Abstract

Text-to-SQL parsers are crucial in enabling non-experts to effortlessly query relational data. Training such parsers, by contrast, generally requires expert annotation of natural language (NL) utterances paired with corresponding SQL queries. In this work, we propose a weak supervision approach for training text-to-SQL parsers. We take advantage of the recently proposed question meaning representation called QDMR, an intermediate between NL and formal query languages. We show that given questions, their QDMR structures (annotated by non-experts or automatically predicted), and the answers, we can automatically synthesize SQL queries that are then used to train text-to-SQL models. Extensive experiments test our approach on five benchmark datasets. The results show that our models perform competitively with those trained on annotated NL-SQL data. Overall, we effectively train text-to-SQL parsers, using zero SQL annotations.

1 Introduction

The development of natural language interfaces to databases has been extensively studied in recent years (Affolter et al., 2019; Kim et al., 2020; Thorne et al., 2021). The current standard is Machine Learning (ML) models which map utterances in natural language (NL) to executable SQL queries (Wang et al., 2020; Rubin and Berant, 2021). These models rely on supervised training examples of NL questions labeled with their corresponding SQL queries. Labeling copious amounts of data is cost-prohibitive as it requires experts that are familiar both with SQL and with the underlying database structure (Yu et al., 2018). Furthermore, it is often difficult to re-use existing training data in one domain in order to generalize to new ones (Suhr et al., 2020). Adapting the model to a new domain requires new NL-SQL training examples, which results in yet another costly round of annotation.

In this paper we propose a weak supervision approach for training text-to-SQL parsers. We avoid the use of manually labeled NL-SQL examples and rely instead on data provided by non-expert users. Fig. 1 presents a high-level view of our approach. The input (left corner, in red) is used to automatically synthesize SQL queries (step 3, in green) which, in turn, are used to train an NL-to-SQL model. The supervision signal consists of the question’s answer and uniquely, a structured representation of the question decomposition, called QDMR. The annotation of both these supervision sources can be effectively crowdsourced to non-experts (Berant et al., 2013; Pasupat and Liang, 2015; Wolfson et al., 2020). In a nutshell, QDMR is a series of computational steps, expressed by semi-structured utterances, that together match the semantics of the original question. The bottom left corner of Fig. 1 shows an example QDMR of the question “Which authors have more than 10 papers in the PVLDB journal?”. The question is broken into five steps, where each step expresses a single logical operation (e.g., select papers, filter those in PVLDB) and may refer to previous steps. As QDMR is derived entirely from its question, it is agnostic to the underlying form of knowledge representation and has been used for questions on images, text and databases (Subramanian et al., 2020; Geva et al., 2021; Saparina and Osokin, 2021). In our work, we use QDMR as an intermediate representation for SQL synthesis. Namely, given an input QDMR we implement an automatic procedure mapping it to SQL. The QDMR can either be manually annotated or effectively predicted by a trained model, as shown in our experiments.

We continue to describe the main components of our system, using the aforementioned supervision (Fig. 1). The SQL Synthesis component (step 1) attempts to convert the input QDMR into a corresponding SQL query. To this end, Phrase DB linking matches phrases in the QDMR with rele-
SQL join paths are automatically inferred given the database schema structure. Last, the QDMR, DB-linked columns and inferred join paths are converted to SQL by the SQL Mapper. In step 2, we rely on question-answer supervision to filter out incorrect candidate SQL. Thus, our Execution-guided SQL Search returns the first candidate query which executes to the correct answer.

Given our synthesis procedure, we evaluate its ability to produce accurate SQL using weak supervision. To this end, we run our synthesis on 9,313 examples of questions, answers and QDMRs from five standard NL-to-SQL benchmarks (Zelle and Mooney, 1996; Li and Jagadish, 2014; Yaghmazadeh et al., 2017; Yu et al., 2018). Overall, our solution successfully synthesizes SQL queries for 77.8% of examples, thereby demonstrating its applicability to a broad range of target databases.

Next, we show our synthesized queries to be an effective alternative to training on expert annotated SQL. We compare a text-to-SQL model, trained on the queries synthesized from questions, answers and QDMRs, to one trained using gold SQL. As our model of choice we use T5-large, which is widely used for sequence-to-sequence modeling tasks (Raffel et al., 2020). Following past work (Shaw et al., 2021; Herzig et al., 2021), we fine-tune T5 to map text to SQL. We experiment with the SPIDER and GEO880 datasets (Yu et al., 2018; Zelle and Mooney, 1996) and compare model performance based on the training supervision. When training on manually annotated QDMRs, the weakly supervised models achieve 91% to 97% of the accuracy of models trained on gold SQL. We further extend our approach to use automatically predicted QDMRs, requiring zero annotation of in-domain QDMRs. Notably, when training on predicted QDMRs models still reach 86% to 93% of the fully supervised versions accuracy. In addition, we evaluate cross-database generalization of models trained on SPIDER (Suhr et al., 2020). We test our models on four additional datasets and show that the weakly supervised models are generally better than the fully supervised ones in terms of cross-database generalization. Overall, our findings show that weak supervision, in the form of question, answers and QDMRs (annotated or predicted) is nearly as effective as gold SQL when training text-to-SQL parsers.

Our codebase and data are publicly available.

