and slot information sharing module for better utilizing the slot information.

In this paper, we bring the idea of dialog state tracking to the multi-turn text-to-SQL task. This task always studies open-domain datasets, which means that interaction states (similar to dialog states) should be grounded to different domains and can not be set universally for all interactions. In this paper, we define two types of interaction states, schema-states and SQL-states. Schema-states record domain-specific information and are changed with the grounded database. SQL-states are used to record the SQL information at each turn for better understanding the interaction process.

Preliminary

Dataset

We evaluated our model on CoSQL (Yu et al. 2019a), which is a large-scale conversational and multi-turn text-to-SQL semantic parsing dataset. An interaction example in CoSQL is shown in Figure 1. CoSQL consists of 30k+ turns plus 10k+ annotated SQL queries, obtained from a Wizard-of-Oz(WOZ) collection of 3k conversations querying 200 complex databases spanning 138 domain. It has been separated into 2164 interactions with 140 databases for training, 292 interactions with 20 databases for development and 551 interactions with 40 databases for testing.

We also evaluate our model on SParC (Yu et al. 2019b), which is another multi-turn text-to-SQL task. SParC consists of 3034 interactions with 140 databases for training, 422 interactions with 20 databases for development and 841 interactions with 40 databases for testing.

In both datasets, several interactions may be grounded to the same database, but each database can only appear in one of the training, development and testing sets for cross-domain evaluation.

Task Formulation

Let U denote a natural language user utterance and Q denote its corresponding SQL query. A multi-turn text-to-SQL task considers an interaction I, which consists of a sequence of (U_t, Q_t) pairs. Each interaction I is grounded to a database S, which consists of several tables T and each table T consists of several column names C. In our experiments, we concatenated each column name with its corresponding table name, thus the set of table names T can be omitted.

The goal of multi-turn text-to-SQL task is to generate $\{Q_t\}_{t=1}^{N(I)}$ in order given $\{U_t\}_{t=1}^{N(I)}$ and S as follows,

$$\{\{U_t\}_{t=1}^{N(I)}, S\} \xrightarrow{map} \{Q_t\}_{t=1}^{N(I)}. \quad (1)$$

The function $N(*)$ in this paper stands for the element number of $*$. In this paper, as we utilize the last predicted SQL query for extracting interaction states and for copying tokens when decoding, the goal at the t-th turn of this task is to generate the t-th SQL query Q_t given the current utterance U_t, all the history utterance $\{U_1, \cdots, U_{t-1}\}$, the table schema S and the last predicted query Q_{t-1}, i.e.,

$$\{U_t, S, \{U_1, \cdots, U_{t-1}\}, Q_{t-1}\} \xrightarrow{map} Q_t. \quad (2)$$

Proposed Method

As shown in Figure 2, our IST-SQL model is built based on the sequence-to-sequence encoder-decoder architecture. It consists of (1) a BERT embedder to embed current utterance tokens and schema column names into embedding vectors, (2) an utterance encoder to integrate history utterances into the representations of current utterance, (3) an interaction states tracking module with state updaters and state encoders, and (4) a decoder with copy mechanism to predicted SQL query tokens in order.

In our model, the BERT embedder is the same as one used in (Zhang et al. 2019) and more details can be found in that paper.

