XRICL: Cross-lingual Retrieval-Augmented In-Context Learning for Cross-lingual Text-to-SQL Semantic Parsing

Peng Shi[♠], **Rui Zhang**[♣], **He Bai**[♠], and **Jimmy Lin**[♠]

[♠] University of Waterloo [♣] Penn State University

{peng.shi,he.bai,jimmylin}@uwaterloo.ca, rmz5227@psu.edu

Abstract

In-context learning using large language models has recently shown surprising results for semantic parsing tasks such as Text-to-SQL translation. Prompting GPT-3 or Codex using several examples of question-SQL pairs can produce excellent results, comparable to state-of-the-art finetuning-based models. However, existing work primarily focuses on English datasets, and it is unknown whether large language models can serve as competitive semantic parsers for other languages. To bridge this gap, our work focuses on cross-lingual Text-to-SQL semantic parsing for translating non-English utterances into SQL queries based on an English schema. We consider a zero-shot transfer learning setting with the assumption that we do not have any labeled examples in the target language (but have annotated examples in English). This work introduces the XRICL framework, which learns to retrieve relevant English exemplars for a given query to construct prompts. We also include global translation exemplars for a target language to facilitate the translation process for large language models. To systematically evaluate our model, we construct two new benchmark datasets, XSPIDER and XKAGGLE-DBQA, which include questions in Chinese, Vietnamese, Farsi, and Hindi. Our experiments show that XRICL effectively leverages large pre-trained language models to outperform existing baselines. Data and code are publicly available at https://github.com/Impavidity/XRICL.

1 Introduction

Semantic parsing is the task of translating natural language questions into meaning representations such as Lambda CDS (Liang, 2013), Python code (Yin et al., 2018), and SQL (Yu et al., 2018). More recently, Text-to-SQL semantic parsing has attracted attention from academia and industry due to its challenging setup and practical applications. Cross-lingual Text-to-SQL semantic parsing (Sherborne and Lapata, 2022b; Min et al., 2019; Sherborne et al., 2020) aims to translate non-English utterances into SQL queries based on an English schema (assuming we have an internationalized database), enabling users to query databases in non-English languages. For example, such a system could help people from around the world access the US government's open data[1] with natural language questions in different languages.

State-of-the-art approaches for Text-to-SQL semantic parsing have been greatly improved by finetuning pre-trained language models as a sequence-to-sequence problem (Scholak et al., 2021; Yin et al., 2020; Herzig et al., 2020; Yu et al., 2021a,b; Shi et al., 2021a). More recently, in-context learning with large language models (LLMs), such as GPT-3 (Brown et al., 2020) and Codex (Chen et al., 2021), has emerged as a new learning paradigm. This paradigm enables effective few-shot learning without model finetuning, showing its practical and scientific value (Beltagy et al., 2022). Recent papers also have shown promising results applying in-context learning to the Text-to-SQL task. Rajkumar et al. (2022) studied if LLMs are already competitive Text-to-SQL semantic parsers without further finetuning on task-specific training data. Additionally, Poesia et al. (2022) and Rubin et al. (2022) investigated the exemplar retrieval problem for the semantic parsing task.

However, previous work mostly focused on English utterances, leaving other languages behind. It is unclear if LLMs are competitive for cross-lingual Text-to-SQL with English exemplars using in-context learning. Even in the mono-lingual setting (where the exemplars and the query are in the same language), many approaches are not practical beyond English due to the paucity of target language query-SQL exemplars.

To bridge this gap, we propose XRICL, a novel framework based on LLMs with in-context learn-