2 Background

Weakly Supervised ML. The performance of supervised ML models hinges on the quantity and quality of their training data. In practice, labeling large-scale datasets for new tasks is often cost-prohibitive. This problem is further exacerbated in semantic parsing tasks (Zettlemoyer and Collins, 2005), as utterances need to be labeled with formal queries. Weak supervision is a broad class of methods aimed at reducing the need to manually label large training sets (Hoffmann et al., 2011; Ratner et al., 2017; Zhang et al., 2019). An influential line of work has been dedicated to weakly supervised semantic parsing, using question-answer pairs, referred to as learning from denotations (Clarke et al., 2010; Liang et al., 2011). Past work has shown that non-experts are capable of annotating answers over knowledge graphs (Berant et al., 2013) and tabular data (Pasupat and Liang, 2015). This approach could potentially be extended to databases by sampling subsets of its tables, such that question-answer examples can be manually annotated. A key issue in learning text-to-SQL parsers from denotations is the vast search space of potential candidate queries. Therefore, past work has focused on con-
straining the search space, which limited applicability to simpler questions over single tables (Wang et al., 2019). Here, we handle arbitrary SQL by using QDMR to constrain the search space.

Question Decomposition QDMR expresses the meaning of a question by breaking it down into simpler sub-questions. Given a question \( x \), its QDMR \( s \) is a sequence of reasoning steps \( s^1, \ldots, s^{|s|} \) required to answer \( x \). Each step \( s^k \) is an intermediate question which represents a relational operation, such as projection or aggregation. Steps may contain phrases from \( x \), tokens signifying a query operation (e.g., “for each”) and references to previous steps. Operation tokens indicate the structure of a step, while its arguments are the references and question phrases. A key advantage of QDMR is that it can be annotated by non-experts and at scale (Wolfsen et al., 2020). Moreover, unlike SQL, annotating QDMR requires zero domain expertise as it is derived entirely from the original question.

3 Weakly Supervised SQL Synthesis

Our input data contains examples of question \( x_i \), database \( D_i \), the answer \( a_i \), and \( s_i \), which is the QDMR of \( x_i \). The QDMR is either annotated or predicted by a trained model \( f \), such that \( f(x_i) = s_i \). For each example, we attempt to synthesize a SQL query \( \hat{Q}_i \), that matches the intent of \( x_i \) and executes to its answer, \( \hat{Q}_i(D_i) = a_i \). The successfully synthesized examples \( \langle x_i, \hat{Q}_i \rangle \) are then used to train an NL-to-SQL model.

3.1 Synthesizing SQL from QDMR

Given QDMR \( s_i \) and database \( D_i \), we automatically map \( s_i \) to SQL. Alg. 1 describes the synthesis process, where candidate query \( \hat{Q}_i \) is incrementally synthesized by iterating over the QDMR steps. Given step \( s^k \), its phrases are automatically linked to columns and values in \( D_i \). Then, relevant join paths are inferred between the columns. Last, each step is automatically mapped to SQL based on its QDMR operator and its arguments (see Table 1).

3.1.1 Phrase DB Linking

As discussed in §2, a QDMR step may have a phrase from \( x_i \) as its argument. When mapping QDMR to SQL these phrases are linked to corresponding columns or values in \( D_i \). For example, in Table 1 the phrases “ships” and “injuries” are linked to columns ship.id and death.injured respectively. We perform phrase-column linking automatically by ranking all columns in \( D_i \) and returning the top one. The ranked list of columns is later used in §3.2 when searching for a correct assignment to all phrases in the QDMR. To compute phrase-column similarity, we tokenize both the phrase and column, then lemmatize their tokens using the WordNet lemmatizer. The similarity score is the average GloVe word embeddings similarity (Pennington et al., 2014) between the phrase and column tokens. All columns in \( D_i \) are then ranked based on their word overlap and similarity with the phrase: (1) we return columns whose lemmatized tokens are identical to those in the phrase; (2) we return columns who share (non stop-word) tokens with the phrase, ordered by phrase-column similarity; (3) we return the remaining columns, ordered by similarity.

We assume that literal values in \( D_i \), such as strings or dates, appear verbatim in the database as they do in the question. Therefore, using string matching, we can identify the columns containing all literal values mentioned in \( s_i \). If a literal value appears in multiple columns, they are all returned as potential links. This assumption may not always hold in practice due to DB-specific language, e.g., the phrase “women” may correspond to the condition | gender = ’F’ . Consequently, we measure the effect of DB-specific language in §4.2.

3.1.2 Join Path Inference

In order to synthesize SQL (Codd, 1970), we infer join paths between the linked columns returned in §3.1.1. Following past work (Guo et al., 2019; Suhr et al., 2020), we implement a heuristic returning the shortest join path connecting two sets of columns. To compute join paths, we convert \( D_i \) into a graph where the nodes are its tables and edges exist for every foreign-key constraint connecting two tables. The JOINP procedure in Alg. 1 joins the tables of

\begin{algorithm}[h]
\caption{SQL Synthesis}
\begin{algorithmic}[1]
\State \textbf{procedure} SQLSYNTH\ \( (\text{QDMR}, D_i ) \rightarrow \text{Q} \)
\State \textbf{for} \( s^k \in s = \langle s^1, \ldots, s^{|s|} \rangle \) do
\State \textbf{cols} ← \text{PHRASECOLUMNLINK}(D, s^k)
\State \textbf{refs} ← \text{REFERENCEDSTEPS}(s^k)
\State \textbf{join} ← [\]
\State \textbf{for} \( s^k \in \text{refs} \) do
\State \textbf{other_cols} ← \text{synth}[j].cols
\State \textbf{join} ← \textbf{join} \cup \text{JOINP}(D, cols, other_cols)
\State \textbf{op} ← \text{OPTYPE}(s^k)
\State \textbf{Q} ← \text{MAXSQL}(\text{op},.cols, join, refs, mapped)
\State \text{mapped}[\text{key}] ← \langle s^k, cols, Q \rangle
\State \textbf{return} \text{mapped[n]}, \text{Q}
\end{algorithmic}
\end{algorithm}

2\url{https://www.nltk.org/}
At this point we have while using parts of the mapped SQL of previous steps, (second column of Table 1). Mapping is performed incrementally for each step, k ∈ P. Each operation has a corresponding SQL using the O-template using the OP-TYPE procedure of Wolfson et al. (2020). Then, following the previous DB linking phase, the arguments of each step are either the linked columns and values or references to previous steps (second column of Table 1). MapSQL uses the step operation type and arguments to automatically map s_k to SQL query $\hat{Q}_k$. Each operation has a unique mapping to SQL as shown in Table 2. The mapping is performed incrementally for each step, while using parts of the mapped SQL of previous steps (stored in the mapped array) when these are referenced. Furthermore, our mapping procedure is able to handle complex SQL that may involve nested queries (Fig. 2) and self-joins (Fig. 3).