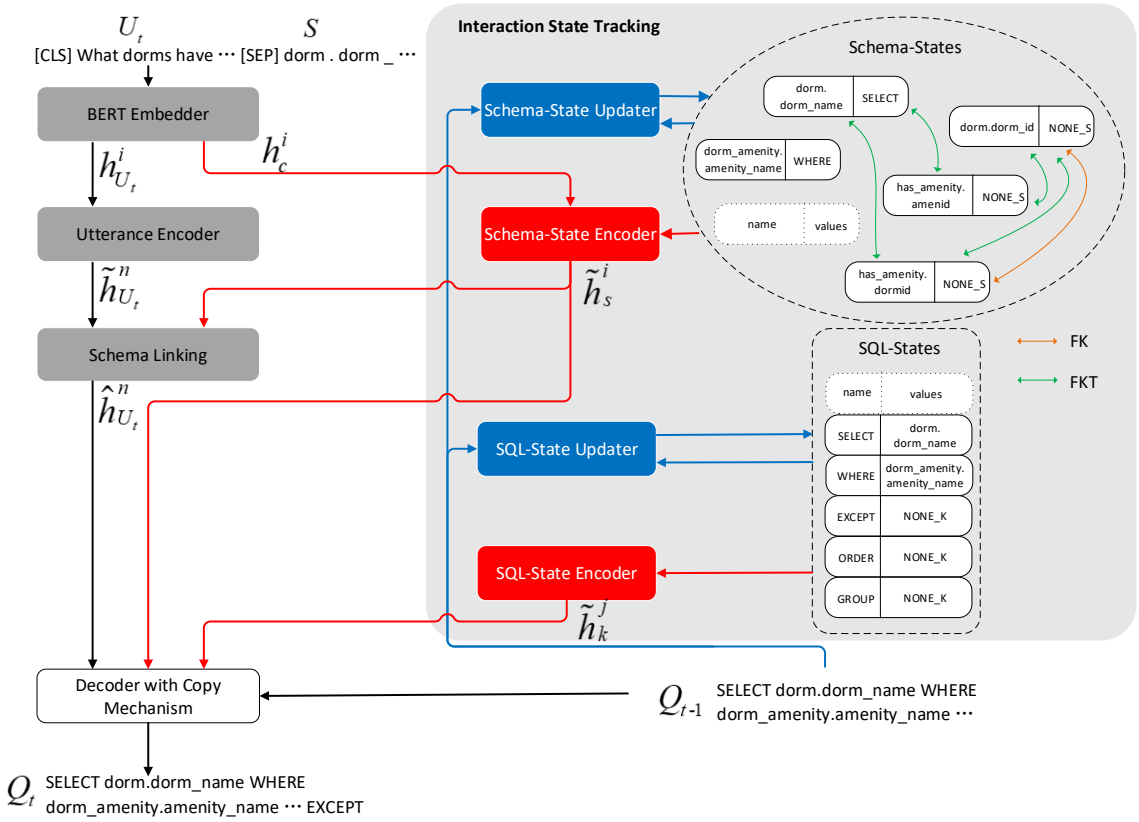

Figure 2: The architecture of proposed IST-SQL model. "FK" and "FKT" stand for Foreign-Key and Foreign-Key-Table relations between schema-states respectively.

Interaction State Tracking

State Updaters As introduced in the Introduction section, two kinds of interaction states are designed in this paper. For SQL-states, their names are SQL keywords and their values come from the schema column names corresponding to the interaction. For schema-states, their names are column names in the schema and their values are SQL keywords. As shown in Figure 2, the SQL-state updater and the schema-state updater extracted state values from the last previous predicted SQL query at each turn.

To extract the SQL-state values at each turn, we separate the last predicted SQL query with all the column names in it. Q_t and Q_{t-1} in Figure 2 correspond to Q_3 and Q_2 in Figure 1 respectively. The SQL-state updater first separates Q_{t-1} into a set of pieces {"SELECT dorm.dorm", "WHERE dorm_amenity.amenity_name"}[1]. If an SQL-state appears in one of these pieces, the column name appears at the end of this piece is added to the SQL-state as its value. As shown in Figure 2, the value of SQL-state "SELECT" is "dorm.dorm" and the value of "WHERE" state is "dorm_amenity.amenity_name". It should be noticed that an SQL-state can have multiple non-repetitive values. The values of those SQL-states that do not appear in the last predicted SQL query are set as a fixed token "NONE_K" .

In a dual way, if the name of a schema-state appears in one of the SQL-state values, the schema-state updater adds the name of this SQL-state to the schema-state as its value. In order to describe the nesting structure of SQL queries, the SQL keywords that occur more than once in the last query are all kept as schema-state values. In Figure 2, the value of schema-state "dorm.dorm" is "SELECT" while the value of "dorm_amenity.amenity_name" state is "WHERE". The values of those schema-states that do not appear in the last predicted SQL query are set as a fixed token "NONE_S" .

State Encoders The function of the schema-state encoder and the SQL-state encoder in Figure 2 is to convert the values of both states into embedding vectors at each turn.

