


Article

DIR: A Large-Scale Dialogue Rewrite Dataset for Cross-Domain Conversational Text-to-SQL

Jieyu Li, Zhi Chen, Lu Chen *, Zichen Zhu , Hanqi Li, Ruisheng Cao and Kai Yu *

X-LANCE Lab, MoE Key Lab of Artificial Intelligence, AI Institute, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

* Correspondence: chenlusz@sjtu.edu.cn (L.C.); kai.yu@sjtu.edu.cn (K.Y.)

Abstract: Semantic co-reference and ellipsis always lead to information deficiency when parsing natural language utterances with SQL in a multi-turn dialogue (i.e., conversational text-to-SQL task). The methodology of dividing a dialogue understanding task into dialogue utterance rewriting and language understanding is feasible to tackle this problem. To this end, we present a two-stage framework to complete conversational text-to-SQL tasks. To construct an efficient rewriting model in the first stage, we provide a large-scale dialogue rewrite dataset (DIR), which is extended from two cross-domain conversational text-to-SQL datasets, SParC and CoSQL. The dataset contains 5908 dialogues involving 160 domains. Therefore, it not only focuses on conversational text-to-SQL tasks, but is also a valuable corpus for dialogue rewrite study. In experiments, we validate the efficiency of our annotations with a popular text-to-SQL parser, RAT-SQL. The experiment results illustrate 11.81 and 27.17 QEM accuracy improvement on SParC and CoSQL, respectively, when we eliminate the semantic incomplete representations problem by directly parsing the golden rewrite utterances. The experiment results of evaluating the performance of the two-stage frameworks using different rewrite models show that the efficiency of rewrite models is important and still needs improvement. Additionally, as a new benchmark of the dialogue rewrite task, we also report the performance results of different baselines for related studies. Our dataset will be publicly available once this paper is accepted.



Citation: Li, J.; Chen, Z.; Chen, L.; Zhu, Z.; Li, H.; Cao, R.; Yu, K. DIR: A Large-Scale Dialogue Rewrite Dataset for Cross-Domain Conversational Text-to-SQL. *Appl. Sci.* **2023**, *13*, 2262. <https://doi.org/10.3390/app13042262>

Academic Editor: Dimitrios G. Aggelis

Received: 5 January 2023

Revised: 30 January 2023

Accepted: 7 February 2023

Published: 9 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: dialogue rewrite; conversational text-to-SQL; two-stage framework

1. Introduction

Structured query language (SQL) is an executable machine language that can represent more complex intentions. Text-to-SQL tasks aim to represent the natural language with SQL so that target results can be queried from the database precisely, which is important for constructing an advanced language-based human-machine interactive system [1]. The conversational text-to-SQL task is a kind of text-to-SQL task considering the scene of task-oriented multi-turn dialogue.

In conversational text-to-SQL tasks, the parsing results always contain complicated structures and a wealth of information. It is difficult to protect the completeness of the information. Some works concatenate the dialogue history and current utterance as the input to complement the historical information. However, there is not only historical information, but also relationship information across different turns. Co-reference and ellipsis of semantics are two common language phenomena in multi-turn dialogues which represent the relationship information. The method of concatenating cannot tackle the problem of relationship information deficiency. Therefore, a line of research [2,3] divides a dialogue understanding task into two parts—dialogue utterance rewriting and language understanding—to inhibit the influence of co-reference and ellipsis. In the rewriting stage, dialogue history and the current utterance will be rewritten as a single utterance. In this case, the elliptical semantics are restored in text forms. Then, the corresponding single-turn model will predict the results with the rewritten utterance in the understanding stage.

The previous empirical performance demonstrates the feasibility of the two-stage framework in various dialogue understanding tasks. To this end, we attempt to apply the two-stage framework to tackle conversational text-to-SQL tasks in this work. To construct the rewriting model in the first stage, we propose a large-scale dialogue rewrite (DIR) dataset in this work. The dataset is expanded from two typical conversational text-to-SQL datasets—SPaC [4] and CoSQL [5]—containing 5908 dialogues and 160 domains. We collect the annotations by crowd-sourcing. To avoid the problem of spelling mistakes and missing words, we built a click-based graph user interface for annotators. To guarantee the data quality, we set up three checking procedures and manually correct the annotation mistakes. DIR is not an expert dataset focusing on conversational text-to-SQL tasks. It is also a quality dataset for dialogue rewrite research, especially in cross-domain dialogue setups. Apart from a large number of samples and domains, DIR also provides rewriting action category tags and operation details to track the rewriting process, which is useful for future interpretation studies.

In the experiments, we first use the typical text-to-SQL model RAT-SQL to parse the oracle rewritten utterances and the concatenated utterances and compare the question exact match (QEM) accuracy of them. The experiment results illustrate that the performance obtains 12% and 27% improvement on SPaC and CoSQL, respectively. We further train different rewrite models in the first stage to assess the robustness of the two-stage framework. The experiment results demonstrate that the dialogue format and efficient rewrite methods are crucial for the two-stage framework. Finally, we summarize the error cases and provide in-depth analysis.

The contributions of this work are as follows:

- We collect a large-scale cross-domain dialogue rewriting dataset DIR for conversational text-to-SQL. We also provide the additional action category labels and the rewrite process tracking annotations for interpretation research.
- We evaluate the effectiveness of the two-stage framework with DIR.
- We provide in-depth analysis of the two-stage framework in conversational text-to-SQL.