### 3.2 Execution-guided SQL Candidate Search

At this point we have $\hat{Q}_1$, which is a potential SQL candidate. However, this candidate may be incorrect due to a wrong phrase-column linking or due to

<table>
<thead>
<tr>
<th>QDMR Step</th>
<th>Phrase-DB Linking</th>
<th>SQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ships 1. SELECT(ship.id)</td>
<td>SELECT ship.id FROM ship;</td>
<td></td>
</tr>
<tr>
<td>2. injuries 2. SELECT(death.injured)</td>
<td>SELECT death.injured FROM death;</td>
<td></td>
</tr>
<tr>
<td>3. number of #2 for each #1 3. GROUP by #2, #1</td>
<td>SELECT COUNT(death.injured) FROM ship, death WHERE death.caused_by_ship_id = ship.id GROUP BY ship.id;</td>
<td></td>
</tr>
<tr>
<td>4. #1 where #3 is highest 4. SUPERLATIVE(max, #1, #3)</td>
<td>SELECT ship.id FROM ship, death WHERE death.caused_by_ship_id = ship.id GROUP BY ship.id ORDER BY COUNT(death.injured) DESC LIMIT 1;</td>
<td></td>
</tr>
<tr>
<td>5. the name of #4 5. PROJECT(ship.name, #4)</td>
<td>SELECT ship.name FROM ship, death WHERE death.caused_by_ship_id = ship.id AND ship.id IN (#4);</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Mapping the QDMR of the question “What is the ship name that caused most total injuries?” to SQL.
query would need to count the number of enrolled students. We account for these structural mismatches by implementing three additional search heuristics which modify the structure of a candidate \( \hat{Q} \). If the candidate executes to the correct result following modification, it is returned by the search process. These modifications are described in detail in Appendix B. Namely, they include the addition of a DISTINCT clause, converting QDMR FILTER steps into SUPERLATIVE and switching between the COUNT and SUM operations.

4 Experiments

Our experiments target two main research questions. First, given access to weak supervision of question-answer pairs and QDMRs, we wish to measure the percentage of SQL queries that can be automatically synthesized. Therefore, in §4.2 we measure the SQL synthesis coverage using 9,313 examples taken from five benchmark datasets. Second, in §4.3 we use the synthesized SQL to train NL-to-SQL models and compare their performance to those trained on gold SQL annotations.

4.1 Setting

Datasets We evaluate both the SQL synthesis coverage and NL-to-SQL accuracy using five NL-to-SQL datasets (see Table 3). The first four datasets contain questions over a single database: ACADEMIC (Li and Jagadish, 2014) has questions over the Microsoft Academic Search database; GEO880 (Zelle and Mooney, 1996) concerns US geography; IMDB and YELP (Yaghmazadeh et al., 2017) contain complex questions on a film and restaurants database, respectively. The SPIDER dataset (Yu et al., 2018) measures domain generalization between databases, and therefore contains questions over 160 different databases. For QDMR data we use the BREAK dataset (Wolfson et al., 2020). The only exception is 259 questions of IMDB and YELP, outside of BREAK, which we (authors) annotate with corresponding QDMRs and release with our code. See Appendix C for license.

Training We fine-tune the T5-large sequence-to-sequence model (Raffel et al., 2020) for both NL-to-SQL and QDMR parsing (§4.2). Namely, for each task we fine-tune the pre-trained model on its specific data. For NL-to-SQL, we fine-tune on mapping utterances \( x_i \); \( cols(D_i) \) to SQL, where sequence \( cols(D_i) \) is a serialization of all columns in database \( D_i \), in an arbitrary order. In QDMR parsing, input questions are mapped to output QDMR strings. We use the T5 implementation by Hugging-Face (Wolf et al., 2020) and train using the Adam optimizer (Kingma and Ba, 2014). Following fine-tuning on the dev sets, we adjust the batch size to 128 and the learning rate to 1e-4 (after experimenting with 1e-5, 1e-4 and 1e-3). All models were trained on an NVIDIA GeForce RTX 3090 GPU.

4.2 SQL Synthesis Coverage

Our first challenge is to measure our ability to synthesize accurate SQL using weak supervision. To evaluate SQL synthesis we define its coverage as the percentage of examples where it successfully produces SQL \( \hat{Q} \) which executes to the correct answer. To ensure our synthesis procedure is domain independent, we test it on examples from five different datasets, spanning across 164 DBs (Table 3).

Annotated QDMRs The upper rows of Table 3 list the SQL synthesis coverage when using manually annotated QDMRs from BREAK. Overall, we evaluate on 9,313 QDMR annotated examples, reaching a coverage of 77.8%. Synthesis coverage for single-DB datasets tends to be slightly higher than that of SPIDER, which we attribute to its larger size and diversity. To further ensure the quality of synthesized SQL, we manually validate a random
subset of 100 queries out of the 7,249 that were synthesized. Our analysis reveals 95% of the queries to be correct interpretations of their original question. In addition, we evaluate synthesis coverage on a subset of 8,887 examples whose SQL denotations are not the empty set. As SQL synthesis relies on answer supervision, discarding examples with empty denotations eliminates the false positives of spurious SQL which incidentally execute to an empty set. Overall, coverage on examples with non-empty denotations is nearly identical, at 77.6% (see Appendix D). We also perform an error analysis on a random set of 100 failed examples, presented in Table 4. SQL synthesis failures are mostly due to QDMR annotation errors or implicit database-specific conditions. E.g., in Geo880 the phrase “major river” should implicitly be mapped to the condition river.length > 750. As our SQL synthesis is domain-general, it does not memorize any domain-specific jargon or rules.