[1] https://data.gov

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5248–5259

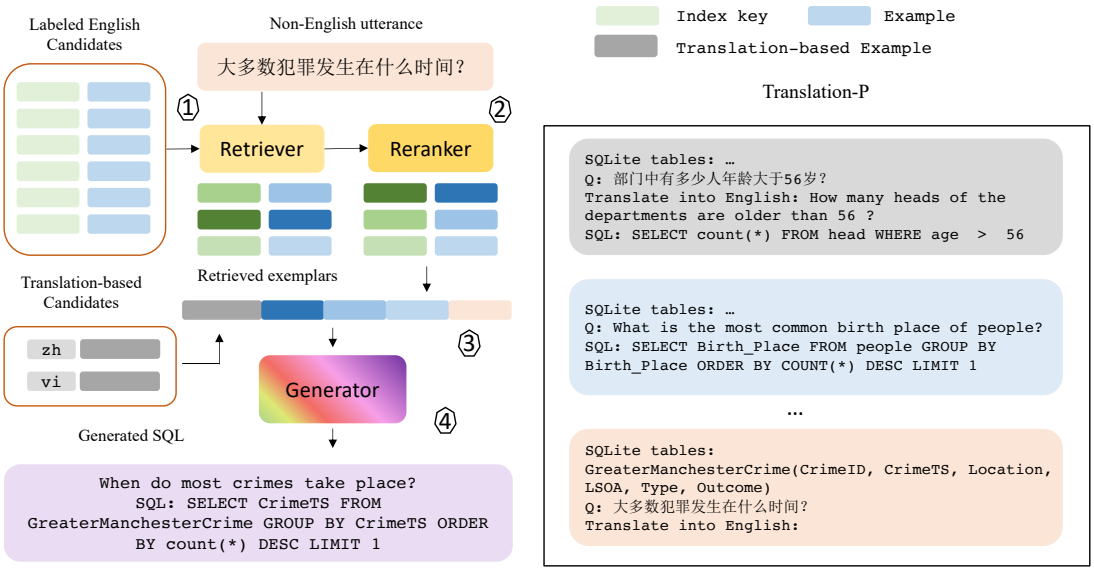

Figure 1: Overview of our proposed XRICL framework. Given a labeled English question-SQL candidate pool and the non-English question as input, our framework uses in-context learning with a large pre-trained language model (e.g., Codex) to generate SQL queries in four steps: (1) Cross-lingual Exemplar Retrieval, (2) Exemplar Reranking, (3) Prompt Construction with Translation as Chain-of-Thought, and (4) Inference.

ing for cross-lingual Text-to-SQL semantic parsing. Specifically, the task is to generate SQL queries for non-English queries based on an English schema and an English query-SQL candidate pool. Our framework first constructs the context prompt by retrieving the most relevant English query-SQL exemplars for each target language query. Since we do not have any training data in the target language, we cannot train a retriever for target queries directly. Our solution is to train an English exemplar retriever with mT5 (Xue et al., 2021) and adopt a model-based cross-lingual transfer method for cross-lingual retrieval. The English exemplar retriever is trained with feedback from the LLM itself by distilling soft labels (likelihood).

Our framework introduces an additional exemplar into the LLM's input context, to instruct the model to translate the target query into English and then to translate the English query into SQL; this approach is inspired by recent work on chain-of-thought prompting (Wei et al., 2022; Shi et al., 2022). However, in our framework, this additional exemplar is identical for all test queries, which means that we only need a single pair of translations for any English-target language pair, requiring minimal translation effort.

During the inference process, the language model is expected to generate the English translation first and then the SQL query. In our experiments, we find that our proposed retriever and reranker can improve the LLMs' cross-lingual few-shot in-context learning performance by a large margin, and further improvements can be observed by adding an additional translation exemplar.

We further construct two benchmarks, XSPIDER and XKAGGLE-DBQA, to systematically evaluate the proposed framework in many languages. For XSPIDER, besides adopting existing work, including CSPIDER (Min et al., 2019) and VSPIDER (Tuan Nguyen et al., 2020), we further translate the SPIDER dataset into Farsi and Hindi for evaluation. For XKAGGLE-DBQA, we translate the English KAGGLE-DBQA dataset into Chinese, Farsi, and Hindi. Experimental results show that our proposed framework improves effectiveness compared to baseline systems.

Our contributions are summarized as follows: (1) We propose a novel retrieve-rerank framework to improve the exemplar selection process for in-context learning for cross-lingual Text-to-SQL semantic parsing. To the best of our knowledge, we are the first to explore the effectiveness of large pre-trained language models for cross-lingual Text-to-SQL semantic parsing. (2) We propose to use translation as a chain-of-thought prompt in the inference process, bridging the cross-lingual gap for large language models. (3) Last, we construct two new benchmarks, XSPIDER and XKAGGLE-DBQA, to facilitate evaluation of cross-lingual Text-to-SQL semantic parsing.

2 Task Formulation

Given a database where the schema s is in English (denoted as the source language), our task is to translate a non-English (denoted the target language) example x (x includes utterance u and schema s) into a SQL query a. In this work, we explore large pre-trained language models such as Codex for this Text-to-SQL task with in-context learning. To support in-context learning, labeled candidates of (utterance, schema, SQL) triples are required. Since more annotated resources are available in English, we assume that the labeled candidate set D is in English. Overall, in-context learning is an efficient method to leverage large pre-trained language models without expensive parameter fine-tuning. Furthermore, the candidate pool can be easily expanded for better generalization to new domains.

3 The XRICL Framework

Our XRICL framework is shown in Figure 1, consisting of four steps:

(1) *Cross-lingual Exemplar Retrieval*: Retrieve a list of N English exemplars that are relevant to the input non-English example x.

(2) *Exemplar Reranking*: Rerank the retrieved N exemplars and use the top K exemplars to construct prompts.

(3) *Prompt Construction with Translation as Chain of Thought*: Construct a prompt consisting of the translation exemplar as a chain of thought, the selected K exemplars, and the input example.

(4) *Inference*: Feed the prompt into a pre-trained language model to generate SQL.

