CoSQL: A Conversational Text-to-SQL Challenge Towards Cross-Domain Natural Language Interfaces to Databases

Tao Yu† Rui Zhang† He Yang Er† Suyi Li† Eric Xue† Bo Pang† Xi Victoria Lin‡ Yi Chern Tan† Tianze Shi‡ Zihan Li‡ Youxuan Jiang‡ Michihiro Yasunaga† Sungrok Shim† Tao Chen† Alexander Fabbri† Zifan Li† Luyao Chen† Yuwen Zhang‡ Shreya Dixit† Vincent Zhang‡ Caiming Xiong‡ Richard Socher‡ Walter S. Lasecki† Dragomir Radev†
†Yale University ‡Salesforce Research §University of Michigan ¶Cornell University
{tao.yu, r.zhang, dragomir.radev}@yale.edu
{xilin, cxiong, rsocher}@salesforce.com

Abstract

We present CoSQL, a corpus for building cross-domain, general-purpose database (DB) querying dialogue systems. It consists of 30k+ turns plus 10k+ annotated SQL queries, obtained from a Wizard-of-Oz (WOZ) collection of 3k dialogues querying 200 complex DBs spanning 138 domains. Each dialogue simulates a real-world DB query scenario with a crowd worker as a user exploring the DB and a SQL expert retrieving answers with SQL, clarifying ambiguous questions, or otherwise informing of unanswerable questions. When user questions are answerable by SQL, the expert describes the SQL and execution results to the user, hence maintaining a natural interaction flow. CoSQL introduces new challenges compared to existing task-oriented dialogue datasets: (1) the dialogue states are grounded in SQL, a domain-independent executable representation, instead of domain-specific slot-value pairs, and (2) because testing is done on unseen databases, success requires generalizing to new domains. CoSQL includes three tasks: SQL-grounded dialogue state tracking, response generation from query results, and user dialogue act prediction. We evaluate a set of strong baselines for each task and show that CoSQL presents significant challenges for future research. The dataset, baselines, and leaderboard will be released at https://yale-lily.github.io/cosql.

1 Introduction

Natural language interfaces to databases (NLIDB) have been studied extensively, with a multitude of different approaches introduced over the past few decades. To this end, considerable progress has been made in querying data via natural language (NL). However, most NL query systems expect the query to be well-formed and stated in a single sentence (Zelle and Mooney, 1996; Li and Jagadish, 2014; Yaghmazadeh et al., 2017; Iyer et al., 2017; Zhong et al., 2017; Xu et al., 2017; Shi et al., 2018;
Wang et al., 2018; Yu et al., 2018b,c). In reality, complex questions are usually answered through interactive exchanges (Figure 1). Even for simple queries, people tend to explore the database by asking multiple basic, interrelated questions (Hale, 2006; Levy, 2008; Frank, 2013; Iyyer et al., 2017). This requires systems capable of sequentially processing conversational requests to access information in relational databases. To drive the progress of building a context-dependent NL query system, corpora such as ATIS (Hemphill et al., 1990; Dahl et al., 1994) and SParC (Yu et al., 2019) have been released. However, these corpora assume all user questions can be mapped into SQL queries and do not include system responses. Furthermore, in many cases, multi-turn interaction between users and NL systems is needed to clarify ambiguous questions (e.g., \(Q_3\) and \(R_3\) in Figure 1), verify returned results, and notify users of unanswerable or unrelated questions. Therefore, a robust dialogue-based NL query agent that can engage with users by forming its own responses has become an increasingly necessary component for the query process. Such systems have already been studied under task-oriented dialogue settings by virtue of continuous effort of corpus creation (Seneff and Polifroni, 2000; Walker et al., 2002; Raux et al., 2005; Mrksic et al., 2015; Asri et al., 2017; Budzianowski et al., 2018) and modelling innovation (Artzi and Zettlemoyer, 2011; Henderson et al., 2013; Lee and Dernoncourt, 2016; hao Su et al., 2016; Dhingra et al., 2016; Li et al., 2016; Mrksic et al., 2017). The goal of these systems is to help users accomplish a specific task, such as flight or hotel booking or transportation planning. However, to achieve these goals, task-oriented dialogue systems rely on pre-defined slots and values for request processing (which can be represented using simple SQL queries consisting of \texttt{SELECT} and \texttt{WHERE} clauses). Thus, these systems only operate on a small number of domains and have difficulty capturing the diverse semantics of practical user questions.

In contrast, the goal of dialogue-based NLIDB systems is to support general-purpose exploration and querying of databases by end users. To do so, these systems must possess the ability to (1) detect questions answerable by SQL, (2) ground user questions into executable SQL queries if possible, (3) return results to the user in a way that is easily understood and verifiable, and (4) handle unanswerable questions. The difficulty of constructing dialogue-based NLIDB systems stems from these requirements. To enable modeling advances in this field, we introduce CoSQL, the first large-scale cross-domain Conversational text-to-SQL corpus collected under the WOZ setting (Budzianowski et al., 2018). CoSQL contains 3,007 dialogues (more than 30k turns with annotated dialogue acts and 10k expert-labeled SQL queries) querying 200 complex DBs spanning across 138 different domains. For each dialogue, we follow the WOZ setup that involves a crowd worker as a DB user and a college computer science student who is familiar with SQL as an expert (§3).