### Predicted QDMRs

While QDMR annotation can be crowdsourced to non-experts (Wolfson et al., 2020), moving to a new domain may incur annotating new in-domain examples. Our first step to address this issue is to evaluate the coverage of SQL synthesis on predicted QDMRs, for out-of-domain questions. As question domains in SPIDER dev differ from those in its training set, it serves as our initial testbed. We further explore this setting in §4.3.4. Our QDMR parser (§4.1) is fine-tuned on BREAK for 10 epochs and we select the model with highest exact string match (EM) on BREAK dev. Evaluating on the hidden test set reveals our model scores 42.3 normalized EM, on par with the state-of-the-art on BREAK. The predicted QDMRs are then used in SQL synthesis together with examples \(\langle x_i, a_i, D_i \rangle\). In Table 3, the last row shows that coverage on SPIDER dev is nearly identical to that of manually annotated QDMRs (77.6% to 77.2%).

### 4.3 Training NL to SQL Models

Next, we compare NL-to-SQL models trained on our synthesized data to training on expert annotated SQL. Given examples \(\langle x_i, D_i \rangle\) we test the following settings: (1) A fully supervised training set, that uses gold SQL annotations \(\langle \langle x_i, Q_i, D_i \rangle \rangle_{i=1}^n\). (2) A weakly supervised training set, where given answer \(a_i\) and QDMR \(s_i\), we synthesize queries \(Q_i\). As SQL synthesis coverage is less than 100%, the process returns a subset of \(m < n\) examples \(\langle \langle x_i, Q_i, D_i \rangle \rangle_{i=1}^m\) on which the model is trained.°

### 4.3.1 Training Data

We train models on two NL-to-SQL datasets: SPIDER (Yu et al., 2018) and Geo880 (Zelle and Mooney, 1996). As our weakly supervised training sets, we use the synthesized examples \(\langle x_i, \hat{Q}_i, D_i \rangle\), described in §4.2, (using annotated QDMRs). We successfully synthesized 5,349 training examples for SPIDER and 547 examples for Geo880 train.

### 4.3.2 Models and Evaluation

#### Models

We fine-tune T5-large for NL-to-SQL, using the hyperparameters from §4.1. We choose T5 as it is agnostic to the structure of its input sequences. Namely, it has been shown to perform competitively on different text-to-SQL datasets, regardless of their SQL conventions (Shaw et al., 2021; Herzig et al., 2021). This property is particularly desirable in our cross-database evaluation (§4.3.3), where we test on multiple datasets.

We train and evaluate the following models:

- **T5-SQL-G** trained on \(\langle \langle x_i, Q_i, D_i \rangle \rangle_{i=1}^n\), using gold SQL, annotated by experts
- **T5-QDMR-G** trained on \(\langle \langle x_i, \hat{Q}_i, D_i \rangle \rangle_{i=1}^m\) with \(\hat{Q}_i\) that were synthesized using weak supervision
- **T5-SQL-Gpart** trained on \(\langle \langle x_i, Q_i, D_i \rangle \rangle_{i=1}^m\), using gold SQL. This models helps us measure the degree to which the smaller size of the synthesized training data and its different query structure (compared to gold SQL) affects performance

°The metric is a strict lower bound on performance.

4https://leaderboard.allenai.org/break

5In practice, we do not train directly on \(\hat{Q}\), but on \(s_i\), following its phrase-column linking. This representation is then automatically mapped to SQL to evaluate its execution.
We begin by comparing the models trained using the following section.

### Evaluation Metric

Due to our SQL being automatically synthesized, its syntax is often different from that of the gold SQL (see Appendix E.2). As a result, the ESM metric of Yu et al. (2018) does not fit our evaluation setup. Instead, we follow Suhr et al. (2020) and evaluate NL-to-SQL models using the execution accuracy of output queries. We define execution accuracy as the percentage of output queries which, when executed on the database, result in the same set of tuples (rows) as $a_t$.

**4.3.3 Training on Annotated QDMRs**

We begin by comparing the models trained using annotated QDMRs to those trained on gold SQL. Meanwhile, the discussion of T5-QDMR-P, trained using predicted QDMRs, is left for §4.3.4. The results in Tables 5-7 list the average accuracy and standard deviation of three model instances, trained using separate random seeds.

#### SPIDER & XSP Evaluation

Tables 5-6 list the results of the SPIDER trained models. All models were trained for 150 epochs and evaluated on the dev set of 1,034 examples. When comparing T5-QDMR-G to the model trained on gold SQL, it achieves 96.8% of its performance (65.8 to 68.0). The T5-SQL-Gpart model, trained on the same 5,349 examples as T5-QDMR-G, performs roughly on par, scoring +0.6 points (66.4 to 65.8).

As SPIDER is used to train cross-database models, we further evaluate our models performance on cross-database semantic parsing (XSP) (Suhr et al., 2020). In Table 6 we test on four additional NL-to-SQL datasets (sizes in parenthesis): ACADEMIC (183), GEO880 (877), IMDB (113) and YELP (66). For ACADEMIC, IMDB and YELP we removed examples whose execution result in an empty set. Otherwise, the significant percentage of such examples would result in false positives of predictions which incidentally execute to an empty set. In practice, evaluation on the full datasets remains mostly unchanged and is provided in Appendix E. Similarly to Suhr et al. (2020), the results in Table 6 show that SPIDER trained models struggle to generalize to XSP examples. However, T5-QDMR-G performance is generally better on XSP examples, which further indicates that QDMR and answer supervision is effective, compared to gold SQL. Example predictions are shown in Appendix E.2.