3.1 Cross-lingual Exemplar Retriever

Given a non-English question, the goal of the cross-lingual exemplar retriever is to find *relevant* exemplars from the English candidate pool efficiently that can improve the predictions of the generators. Considering that we use labeled examples in English (a high-resource language) as candidates, we formulate this step as a cross-lingual retrieval problem, where the test question is in a non-English language. In this case, traditional term matching methods such as BM25 (Robertson and Zaragoza, 2009) or BM25 + RM3 query expansion (Lin, 2018) cannot be applied due to token mismatch. Instead, we propose to use a bi-encoder for cross-lingual semantic retrieval with model-based zero-shot transfer. We further improve the retriever with distillation-based training.

Model. Here, we leverage the popular bi-encoder architecture known as dense passage retriever (DPR) (Karpukhin et al., 2020), where the query and candidates are mapped into representation vectors *independently*. The retriever uses a dense encoder $\mathrm{E}_u(\cdot)$ that converts an utterance into a d-dimensional vector and builds an index over the candidate pool that is used for retrieval.

For a test instance x, we use the same dense encoder to map the utterance into a d-dimensional vector (denoted the query vector). Based on the query vector, the closest top N exemplars are retrieved from the pre-built index based on the predefined distance function. Following Karpukhin et al. (2020), we define the distance function as

$$\mathrm{sim}(x, z) = \mathrm{E}_u(x)^\top \mathrm{E}_u(z) \quad (1)$$

where Z is the set of candidate exemplars and $z \in Z$. We use a transformer as the dense encoder, and the average of the contextual embeddings of the utterance tokens is taken as the representation of the encoded text.

Model-based Cross-lingual Transfer. Considering that we do not have training data in target languages, we adopt a model-based cross-lingual transfer method, where we leverage the zero-shot cross-lingual transfer ability of multilingual pre-trained transformers such as mBERT (Devlin et al., 2019), XLM-Roberta (Conneau et al., 2020), mBART (Liu et al., 2020), and mT5 (Xue et al., 2021). Specifically, we train the dense retriever in the source language, where both the query utterance and candidate utterances are in English (in our case), and apply inference directly on query utterances in the target language and retrieve English exemplars in a cross-lingual manner.

Distillation-based Training. One common practice for bi-encoder training is contrastive learning. Given a query, positive examples and negative examples are required. The model is optimized such that examples from the positive class have similar representations and examples from the negative class have different representations.

The key here is how to define positive and negative examples for the semantic parsing task. Recently, Hu et al. (2022) used the similarity of target meaning representations to first rank the candidates and choose the top-k as positive examples and the bottom-k as negative examples. Instead

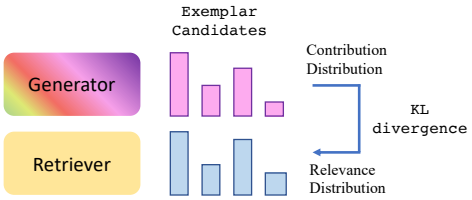

Figure 2: Illustration of distillation-based training. The contribution distribution is the likelihood distribution of the top-N exemplars produced by the LLM. The relevance distribution is the ranking score distribution produced by the retriever.

of using human-designed relevance metrics, Rubin et al. (2022) proposed to use a language model to label positive and negative examples for contrastive learning; similar to Hu et al. (2022), hard labels are used. Another way to train the bi-encoder is to use a regression-based loss function. Poesia et al. (2022) proposed to retrieve exemplars that have relevant program structures (tree edit distance of SQL abstract syntax trees is used as the relevance metric) for the test utterances and the model is optimized with mean-squared error loss for predicting the similarity score.

As an alternative to these above approaches, we train our retriever by distilling the LLM's scoring function. This scoring function calculates the ground-truth SQL query's likelihood given an English exemplar z_k and the input utterance x, which estimates the importance of this exemplar for parsing the given input utterance. Hence, we score the retrieved English exemplars with an LLM and optimize the KL divergence between the LLM's ranking scores and the retriever's ranking scores to update the retriever, as shown in Figure 2. This retriever is denoted DE-Retriever (Distillation-based Exemplar Retriever). Intuitively, with the KL divergence loss function, the model tries to match the probability of retrieving an exemplar z_k with the contribution of that exemplar to the generated SQL query a.

We first obtain N exemplars with the highest scores based on Equation (1), denoted as Z_{top-N}. Our loss function is defined as:

$$\mathcal{L}_{\texttt{distill}} = \texttt{KL}(\ \texttt{SG}(p(z_n \mid x, a, Z_{top-N}; G)) \\ \| \ p(z_n \mid x, Z; E)), \quad (2)$$

where SG denotes the stop gradient operation, G denotes the generator, and E denotes the retriever encoder. We further compute $p(z_n \mid x, a, Z_{top-N}; G)$ as follows:

$$p(z_n \mid x, a, Z_{top-N}) \propto \\ p(a \mid x, z_n, Z_{top-N}; G)\ p(z_n \mid x, Z_{top-N}) \quad (3)$$

We approximate the posterior under the assumption that we have a uniform prior over the set of retrieved exemplars, so $p(z_n \mid x, Z_{top-N})$ is approximated as $\frac{1}{N}$. We further compute $p(a \mid x, z_n, Z_{top-N}; G)$ as:

$$\frac{\exp(p(a \mid x, z_n))}{\sum_{j=1}^{N} \exp(p(a \mid x, z_j))} \quad (4)$$

where $p(a \mid x, z_j)$ is computed with the generator.