Like Spider\(^2\) (Yu et al., 2018c) and SParC (Yu et al., 2019), the cross-domain setting in CoSQL enables us to test the ability of systems to generalize on querying different domains via dialogues. We split the dataset in a way that each database only appears in one of train, development, or test set. This setting requires systems to generalize to new domains without additional annotation.

More importantly, unlike most prior work in text-to-SQL systems, CoSQL demonstrates greater language diversity and more frequent user focus changes. It also includes a significant amount of questions that require user clarification and questions that cannot be mapped to SQL queries, introducing the potential to evaluate text-to-SQL dialog act prediction. These features pose new challenges for text-to-SQL systems. Moreover, CoSQL includes system responses that describe SQL queries and the returned results in a way that is easy for users with different backgrounds to understand and verify, as faithful and comprehensible presentation of query results is a crucial component of any NLIDB system.\(^3\)

We introduce three challenge tasks on CoSQL: (1) \textit{SQL-grounded dialogue state tracking} to map user utterances into SQL queries if possible given the interaction history (§5.1), (2) \textit{natural language response generation} based on an executed SQL and its results for user verification

\(^2\)Spider task is available at \url{https://yale-lily.github.io/spider}

\(^3\)The DB community has developed query visualization (Li and Jagadish, 2014) and other techniques to provide faithful explanation of a SQL query. These explanations are complementary to the NL ones used in our work and future NLIDB systems could integrate them.
and (3) **user dialogue act prediction** to detect and resolve ambiguous and unanswerable questions (§5.3). We provide detailed data analysis and qualitative examples (§4). For each of the three tasks, we benchmark several competitive baseline models (§6). The performances of these models indicate plenty of room for improvement.

2 Related Work

**Text-to-SQL generation** Text-to-SQL generation has been studied for decades in both DB and NLP communities. (Warren and Pereira, 1982; Zettlemoyer and Collins, 2005; Popescu et al., 2003; Li et al., 2006; Li and Jagadish, 2014; Iyer et al., 2017; Zhong et al., 2017; Xu et al., 2017; Yu et al., 2018a; Dong and Lapata, 2018; Finegan-Dollak et al., 2018; Guo et al., 2019; Bogin et al., 2019). However, the majority of previous work focuses on converting a single, complex question into its corresponding SQL query. Only a few datasets have been constructed for the purpose of mapping context-dependent questions to structured queries. Price (1990); Dahl et al. (1994) collected ATIS that includes series of questions from users interacting with a flight database. Yu et al. (2019) introduced SPArC, a large cross-domain semantic parsing in context dataset, consisting of 4k question sequences with 12k questions annotated with SQL queries over 200 complex DBs. Similar to ATIS, SPArC includes sequences of questions instead of conversational interactions. An NL question and its corresponding SQL annotation in SPArC are constructed by the same expert. Recent works (Suhr et al., 2018; Zhang et al., 2019) have built context-dependent text-to-SQL systems on top of these datasets.

In contrast, CoSQL was collected under a WOZ setting involving interactions between two parties, which contributes to its diverse semantics and discourse covering most types of conversational DB querying interactions (e.g. the system will ask for clarification of ambiguous questions, or inform the user of unanswerable and irrelevant questions). Also, CoSQL includes a natural language system response for the user to understand and verify the systems actions.

**Task-oriented dialog systems** Task-oriented dialog systems (Henderson et al., 2014; Wen et al., 2016; Mrkšić et al., 2017; Budzianowski et al., 2018) have attracted increasing attention especially due to their commercial values. The goal is to help users accomplish a specific task such as hotel reservation, flight booking, or travel information. These systems (Bordes and Weston, 2017; Zhong et al., 2018; Wu et al., 2019) often pre-define slot templates grounded to domain-specific ontology, limiting the ability to generalize to unseen domains. In comparison, our work is to build a system for general-purpose DB exploration and querying. The domain-independent intent representation (SQL query) enables the trained system to work on unseen domains (DB schemas).

While most task-oriented dialog systems need to actively poke the user for information to fill in pre-defined slot-value pairs, the primary goal of system responses in CoSQL is to offer users a reliable way to understand and verify the returned results. If a question can be converted into a SQL query, the user is shown the execution result and the system will describe the SQL query and the result in natural language. In case the user questions are ambiguous or unanswerable by SQL, the system either requests the user to rephrase or informs them to ask other questions.

**Data-to-Text generation** Response generation in CoSQL takes a structured SQL query and its corresponding result table to generate an NL description of the system’s interpretation of the user request. Compared to most dialogue-act-to-text generation tasks, the richer semantics of SQL queries makes our task more challenging – besides generating natural and coherent descriptions, faithfully preserving the logic of a SQL query in an NL response is also crucial in our task. Furthermore, this component is related to previous work on text generation from structured data (McKeown, 1985; Iyer et al., 2016; Wiseman et al., 2017).