Considering that the names of schema-states are column names in a relational table-based database and different column names are related to each other, a relational graph neural network (RGNN) is designed to represent such relations and to propagate node information to nearby nodes. Here, the RGNN model uses schema-states as nodes and builds edges according to their relations. Based on foreign keys [2], two kinds of relations between two schema-states

[1]Following previous methods, "FROM", "JOIN" and "ON" clauses are removed when model predicting because they can be easily filled with some simple rules after all the other clauses are predicted.

[2]Foreign key is a pre-defined relation between column names

are considered as follows.

- **Foreign-Key (FK)** If two column names are a foreign key pair in the database, we build an edge with "IN" and "OUT" directions between them, as shown by the yellow line and arrows between schema-states in Figure 2.
- **Foreign-Key-Table (FKT)** If two column names are in different tables and the two tables have one or more foreign key pairs in their columns, we build an edge with "IN" and "OUT" directions between them, as shown by the green lines and arrows between schema-states in Figure 2.

The representation of each node, i.e., schema-state, is initialed with the representations of its name and values. Because the schema-state names are all the schema column names, we directly use the column name representations $\{\boldsymbol{h}_c^i\}_{i=1}^{N(c)}$ generated by the BERT embedder in Figure 2, where $N(c)$ is the total number of schema columns. The value embedding vectors $\{\boldsymbol{h}_k^j\}_{j=1}^{N(k)}$ are initialed randomly, where $N(k)$ is the total number of SQL keywords. The initial representation of each schema-state is calculated as

$$\boldsymbol{h}_s^i = tanh(\boldsymbol{W_1}(\boldsymbol{h}_c^i + \sum_{j \in V_s^i} \boldsymbol{h}_k^j)), \tag{3}$$

where V_s^i is the value index set of the i-th schema-state, $\boldsymbol{W}_1 \in \mathbb{R}^{d \times d}$ is a trainable matrix and d is the hidden state dimension.

Then, $\{\boldsymbol{h}_s^i\}_{i=1}^{N(c)}$ are fed into the RGNN and the final representations are calculated as

$$\boldsymbol{h}_{out}^i = \sum_{e \in \{FK, FKT\}} \sum_{j \in \varepsilon_{out}^i(e)} \boldsymbol{W}_{out}(\boldsymbol{z}_{out}^e * \boldsymbol{h}_s^j), \tag{4}$$

$$\boldsymbol{h}_{in}^i = \sum_{e \in \{FK, FKT\}} \sum_{j \in \varepsilon_{in}^i(e)} \boldsymbol{W}_{in}(\boldsymbol{z}_{in}^e * \boldsymbol{h}_s^j), \tag{5}$$

$$\widetilde{\boldsymbol{h}}_s^i = \boldsymbol{h}_s^i + 0.5 * \boldsymbol{h}_{out}^i + 0.5 * \boldsymbol{h}_{in}^i, \tag{6}$$

where $\boldsymbol{W}_{out}$ and $\boldsymbol{W}_{in}$ are trainable matrices for "IN" and "OUT" directions, $\boldsymbol{z}_{out}^e$ and $\boldsymbol{z}_{in}^e$ stand for the "IN" and "OUT" embedding vectors of edge e, which are randomly initialized and are updated during training. $*$ stands for element-wise product. $\varepsilon_{in,out}^i(e)$ denotes the set of schema-state indices that have edge e connecting with state i. $\widetilde{\boldsymbol{h}}_s^i$ is the final schema-state representation for the i-th schema-state which is further used by schema linking and decoding.

Considering that the names of SQL-states are all SQL keywords without explicit relations, we simply build a network with one non-linear layer to derive their representations as

$$\widetilde{\boldsymbol{h}}_k^j = tanh(\boldsymbol{W_2}(\boldsymbol{h}_k^j + \sum_{i \in V_k^j} \boldsymbol{h}_s^i)), \tag{7}$$

where $\boldsymbol{h}_k^j$ is the representation of the j-th SQL-state name, V_k^j is the value index set of the j-th SQL-state, $\boldsymbol{W}_2 \in \mathbb{R}^{d \times d}$ is a trainable matrix. The final SQL-state representation $\widetilde{\boldsymbol{h}}_k^i$ is further fed into the decoder as Figure 2 shows.

and stands for that these two columns has the same meanings and values but are in different tables. More details can be found in (Yu et al. 2018c).