More specifically, we use example z_j as the prompt and concatenate it with test instance u and target SQL a. Then we feed it to the generator to compute the log probability of each token $\log(p(a_i))$ in the target SQL query a; $p(a \mid x, z_j)$ can be computed as $\exp(\sum \log(p(a_i)))$.

3.2 Exemplar Reranking

For tasks such as information retrieval and open-domain question answering, reranking is widely adopted to further improve retrieval results by incorporating a reranker. Such a two-stage procedure is also useful in a variety of natural language processing tasks. In this work, following the retrieve-and-rerank idea, we propose to incorporate an exemplar reranker in our framework. This reranker can leverage token-level interactions between the utterances to better rank the exemplars.

More specifically, the query utterance u and the candidate utterance u_z are concatenated together with special tokens: `[CLS]` u `[SEP]` u_z `[SEP]`. The tokenized input is fed into a transformer model. An MLP with sigmoid activation is applied on top of the contextual embedding of the `[CLS]` token to obtain the relevance score of the candidate example (Lin et al., 2021). Sigmoid cross-entropy loss is used and the model is optimized to produce a relevance score as $p(a|x, z_n, Z_{top-N}; G)$. This reranker is denoted DE-Reranker (Distillation-based Exemplar Reranker).

3.3 Prompt Construction with Translation as Chain of Thought

From the input instance x and the list of retrieved-and-reranked exemplars Z, we construct the augmented query by concatenating exemplars with the input instance following previous work (Hu et al., 2022; Rubin et al., 2022; Poesia et al., 2022; Liu

et al., 2022; Brown et al., 2020; Pasupat et al., 2021). For the exemplar, we linearize the table schema, the question, and the SQL query. The exemplars are sorted by relevance score in descending order. For the test instance, only the table schema and the question are linearized. We denote this prompting approach Vanilla-P.

Translation as Chain of Thought: Recent work on chain-of-thought prompting is designed to solve the multi-step reasoning problem by providing intermediate reasoning steps before the final answer in the prompt (Wei et al., 2022). Inspired by this, we use the translation pair (from non-English to English in our case) as an intermediate step for cross-lingual semantic parsing inference.

Specifically, a translation-based exemplar is inserted in front of Z. For example, in the right part of Figure 1, the grey box contains the Chinese version of the translation as a chain-of-thought prompt. The question in the prompt is in the target language, followed by an instruction `Translate into English` and the English translation of the question. Note that this translation-based exemplar is shared among all the test instances in that language, as shown in the left part of Figure 1. The translation-based examples are indexed by the language code, such as `zh` and `vi`. In this way, it only requires minimal translation effort to build the global translation-based exemplar. We denote this prompting approach Translation-P.

3.4 Inference

For inference, we feed the constructed prompt to a large pre-trained language model to generate the target SQL query with greedy decoding. In this work, we consider **Codex** (Codex-Davinci-001) (Chen et al., 2021) because it has shown superior performance for the English Text-to-SQL task (Poesia et al., 2022).

4 Experimental Settings

In this section, we describe the datasets, implementation details, and baselines for our experiments.

4.1 Datasets

We create two benchmarks, XSPIDER and XKAGGLE-DBQA, by translating existing English Text-to-SQL datasets into other languages and evaluate our methods on these two benchmarks.

XSPIDER: CSPIDER (Min et al., 2019) and VSPIDER (Tuan Nguyen et al., 2020) are **C**hinese (zh) and **V**ietnamese (vi) cross-domain Text-to-SQL datasets translated from SPIDER (Yu et al., 2018). More specifically, we use the English SPIDER training set as the candidate pool and training data for retriever-reranker models. We use the development sets of CSPIDER and VSPIDER for cross-lingual evaluation. We further translate the SPIDER development set into Farsi (fa) and Hindi (hi) for a more comprehensive evaluation.

XKAGGLE-DBQA: This is a recently constructed dataset for more realistic and challenging Text-to-SQL evaluation. The dataset is based on 8 databases from Kaggle. We translate the questions into Chinese (zh), Farsi (fa), and Hindi (hi) for cross-lingual evaluation. We use the English SPIDER training set as the candidate pool.

4.2 Experimental Details

For the exemplar retriever, we use 24-layer transformers initialized with the parameters of the mT5 encoder that is then fine-tuned on the English SPIDER dataset for the Text-to-SQL task. For the exemplar reranker, we use InfoXLM (Chi et al., 2021) as the starting point. We train the retriever and reranker on the English SPIDER dataset and then apply both models to cross-lingual retrieval and reranking in a zero-shot fashion. For the Codex configuration, we use greedy decoding by setting the temperature to zero. We use $N = 16$ and $K = 8$ for all experiments, which means that the DE-Retriever first retrieves 16 exemplars from the candidate pool and the DE-Reranker produces the top 8 exemplars for prompt construction.