3 Data Collection

We follow the Wizard-of-Oz setup which facilitates dialogues between DB users and SQL experts to create CoSQL. We recruited Amazon Mechanical Turkers (AMT) to act as DB users and trained 25 graduate- and undergraduate-level computer science students proficient in SQL to act as DB experts. The collection interface (Lasecki et al., 2013) is designed to be easy-to-operate for the experts and intuitive for the users. Detailed explanations of the data collection process is provided below.
Reference goal selection  We pre-select a reference goal for each dialogue to ensure the interaction is meaningful and to reduce redundancy within the dataset. Users are asked to explore the given DB content to come up with questions that are likely to naturally arise in real-life scenarios and reflect their query intentions as specified by the reference goals. Following Yu et al. (2019), we selected the complex questions classified as medium, hard, and extra hard in Spider (Yu et al., 2018c) as the reference goals.4 In total, 3,783 questions were selected on 200 databases. After annotation and reviewing, 3,007 of them were finished and kept in the final dataset.

User setup  We developed online chatting interfaces to pair the user with the expert (Figure 6 and 7 in Appendix). When a data collection session starts, the user is first shown multiple tables from a DB to which a reference goal is grounded and is required to read through them. Once they have examined the data stored in the tables, the reference goal question will be revealed on the same screen. The user is encouraged to use the goal question as a guide to ask interrelated questions, but is also allowed to ask other questions exploring the DB. We require the user to ask at least 3 questions.5 In each turn, if the user question can be answered by a SQL query, they will be shown the result table, and the expert will write an NL response interpreting the executed SQL query based on their understanding of the user’s query intent (Figure 8 Appendix). If the user question is ambiguous or cannot be answered with SQL, they will receive clarification questions or notice to rephrase from the expert (detailed in expert setup).

Expert setup  Within each session, the expert is shown the same DB content and the reference goal as the user (Figure 8 in Appendix). For each dialogue turn, the expert first checks the user question and labels it using a set of pre-defined user dialog action types (DATs, see Table 4). Then the expert sets the DAT of his response according to the user DAT. Both the user and the expert can have multiple DATs labels in each turn. If the user question is answerable in SQL (labeled as INFORM_SQL, e.g. \( Q_1 \) in Figure 1), the expert writes down the SQL query, executes it, checks the result table, and sends the result table to the user. The expert then describes the SQL query and result table in natural language and sends the response. If the user question is ambiguous, the expert needs to write an appropriate response to clarify the ambiguity (labeled as AMBIGUOUS, e.g. \( Q_3 \) in Figure 1). Some user questions require the expert to infer the answer based on their world knowledge (labeled as INFER_SQL, e.g. \( Q_3 \) in Figure 10). If the user question cannot be answered by SQL, the expert will inform them to ask well-formed questions (labeled as NOTRELATED, CANNOTUNDERSTAND, or CANNOTANSWER). In other cases (labeled as GREETING, THANKYOU, etc.), the expert responds with general dialogue expressions (\( Q_8 \) in Figure 1).

User quality control  Because of the real-time dialogue setting and the expensive annotation procedures on the expert side, conducting quality control on user is crucial for our data collection. We use LegionTools7 (Lasecki et al., 2014) to post our tasks onto AMT and to recruit and route AMT workers for synchronous real time crowd sourcing tasks. We specify that only workers from the U.S. with 95% approval rates are allowed to accept our task. Before proceeding to the chat room, each AMT worker has to go through a tutorial and pass two short questions8 to test their knowledge about our task. Only the user who passes the quiz proceeds to the chat room. Throughout the data collection, if a user ignores our instructions in a specific turn, we allow the experts to alert the user through chat and label the corresponding turn as DROP. If a user’s actions continue to deviate from instructions, the expert can terminate the dialog before it ends. After each dialogue session terminates, we ask the expert to provide a score from 1 to 5 as an evaluation of the user’s performance. Dialogues with a score below 3 are dropped and the user will be blocked from future participation.

4Yu et al. (2019) also includes 12.9% of the easy questions in Spider in order to increase dataset diversity. In this work we prioritize the complex questions that trigger more interesting interactions and do not include any easy questions.
5The worker is paid $1.20 USD for each dialog. To encourage interaction, we offer $0.50 USD bonus for each dialogue if the user asks more than 4 interesting, interrelated questions.
6We use the same SQL annotation protocol as Spider (Yu et al., 2018c) to ensure the same SQL pattern was chosen when multiple equivalent queries were available.
7https://www.cromalab.net/LegionTools/
8One is on how to ask interrelated questions and the other is on how to read multiple tables with reference keys.
Data review and post-process We conduct a multi-pass data reviewing process. Two student conducted a first-round review. They focus on correcting any errors in the DATs of the users and the experts, checking if the SQL queries match the user’s questions, and modifying or rewriting the expert’s responses to contain necessary information in the SQL queries in case they miss any of them. Also, they re-evaluate all dialogues based on the diversity of user questions and reject any dialogues that only contain repeated, simple, and thematically-independent user questions (about 6% of the dialogues). After the first-round review, another two student experts reviewed the refined data to double check the correctness of the DATs, the SQL queries, and the expert responses. They also corrected any grammar errors, and rephrased the user’s questions and the expert’s responses in a more natural way if necessary. Finally, we ran and parsed all annotated SQL queries to make sure they were executable, following the same annotation protocol as the Spider dataset.