Utterance Encoder

Integrating history utterance information into the representation of current utterance is important for multi-turn text-to-SQL. Unlike previous methods that utilized history utterances at sentence-level, we construct a token-level utterance encoder based on multi-head attention mechanism to enrich utterance representations. Let $\boldsymbol{h}_{U_t}^n$ stand for the embedding vector given by the BERT embedder for the n-th token in utterance U_t. We collect the history information as

$$\boldsymbol{h}_{\{1,\cdots,t-1\}\to t}^n = \frac{1}{t-1} \sum_{m=1}^{t-1} \boldsymbol{h}_{m\to t}^n, \tag{8}$$

where

$$\boldsymbol{h}_{m\to t}^n = \sum_{k=1}^{K} \sum_{l=1}^{N(U_m)} \alpha_m^{nkl} \boldsymbol{h}_{U_m}^l, \tag{9}$$

$$\alpha_m^{nkl} = softmax(s_m^{nkl}), \tag{10}$$

$$s_m^{nkl} = (\boldsymbol{h}_{U_t}^n * \boldsymbol{a}_u^k) \cdot \boldsymbol{h}_{U_m}^l. \tag{11}$$

Here, $*$ stands for element-wise product and $\cdot$ stands for dot product, the $softmax$ function is executed over the index of l, $\boldsymbol{a}_u^k$ is a trainable vector and K is the head number. The embedding vectors of current utterance tokens after utterance encoder are calculated as

$$\widetilde{\boldsymbol{h}}_{U_t}^n = BiLSTM([\boldsymbol{h}_{U_t}^n; \boldsymbol{h}_{\{1\cdots t-1\}\to t}^n]), \tag{12}$$

and are further fed into the schema linking module.

Schema Linking

Similar to previous studies (Guo et al. 2019; Bogin, Gardner, and Berant 2019b), we also build an schema linking module, which integrates database information into utterance representations to deal with unseen-domain problem. Previous methods usually treated schema linking as a pre-processing step to generate some features linking utterance tokens and schema items. In our IST-SQL model, the schema linking module is similar to the utterance encoder. There are two main differences. First, $\{\boldsymbol{h}_{U_m}^l\}_{m=1}^{t-1}$ in Eq. (9) and (11) are replaced with schema-state representation $\widetilde{\boldsymbol{h}}_s^i$ for calculating its relevance with utterance representation $\widetilde{\boldsymbol{h}}_{U_t}^n$, and the averaging operation in Eq. (8) is not conducted. Second, considering $\widetilde{\boldsymbol{h}}_{U_t}^n$ and $\widetilde{\boldsymbol{h}}_s^i$ are embedded in different spaces, a transform vector $\boldsymbol{b}_s^k$ is applied to Eq. (9) for the k-th head as

$$\boldsymbol{h}_{S\to U_t}^n = \sum_{k=1}^{K} \sum_{l=1}^{N(S)} \boldsymbol{b}_s^k * \alpha_t^{nkl} \widetilde{\boldsymbol{h}}_s^n. \tag{13}$$

Following Eq. (12), a BiLSTM is built upon $\boldsymbol{h}_{S\to U_t}^n$ to obtain the final utterance representations $\hat{\boldsymbol{h}}_{U_t}^n$ which is sent into the decoder.

Decoder with Copy Mechanism

Our decoder is similar with that proposed by Zhang et al. (2019). The representations of current utterance, schema-states, SQL-states and last predicted SQL query are used as the inputs of the decoder. It should be noticed that all

representations of schema column names and SQL keywords used in (Zhang et al. 2019) are replaced by the schema-state representations and SQL-state representations in our model. Moreover, when calculating the scores of SQL keywords, we measure the similarity between the decoder hidden states and SQL-states directly without introducing additional module like the MLP used by Zhang et al. (2019).