2. Related Work

2.1. Conversational Text-to-SQL

To construct an advanced human–machine interaction system, researchers attempt to build a text-to-SQL parser that can automatically translate the human language to SQL so that the system can finish complicated querying. Previous works in text-to-SQL studies achieved remarkable progress. In encoding method studies, GNN [6] is always used to encode the schema linking [7–12], which enhance the relationship between human language and database schema. In decoder algorithm research, grammar-based decoding methods [13–15] and token-based constrained decoding methods [8,16] are considered to parse precise SQL queries. Nowadays, researchers pay more attention to the application in dialogue scenes and provide a conversational text-to-SQL dataset, SPaC [4]. In each turn of the dialogue in SPaC, the system responses are not represented in human language, which is unusual in the real world. To this end, researchers further proposed another conversational text-to-SQL dataset, CoSQL [5]. Different from SPaC, the system responses in CoSQL are mainly rule-based synthetic natural language utterances. Recently, most of the published studies focused on modeling the incremental semantic information between different turns. R²-SQL [17] encodes the dynamic contextualized schema graph for each dialogue turn and achieves 55.8 QEM accuracy in SPaC. TC [18] directly predicts the incremental SQL clause and achieves 65.7 QEM in SPaC when using a task-specific pre-trained language model GAP [19] as the backbone. SCoRe [20] provides an efficient task-specific pre-trained language model which achieves 62.4 QEM accuracy in SPaC. HIE-SQL [21] enhances the history information via encoding the relationships among natural language, last-turn system responses, and database schema. Finally, with the help of a task-specific pre-trained language model, GRAPPA [22], HIE-SQL achieves 64.6 QEM

accuracy in SParC and 53.9 QEM accuracy in CoSQL. STAR [23] additionally considers the previous SQL and achieves 67.4 QEM accuracy in SParC and 57.8 QEM accuracy in CoSQL. Although these methods consider the incremental information across the dialogue, all of them ignore the problem of incomplete semantic representation.

2.2. Dialogue Rewrite

Dialogue rewriting is always used to reduce a multi-turn dialogue task into a single-turn NLP task via co-reference resolution and ellipsis complementing. Several corresponding benchmarks have been published for different dialogue tasks. For example, MULTI [24] and REWRITE [25] provide rewrite annotations for chat-bot scenes. CANARD [26], expanded from QuAC, considers tackling the sequential question-answering tasks. Actually, the present conversational text-to-SQL tasks focus on cross-domain task-oriented dialogue. However, existent task-oriented dialogue rewrite datasets are not desirable. TASK [2] is collected, which is expanded from a typical dialogue dataset, CamRest676, by modifying a semantic-complete utterance to an utterance containing co-reference or ellipsis phrases. However, it focuses on a single domain. CQR [27], considered the multi-domain dialogue scene, which is modified from another corpus [28]. It encourages annotators to rewrite utterances with their own words, which increases the difficulty of rewriting when confronting complicated utterances. Based on a large-scale multi-domain dialogue dataset, MultiWOZ, the researcher collected the dataset MultiWOZ2.3 [29] by adding co-reference annotations. However, the number of annotations is low, and the complement phrases are less varied. To this end, we propose a large-scale cross-domain task-oriented dialogue rewrite dataset. It provides specific rewrite annotations for conversational text-to-SQL. Moreover, it is also a high-quality data resource for task-oriented dialogue rewrite studies. More details of the differences between the present datasets and DIR are shown in Table 1.

Table 1. Examples of all the semantic states.

Utterance	Category
History: Show ids of all employees. Current: Show ids of all employees who have destroyed a document.	Semantically Complete
History: We went to see a concert last night. Current: The tickets were really expensive.	Bridging Anaphora
History: What school has the most number of students? Current: How many teachers are in that school?	Definite Noun Phrases
History: Show me the age of all pilots! Current: what is the name of the oldest one?	One Anaphora
History: List all the shop names. Current: Which shop has the least quantity of devices of those?	Demonstrative Pronoun
History: Show the name for all employees. Current: What's their age?	Possessive Determiner
History: Show the name of all teachers. Current: Who teach math?	Continuation
History: Show the director for all movies. Current: How about the name?	Substitution-Explicit
History: Find all students born before 1998? Current: How about after?	Substitution-Implicit
History: What are the distinct last names of staff? Current: Of customers?	Substitution-Schema
History: Show me the number of invoices by country! Current: What is the total invoice for each!	Substitution-Operator

3. Dialogue Rewrite Task

The dialogue rewrite task aims to merge the dialogue history and the current utterance into a single utterance. In this case, elliptical semantics caused by co-reference and ellipsis can be restored. We formulate the task as below. Utterance set $U = \{u_1, u_2, \dots, u_n\}$ is a continuous fragment of the dialogue \mathcal{D} . The dialogue rewrite task aims to learn a function

$$\mathcal{F}(U) \rightarrow U', \quad (1)$$

where U' is also a set of utterances but does not contain any co-reference or ellipsis phenomena in each of its utterances.

As Table 1 shows, the semantic states of a utterance can be split into three types [30]. **Semantically complete** utterances contain complete semantics, and they are retained during rewriting.

Co-reference is a language phenomenon for which there is an *anaphor* in the utterance which refers to an *antecedent* in another utterance. There are five categories of co-reference states:

1. Bridging anaphora and their antecedents are linked via various lexico-semantic frames or encyclopedic relations.
2. Definite noun phrase is a determined noun phrase whose head is a noun with definiteness.
3. One anaphora is an anaphoric noun phrase headed by *one*.
4. Demonstrative pronoun is a pronoun used to point to specific people or things.
5. Possessive determiner is one of the words *my, your, his, her, its, our, and their*.