4 Data Statistics and Analysis

We report the statistics of CoSQL and compare it to other task-oriented dialog and context-dependent text-to-SQL datasets. We also conduct detailed analyses on its contextual, cross-domain nature, and question diversity.

Data statistics Table 1 and 2 summarize the statistics of CoSQL. CoSQL contains 3k+ dialogues in total (2,164 in training), which is comparable to or bigger than most commonly used task-oriented dialogue datasets. Figure 2 shows the distribution of dialogue length in the corpus, approximately 80% of dialogues involve 8 or more turns, with a total of 31,148 turns. The average number of tokens in each turn is 11.21. Noticeably, the domain of CoSQL spans over 200 complex databases, overshadowing most other task-oriented dialogue datasets. Comparing to existing context-dependent text-to-SQL datasets, CoSQL contains significantly more turns, out of which 11,039 user utterances are convertible to SQL. In contrast, all NL utterances in ATIS and SPARC can be mapped to SQL. CoSQL also has a much larger NL vocabulary.

Dialogue act distribution As shown in Figure 3, CoSQL contains a fairly diverse set of user dialogue action types (DATs). Unsurprisingly, INFORM_SQL and THANK_YOU are the two most commonly seen DATs. Among the rest of DATs, approximately 40% are AMBIGUOUS, demonstrating the paramount importance of system clarification in the DB querying process in general. Another 20% of this subgroup is INFER_SQL, which signifies questions that cannot be answered without the aid of human inference.

Semantic complexity As shown in Table 1, if we consider the column names of the 200 DBs of CoSQL as slots and their entries as values, the number of slot-value pairs far exceed those defined in other task-oriented dialogues. Figure 4 shows the total number of occurrences of different SQL keywords in the SQL queries corresponding to these questions. The SQL queries in CoSQL cover all common SQL keywords as well as complicated syntactic structure such as nesting.

The review interface is shown in Figure 9 (Appendix).
Table 1: Comparison of CoSQL to some commonly used task-oriented dialogue datasets. The numbers are computed for the training part of data in consistency with previous work (Budzianowski et al., 2018).

<table>
<thead>
<tr>
<th></th>
<th>DSTC2</th>
<th>WOZ 2.0</th>
<th>KVRET</th>
<th>MultiWOZ</th>
<th>CoSQL</th>
</tr>
</thead>
<tbody>
<tr>
<td># dialogs</td>
<td>1,612</td>
<td>600</td>
<td>2,425</td>
<td>8,438</td>
<td>2,164</td>
</tr>
<tr>
<td>Total # turns</td>
<td>23,354</td>
<td>4,472</td>
<td>12,732</td>
<td>115,424</td>
<td>22,422</td>
</tr>
<tr>
<td>Total # tokens</td>
<td>199,431</td>
<td>50,264</td>
<td>102,077</td>
<td>1,520,970</td>
<td>22,8197</td>
</tr>
<tr>
<td>Avg. # turns/dialog</td>
<td>14.49</td>
<td>7.45</td>
<td>5.25</td>
<td>13.68</td>
<td>10.36</td>
</tr>
<tr>
<td>Avg. # tokens/turn</td>
<td>8.54</td>
<td>11.24</td>
<td>8.02</td>
<td>13.18</td>
<td>11.34</td>
</tr>
<tr>
<td>Total # unique tokens</td>
<td>986</td>
<td>2,142</td>
<td>2,842</td>
<td>24,071</td>
<td>7,502</td>
</tr>
<tr>
<td># databases</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>140</td>
</tr>
<tr>
<td># Slots #</td>
<td>8</td>
<td>4</td>
<td>13</td>
<td>25</td>
<td>3,696</td>
</tr>
<tr>
<td># Values #</td>
<td>212</td>
<td>99</td>
<td>1,363</td>
<td>4,510</td>
<td>&gt;1,000,000</td>
</tr>
</tbody>
</table>

Table 2: Comparison of CoSQL with other context-dependent text-to-SQL datasets. The number are computed over the entire datasets. *For CoSQL we count the total # user utterances.

<table>
<thead>
<tr>
<th></th>
<th>CoSQL</th>
<th>SParC</th>
<th>ATIS</th>
</tr>
</thead>
<tbody>
<tr>
<td># Q sequence</td>
<td>3,007</td>
<td>4298</td>
<td>1658</td>
</tr>
<tr>
<td># user questions</td>
<td>15,598*</td>
<td>12,726</td>
<td>11,653</td>
</tr>
<tr>
<td># databases</td>
<td>200</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td># tables</td>
<td>9,585</td>
<td>3794</td>
<td>1582</td>
</tr>
<tr>
<td>Avg. Q len</td>
<td>11.2</td>
<td>8.1</td>
<td>10.2</td>
</tr>
<tr>
<td>Avg. # Q turns</td>
<td>5.2</td>
<td>3.0</td>
<td>7.0</td>
</tr>
<tr>
<td>Unanswerable Q</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>User intent</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>System response</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

Figure 4: SQL keyword counts.