Experiments

Implementation Details

All hidden states in our proposed IST-SQL model had 300 dimensions except the BERT embedder with 768 hidden dimensions. The head number K was set as 3 heuristically. When training model parameters, in addition to the average token level cross-entropy loss for all the SQL tokens in an interaction, regularization terms were added to encourage the diversity of the multi-head attentions used in the utterance encoder and schema linking modules.

The model was implemented using PyTorch (Paszke et al. 2017). We used ADAM optimizer (Kingma and Ba 2014) to minimize the loss function. The BERT embedder was initialized with a pre-trained small uncased BERT model[3]. All the other parameters were randomly initialized from a uniform distribution between [-0.1, 0.1]. The BERT embedder was fine-tuned with learning rate of $1e-5$ while the other parameters was trained with learning rate of $1e-3$. An early stop mechanism was used with patient number 10 on the development set. The best model was selected based on the token-level string matching accuracy on the development set. The golden last SQL query was fed into our interaction states tracking module at the training stage. The final IST-SQL model has 448M parameters while the baseline model Edit-Net (Zhang et al. 2019) has 468M parameters. All code is published to help replicate our results[4].

Metrics

Two metrics, question match accuracy (QM) and interaction match accuracy (IM) (Yu et al. 2019b), were used in our evaluation. QM is the percentage of the queries corresponding to all evaluated questions that are correctly predicted. While IM is the percentage of interactions with all queries correctly predicted. In order to avoid ordering issues, instead of using simple string matching on each predicted query, we followed the method proposed by Yu et al. (2018c) which decomposed predicted queries into different SQL components such as $SELECT$, $WHERE$, $GROUPBY$, and $ORDERBY$, and computed accuracy for each component using set match separately.

Overall Results

We compared our IST-SQL model with three baseline models, SyntxtSQL-con (Yu et al. 2019c), CD-Seq2Seq (Yu et al. 2019c) and Edit-Net (Zhang et al. 2019). Because the test set of CoSQL is not publicly available and there are limitations on the times of online submissions, we only evaluated our model which achieved the best QM performance on the development set. The results are shown in Table 1. We can see that our IST-SQL model achieved the best development and test performance among all the models on both QM and IM. Comparing with the state-of-the-art method Edit-Net, our model improved its QM from 39.9% to 44.4% and IM from 13.7% to 15.2%. Furthermore, our methods achieves the second place on the leaderboard of CoSQL[5], while the first place method has not been published yet.

Model	QM (%)		IM (%)	
	Dev	Test	Dev	Test
SyntxtSQL-con (Yu et al. 2019c)	15.1	14.1	2.7	2.2
CD-Seq2Seq (Yu et al. 2019c)	13.8	13.9	2.1	2.6
Edit-Net (Zhang et al. 2019)	39.9	40.8	12.3	13.7
IST-SQL (Ours)	**44.4**	**41.8**	**14.7**	**15.2**

Table 1: Results of different models on CoSQL. Due to submission limitations, we only report the results of our model which achieved the best QM performance on the development set. The results of other models are copied from (Yu et al. 2019a) and (Zhang et al. 2019).

Model	QM (%)	IM (%)
Edit-Net (Zhang et al. 2019)	45.35 ± 0.62	26.60 ± 0.70
IST-SQL (Ours)	$\mathbf{47.55 \pm 0.80}$	$\mathbf{29.93 \pm 0.93}$

Table 2: Results of different models on SParC development set. Each model was trained four times. The mean and standard deviation of each metric are reported.

We also compared our model with Edit-Net on the SParC dataset. Due to time limitation, we haven't received the online test set results till paper submission. Thus, we only report the QM and IM results on the development set. Each model was trained four times, and the mean and standard deviation are shown in Figure 2. We can find that our model improved the Edit-Net performance from 45.35% to 47.55% on QM and from 26.6% to 29.93% on IM, while both of the improvements were larger than their standard deviations.

Ablation Study

We further investigated the effects of proposed schema-states and SQL-states in our model. Two ablation experiments were performed by removing each type of interaction states from the full model. When removing schema-states, all schema-state representations used in downstream modules were replaced with the representations of schema column names generated by the BERT embedder. When removing SQL-states, all SQL-state representations used in downstream modules were replaced with the embedding vectors estimated for all SQL keywords. For better model comparison, we trained each model four times and calculated the mean and standard deviation of its QM and IM.