Ellipsis are the expressions that are not syntactically sentential but nevertheless have characters that yield propositional contents given a context [31]. There are five categories of ellipsis states:

1. Continuation is the subsequent utterance that omits previous query conditions.
2. Substitution-explicit exists when the current utterance contains the same structure of querying as the previous utterance. The repeating parts are omitted in the current utterance, and substituting the different parts will restore the complete query. The substitution phrase is the query target and is explicit in this category.
3. Substitution-implicit is where the substitution phrase is the query target and is implicit in this category.
4. Substitution-schema is where the substitution phrase is the query condition.
5. Substitution-operator is where the substitution phrase involves a mathematical operation.

All of the aforementioned category labels are provided in DIR. We will further introduce the details of our dataset in the next section.

4. Dataset

In this section, we introduce the annotations provided by DIR in Section 4.1. Then, we illustrate the process of data collection and quality evaluation in Section 4.2. Finally, we show the statistics information of DIR in Section 4.3.

4.1. Annotation

Figure 1 illustrates an annotation example. We formulate the rewriting process of dialogue as a sequence of rewriting actions. If an utterance needs to be rewritten, the corresponding action will contain the semantic state tags and operation details. Otherwise, the action only contains a *none* operation tag as the example in Figure 1 shown. In dialogue rewrite (DIR), we encourage that the rewritten utterances only consist of the spans of history utterances and predefined keywords (e.g., *the, and* and). We provide three annotations in total for the utterances in each dialogue.

Complemented annotations are the rewriting results of each utterance. The utterances containing complete semantics will be concatenated with the rewritten utterances of the last turn.

Semantic state tags contain action type and action category as shown in Figure 1. They represent the semantic states of the current utterance. We provide three kinds of action type, *semantically complete*, *co-reference*, and *ellipsis*, which are the same as we introduced in Section 4. We label the action category according to the categories introduced in Section 4.

Operation details contain three annotations—operation span, action operation, and operation position—used for recording the rewrite trace. The operation span is the position of the phrases used to complement the current utterance. To improve fluency, we provide some keywords for annotators to construct readable spans. The operation span labels also record the position of the first time each phrase appears in the dialogue. For example, the span *the students with dogs and older than 10*, used for rewriting the third turn of the dialogue in Figure 1, contains keywords *the* and *and*. It also contains two spans from the first and the second turns, respectively. For the span *students with dogs*, we use *turn 0* to mark its position, although it also in the second rewritten utterance. The action operation is one of *insert* and *replace*, representing the corresponding operation apply on the operation span. The operation position notes the target position of the operation span. In this case, we can track the whole rewriting process using these three annotations.

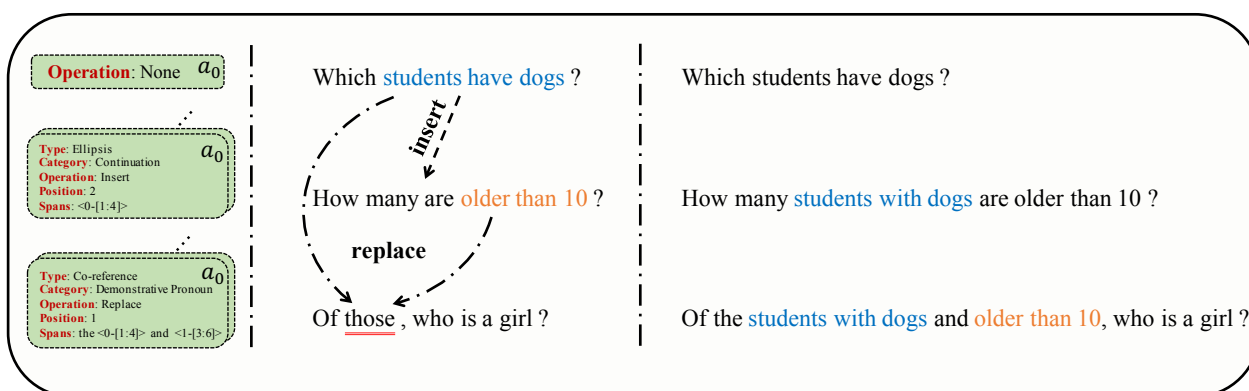


Figure 1. Overview of the annotations in DIR.

4.2. Data Collection

In this section, we will introduce the process of data collection. The overview is illustrated in Figure 2. We collect the annotations via crowd-sourcing. During the process, we apply sampling inspection as the first round of quality checking (Section 4.2.1). Then we detect and correct the bad cases by keyword recognizing as the second round of quality checking (Section 4.2.2). Finally, we utilize a pre-trained text-to-SQL parser to parse the rewritten utterances. Meanwhile, we screen out the unsuccessful samples and manually correct the mistakes, which is the last round of quality checking (Section 4.2.3).