Semantic changes by turns We compute the frequency of occurrences of common SQL keywords in different turns for both CoSQL and SParC and compare them in Figure 5 (upper: CoSQL, lower: SParC). Here we count the turn # based on user utterance only. Since CoSQL and SParC span the same domains, Figure 5 reveals a comparison of semantic changes between context-dependent DB questions issued by end users (CoSQL) and expert users (SParC). For CoSQL, the frequencies of all keywords except for WHERE do not change significantly throughout the conversation, and the average frequencies of these keywords are in general lower than those of SParC. In addition, WHERE occurs slightly more frequently in CoSQL than in SParC. We believe this indicates the exploratory nature of the dialogues we collected, as the users switch their focus more frequently instead of building questions upon previous ones. For example, SQL AGG components occur most frequently in the beginning of dialogues, as a result of users familiarizing themselves with the amount of data in the DB or other statistical measures. In contrast, the frequencies of almost all SQL components in SParC increase as the question turn increases. This sug-
gests that questions in SPARC have stronger interdependency, as the purpose of this corpus is to study text-to-SQL in context.

**Cross domain** As shown in Table 3, the dialogues in CoSQL are randomly split into train, development and test sets by DB with a ratio of 7:1:2 (the same split as SPARC and Spider).

<table>
<thead>
<tr>
<th># Dialogs</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># Databases</td>
<td>140</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>

Table 3: Dataset Split Statistics

### 5 Tasks and Models

CoSQL is meant to be used as the first benchmark for building general-purpose DB querying dialogue systems in arbitrary domains. Such systems take a user question and determine if it can be answered by SQL (user dialogue act prediction). If the question can be answered by SQL, the system translates it into the corresponding SQL query (SQL-grounded dialogue state tracking), executes the query, returns and shows the result to the user. To improve interpretability and trustworthiness of the result, the system describes the predicted SQL query and result tables to the user for their verification (response generation from SQL and result). Finally, the user checks the results and the system responses and decides if the desired information is obtained or additional questions shall be asked.

Some components relevant to the process above are beyond the scope of our work. First, our response generation task only includes turns where the system’s dialogue act is **CONFORM_SQL**. In case the system cannot understand the user’s question (the system dialogue act is **CANNOT_ANSWER**) or considers it as unanswerable (**CANNOT_ANSWER**), the system will reply in a standard way to inform the user that it needs clarification or cannot answer that question. The same applies to questions that require human inference (e.g., the system confirms with the user which types of dorms he or she was talking about by asking **R3** instead of immediately translating **Q3** in Figure 1). Currently we do not have a task setup to evaluate the quality of system clarifications. Second, some user questions cannot be directly answered by SQL but are possible to be answered with other type of logical reasoning (e.g., **Q3** in Figure 1). We exclude these questions from our task design and leave them for future research.

#### 5.1 SQL-Grounded dialogue State Tracking

In CoSQL, user dialogue states are grounded in SQL queries. Dialogue state tracking (DST) in this case is to predict the correct SQL query for each user utterance with **INFORM_SQL** label given the interaction context and the DB schema. In our setup, the system does not have access to gold SQL queries from previous turns, which is different from the traditional DST settings in dialogue management where the history of ground-truth dialogue states is given. Comparing to other context-dependent text-to-SQL tasks such as SPARC and ATIS, the DST task in CoSQL also include the ambiguous questions if the user affirms the system clarification of them (e.g., **Q4** in Figure 1). In this case, the system clarification is also given as part of the interaction context to predict the SQL query corresponding to the question.\(^\text{11}\)\(^\text{12}\)

For instance, to generate **S4** in Figure 1, the input consists of all previous questions (**Q1**, **Q2**, **Q3**), the current user question (**Q4**), the DB schema, and the system response **R3**.

We benchmark the performance of two strong context-dependent neural text-to-SQL models on this the task, which are the baseline models reported on SPARC by Yu et al. (2019).

**Context-dependent Seq2Seq (CD-Seq2Seq)**

The model is originally introduced by (Suhr et al., 2018) for the ATIS task. It incorporates interaction history and is able to copy segments of previous generated SQL queries. Yu et al. (2019) extends it to encode DB schema information such that it works for the cross-domain setting in SPARC. We apply the model to our task without any changes.

**SyntaxSQL-con**

SyntaxSQLNet is a SQL-specific syntax-tree based model introduced for Spider (Yu et al., 2018b). Yu et al. (2019) extends it to take previous questions as input when predicting SQL for the current question. We apply the model to our task without any changes.

\(^\text{11}\)If a dialogue contains multiple ambiguous questions, the system clarification to all ambiguous questions will be given as input.

\(^\text{12}\)The ambiguous questions not confirmed by the user and their system responses are given as part of the conversation history but we do not require a system to predict SQL queries for them.
5.2 Response Generation from SQL and Query Results

This task requires generating a natural language description of the SQL query and the result for each system response labeled as CONFORM_SQL. It considers a SQL query, the execution result, and the DB schema. Preserving logical consistency between SQL and NL response is crucial in this task, in addition to naturalness and syntactic correctness. Unlike other SQL-to-text generation tasks (Xu et al., 2018), our task maps the SQL query to a statement and summarizes the result in that statement (instead of just mapping it back to the user question).