**4.3.4 Training on Predicted QDMRs**

We extend our approach by replacing the annotated QDMRs with the predictions of a trained QDMR parser (a T5-large model, see §4.1). In this setting, we now have two sets of questions: (1) questions used to train the QDMR parser; (2) questions used to synthesize NL-SQL data. We want these two sets to be as separate as possible, so that training the QDMR parser would not require new in-domain annotations. Namely, the parser must generalize to questions in the NL-SQL domains while being trained on as few of these questions as possible.

#### SPIDER

Unfortunately, SPIDER questions make up a large portion of the BREAK training set, used to train the QDMR parser. We therefore experiment with two alternatives to minimize the in-domain QDMR annotations of NL-SQL questions. First, is to train the parser using few-shot QDMR annotations for SPIDER. Second, is to split SPIDER to...
questions used as the NL-SQL data, while the rest are used to train the QDMR parser.

In Table 5, T5-QDMR-P is trained on 5,075 queries, synthesized using predicted QDMRs (and answer supervision) for SPIDER train questions. The predictions were generated by a QDMR parser trained on a subset of BREAK, excluding all SPIDER questions save 700 (10% of SPIDER train). Keeping few in-domain examples minimizes additional QDMR annotation while preserving the predictions quality. Training on the predicted QDMRs, instead of the annotated ones, resulted in accuracy being down 2.9 points (65.8 to 62.9) while the model achieves 92.5% of T5-SQL-G performance on SPIDER dev. On XSP examples, T5-QDMR-P is competitive with T5-QDMR-G (Table 6).

In Table 8, we experiment with training T5-QDMR-P without in-domain QDMR annotations. We avoid any overlap between the questions and domains used to train the QDMR parser and those used for SQL synthesis. We randomly sample 30-40 databases from SPIDER and use their corresponding questions exclusively as our NL-SQL data. For training the QDMR parser, we use BREAK while discarding the sampled questions. We experiment with 3 random samples of SPIDER train, numbering 1,348, 2,028 and 2,076 examples, with synthesized training data of 1,129, 1,440 and 1,552 examples respectively. Results in Table 8 show that, on average, T5-QDMR-P achieves 95.5% of the performance of T5-SQL-G. This indicates that even without any in-domain QDMR annotations, data induced from answer supervision and out-of-domain QDMRs is effective in training NL-to-SQL models, compared to gold SQL.

**GEO880** For predicted QDMRs on GEO880, we train the QDMR parser on BREAK while discarding all of its 547 questions. Therefore, the parser was trained without any in-domain QDMR annotations for GEO880. SQL synthesis using the predicted QDMRs resulted in 432 queries. In Table 7, T5-QDMR-P reaches 85.7% of T5-SQL-G performance while being trained using question-answer supervision and no in-domain QDMR annotations.

### Table 8: SPIDER models results on the dev set. T5-QDMR-P is trained without using any QDMR annotations for training set questions. We train separate models on the three randomly sampled training sets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Supervision</th>
<th>Train. set</th>
<th>DB #</th>
<th>Exec. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>T5-SQL-G</td>
<td>(x_i, Q_i, D_i)</td>
<td>1,348</td>
<td>30</td>
<td>48.4</td>
</tr>
<tr>
<td>T5-SQL-G___</td>
<td>(x_i, Q_i, D_i)</td>
<td>1,129</td>
<td>30</td>
<td>47.4</td>
</tr>
<tr>
<td>T5-QDMR-P</td>
<td>(x_i, Q_i, D_i)</td>
<td>1,129</td>
<td>30</td>
<td>46.2</td>
</tr>
<tr>
<td>T5-SQL-G</td>
<td>(x_i, Q_i, D_i)</td>
<td>2,028</td>
<td>40</td>
<td>54.7</td>
</tr>
<tr>
<td>T5-SQL-G___</td>
<td>(x_i, Q_i, D_i)</td>
<td>1,440</td>
<td>40</td>
<td>51.3</td>
</tr>
<tr>
<td>T5-QDMR-P</td>
<td>(x_i, Q_i, D_i)</td>
<td>1,440</td>
<td>40</td>
<td>52.1</td>
</tr>
<tr>
<td>T5-SQL-G</td>
<td>(x_i, Q_i, D_i)</td>
<td>2,076</td>
<td>40</td>
<td>56.2</td>
</tr>
<tr>
<td>T5-SQL-G___</td>
<td>(x_i, Q_i, D_i)</td>
<td>1,552</td>
<td>40</td>
<td>53.7</td>
</tr>
<tr>
<td>T5-QDMR-P</td>
<td>(x_i, Q_i, D_i)</td>
<td>1,552</td>
<td>40</td>
<td>53.8</td>
</tr>
</tbody>
</table>

6 Conclusions
This work presents a weakly supervised approach for generating NL-SQL training data, using answer and QDMR supervision. We implemented an automatic SQL synthesis procedure, capable of generating effective training data for dozens of target databases. Experiments on multiple NL-to-SQL benchmarks demonstrate the efficacy of our synthesized training data. Namely, our weakly-supervised models achieve 91%-97% of the performance of fully supervised models trained on annotated SQL. Further constraining our models supervision to few or zero in-domain QDMRs still reaches 86%-93% of the fully supervised models performance. Overall, we provide an effective solution to train text-to-SQL parsers while requiring zero SQL annotations.
References


Xiang Deng, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. 2021. Structure-grounded pretraining for text-to-sql. In NAACL.