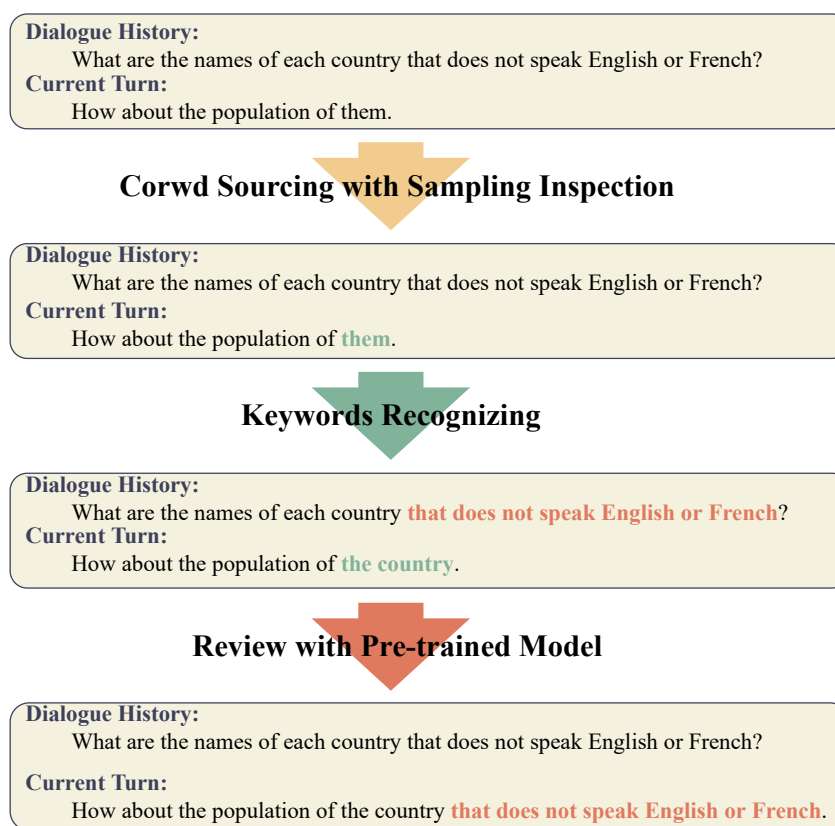


Figure 2. The process of data collection and data cleaning.

4.2.1. Crowd Sourcing with Sampling Inspection

We collect the annotations by crowd sourcing. For each utterance, we encourage the annotators to rewrite using the words of the current utterance and dialogue history. To avoid spelling mistakes during the rewriting, we construct a clicking-based annotation platform. The annotators create the rewritten utterance by clicking the words they need. Meanwhile, the rewrite trace, which we mentioned in Section 4.1, can also be recorded. The annotators also need to label the semantic state tags for each utterance. For some special cases, we allow the annotators to rewrite by hand. During the annotating, we randomly sample and check the data quality once the number of annotations achieves a predefined threshold value. Then we will correct the wrong annotations to ensure the data quality. Finally, we randomly review 6.33% annotations when the process of annotation is finished.

4.2.2. Keywords Recognizing

For each rewritten utterance, we recognize whether a predefined keyword exists. The predefined keywords including typical anaphora (e.g., *that* and *this*) and possessive determiners (e.g., *their* and *his*). If they exist, we will check the annotations and correct the mistakes.

4.2.3. Review with Pre-Trained Model

For each rewritten utterance, we attempt to parse it using a pre-trained text-to-SQL parser. Then, we compare the parsing result and its golden SQL, which is provided by the original conversational text-to-SQL datasets. If a SQL keyword (e.g., *WHERE* and *avg*) or a schema element (e.g., *table* and *column*) is missing, we will check the annotation and correct the mistakes.

4.3. Statistics

We collect a total of 5193 dialogues for the training set and 715 dialogues for the development set in DIR. To estimate the complexity of a dialogue rewrite dataset, we introduce three metrics in this section. All of them are positively correlated with complexity. **Span per utterance (S/U)** represents the average number of spans in dialogue history, used to complement the semantics of the current utterance.

Expansion ate (ER) represents the ratio of the rewritten utterance length and its original length.

Accumulation rate (AR) represents the average number of turns used to complement the semantics of the current utterance.

Table 2 illustrates the statistics information of DIR and other published dialogue rewrite datasets. For TASK [2] and CQR [32], the labels marking the source of the complement span are deficient. Therefore, we do not calculate the S/U and the AR for them to avoid the misestimate caused by text matching. The statistical results demonstrate that DIR involves the largest number of domains and is also challenging. In Section 6, we will illustrate the efficiency of the two-stage framework training with DIR.

Table 2. Comparison of DIR and other analogous dialogue rewrite datasets. DIR-SParC and DIR-CoSQL are the two parts of DIR, split according to the data source. **Num. Dom** refers to the number of domains. **R_D** refers to the ratio of the dialogue with at least one semantic incomplete utterance. **R_T** refers to the ratio of semantic incomplete utterances. **S/U**, **ER** and **AR** are introduced in Section 4.3.

Dataset	Num. Dom	R _D	R _T	S/U	ER	AR
TASK	1	92.75	24.25	-	1.27	1.11
CQR	3	62.17	47.34	-	2.18	-
MultiWOZ 2.3	7	20.16	4.75	1.23	0.96	1.18
DIR-SParC	160	89.35	50.02	1.41	1.57	1.23
DIR-CoSQL	160	49.39	24.16	1.29	1.57	1.14
DIR	160	72.77	38.40	1.38	1.57	1.21

We further statistic the average S/U and the ER of the utterances with different accumulation rates (AR). The results are depicted in Figure 3. The S/U scores show that most of the history utterances only provide 1 span for complementing. The ER scores demonstrate that the utterances which are rewritten involving multiple turns obtain a long length.