We experiment with three baseline methods for this task.

**Template-based** Given the SQL and NL response pairs in the training set, we masked variable values in both the SQL and NL response to form parallel SQL-response templates. Given a new SQL query, we employ rule-based approach to select the closest SQL-response template pair from the set. After that, we fill in the selected response template with the columns, tables, and values of the SQL query and the result to generate the final response (see more in Appendix).

**Seq2Seq** We experiment with a vanilla Seq2Seq model (Sutskever et al., 2014) with attention (Bahdanau et al., 2015), a standard baseline for text generation tasks.

**Pointer-generator** Oftentimes the column or table names in the NL response are copied from the input SQL query. To capture this phenomenon, we experiment with a pointer-generator network (See et al., 2017), which addresses the problem of out-of-vocabulary word generation in summarization and other text generation tasks. We use a modified version of the implementation from Chen and Bansal (2018).

5.3 User dialogue Act Prediction

For a real-world DB querying dialogue system, it has to decide if the user question can be mapped to a SQL query or if special actions are needed. We define a series of dialogue acts for the DB user and the SQL expert (Table 4). For example, if the user question can be answered by a SQL query, the dialogue act of the question is INFORM_SQL.

Since the system DATs are defined in response to the user DATs, our task does not include system dialogue acts prediction.

We experiment with two baseline models for this task.

**Majority** The dialogue acts of all the user questions are predicted to be the majority dialogue act INFORM_SQL.

**TBCNN-pair** We employ TBCNN-pair (Mou et al., 2016), a tree-based CNN model with heuristics for predicting entailment and contradiction between sentences. We change the two sentence inputs for the model to a user utterance and the DB schema, and follow the same method in SQL-Net (Xu et al., 2017) to encode each column name.

6 Results and Discussion

**SQL-grounded dialog state tracking** We use the same evaluation metrics used by the SPARC dataset (Yu et al., 2019) to evaluate the model’s performance on all questions and interactions (dialogs). The performances of CD-Seq2Seq and SyntaxSQL-con are reported in Table 5. The two models achieve less than 16% question-level accuracy and less than 3% on interaction-level accuracy. Since the two models have been benchmarked on both CoSQL and SPARC, we cross-compare their performance on these two datasets. Both models perform significantly worse on CoSQL DST than on SPARC. This indicates that CoSQL DST is more difficult than SPARC. The possible reasons is that the questions in CoSQL are generated by a more diverse pool of users (crowd workers instead of SQL experts), the task includes ambiguous questions and the context contains more complex intent switches.

**Response generation** Table 6 shows the results of three different baselines on three metrics: BLEU score (Papineni et al., 2002), logic correctness rate (LCR), and grammar. To compute LCR and grammar score, we randomly sampled 100 descriptions generated by each model. Three students proficient in English participated in the evaluation. They were asked to choose a score 0 or 1 for LCR, and 1 to 5 for grammar check (the larger, the better). For LCR, the final score was decided by majority vote. We computed the average grammar score.

\[^{13}\text{§A.1 defines the complete set of dialogue action types.}\]
Groups | Dialog acts
---|---
DB user | inform_sql, infer_sql, ambiguous, affirm, negate, not_related, cannot_understand, cannot_answer, greeting, goodbye, thank_you
DB expert | conform_sql, clarify, reject, request_more, greeting, sorry, welcome, goodbye

Table 4: Dialog acts in CoSQL. See § A.1 for the comprehensive definition of each dialogue act.

<table>
<thead>
<tr>
<th>Model</th>
<th>Question Match</th>
<th>Interaction Match</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>CD-Seq2Seq</td>
<td>13.8</td>
<td>13.9</td>
</tr>
<tr>
<td>SyntaxSQL-con</td>
<td>15.1</td>
<td>14.1</td>
</tr>
</tbody>
</table>

Table 5: Performance of various methods over all questions (question match) and all interactions (interaction match).

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>LCR (%)</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
<td>Test</td>
</tr>
<tr>
<td>Template</td>
<td>9.5</td>
<td>9.3</td>
<td>41.0</td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>15.3</td>
<td>14.1</td>
<td>27.0</td>
</tr>
<tr>
<td>Pointer-generator</td>
<td>16.4</td>
<td>15.1</td>
<td>35.0</td>
</tr>
</tbody>
</table>

Table 6: BLEU scores on the development and test sets, and human evaluations of logic correctness rate (LCR) and grammar check on the 100 examples randomly sampled from the test set.

Interestingly, the human evaluation and BLEU scores do not completely agree. While the template-based method is brittle and requires manual effort, it performs significantly better than the two end-to-end neural models in the human evaluation. Because the SQL-question templates provide natural and grammatical sketch of the output, it serves as an advantage in our human evaluation. However, this approach is limited by the small coverage of the training templates and its LCR is only around 40%. On the other hand, the neural models achieve better BLEU scores than the template-based approach. A possible reason for this is that they tend to generate words frequently associated with certain SQL queries. However, the neural models struggle to preserve the SQL query logic in the output. Unsurprisingly, pointer-generator performs better than basic Seq2Seq in terms of both BLEU and human evaluation. The low performances of all methods on LCR show that the task is indeed very challenging.