The results are shown in Table 4. When removing schema-states from the model, the performance of QM dropped from 43.05% to 40.88% while the IM dropped from 15.33% to 12.6%. When removing SQL-states, the

[3]https://github.com/google-research/bert

[4]https://github.com/runzewang/IST-SQL

[5]https://yale-lily.github.io/cosql

		SELECT	WHERE	GROUP	ORDER	AND/OR	IUEN
Clause Number		1004	573	128	165	20	19
F1(%)	Edit-Net	73.6	57.6	40.7	64.5	96.7	14.8
	IST-SQL	73.7	60.7	51.2	66.0	97.2	15.6

Table 3: The development set F1 (%) results of two models on different SQL clauses.

Model	**QM** (%)	**IM** (%)
Ours	**43.05 ± 0.85**	**15.33 ± 0.51**
w/o schema-states	40.88 ± 0.83	12.60 ± 0.79
w/o SQL-states	41.73 ± 0.89	13.70 ± 1.20

Table 4: Results of ablation studies on the two types of interaction states in our model. Each model was trained four times and was evaluated on the development set.

		Easy	Medium	Hard	Extra-Hard
Utterance Number		415	320	162	107
QM(%)	Edit-Net	62.9	33.1	22.2	10.3
	IST-SQL	66.0	36.2	27.8	10.3

Table 5: The development set QM (%) results of two models on utterances with different difficulty levels.

performance on OM dropped from 43.5% to 41.73% while the IM dropped from 15.33% to 13.7%. Furthermore, all performance degradations were larger than their corresponding standard deviations. These results indicate that both schema-states and SQL-states contributed to achieve the overall performance of our IST-SQL model.

Analysis

In order to better understand the advantages of our IST-SQL model, three analysis experiments were further conducted to compare IST-SQL model with the baseline Edit-Net model. We re-trained the Edit-Net model four times and chose the best one with 41.2% QM and 13.7% IM on the development set, which was better than the model shared by its authors.

Difficulty Level We first evaluated the two models, IST-SQL and Edit-Net, on the utterances with different difficulty levels. Here, the difficulty levels were determined based on the components of target SQL queries, and the decision rules followed previous work (Yu et al. 2018c). The results are shown in Table 5. We can see that our IST-SQL model outperformed Edit-Net on "Easy", "Medium" and "Hard" levels, while obtained the same QM results on "Extra-Hard" level whose target SQL queries are most complex and usually contain nesting structures.

Turn Index Then, we evaluated the two models on the utterances with different turn indices. We split all the interactions in the development set and regrouped them based on their turn indices in interactions for evaluation. The results are shown in Table 6. We can find that our model achieved better performance than Edit-Net on utterances with all turn indices. For our IST-SQL model, the QM of the utterances at the fourth turn was comparable with that of the utterances at the second turn. These results demonstrate the ability of our model on dealing with deep turns by tracking the interaction process with schema-states and SQL-states.

		1	2	3	4	>4
Utterance Number		292	283	244	114	71
QM(%)	Edit-Net	52.1	39.6	37.3	36.0	25.4
	IST-SQL	56.2	41.0	41.0	41.2	26.8

Table 6: The development set QM (%) results of two models on utterances with different turn indices.

SQL Clause As last, we evaluated the development set performance of these two models on different SQL clauses. The true SQL queries and predicted SQL queries were separated into several clauses with SQL keywords. Each clause had a value set that was organized following previous study (Yu et al. 2018c). For each clause, we calculated its F1 score of predicted SQL queries according to true SQL queries. The results are shown in Table 3. We can find that our model outperformed Edit-Net on all SQL keyword clauses. The "AND/OR" and "SELECT" clauses were the two easiest ones and the advantages of IST-SQL over Edit-Net were slight on these two clauses. "GROUP" and "WHERE" were the two clauses that achieved the largest improvements. One reason is that these two clauses usually contain more schema column names than other clauses such as "ORDER" and "AND/OR". Another reason is that these two clauses usually appear at later turns than the clauses such as "SELECT". Therefore, these two clauses can benefit most from tracking interaction states in the IST-SQL.