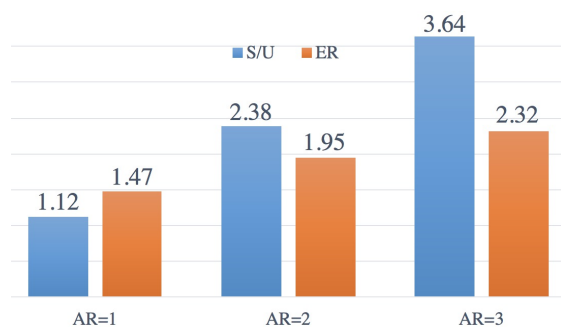


Figure 3. S/U metric and ER metric for the utterances with different accumulation rate.

5. Dialogue Rewrite

In this section, we first introduce present dialogue rewrite baselines in Section 5.1. Then we demonstrate the challenging of DIR in Section 5.2.

5.1. Baseline

In this work, we consider three rewriting baselines.

Concat is the method that directly concatenates the whole dialogue into a long sequence rather than applying a rewriting algorithm. Combining the dialogue history does not lose any historical information. However, the great length will leverage the performance. This baseline illustrates the performance of inputting complete historical information but incomplete relational information.

Seq2Seq [33] is a vanilla model, using LSTM to encode the input and using GRU for decoding. The attention mechanism is utilized to enhance the encoding process.

RUN [34] is a state-of-the-art method that reformulates the rewriting task as a figure semantic segmentation task. It uses an edit matrix to describe the process of rewriting. Then, it uses a U-type CNN [35] to predict the edit operations. Finally, it restores the utterance according to the edit operations.

5.2. DIR as a Challenging Benchmark

The metrics to estimate rewrite performance are F-score [36], BLEU [37], and ROUGE [38], which are the main metrics in related studies. The results shown in Table 3 demonstrate that the RUN achieves the best performance.

Table 3. The rewrite performance of three different methods.

Methods	F-Score	BLEU			ROUGE		
		1	2	4	1	2	L
Concat	57.54	51.55	51.18	50.15	83.02	72.03	75.18
Seq2Seq	23.11	69.38	61.27	52.35	57.06	43.02	55.06
RUN	65.03	89.80	87.38	83.44	93.90	88.16	92.45

We further investigate the influence of the accumulation rate (AR). The following analysis is based on RUN. The considered lengths of the high AR utterances are much longer than the low AR utterances. Meanwhile, the BLEU-4 scores are sensitive to the utterance length. Therefore, we introduce the relative BLEU4 (R-BLEU4) score in this experiment. The metric R-BLEU4 is the production of the BLEU4 score and the length of the golden rewritten utterance. Figure 4 illustrates the BLEU4 score and R-BLEU4 score of the rewritten utterances with different AR. Both of the metrics demonstrate that rewriting the high AR utterances is challenging.

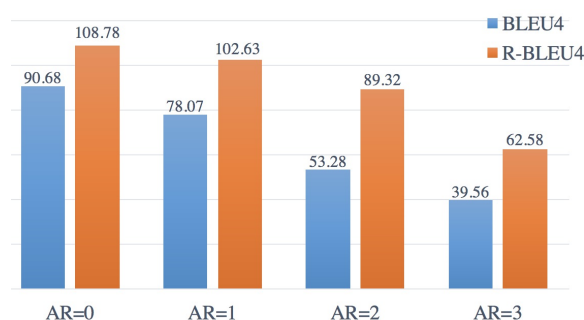


Figure 4. BLEU4 score and R-BLEU4 score of the utterances with different accumulate rates.

6. Conversational Text-to-SQL

Conversational text-to-SQL tasks aim to parse user intentions with SQL in a human-machine dialogue. In traditional dialogue state tracking tasks, the system understands human words via simple structural representations. Typically, slot-value pair is the popular option. However, the queries represented by slot-value pairs always lack diversity. Only simple queries are allowed. Different from slot-value pair, SQL can represent complex querying which is beneficial to construct an advanced dialogue system. On the other hand, the parsing of SQL is much more challenging than slot-value pairs.

Different from the single-turn text-to-SQL tasks, a conversational text-to-SQL parser always needs to understand the information of multiple sentences. Apart from that, parsing SQL also needs complete information to finish structural inference. In this case, co-reference and ellipsis phenomena among multiple utterances significantly increase the parsing difficulty. However, the rewriting results of these utterances always contain subordinate clauses, which is textually similar to a SQL query with a nested structure. Therefore, the two-stage framework is suitable for tackling conversational text-to-SQL tasks. We will introduce the framework in Section 6.1 and provide further experimental analysis in Section 6.2.

6.1. Two-Stage Framework

The two-stage framework consists of the rewriting stage and the understanding stage. In the rewriting stage, a rewriting model is used to merge the dialogue history and resolve the co-reference and ellipsis phenomena of the current utterance. In the understanding stage, the downstream single-turn model finishes the task.

6.1.1. Metrics

Question exact match (QEM) is the metric used to assess the accuracy of the results in conversational text-to-SQL tasks. It compares the contents of each clause of the predicted SQL and the golden SQL. The case is treated as correct if and only if all the clauses match with the golden clauses. QEM represents the accuracy.

Question execution (QEX) is the metric used to assess the accuracy of the results by comparing the execution results of the ground truth SQL and the predicted SQL. The case is correct if and only if the querying results are completely the same. However, there are some pseudo programs in which the execution results are always empty in SPaC and CoSQL. In this case, the execution results cannot always precisely estimate whether the predicted SQL is correct. As a result, the QEX accuracy is always much higher than the QEM accuracy for each model.