In terms of evaluation metrics, we use **E**xact **M**atch (EM) accuracy for both the XSPIDER benchmark and the XKAGGLE-DBQA benchmark. Following Zhong et al. (2020), we report the **T**est-**s**uite (TS) accuracy. Only the datasets that are aligned with the SPIDER dev set can be evaluated with TS accuracy, so the XKAGGLE-DBQA benchmark is not applicable. Because the CSPIDER dev set is only partially aligned to the SPIDER dev set, the full CSPIDER (zh-full) dev set can be only evaluated with EM accuracy. We collect a subset of the CSPIDER dev set (zh) whose queries are aligned with the English SPIDER dev set, and further evaluate these using TS accuracy.

4.3 Baselines

mT5 zero-shot transfer is a baseline model that is trained with the English SPIDER training set.

Model	zh-full	zh		vi		fa		hi	
	EM	EM	TS	EM	TS	EM	TS	EM	TS
(1) mT5 zero-shot	39.7	47.9	48.4	42.1	40.1	41.3	39.5	**41.2**	**39.7**
(2) mUSE	38.4	43.0	46.8	31.8	33.4	28.9	31.1	22.2	23.7
(3) mSBERT	37.9	41.3	47.1	34.6	33.5	29.3	31.8	22.0	22.3
(4) mT5-encoder	44.4	48.1	51.4	41.3	39.5	38.4	38.5	28.6	27.0
(5) DE-Retriever	46.0	50.4	53.9	42.2	40.7	38.2	40.0	29.9	27.9
(6) DE-R^2	46.4	52.1	55.3	**44.4**	41.9	40.0	40.6	30.0	28.2
(7) + Translation-P	**47.4**	**52.7**	**55.7**	43.7	**43.6**	**43.2**	**45.1**	**32.6**	**32.4**

Table 1: Results on the XSPIDER dev set. "zh-full" and "zh" are two different splits from CSPIDER (Min et al., 2019). EM and TS are exact match accuracy and test suite accuracy, respectively. Entry (5) is based on the DE-Retriever with Vanilla-P. Entry (6) is based on the DE-Retriever and DE-Reranker (denoted as DE-R^2) with Vanilla-P. Entry (7) is based on DE-R^2 with Translation-P.

The model is based on the pre-trained sequence-to-sequence multilingual language model mT5-large (Xue et al., 2021). This model has zero-shot cross-lingual transfer ability, with which the model can directly handle non-English utterances.

mUSE and mSBERT are baselines that use unsupervised retrievers to obtain exemplars: multilingual Universal Sentence Encoder (Yang et al., 2020) and multilingual Sentence-BERT (Reimers and Gurevych, 2019). Prompts are then constructed for in-context learning with Codex.

5 Results

5.1 Results on XSPIDER

Results on XSPIDER are shown in Table 1. We report the EM and TS accuracy. For the full CSPIDER dataset (zh-full), since TS Accuracy is not supported, we only report EM accuracy. We report both TS and EM accuracy on the subset of CSPIDER. Entry (1) reports the zero-shot performance of the mT5 model that is trained on the English SPIDER dataset. On `zh-full`, `vi`, `fa`, and `hi`, the mT5 zero-shot method obtains on average 41.1 EM accuracy and 39.8 TS accuracy (average TS accuracy is computed without `zh-full` because the metric cannot be computed on the full CSPIDER).

From entry (2) to entry (7), the methods are based on in-context few-shot learning. For entries (2–6), the prompting method is Vanilla-P. For entry (7), prompting with Translation-P is applied.

With unsupervised exemplar retrievers such as mUSE and mSBERT, shown in entries (2) and (3), Codex performs worse than mT5 zero-shot transfer, especially for Farsi (39.5→31.1/31.8 on TS accuracy) and Hindi (39.7→23.7/22.3 on TS accuracy). By switching the unsupervised exemplar retriever to the mT5-encoder, which is the encoder component of the fine-tuned mT5 model, the effectiveness of Codex improves by a large margin. For example, on the CSPIDER subset, TS accuracy improves to 51.4 from 47.1, outperforming mT5 zero-shot performance by 3 points. This indicates that the exemplar retrieval component is essential to take advantage of the competitive performance of LLMs such as Codex. For languages such as Vietnamese and Farsi, Codex is comparable to mT5 zero-shot transfer, while for Hindi, there is still a large gap (39.7 vs. 27.0 on TS accuracy).

By applying our proposed distillation based retriever-reranker pipeline (denoted as DE-R^2) for retrieving exemplars, impressive improvements can be observed in all four languages by comparing entry (6) with entry (4). Our end-to-end results are shown in entry (7), where we see that our proposed framework achieves the best results for most of the languages (except Vietnamese EM accuracy) in the in-context learning setting.