Conclusion

In this paper, we have proposed an IST-SQL model to deal with the multi-turn text-to-SQL task. Inspired by the dialogue state tracking component in task-oriented dialogue systems, this model defines two types of interaction states, schema-states and SQL-states, and their values are updated according to last predicted SQL query at each turn. Furthermore, a relational graph neural network is employed to calculate schema-state representations while a non-linear layer is adopted to obtain SQL-state representations. These state representations are combined with utterance representations for decoding the SQL query at current turn. Our evaluation results on CoSQL and SParC datasets have demonstrated the effectiveness of the IST-SQL model. To improve current definitions and tracking modules of interaction states for conversational text-to-SQL with response generation will be a task of our future work.

References

Artzi, Y.; and Zettlemoyer, L. 2011. Bootstrapping semantic parsers from conversations. In *Proceedings of the conference on empirical methods in natural language processing*, 421–432. Association for Computational Linguistics.

Bogin, B.; Gardner, M.; and Berant, J. 2019a. Global Reasoning over Database Structures for Text-to-SQL Parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3650–3655.

Bogin, B.; Gardner, M.; and Berant, J. 2019b. Representing Schema Structure with Graph Neural Networks for Text-to-SQL Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics1*.

Budzianowski, P.; Wen, T.; Tseng, B.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gasic, M. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 5016–5026. Association for Computational Linguistics.

Clarke, J.; Goldwasser, D.; Chang, M.-W.; and Roth, D. 2010. Driving semantic parsing from the world's response. In *Proceedings of the fourteenth conference on computational natural language learning*, 18–27. Association for Computational Linguistics.

Dahl, D. A.; Bates, M.; Brown, M.; Fisher, W.; Hunicke-Smith, K.; Pallett, D.; Pao, C.; Rudnicky, A.; and Shriberg, E. 1994. Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Proceedings of the workshop on Human Language Technology*, 43–48. Association for Computational Linguistics.

Dong, L.; and Lapata, M. 2016. Language to Logical Form with Neural Attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 33–43.

Dong, L.; and Lapata, M. 2018. Coarse-to-Fine Decoding for Neural Semantic Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 731–742.

Finegan-Dollak, C.; Kummerfeld, J. K.; Zhang, L.; Ramanathan, K.; Sadasivam, S.; Zhang, R.; and Radev, D. 2018. Improving Text-to-SQL Evaluation Methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 351–360.

Gu, J.-C.; Ling, Z.-H.; and Liu, Q. 2019. Interactive matching network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2321–2324.

Gu, J.-C.; Ling, Z.-H.; Zhu, X.; and Liu, Q. 2019. Dually Interactive Matching Network for Personalized Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1845–1854.

Guo, J.; Zhan, Z.; Gao, Y.; Xiao, Y.; Lou, J.-G.; Liu, T.; and Zhang, D. 2019. Towards Complex Text-to-SQL in Cross-Domain Database with Intermediate Representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4524C4535.

Hemphill, C. T.; Godfrey, J. J.; and Doddington, G. R. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Henderson, M.; Thomson, B.; and Williams, J. D. 2014. The Second Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*, 263–272. The Association for Computer Linguistics.

Hu, J.; Yang, Y.; Chen, C.; He, L.; and Yu, Z. 2020. SAS: Dialogue State Tracking via Slot Attention and Slot Information Sharing. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 6366–6375. Association for Computational Linguistics.

Iyyer, M.; Yih, W.-t.; and Chang, M.-W. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1821–1831.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Krishnamurthy, J.; Dasigi, P.; and Gardner, M. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1516–1526.

Miller, S.; Stallard, D.; Bobrow, R.; and Schwartz, R. 1996. A fully statistical approach to natural language interfaces. In *34th Annual Meeting of the Association for Computational Linguistics*, 55–61.

Mrksic, N.; Séaghdha, D. Ó.; Wen, T.; Thomson, B.; and Young, S. J. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1777–1788. Association for Computational Linguistics.