6.1.2. Rewrite Stage Models

Concat represents that the user question in the understanding stage is the concatenation of the current utterance and the dialogue history. In this case, the input contains all the explicit historical information. However, the implicit relational information is still neglected.

Oracle represents that the utterance part is the rewritten utterance, in other words, the annotation of DIR. Compared with the method of concatenating all the utterances, this method explicitly encodes both historical and relational information.

6.1.3. Understanding Stage Model

RAT-SQL is a popular text-to-SQL encoding method that models the relationships between user questions and database schema with a GNN, which is known as the RAT layer. In this work, we use BERT-base to encode the user question tokens and database schema items (tables and columns). Then, we build the relationships with the RAT layer. Finally, we decode an action sequence to reconstruct the abstract syntax tree of the target SQL. In this section, we utilize RAT-SQL as the understanding model.

The experiment results in Table 4 illustrate that resolving co-reference and ellipsis achieves 21.81 QEM accuracy on SPaC and 27.17 QEM accuracy on CoSQL. The improvements demonstrate that the deficiency of relational information leads to parsing errors. A two-stage framework system will achieve significant improvement if we complement the semantic information. Additionally, the experiment assumption of Oracle is that the rewriting model exactly outputs the correct rewritten utterances. In this case, the experiment results are the upper bound of the two-stage framework with RAT-SQL as the understanding model. We further investigate the influence of the rewriting model in the next section.

Table 4. Comparison of the two-stage framework and existing methods. **Concat RAT-SQL** represents the method using concatenated dialogue utterances as the inputs. **Oracle RAT-SQL** represents the method using the golden rewritten utterances as the inputs. The metric used in this experiment is QEM accuracy.

Models	SParC	CoSQL
CD-S2S [4]	21.9	13.8
SyntaxSQL [4]	18.5	15.1
EditSQL [39]	47.2	39.9
IGSQL [40]	50.7	44.1
RichContext [34]	52.1	41.0
R ² SQL [17]	54.1	46.8
Concat RAT-SQL	47.46	25.81
Oracle RAT-SQL	59.27	52.98

6.2. Efficient Rewrite Model is Necessary

In this section, we study the influence of the rewriting model in the two-stage framework. For the understanding stage, we use RAT-SQL as the single-turn text-to-SQL parser, which is the same as that used in Section 6.1. For the rewrite stage, we compare three different baseline models introduced in Section 5.1. Additionally, we also report the results of using the ground truth rewrite utterances as the rewrite model outputs (Oracle RAT-SQL), which is the upper bound performance of the two-stage framework with RAT-SQL as the understanding model. Table 5 illustrates the QEM of the systems using different rewriting models. The results illustrate that the performance of the whole system positively corresponds with the rewrite model performance. Moreover, there is still a large margin between present models and the upper bound. In this case, an efficient rewrite model is necessary to build an advanced two-stage text-to-SQL parser.

Table 5. QEM accuracy and QEX accuracy of the two-stage models with different rewriting models.

Models	BLEU-4	ROUGE-L	F-Score	SParC		CoSQL	
				QEM	QEX	QEM	QEX
Concat RAT-SQL	50.15	75.18	57.54	47.46	70.38	25.81	45.18
Seq2Seq RAT-SQL	52.35	55.06	23.11	24.02	47.72	27.04	48.22
RUN RAT-SQL	83.44	92.45	65.03	49.79	72.87	50.20	71.24
Oracle RAT-SQL	-	-	-	59.27	81.01	52.98	74.82

Furthermore, we analyze the influence of each component of the two-stage framework. The rewrite model is RUN in this experiment. We screen out the samples that are parsed successfully using golden rewritten utterances. Then, we further filter out the samples which are not parsed accurately using the two-stage framework in these samples. That portion refers to the cases that are parsed unsuccessfully caused by the wrong rewriting. Results in Figure 5 show that the rewrite-caused error makes up the majority, and it further demonstrates that there is still a large margin to improve.

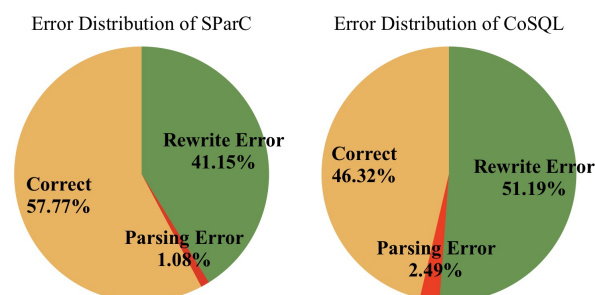


Figure 5. Distribution of the error type.

In Table 5, we observe that the performance results on CoSQL relate to the BLEU score of the rewriting models. However, the performance results on SParC relate to the other metrics. We assume that the phenomena are led by the diversity of datasets, which we will discuss in the next section.

6.3. Influence of Dialogue Formulation

To verify the hypothesis, we split our dataset into two parts, DIR-SParC and DIR-CoSQL, according to the data source. Then, we train RUN with different training data and evaluate these two parts. Table 6 illustrates that the model cannot automatically adapt novel dialogue from other datasets. We assume that the reason is that the dialogue formulations of SParC and CoSQL are different. In SParC, there is no natural language system response at each turn. However, CoSQL provides the system response and system act labels in each dialogue. Additionally, the fluency in CoSQL is better than SParC. Therefore, it is difficult for models to adapt the novel dialogue formulation without any annotations.

Table 6. Rewriting performance on RUN with different parts of DIR as the training set.