User dialog act prediction Table 7 shows the accuracy of the two baselines on predicting user dialog acts. The result of Majority indicates that about 40% of user questions cannot be directly converted into SQL queries. This confirms the necessity of considering a larger set of dialogue actions for building a practical NLIDB system. Even though TBCNN can predict around 85% of user intents correctly, most of the correct predictions are for simple classes such as inform_sql, thank_you, and goodbye etc. The F-scores for more interesting and important dialog acts such as infer_sql and ambiguous are around 10%. This indicates that improving the accuracy on user DAT prediction is still important.

7 Conclusion and Future Work

In this paper, we introduce CoSQL, the first large-scale cross-domain conversational text-to-SQL corpus collected under a Wizard-of-Oz setup. Its language and discourse diversity and cross-domain setting raise exciting open problems for future research. Especially, the baseline model performances on the three challenge tasks suggest plenty space for improvement. The data and challenge leaderboard will be publicly available at https://yale-lily.github.io/cosql.

Future Work As discussed in Section 5, some examples in CoSQL include ambiguous and unanswerable user questions and we do not study how a system can effectively clarify those questions or guide the user to ask questions that are answerable. Also, some user questions cannot be answered with SQL but by other forms of logical reasoning the correct answer can be derived. We urge the community to investigate these problems in future work in order to build practical, robust and reliable conversational natural language interfaces to databases.

8 Acknowledgement

This work was supported in part by IBM under the Sapphire Project at the University of Michigan.
References


Pei hao Su, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. Online active reward learning for policy optimisation in spoken dialogue systems. *ACL*.


Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Irene Li Heyang Er, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Vincent Zhang


### Appendices

This section provides description of dialog actions in A.1, more details on baseline modifications and hyperparameters in A.2, system response guides in A.3, additional dialog examples in CoSQL dataset in Figure 10 and 11, and the DB user (AMT turkers) and the SQL expert (college computer science students) annotation interfaces in Figure 6, 7, 8, and 9.

#### A.1 Description of Dialog Acts

For the DB user, we define the following dialog acts:

- **INFORM_SQL** The user informs his/her request if the users question can be answered by SQL. The system needs to write SQL.

- **INFER_SQL** If the users question must be answered by SQL+human inference. For example, users questions are are they...? (yes/no question) or the 3rd oldest.... SQL cannot directly (or unnecessarily complicated) return the answer, but we can infer the answer based on the SQL results.

---

**Appendix A**

This section provides description of dialog actions in A.1, more details on baseline modifications and hyperparameters in A.2, system response guides in A.3, additional dialog examples in CoSQL dataset in Figure 10 and 11, and the DB user (AMT turkers) and the SQL expert (college computer science students) annotation interfaces in Figure 6, 7, 8, and 9.
The users question is ambiguous, the system needs to double check the user’s intent (e.g. what/did you mean by...?) or ask for which columns to return.

Affirm: Affirm something said by the system (user says yes/agree).

Negate: Negate something said by the system (user says no/deny).

Not Related: The users question is not related to the database, the system reminds the user.

Cannot Understand: The users question cannot be understood by the system, the system asks the user to rephrase or paraphrase question.

Cannot Answer: The users question cannot be easily answered by SQL, the system tells the user its limitation.

Greeting: Greet the system.

Good Bye: Say goodbye to the system.

Thank You: Thank the system.

For the system, we define the following dialog acts:

Confirm SQL: The system creates a natural language response that describes SQL and result table, and asks the user to confirm if the system understood his/her intention.

Clarify: Ask the user to double check and clarify his/her intention when the users question is ambiguous.

Reject: Tell the user you did not understand/cannot answer his/her question, or the user question is not related.

Request More: Ask the user if he/she would like to ask for more info.

Greeting: Greet the system.

Sorry: Apologize to the user.

Welcome: Tell the user he/she is welcome.

Good Bye: Say goodbye to the user.

A.2 Modifications and Hyperparameters for Baselines

CD-Seq2Seq: We apply the model with the same settings used in SPARC without any changes.

SyntaxSQL-con: We apply the model with the same settings used in SPARC without any changes.

Template-based: We first create a list of SQL query patterns without values, column and table names that cover the most cases in the train set of CoSQL. And then we manually changed the patterns and their corresponding responses to make sure that table, column, and value slots in the responses have one-to-one map to the slots in the SQL query. Once we have the SQL-response mapping list, during the prediction, new SQL statements are compared with every templates to find the best template to use. A score will be computed to represent the similarity between the SQL and each template. The score is computed based on the number of each SQL key components existing in the SQL and each template. Components of the same types are grouped together to allow more flexible matching, like count, max, min are grouped to aggregate. A concrete example of templates is shown: SELECT column0 FROM table0 WHERE column1 comparison0 value0. column0,1 and table0 represent column name and table name respectively. comparison0 represents one of the comparison operator including >=, <=, <, >, =, !=, and like. value0 represents a value the user uses to constrain the query result.