Comparing the best results of in-context learning with mT5 zero-shot results, we can see that Codex can achieve better performance in Chinese, Vietnamese, and Farsi. For example, XRICL outperforms mT5 zero-shot by 7.7 EM accuracy on the full dev set of CSPIDER. One exception is Hindi, where the best in-context learning performance cannot match mT5 zero-shot transfer. One possible explanation is that Codex has weaker modeling ability in Hindi because less Hindi data were accessible during the training.

5.2 Results on XKAGGLE-DBQA

There is agreement by researchers today that XKAGGLE-DBQA is a more realistic evaluation for the Text-to-SQL parsing task. The databases are real-world databases with abbreviated column

Model	zh	fa	hi
(1) mT5 zero-shot	9.7	8.1	7.6
(2) mUSE	20.7	12.4	16.2
(3) mSBERT	14.7	13.0	11.9
(4) mT5-Encoder	22.2	16.8	16.2
(5) DE-Retriever	26.5	18.4	16.8
(6) DE-R^2	27.0	18.4	17.8
(7) + Translation-P	**28.1**	**20.0**	**19.5**

Table 2: Results on the XKAGGLE-DBQA test set. We report exact match (EM) accuracy.

Model	EM	EX	TS
Rubin et al. (2022) (our impl.)	48.5	53.5	50.3
Poesia et al. (2022)	-	60.0	-
Rajkumar et al. (2022)	-	**67.0**	55.1
DE-Retriever (Ours)	53.5	60.3	**56.3**

Table 3: Results on the English SPIDER development set. Our system achieves results comparable to other state-of-the-art in-context learning methods for English Text-to-SQL. EM: Exact Match Accuracy. EX: Execution Accuracy. TS: Test-suite Accuracy (Zhong et al., 2020).

names. We use the training set of English SPIDER as the candidate pool. In this case, both the model's generalization ability and its cross-lingual transfer capability can be tested.

The XKAGGLE-DBQA results are shown in Table 2. Entry (1) shows the zero-shot cross-lingual cross-domain transfer performance of the mT5 model trained on the English SPIDER dataset. For example, on Chinese KAGGLE-DBQA, mT5 only obtains 9.7 EM accuracy. For comparison, mT5 reach 20.0 EM accuracy on the English test set in a zero-shot fashion, outperforming the previous state of the art obtained by RAT-SQL (Wang et al., 2020) with 18.4 EM accuracy (Lee et al., 2021) using column descriptions and model adaptation. This indicates that the mT5 model is more robust than RAT-SQL on domain transfer. However, the effectiveness degrades drastically when mT5 is applied to non-English languages. The mT5 zero-shot method on average obtains only 8.5 EM accuracy in the three languages.

For the Codex-based in-context learning methods, the results are shown in entries (2–7). With unsupervised retrieval methods such as mUSE, Codex can reach 20.7 EM accuracy in Chinese, improving over the zero-shot mT5 baseline. Comparing entries (2) and (3), there is no clear winner for these two unsupervised retrieval methods. Our end-to-end results are shown in entry (7), which achieves state-of-the-art performance on the XKAGGLE-DBQA benchmark, with 22.5 EM accuracy on average, which is better than the mT5 zero-shot method. For example, on Chinese KAGGLE-DBQA, our framework obtains an 18.4 point improvement over mT5 zero-shot transfer.

6 Analysis

6.1 Effectiveness on English Text-to-SQL

We show that our model is comparable to other in-context learning methods for English semantic parsing. Through this comparison, we show that our framework is built on a competitive backbone for Text-to-SQL. We use the DE-Retriever as the backbone model in the ablation study and compare with three recent methods, described as follows: Rubin et al. (2022) used hard labels obtained from the generator to train the retriever. Poesia et al. (2022) used the tree edit distance of SQL queries as a similarity function: a smaller distance means better exemplar quality for the specific test instance. The ranking model is optimized to predict the target SQL pair tree edit distance based on the utterance pair. Rajkumar et al. (2022) designed an efficient prompt that leverages table contents for zero-shot Text-to-SQL. We refer the reader to the original papers for more details.

Table 3 shows the results on the SPIDER development set. Our backbone system (DE-Retriever + Codex Generator) obtains 53.5 EM accuracy and 60.3 EX accuracy, which is comparable to the 60.0 EX accuracy reported by Poesia et al. (2022). Comparing to Rajkumar et al. (2022), our system obtains comparable TS accuracy (56.3 vs. 55.1).

6.2 Effectiveness of DE-R^2

We analyze the effectiveness of DE-R^2 on the XSPIDER benchmark and the XKAGGLE-DBQA benchmark. By comparing entries (5) and (4) in Table 1 and Table 2, we can observe that the DE-Retriever can improve over the mT5-encoder baseline in most of the languages (except EM accuracy in Farsi). Comparing entries (6) and (5), we find that the reranker can further improve the EM accuracy and the TS accuracy. This indicates that our XRICL framework is effective in selecting good exemplars as prompts.