Wenqiang Lei, Weixin Wang, Zhixin Ma, Tian Gan, Wei Lu, Min-Yen Kan, and Tat-Seng Chua. 2020. Re-examining the role of schema linking in text-to-sql. In EMNLP.


Irina Saparina and Anton Osookin. 2021. Sparqling database queries from intermediate question decompositions. In EMNLP.


Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In ACL/IJCNLP.


Bailin Wang, Ivan Titov, and Mirella Lapata. 2019. Learning semantic parsers from denotations with latent structured alignments and abstract programs. In EMNLP.


Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir R. Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In Empirical Methods in Natural Language Processing (EMNLP).


A QDMR to SQL Mapping Rules

Table 9 lists all of the QDMR operations along with their mapping rules to SQL. For a thorough description of QDMR semantics please refer to Wolfson et al. (2020).

B SQL Candidate Search Heuristics

We further describe the execution-guided search process for candidate SQL queries, that was introduced in §3.2. Given the search space of candidate queries, we use four heuristics to find candidates $Q_i$ which execute to the correct answer, $a_i$.

1. Phrase linking search: We avoid iterating over each phrase-column assignment by ordering them according to their phrase-column ranking, as described in §3.1.1. The query $Q_i^{(1)}$ is induced from the top ranked assignment, where each
<table>
<thead>
<tr>
<th>QDMR Operation</th>
<th>SQL Mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELECT(t.col)</td>
<td>SELECT t.col FROM t;</td>
</tr>
<tr>
<td>SELECT(val)</td>
<td>SELECT t.col FROM t WHERE t.col = val;</td>
</tr>
<tr>
<td>FILTER(#x, =, val)</td>
<td>SELECT #x(SELECT) FROM #x[FROM] WHERE #x[WHERE] AND t.col = val;</td>
</tr>
<tr>
<td>PROJECT(t.col, #x)</td>
<td>SELECT t.col FROM t, #x[FROM] WHERE Join(t, #x[FROM]) AND #x[SELECT] IN (#x);</td>
</tr>
<tr>
<td>AGGREGATE(count, #x)</td>
<td>SELECT COUNT(#x[SELECT]) FROM #x[FROM] WHERE #x[WHERE];</td>
</tr>
<tr>
<td>GROUP(avg, #x, #y)</td>
<td>SELECT AVG(#x[SELECT]) FROM #x[FROM], #y[FROM] WHERE Join(#x[FROM], #y[FROM]) AND #x[WHERE] AND #y[WHERE] GROUP BY #y[SELECT];</td>
</tr>
<tr>
<td>SUPERLATIVE(max, k, #x, #y)</td>
<td>SELECT #x(SELECT) FROM #x[FROM], #y[FROM] WHERE Join(#x[FROM], #y[FROM]) AND #x[WHERE] GROUP BY #y[SELECT] DESC k;</td>
</tr>
<tr>
<td>COMPARATIVE(#x, #y, &gt;, val)</td>
<td>SELECT #x(SELECT) FROM #x[FROM], #y[FROM] WHERE Join(#x[FROM], #y[FROM]) AND #x[WHERE] AND #y[WHERE] ORDER BY #y[SELECT] &gt; val;</td>
</tr>
<tr>
<td>UNION(#x, #y)</td>
<td>SELECT #x(SELECT), #y(SELECT) FROM #x[FROM], #y[FROM] WHERE Join(#x[FROM], #y[FROM]) AND #x[WHERE] OR #y[WHERE];</td>
</tr>
<tr>
<td>INTERSECT(t.col, #x, #y)</td>
<td>SELECT t.col FROM t, #x[FROM], #y[FROM] WHERE Join(t, #x[FROM], #y[FROM]) AND #x[WHERE] AND #y[WHERE] ORDER BY #y[SELECT];</td>
</tr>
<tr>
<td>PROJECT(t.col, #x)</td>
<td>SELECT t.col FROM t, #x[FROM] WHERE Join(t, #x[FROM]) AND #x[SELECT] IN (#x);</td>
</tr>
<tr>
<td>FILTER(#x, =, val)</td>
<td>SELECT #x(SELECT) FROM #x[FROM] WHERE #x[WHERE] AND t.col = val;</td>
</tr>
<tr>
<td>SELECT(val)</td>
<td>SELECT t.col FROM t;</td>
</tr>
<tr>
<td>AGGREGATE(count, #x)</td>
<td>SELECT COUNT(#x[SELECT]) FROM #x[FROM] WHERE #x[WHERE];</td>
</tr>
<tr>
<td>GROUP(avg, #x, #y)</td>
<td>SELECT AVG(#x[SELECT]) FROM #x[FROM], #y[FROM] WHERE Join(#x[FROM], #y[FROM]) AND #x[WHERE] AND #y[WHERE] GROUP BY #y[SELECT];</td>
</tr>
<tr>
<td>SUPERLATIVE(\text{max}, k, #x, #y)</td>
<td>\text{SELECT } #x(\text{SELECT}) \text{ FROM } #x[\text{FROM}], #y[\text{FROM}] \text{ WHERE } \text{Join}(#x[\text{FROM}], #y[\text{FROM}]) \text{ AND } #x[\text{WHERE}] \text{ GROUP BY } #y[\text{SELECT}] \text{ DESC } k;</td>
</tr>
<tr>
<td>COMPARATIVE(#x, #y, &gt;, val)</td>
<td>\text{SELECT } #x(\text{SELECT}) \text{ FROM } #x[\text{FROM}], #y[\text{FROM}] \text{ WHERE } \text{Join}(#x[\text{FROM}], #y[\text{FROM}]) \text{ AND } #x[\text{WHERE}] \text{ AND } #y[\text{WHERE}] \text{ ORDER BY } #y[\text{SELECT}] &gt; val;</td>
</tr>
<tr>
<td>UNION(#x, #y)</td>
<td>\text{SELECT } #x(\text{SELECT}), #y(\text{SELECT}) \text{ FROM } #x[\text{FROM}], #y[\text{FROM}] \text{ WHERE } \text{Join}(#x[\text{FROM}], #y[\text{FROM}]) \text{ AND } #x[\text{WHERE}] \text{ OR } #y[\text{WHERE}];</td>
</tr>
<tr>
<td>INTERSECT(t.col, #x, #y)</td>
<td>\text{SELECT } t.col \text{ FROM } t, #x[\text{FROM}], #y[\text{FROM}] \text{ WHERE } \text{Join}(t, #x[\text{FROM}], #y[\text{FROM}]) \text{ AND } #x[\text{WHERE}] \text{ AND } #y[\text{WHERE}] \text{ ORDER BY } #y[\text{SELECT}];</td>
</tr>
<tr>
<td>PROJECT(t.col, #x)</td>
<td>\text{SELECT } t.col \text{ FROM } t, #x[\text{FROM}] \text{ WHERE } \text{Join}(t, #x[\text{FROM}]) \text{ AND } #x[\text{SELECT}] \text{ IN } (#x);</td>
</tr>
<tr>
<td>FILTER(#x, =, val)</td>
<td>\text{SELECT } #x(SELECT) \text{ FROM } #x[FROM] \text{ WHERE } #x[WHERE] \text{ AND } t.col = val;</td>
</tr>
</tbody>
</table>