Train Dataset	DIR-SParC			DIR-CoSQL		
	BLEU-4	ROUGE-L	EM	BLEU-4	ROUGE-L	EM
DIR-SParC	55.86	78.94	11.22	78.11	89.02	38.97
DIR-CoSQL	39.29	67.91	0.01	83.58	91.39	73.96
DIR	64.08	82.90	18.45	88.31	94.26	74.25

The exact match (EM) accuracy in this experiment represents the rate of the cases where the rewritten utterance is exactly equal to the annotations. We assume that the rewrite rate and the complexity of SParC utterances lead to low EM accuracy on DIR-SParC parts.

7. Conclusions

In this work, we investigate the efficiency of the rewriting–understanding framework on conversational text-to-SQL. To build the rewriting model, we propose a large-scale cross-domain dialogue rewrite dataset, DIR. Apart from the rewrite annotations, we also provide rewrite category labels and rewrite trace annotations for dialogue rewrite studies, especially interpretation studies. On the other hand, DIR is also a large-scale corpus for dialogue rewriting studies. To this end, we propose several baselines for carrying on the corresponding research. In experiments, we first verify the efficiency of the two-stage framework. Experiment results demonstrate that complementing the relational information by dialogue rewriting can achieve remarkable improvement. We further assess the performance of the systems using different rewrite models. Experiment results illustrate that an efficient rewrite algorithm is necessary to construct a two-stage conversational text-to-SQL parser. Finally, we depict the influence of dialogue formulation in rewrite models. We suggest that it is a challenge for models to adapt the novel dialogue formulation without any annotations.

Author Contributions: Conceptualization, J.L. Z.C. and L.C.; Data Curation, J.L., Z.Z. and H.L.; supervision, Z.C., R.C., L.C. and K.Y.; Writing—original draft, J.L.; Writing—review and editing, L.C. and K.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the China NSFC Projects (No. 62120106006 and No. 62106142), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), CCF-Tencent Open Fund and Startup Fund for Youngman Research at SJTU (SFYR at SJTU).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset will be publicly available at <https://github.com/Auracion/DIR> (accessed on 4 January 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Androutsopoulos, I.; Ritchie, G.D.; Thanisch, P. Natural language interfaces to databases-an introduction. *Nat. Lang. Eng.* **1995**, *1*, 29–81. [[CrossRef](#)]
2. Quan, J.; Xiong, D.; Webber, B.; Hu, C. GECOR: An End-to-End Generative Ellipsis and Co-reference Resolution Model for Task-Oriented Dialogue. 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), Hong Kong, China, 3–7 November 2019; pp. 4539–4549.
3. Chen, Z.; Chen, L.; Li, H.; Cao, R.; Ma, D.; Wu, M.; Yu, K. Decoupled Dialogue Modeling and Semantic Parsing for Multi-Turn Text-to-SQL. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Virtual, 1–6 August 2021; pp. 3063–3074.
4. Yu, T.; Zhang, R.; Yasunaga, M.; Tan, Y.C.; Lin, X.V.; Li, S.; Heyang Er, I.L.; Pang, B.; Chen, T.; Ji, E.; et al. SPaC: Cross-Domain Semantic Parsing in Context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
5. Yu, T.; Zhang, R.; Er, H.; Li, S.; Xue, E.; Pang, B.; Lin, X.V.; Tan, Y.C.; Shi, T.; Li, Z.; et al. 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), Hong Kong, China, 3–7 November 2019; pp. 1962–1979.
6. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **2008**, *20*, 61–80. [[CrossRef](#)] [[PubMed](#)]
7. Bogin, B.; Gardner, M.; Berant, J. Global reasoning over database structures for text-to-sql parsing. *arXiv* **2019**, arXiv:1908.11214.
8. Lin, X.V.; Socher, R.; Xiong, C. Bridging textual and tabular data for cross-domain text-to-sql semantic parsing. *arXiv* **2020**, arXiv:2012.12627.
9. Chen, Z.; Chen, L.; Zhao, Y.; Cao, R.; Xu, Z.; Zhu, S.; Yu, K. ShadowGNN: Graph projection neural network for text-to-SQL parser. *arXiv* **2021**, arXiv:2104.04689.
10. Hui, B.; Geng, R.; Wang, L.; Qin, B.; Li, Y.; Li, B.; Sun, J.; Li, Y. S²SQL: Injecting Syntax to Question-Schema Interaction Graph Encoder for Text-to-SQL Parsers. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2022, Dublin, Ireland, 22–27 May 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 1254–1262.
11. Wang, B.; Shin, R.; Liu, X.; Polozov, O.; Richardson, M. Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. *arXiv* **2019**, arXiv:1911.04942.
12. Cao, R.; Chen, L.; Chen, Z.; Zhao, Y.; Zhu, S.; Yu, K. LGE SQL: Line Graph Enhanced Text-to-SQL Model with Mixed Local and Non-Local Relations. *arXiv* **2021**, arXiv:2106.01093.
13. Krishnamurthy, J.; Dasigi, P.; Gardner, M. Neural Semantic Parsing with Type Constraints for Semi-Structured Tables. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 7–11 September 2017; pp. 1516–1526.
14. Yin, P.; Neubig, G. A Syntactic Neural Model for General-Purpose Code Generation. *arXiv* **2017**, arXiv:1704.01696.
15. Guo, J.; Zhan, Z.; Gao, Y.; Xiao, Y.; Lou, J.G.; Liu, T.; Zhang, D. Towards Complex Text-to-SQL in Cross-Domain Database with Intermediate Representation. *arXiv* **2019**, arXiv:1905.0820.
16. Scholak, T.; Schucher, N.; Bahdanau, D. PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models. *arXiv*, **2021**, arXiv:2109.05093.
17. Hui, B.; Geng, R.; Ren, Q.; Li, B.; Li, Y.; Sun, J.; Huang, F.; Si, L.; Zhu, P.; Zhu, X. Dynamic Hybrid Relation Network for Cross-Domain Context-Dependent Semantic Parsing. *arXiv* **2021**, arXiv:2101.01686.
18. Li, Y.; Zhang, H.; Li, Y.; Wang, S.; Wu, W.; Zhang, Y. Pay More Attention to History: A Context Modeling Strategy for Conversational Text-to-SQL. *arXiv* **2021**, arXiv:2112.08735.
19. Shi, P.; Ng, P.; Wang, Z.; Zhu, H.; Li, A.H.; Wang, J.; dos Santos, C.N.; Xiang, B. Learning contextual representations for semantic parsing with generation-augmented pre-training. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 13806–13814.
20. Yu, T.; Zhang, R.; Polozov, O.; Meek, C.; Awadallah, A.H. SCoRE: Pre-Training for Context Representation in Conversational Semantic Parsing. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
21. Zheng, Y.; Wang, H.; Dong, B.; Wang, X.; Li, C. HIE-SQL: History Information Enhanced Network for Context-Dependent Text-to-SQL Semantic Parsing. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2022, Dublin, Ireland, 22–27 May 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 2997–3007. [[CrossRef](#)]
22. Yu, T.; Wu, C.S.; Lin, X.V.; Wang, B.; Tan, Y.C.; Yang, X.; Radev, D.; Socher, R.; Xiong, C. Grappa: Grammar-augmented pre-training for table semantic parsing. *arXiv* **2020**, arXiv:2009.13845.
23. Cai, Z.; Li, X.; Hui, B.; Yang, M.; Li, B.; Li, B.; Cao, Z.; Li, W.; Huang, F.; Si, L.; et al. STAR: SQL Guided Pre-Training for Context-dependent Text-to-SQL Parsing. *arXiv* **2022**, arXiv:2210.11888.
24. Pan, Z.; Bai, K.; Wang, Y.; Zhou, L.; Liu, X. Improving open-domain dialogue systems via multi-turn incomplete utterance restoration. 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), Dublin, Ireland, 22–27 May 2019; pp. 1824–1833.