Seq2Seq: We train a word2vec embedding model on the concatenation of the SQL query and response output of the training data for the embedding layer of our Seq2Seq model. We use an embedding dimension of 128, hidden dimension of 256, a single-layer bi-directional LSTM encoder and uni-directional LSTM decoder with attention. We use a batch size of 32, clip the norm of the gradient at 2.0, and do early stopping on the validation loss with a patience of 5. We perform decoding with greedy search.

Pointer-generator: We follow the same settings as in the Seq2Seq case with the addition of the copy mechanism during training and testing.

TBCNN-pair: The model is modified mainly on the sentence embedding part and classifier part. The input of the modified model is a user utterance and the related column names. Therefore, we replace one of the two sentence embedding modules with a database column name encoding module, which generates representations of the col-
umn names related to the sentence. The classifier is modified by adding a label(user dialogue act) number predicting module, which predicts the number of the labels(user dialogue acts) of the user utterance. The label number prediction module is similar to the column number prediction module in SQLNet.

A.3 System Response Guide

System response should be standard and professional. We follow the rules below to write responses for different system dialog action type:

**CLARIFY**  "Did you mean...?", "What did you mean by...?", or anything similar.

**REJECT**  "Sorry, I don’t have the answer to your question..." or anything similar.

**REQUEST_MORE**  "Do you want anything else?" or anything similar.

**CONFORM.SQL**  We convert SQL written by us back to natural language. (We should use the column names and values in the SQL). Our response has to describe all information in the SQL independently instead of referring to any previous context or subject.

1. If the returned result can be combined with the SQL description, combine them together to generate the response. For example:

   Given SQL:
   ```sql```
   SELECT AVG(SALARY) FROM INSTRUCTOR
   ```
   Result Returned: 200k
   Your Response: “The average salary of all instructors is 200k.”

2. If the returned result is too large and cannot be combined with the SQL description, describe them separately. For example:

   Given SQL:
   ```sql```
   SELECT AVG(T1.SALARY),T1.DEPARTMENT_ID FROM INSTRUCTOR AS T1 JOIN DEPARTMENT AS T2 ON T1.DEPARTMENT_ID = T2.ID GROUP BY T1.DEPARTMENT_ID
   ```
   Result Returned: a long table
   Your Response: “Here is the result table that shows the average salary in each department. For example, the average of CS professors is 250k.”
1. Read the tables below, which will be used by the assistant to answer your questions. Once complete, click "Next" to proceed.

**Table Name: Product_Suppliers**

<table>
<thead>
<tr>
<th>product_id</th>
<th>supplier_id</th>
<th>date_supplied_from</th>
<th>date_supplied_to</th>
<th>total_amount_purchased</th>
<th>total_value_purchased</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>2017-07-02 00:03:12</td>
<td>2018-03-31 07:30:49</td>
<td>56985.67</td>
<td>36974.66</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2017-10-16 16:10:37</td>
<td>2018-03-24 03:29:44</td>
<td>15752.45</td>
<td>1373.78</td>
</tr>
</tbody>
</table>

**Table Name: Order_Items**

<table>
<thead>
<tr>
<th>order_item_id</th>
<th>order_id</th>
<th>product_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

2. You are playing the role of a user who wants to know the answer to the question below. Remember:
   1. Ask at least 3 related questions about the data on the tables, you can refer to the question below.
   2. Good questions build up on previous questions. You can either break the given question down into small questions or ask other related questions.
   3. You'll get $0.5 bonus later if you follow the rules and ask more than 4 good related questions.

**QUESTION:** Return the ids of all products that were ordered more than three times or supplied more than 80000.

**Examples**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>How many female students are there?</td>
<td>Show me classes in the spring</td>
<td>Where is the location of the conference?</td>
<td>What is the grade for Eric?</td>
<td>Best college in CT?</td>
</tr>
<tr>
<td>What are their names?</td>
<td>How about in the summer?</td>
<td>What is the time?</td>
<td>And for John?</td>
<td>Answer: Yale</td>
</tr>
</tbody>
</table>

**Figure 6: DB User Interface**

**Figure 7: DB User Related Questions:** a pop-up window when the user clicks highlighted "related questions" in the above interface.
Figure 8: SQL Expert Interface

<table>
<thead>
<tr>
<th>Reference Table: Order_Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>item_id</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference Table: Product_Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>product_id</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

**Final SQL:**

```sql
SELECT COUNT(DISTINCT item_id) FROM SQLs WHERE item_amount > 100000
```
Figure 9: Dialogue Review Interface
D1: Database about soccer containing 3 tables

Q2: What are their names?

INFORM_SQL

S2: SELECT cName FROM College WHERE enr > 18000
A2: OU | FSU
R2: The names of the colleges that have enrollment values greater than 18000 are OU and FSU

Q3: Which one has the higher enrollment?

INFER_SQL

S1: SELECT cName FROM College WHERE enr > 18000 ORDER BY enr DESC LIMIT 1
A1: OU
R1: Among the results, OU has a higher enrollment.

Figure 10: Part of a dialogue example with INFER_SQL user dialog label

Figure 11: A complete dialogue example