Table 9: QDMR to SQL mapping rules for all QDMR operations. #x denotes a previously mapped SQL query while #x[CLAUSE] denotes its relevant SQL clause. For example, #x[FROM] returns all tables in the FROM clause of SQL query #x. Join, denotes the inferred join paths between sets of tables (see §3.1.2). Note that AGGREGATE and GROUP steps may use the operations: min, max, count, sum and avg. SUPERLATIVE steps may use min, max operations and COMPARATIVE steps use the operations: $\leq, \geq, \geq, \leq$. Last, SORT steps sort in either ascending (asc) or descending (desc) order and ARITHMETIC steps use one of the following: $+, -, \times, \div$.

phrase in $s_i$ is assigned its top ranked column. If $\hat{Q}_i^{(1)}(D_i) \neq a_i$ we continue the candidate search using heuristics 2-4 (described below). Assuming that the additional search heuristics failed to find a candidate $\hat{Q}_i^{(1)}$ such that $Q_i^{(1)}(D_i) = a_i$, we return to the phrase linking component and resume the process using the candidate SQL induced from the following assignment $\hat{Q}_i^{(2)}$, and so forth. In practice, we limit the number of assignments and review only those covering the top-k most similar columns for each phrase in $s_i$, where $k = 20$. Our error analysis (Table 4) reveals that only a small fraction of failures are due to limiting $k$. Step 2 in Fig. 1 represents the iterative process, where $\hat{Q}_i^{(1)}$ executes to an incorrect result while the following candidate $\hat{Q}_i^{(2)}$ correctly links the phrase “authors” to column author.name and executes to $a_i$, thereby ending the search.

2. Distinct modification: Given a candidate SQL $Q_i$ such that $Q_i(D_i) \neq a_i$, we add DISTINCT to its SELECT clause. In Table 10 the SQL executes to the correct result, following its modification.

3. Superlative modification: This heuristic automatically corrects semantic mismatches between annotated QDMR structures and the underlying database. Concretely, steps in $s_i$ that represent PROJECT and FILTER operations may entail an implicit ARGMAX/ARGMIN operation. For example for the question “What is the size of the largest state in the USA?”, in the third row of Table 10. Step (3) of the question’s annotated QDMR is the PROJECT operation, “state with the largest #2”.

While conforming to the PROJECT operation template, the step entails an ARGMAX operation. Using the NLTK part-of-speech tagger, we automatically identify any superlative tokens in the PROJECT and FILTER steps of $s_i$. These steps are then replaced with the appropriate SUPERLATIVE type steps. In Table 10, the original step (3) is modified to the step “#1 where #2 is highest”.

4. Aggregate modification: This heuristic replaces instances of COUNT in QDMR steps with SUM operations, and vice-versa. In Table 10, the question “Find the total student enrollment for different affiliation type schools.”, is incorrectly mapped to a candidate query involving a COUNT operation on university.enrollment. By modifying the aggregate operation to SUM, the new $Q_i$ correctly executes to $a_i$ and is therefore returned as the output.
### C Data License

We list the license (when publicly available) and release details of the datasets used in our paper.

The Break dataset (Wolfson et al., 2020) is under the MIT License. Spider (Yu et al., 2018) is under the CC BY-SA 4.0 License. Geo880 (Zelle and Mooney, 1996) is available under the GNU General Public License 2.0.

The text-to-SQL versions of Geo880 and Academic (Li and Jagadish, 2014) were made publicly available by Finegan-Dollak et al. (2018) in: https://github.com/jkkummerfeld/text2sql-data/.

The IMDB and YELP datasets were publicly released by Yaghmazadeh et al. (2017) in: goo.gl/DbUBMM.

### D SQL Synthesis Coverage

We provide additional results of SQL synthesis coverage. Table 11 lists the coverage results, per dataset, when discarding all examples whose SQL executes to an empty set. Out of the 9,313 original examples, 8,887 examples have non-empty denotations. Coverage scores per dataset remain generally the same as they do when evaluating on all examples. These results further indicate the effectiveness of the SQL synthesis procedure. Namely, this ensures the synthesis results in Table 3 are faithful, despite the potential noise introduced by SQL with empty denotations.