25. Su, H.; Shen, X.; Zhang, R.; Sun, F.; Hu, P.; Niu, C.; Zhou, J. Improving Multi-turn Dialogue Modelling with Utterance ReWriter. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 22–31.
26. Elgohary, A.; Peskov, D.; Boyd-Graber, J. Can You Unpack That? Learning to Rewrite Questions-in-Context. 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), Hong Kong, China, 3–7 November 2019; pp. 5920–5926.
27. Regan, M.; Rastogi, P.; Gupta, A.; Mathias, L. A dataset for resolving referring expressions in spoken dialogue via contextual query rewrites (cqr). *arXiv* **2019**, arXiv:1903.11783.
28. Eric, M.; Manning, C.D. Key-value retrieval networks for task-oriented dialogue. *arXiv* **2017**, arXiv:1705.05414.
29. Han, T.; Liu, X.; Takanobu, R.; Lian, Y.; Huang, C.; Peng, W.; Huang, M. MultiWOZ-coref: A multi-domain task-oriented dataset enhanced with annotation corrections and co-reference annotation. *arXiv* **2020**, arXiv:2010.05594.
30. Liu, Q.; Chen, B.; Guo, J.; Lou, J.G.; Zhou, B.; Zhang, D. How far are we from effective context modeling? An exploratory study on semantic parsing in context. *arXiv* **2020**, arXiv:2002.00652.
31. Stainton, R.J. 80Semantic Ellipsis. In *Words and Thoughts: Subsentences, Ellipsis, and the Philosophy of Language*; Oxford University Press: Cambridge, MA, USA, 2006.;oso/9780199250387.003.0005. Available online: https://academic.oup.com/book/0/chapter/314898219/chapter-ag-pdf/44494641/book_36193_section_314898219.ag.pdf (accessed on 15 June 2007). [[CrossRef](#)]
32. Rastogi, P.; Gupta, A.; Chen, T.; Mathias, L. Scaling Multi-Domain Dialogue State Tracking via Query Reformulation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019.
33. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *arXiv* **2014**, arXiv:1409.3215.
34. Liu, Q.; Chen, B.; Lou, J.G.; Zhou, B.; Zhang, D. Incomplete Utterance Rewriting as Semantic Segmentation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Virtual, 16–20 November 2020; pp. 2846–2857.
35. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
36. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Hobart, Australia, 4–8 December 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1015–1021.
37. Habash, N.; Sadat, F. Arabic preprocessing schemes for statistical machine translation. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York, NY, USA, 4–9 June 2006; pp. 49–52.
38. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 4–10 July 2004; pp. 74–81.
39. Zhang, R.; Yu, T.; Er, H.; Shim, S.; Xue, E.; Lin, X.V.; Shi, T.; Xiong, C.; Socher, R.; Radev, D. 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), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 5338–5349. [[CrossRef](#)]
40. Cai, Y.; Wan, X. IGSQ: Database Schema Interaction Graph Based Neural Model for Context-Dependent Text-to-SQL Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6903–6912. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.