6.3 Effectiveness of Chain-of-Thought Prompt

By comparing entries (7) and (6) in Table 1 and Table 2, we find that Translation-P can further im-

Model	zh-full	zh	
	EM	EM	TS
(1) DE-R^2 + Translation-P	47.4	52.7	55.7
(2) T-Oracle	46.3	52.6	57.6
(3) TG-Oracle	52.5	58.0	62.2

Table 4: Results with oracles: T-Oracle is the Template Oracle and TG-Oracle is the Template+Generator Oracle. EM accuracy and TS accuracy are reported.

prove the semantic parsing ability of Codex on top of DE-R^2, except EM accuracy for Vietnamese.

6.4 Oracle Performance

It is interesting to investigate the upper bound of Codex on cross-lingual Text-to-SQL semantic parsing. We design two pipelines to experiment with the capabilities of Codex when an oracle is available (i.e., the target SQL query is accessible to help the retrieval and reranking). We experiment with two different oracles:

Template Oracle: We retrieve exemplars using the *gold* parse. The template is extracted from the target SQL query and only exemplars with the same SQL template are retrieved. This is based on the assumption that utterances with the same SQL templates share the same query intent and the generator can benefit from these exemplars.

Template Oracle + Codex LM oracle: Here we introduce an oracle from the generator (Codex) into the pipeline. More specifically, we replicate the training process in the testing phase. The exemplars with the same SQL templates are first retrieved. For each retrieved exemplar, we use Codex to compute its contribution to the test instance as the reranking score. We then use the top-k as the exemplars.

The experimental results are shown in Table 4. Comparing entries (1) and (2), we can observe that our XRICL framework can outperform the Template Oracle in terms of EM accuracy on the full dataset and is comparable on the subset. Template Oracle + Codex LM Oracle reaches 52.5 on the full dataset and 58.0 on the subset in terms of EM accuracy. This suggests that signals from the Codex LM are useful and that there is additional room for improvement in our framework.

7 Related Work

In-context Learning: In-context learning is a relatively new paradigm for zero-shot and few-shot learning with large-scale pre-trained language models, first proposed in GPT-3 (Brown et al., 2020). In-context learning for semantic parsing has been intensively investigated recently (Pasupat et al., 2021; Rubin et al., 2022; Shin and Van Durme, 2022; Rajkumar et al., 2022; Hu et al., 2022; Xie et al., 2022; Chen et al., 2021; Poesia et al., 2022). However, most of the work considers only English, without examining the cross-lingual ability of the proposed methods. Winata et al. (2021) evaluated the multilinguality of pre-trained language models on non-English multi-class classification with in-context learning. However, their task is simpler than semantic parsing tasks such as ours. To the best of our knowledge, we are the first to explore cross-lingual Text-to-SQL semantic parsing under the in-context learning setting.

Cross-lingual Semantic Parsing: Cross-lingual semantic parsing aims to handle user utterances from multiple languages and translate them into formal representations. Recent advances can be categorized into two threads: multilingual dataset creation and model development.

For example, Bai et al. (2018) adapted a Chinese dialogue parsing dataset into English. Min et al. (2019) and Tuan Nguyen et al. (2020) adapted the English Text-to-SQL dataset SPIDER (Yu et al., 2018) into Chinese and Vietnamese, which are used in this work for evaluation. Some multilingual datasets with different formal representations have also been created, such as SPARQL (Cui et al., 2022) and TOP (Li et al., 2021).

In terms of model development, Shao et al. (2020) is the most relevant to our work, which leveraged bilingual input for the semantic parsing task. However, they used RNN models and focused on multilingual representation alignment with pre-training. Instead, our work focuses on representation mixup with large multilingual pre-trained models. Improving cross-lingual zero-shot transfer is another direction (Sherborne et al., 2020; Sherborne and Lapata, 2022b,a).

Multilingual and Cross-lingual Retrieval: In multilingual retrieval, the task is to retrieve relevant documents where the user queries and the corpora are in the same language. Recent work takes advantage of cross-language transfer using pre-trained multilingual models (Shi et al., 2020, 2021b; Zhang et al., 2022b, 2021). For example, Shi et al. (2021b) used DPR to retrieve documents based on ad-hoc queries in six languages. On the

other hand, cross-lingual retrievers help users find relevant documents in languages that are different from that of the queries. This task has a long history that goes back several decades (Nie, 2010), but recent work includes Zhang et al. (2022a); Litschko et al. (2022); Sun and Duh (2020). For instance, Asai et al. (2021) created a cross-lingual open-domain question answering dataset where the system is required to retrieve passages from different languages to answer user questions.