### E NL to SQL Models Results

#### E.1 Evaluation on the Full XSP Datasets

We provide additional results of the models trained on Spider. Namely, we evaluate on all examples of the Academic, IMDB and YELP datasets, including examples whose denotations are empty. Table 12 lists the results of all the models trained on the original training set of Spider. In Table 13 we provide the XSP results of the models trained on the random subsets of Spider train, used in §4.3.4. Similar to our previous experiments, T5-QDMR-P is generally better than T5-SQL-G in terms of its cross-database generalization.

#### E.2 Qualitative Results

Table 14 includes some example predictions of our Spider trained models from Tables 5-6. For each example we describe its question and target (gold) SQL annotation, followed by each model’s result.
Table 12: Model execution accuracy on SPIDER and its performance on cross-database (XSP) examples. Evaluation on ACADEMIC, IMDB and YELP is on the full datasets, including examples with empty denotations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Supervision</th>
<th>Training set</th>
<th>SPIDER dev.</th>
<th>ACADEMIC</th>
<th>GEO880</th>
<th>IMDB</th>
<th>YELP</th>
</tr>
</thead>
<tbody>
<tr>
<td>T5-SQL-G</td>
<td>( (x_i, Q_i, D_i) )</td>
<td>7,000</td>
<td>68.0 ± 0.3</td>
<td>7.9 ± 1.3</td>
<td>33.6 ± 2.5</td>
<td>19.1 ± 2.9</td>
<td>25.3 ± 1.7</td>
</tr>
<tr>
<td>T5-SQL-G$_{\text{part}}$</td>
<td>( (x_i, Q_i, D_i) )</td>
<td>5,349</td>
<td>66.4 ± 0.8</td>
<td>4.9 ± 1.7</td>
<td>32.4 ± 1.3</td>
<td>21.1 ± 0.7</td>
<td>26.1 ± 1.0</td>
</tr>
<tr>
<td>T5-QDMR-G</td>
<td>( (x_i, a_i, s_i, D_i) )</td>
<td>5,349</td>
<td>65.8 ± 0.3</td>
<td>11.2 ± 1.0</td>
<td>40.4 ± 1.8</td>
<td>30.3 ± 3.1</td>
<td>25.8 ± 5.1</td>
</tr>
<tr>
<td>T5-QDMR-P</td>
<td>( (x_i, a_i, D_i) )</td>
<td>5,075</td>
<td>62.9 ± 0.8</td>
<td>8.4 ± 0.9</td>
<td>39.7 ± 1.4</td>
<td>27.0 ± 5.1</td>
<td>28.2 ± 2.9</td>
</tr>
</tbody>
</table>

Table 13: Model results on SPIDER dev when trained on predicted QDMRs versus gold SQL. We train separate models on each of the three randomly sampled training sets. Results include the performance on XSP examples where the evaluation on ACADEMIC, IMDB and YELP is on the full datasets, including examples with empty denotations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Supervision</th>
<th>Train. set</th>
<th>DB #</th>
<th>SPIDER dev.</th>
<th>ACADEMIC</th>
<th>GEO880</th>
<th>IMDB</th>
<th>YELP</th>
</tr>
</thead>
<tbody>
<tr>
<td>T5-SQL-G</td>
<td>( (x_i, Q_i, D_i) )</td>
<td>1,348</td>
<td>30</td>
<td>48.4</td>
<td>2.1</td>
<td>29.6</td>
<td>9.9</td>
<td>22.6</td>
</tr>
<tr>
<td>T5-SQL-G$_{\text{part}}$</td>
<td>( (x_i, Q_i, D_i) )</td>
<td>1,129</td>
<td>30</td>
<td>47.4</td>
<td>2.6</td>
<td>26.9</td>
<td>14.5</td>
<td>16.9</td>
</tr>
<tr>
<td>T5-QDMR-P</td>
<td>( (x_i, a_i, D_i) )</td>
<td>1,129</td>
<td>30</td>
<td>46.2</td>
<td>8.4</td>
<td>29.0</td>
<td>16.0</td>
<td>16.9</td>
</tr>
<tr>
<td>T5-SQL-G</td>
<td>( (x_i, Q_i, D_i) )</td>
<td>2,028</td>
<td>40</td>
<td>54.7</td>
<td>6.3</td>
<td>28.3</td>
<td>18.3</td>
<td>21.0</td>
</tr>
<tr>
<td>T5-SQL-G$_{\text{part}}$</td>
<td>( (x_i, Q_i, D_i) )</td>
<td>1,440</td>
<td>40</td>
<td>51.3</td>
<td>3.7</td>
<td>21.2</td>
<td>12.2</td>
<td>19.4</td>
</tr>
<tr>
<td>T5-QDMR-P</td>
<td>( (x_i, a_i, D_i) )</td>
<td>1,440</td>
<td>40</td>
<td>52.1</td>
<td>6.8</td>
<td>27.4</td>
<td>12.2</td>
<td>18.5</td>
</tr>
<tr>
<td>T5-SQL-G</td>
<td>( (x_i, Q_i, D_i) )</td>
<td>2,076</td>
<td>40</td>
<td>56.2</td>
<td>3.2</td>
<td>25.5</td>
<td>13.0</td>
<td>24.5</td>
</tr>
<tr>
<td>T5-SQL-G$_{\text{part}}$</td>
<td>( (x_i, Q_i, D_i) )</td>
<td>1,552</td>
<td>40</td>
<td>53.7</td>
<td>2.3</td>
<td>17.8</td>
<td>10.2</td>
<td>22.8</td>
</tr>
<tr>
<td>T5-QDMR-P</td>
<td>( (x_i, a_i, D_i) )</td>
<td>1,552</td>
<td>40</td>
<td>53.8</td>
<td>6.1</td>
<td>32.3</td>
<td>19.8</td>
<td>21.8</td>
</tr>
</tbody>
</table>

13
Table 14: Example predictions of the SPIDER trained models from Tables 5-6. We denote correct and incorrect predictions by ✓ and ✗.