Ouyang, Y.; Chen, M.; Dai, X.; Zhao, Y.; Huang, S.; and Chen, J. 2020. Dialogue State Tracking with Explicit Slot Connection Modeling. In Jurafsky, D.; Chai, J.; Schluter,

N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 34–40. Association for Computational Linguistics.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch .

Suhr, A.; Iyer, S.; and Artzi, Y. 2018. Learning to Map Context-Dependent Sentences to Executable Formal Queries. In *Proceedings of NAACL-HLT*, 2238–2249.

Wang, K.; Tian, J.; Wang, R.; Quan, X.; and Yu, J. 2020. Multi-Domain Dialogue Acts and Response Co-Generation. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 7125–7134. Association for Computational Linguistics.

Wang, Y.; Berant, J.; and Liang, P. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1332–1342.

Wen, T.; Vandyke, D.; Mrksic, N.; Gasic, M.; Rojas-Barahona, L. M.; Su, P.; Ultes, S.; and Young, S. J. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In Lapata, M.; Blunsom, P.; and Koller, A., eds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, 438–449. Association for Computational Linguistics.

Xu, X.; Liu, C.; and Song, D. 2017. Sqlnet: Generating structured queries from natural language without reinforcement learning. *arXiv preprint arXiv:1711.04436* .

Yin, P.; and Neubig, G. 2017. A Syntactic Neural Model for General-Purpose Code Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 440–450.

Yu, T.; Li, Z.; Zhang, Z.; Zhang, R.; and Radev, D. 2018a. TypeSQL: Knowledge-based Type-Aware Neural Text-to-SQL Generation. In *Proceedings of NAACL-HLT*, 588–594.

Yu, T.; Yasunaga, M.; Yang, K.; Zhang, R.; Wang, D.; Li, Z.; and Radev, D. R. 2018b. SyntaxSQLNet: Syntax Tree Networks for Complex and Cross-Domain Text-to-SQL Task. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 1653–1663. Association for Computational Linguistics.

Yu, T.; Zhang, R.; Er, H.; Li, S.; Xue, E.; Pang, B.; Lin, X. V.; Tan, Y. C.; Shi, T.; Li, Z.; et al. 2019a. CoSQL: A Conversational Text-to-SQL Challenge Towards Cross-Domain Natural Language Interfaces to Databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1962–1979.

Yu, T.; Zhang, R.; Yang, K.; Yasunaga, M.; Wang, D.; Li, Z.; Ma, J.; Li, I.; Yao, Q.; Roman, S.; et al. 2018c. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3911–3921.

Yu, T.; Zhang, R.; Yasunaga, M.; Tan, Y. C.; Lin, X. V.; Li, S.; Er, H.; Li, I.; Pang, B.; Chen, T.; et al. 2019b. SParC: Cross-Domain Semantic Parsing in Context. *arXiv preprint arXiv:1906.02285* .

Yu, T.; Zhang, R.; Yasunaga, M.; Tan, Y. C.; Lin, X. V.; Li, S.; Heyang Er, I. L.; Pang, B.; Chen, T.; Ji, E.; Dixit, S.; Proctor, D.; Shim, S.; Jonathan Kraft, V. Z.; Xiong, C.; Socher, R.; and Radev, D. 2019c. SParC: Cross-Domain Semantic Parsing in Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Zelle, J. M.; and Mooney, R. J. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the thirteenth national conference on Artificial intelligence-Volume 2*, 1050–1055.

Zettlemoyer, L. S.; and Collins, M. 2005. Learning to map sentences to logical form: structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, 658–666.

Zhang, L.; Wang, R.; Zhou, J.; Yu, J.; Ling, Z.; and Xiong, H. 2020. Joint Intent Detection and Entity Linking on Spatial Domain Queries. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 4937–4947.

Zhang, R.; Yu, T.; Er, H.; Shim, S.; Xue, E.; Lin, X. V.; Shi, T.; Xiong, C.; Socher, R.; and Radev, D. 2019. Editing-Based SQL Query Generation for Cross-Domain Context-Dependent Questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5341–5352.

Zhong, V.; Xiong, C.; and Socher, R. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. *arXiv preprint arXiv:1709.00103* .