8 Conclusion

In this work, we proposed the XRICL framework that improves in-context learning for cross-lingual Text-to-SQL semantic parsing. The retrieve-and-rerank models that we propose can learn signals from large pre-trained models (Codex) to improve the quality of selected exemplars, which can further benefit the generator. By integrating prompts inspired by chain of thought, our proposed Translation-P method can bridge the cross-lingual gap for the generator. Extensive experiments on XSPIDER and XKAGGLE-DBQA demonstrate the effectiveness of our framework, which obtains state-of-the-art performance on few-shot in-context learning in most of the datasets, thus unlocking the potential of Codex.

9 Limitations

Our work is based on the large language model Codex, which is not open-sourced. To replicate our experiments, an application to OpenAI for Codex API access is required. Due to annotation costs, we were unable to evaluate on more languages than those described in this paper. In the future, we plan to collect more data to investigate Codex performance on different language families.

Acknowledgements

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, Compute Ontario, and Compute Canada.

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online.

He Bai, Yu Zhou, Jiajun Zhang, Liang Zhao, Mei-Yuh Hwang, and Chengqing Zong. 2018. Source critical reinforcement learning for transferring spoken language understanding to a new language. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3597–3607, Santa Fe, New Mexico, USA.

Iz Beltagy, Arman Cohan, Robert Logan IV, Sewon Min, and Sameer Singh. 2022. Zero- and few-shot NLP with pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 32–37, Dublin, Ireland.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2022. Compositional generalization in multilingual semantic parsing over Wikidata. *Transactions of the Association for Computational Linguistics*, 10:937–955.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. *arXiv preprint arXiv:2203.08568*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online.

Chia-Hsuan Lee, Oleksandr Polozov, and Matthew Richardson. 2021. KaggleDBQA: Realistic evaluation of text-to-SQL parsers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online.

Percy Liang. 2013. Lambda dependency-based compositional semantics. *arXiv preprint arXiv:1309.4408*.

Jimmy Lin. 2018. The neural hype and comparisons against weak baselines. *SIGIR Forum*, 52(2):40–51.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool Publishers.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Qingkai Min, Yuefeng Shi, and Yue Zhang. 2019. A pilot study for Chinese SQL semantic parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3652–3658, Hong Kong, China.

Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan & Claypool Publishers.

Panupong Pasupat, Yuan Zhang, and Kelvin Guu. 2021. Controllable semantic parsing via retrieval augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7683–7698, Online and Punta Cana, Dominican Republic.

Gabriel Poesia, Alex Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2022. Synchromesh: Reliable code generation from pre-trained language models. In *International Conference on Learning Representations*.

Nitarshan Rajkumar, Raymond Li, and Dzmitry Bahdanau. 2022. Evaluating the text-to-SQL capabilities of large language models. *arXiv preprint arXiv:2204.00498*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901, Online and Punta Cana, Dominican Republic.

Bo Shao, Yeyun Gong, Weizhen Qi, Nan Duan, and Xiaola Lin. 2020. Multi-level alignment pretraining for multi-lingual semantic parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3246–3256.

Tom Sherborne and Mirella Lapata. 2022a. Meta-learning a cross-lingual manifold for semantic parsing. *arXiv preprint arXiv:2209.12577*.

Tom Sherborne and Mirella Lapata. 2022b. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland.

Tom Sherborne, Yumo Xu, and Mirella Lapata. 2020. Bootstrapping a crosslingual semantic parser. *arXiv preprint arXiv:2004.02585*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Peng Shi, He Bai, and Jimmy Lin. 2020. Cross-lingual training of neural models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2768–2773.

Peng Shi, Patrick Ng, Zhiguo Wang, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Cicero Nogueira dos Santos, and Bing Xiang. 2021a. Learning contextual representations for semantic parsing with generation-augmented pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13806–13814.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021b. Cross-lingual training of dense retrievers for document retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 251–253, Punta Cana, Dominican Republic.

Richard Shin and Benjamin Van Durme. 2022. Few-shot semantic parsing with language models trained on code. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5417–5425, Seattle, United States.

Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A pilot study of text-to-SQL semantic parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085, Online.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7567–7578, Online.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, et al. 2022. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online.

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. 2018. StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong. 2021a. GraPPa: Grammar-augmented pre-training for table semantic parsing. In *International Conference on Learning Representations*.

Tao Yu, Rui Zhang, Alex Polozov, Christopher Meek, and Ahmed Hassan Awadallah. 2021b. SCoRe: Pre-training for context representation in conversational semantic parsing. In *International Conference on Learning Representations*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing, pages 3911–3921, Brussels, Belgium.

Fuwei Zhang, Zhao Zhang, Xiang Ao, Dehong Gao, Fuzhen Zhuang, Yi Wei, and Qing He. 2022a. Mind the gap: Cross-lingual information retrieval with hierarchical knowledge enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4345–4353.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022b. Towards best practices for training multilingual dense retrieval models. *arXiv preprint arXiv:2204.02363*.

Ruiqi Zhong, Tao Yu, and Dan Klein. 2020. Semantic evaluation for text-to-SQL with distilled test suites. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 396–411, Online.