Text-to-SQL 前沿技术调研报告

日期：2023 年 2 月 7 日

摘　要

先前的对于 Text-to-SQL 的调研主要基于Text-to-SQL 综述等网络推文以及PaperWithCode中提及的 sota 情况等。对最近领域内的发展并未做明确描述，甚至在多轮任务 Text-to-SQL 的近期发展中做出了错误的推断。基于此背景，本文参考arxiv网站在 Text-to-SQL 领域收录的文章，并从单轮任务和对话式多轮任务两个方向，就最近半年行业论文以及 sota 模型展开调研，总结了 Text-to-SQL 领域的困难，并提出了关于领域内可以改进的方面。

关键词： Text-to-SQL ，领域前沿，技术调研

1　技术概览

当前 Text-to-SQL 领域单轮任务和多轮任务的经典模型多是基于 Seq2Seq 框架，使用较多的是[1]预训练模型 T5[**1**] 与 PICARD 约束过滤模型 [**2**]。除此之外还存在许多论文提出的关系注意力机制 [**3**]、图注意力机制 [**4**] 等，这些注意力机制的嵌入在一定程度上能提高模型的效果。也存在部分模型使用预训练模型 BERT 等编码器，随后在解码器部分根据解码器生成隐藏层生成 AST 语法树，并根据语法树解码生成 SQL。但当前单/多轮任务 sota 模型均使用基于 T5[**1**] 与 PICARD[**2**] 的方式 (单轮任务额外加入了图注意力机制 [**4**])。近半年里，除了对模型方面的研究，该领域还出现了对 Text-to-SQL 领域合成数据 [**5**]、模型评估指标[2]Dr.Spider[**6**] 等研究，下面我将对其进行阐述。

1.1　预训练模型 T5

T5 模型是 Transformer 的 Encoder-Decoder 模型，是谷歌在 Text-to-Text 预训练模型超大规模的探索。其为整个 NLP 预训练模型领域提供了一个通用框架，将都有任务都转化为一种形式，能使用同样的模型，同样的损失函数，同样的训练过程，同样的解码过程完成所有的 NLP 任务。在其论文 Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[**1**] 中提到模型规模的扩大，对模型在下游任务的表现居有提升效果，即使模型参数从 3Billion 扩大到 11Billion，性能提升的间隔仍未发生变缓，因此 T5 模型容量继续提升能使模型下游任务表现得到更进一步的提升[3]。

1.2　PICARD 约束过滤模型

在 Text-to-SQL 领域，针对使用大型语言模型中，SQL 语句输出部分因采用自回归的方式生成，生成的语句缺乏 SQL 语句规范约束，也缺乏与数据库的交互，导致生成的 SQL 语句可能无法执行。为了解决这个问题，Scholak 等人 [**2**] 提出了一种通过增量解析约束语言模型的自回归解码器的方法。通过拼接在

[1]该结论基于本次对arxiv网站 Text-to-SQL 领域近半年内收录的文章。

[2]在先前关于 Text-to-SQL 领域入门调研中提到目前领域中模型的评估指标通常有 Execution with Values 与 Exact Set Match without Values，多轮任务则基本只取 Exact Set Match without Values 分交互准确率和问题回答准确率。

[3]本小节关于 T5 模型的介绍参考自知乎推文，并可以在huggingface上找到并使用该预训练模型

解码器后，以解码器的输出作为输入，通过四种 PICARD 模式设置筛选吸取了 AST 方法的优势，剔除不合规范的 SQL 生成结果。实验结果证明，该模型在单轮与多轮任务中均可使原模型性能有较大提升，并且当前单轮与多轮任务 sota 模型均使用该方法 (详见第二节)。

1.3 关系注意力机制与图注意力机制

两机制均为对注意力机制的改进。其中关系注意力机制[4]提出于 RAT-SQL 模型 [**3**]，在自注意机制 QKV 的计算中注入代表先前存在的关系特征 (表与列之间的关系，自然语言文本与列名，表名的关系等)，将编码器模型偏向于此关系。该方法在目前优秀的多轮任务模型 CQR-SQL[**7**] 中仍有使用。而图注意力机制提出自单轮任务 sota 模型 Graphix-T5[**8**], 该部分实现的方式与关系注意力机制相同，但是因为 Graphix-T5 模型中编码器部分需要融合图注意力机制的结果与原本 T5 模型的编码器结果，因此公式与关系注意力机制有所不同。这两个机制主要目的都是增强模型对数据库的结构信息的学习。

1.4 合成数据的重要性

论文 [**9**] 提出了生成优质合成数据的一种新的方法，解决了合成数据集过程中面临的问题。文中谈到通过更贴近生活使用的合成数据再次训练模型，能使模型在下游任务的效果进一步提升，并且该论文中通过实验证明使用合成数据训练的 T5-3B+PICARD 模型可以达到接近单轮任务 sota 模型 Graphix-T5 的效果。相比同领域合成数据的方法，本文克服了很多合成数据方法的弊端，在使用 SQL 生成 NLQ 时加入具有直观性的 IR 中间层，使生辰的 NLQ 更自然，信息损失最小。

1.5 模型评估指标 Dr.Spider[10]

本文关注 Text-to-SQL 领域模型鲁棒性的检验着力于解决领域泛化性上的问题，指出 Text-to-SQL 模型容易受特定与任务的扰动，提出了一个基于 Spider 的综合鲁棒性基准 (跨域文本到 sql 基准) 来诊断模型的鲁棒性。Dr.Spider[**10**] 旨在综合评估模型在文本到 SQL 任务的每个组件 (即 DB、NLQ 和 SQL) 上的扰动的鲁棒性，对此 Dr.Spider 通过 Spider 的开发集创建针对三个组件的基准测试[5]。通过 RATSQL、SMBOP、T5-LARGE、T5-3B、T5-BLK、PICARD 等一系列模型实验，得出结论：1) 更大规模的模型具有更良好的鲁棒性。2) 解码器中使用先生成 AST 再生成的方法中结合自顶向下和自底向上两种生成方法可以提高模型鲁棒性。3) 开发更好的自然语言问题与数据库内容链接方式能有效提升模型鲁棒性。

2 SOTA 介绍

2.1 单轮任务: Graphix-T5 + PICARD

该模型使用预训练模型 T5 与图注意机制，模型着力解决如何有效将关系结构[6]与如何最大限度地利用预训练模型两问题。GRAPHIX-T5 借鉴 RATSQL[**3**] 中的关系注意机制，在编码层同时进行 T5 模型的 encoder 编码与关系注意机制，并将输出加和得到每层编码结果。在解码层，模型采用 T5 模型的 decoder，后拼接使用 PICARD 模型 [**2**] 生成 SQL 语句。模型不同于 GNN-T5[7]将完整的 GNN 穿插在 T5 模型的 encoder 之后继续编码学习数据库 schema，而是将学习数据库 schema 的过程添加到每一层编码中，不易造成 T5 模型的割裂。

[4]该部分在先前的工作中有介绍，详见 Text-to-SQL 领域入门调研

[5]因为 Spider 数据集的测试集并未公布，再此无法使用。

[6]表示数据库中表名与列名的显示关系，以及自然语言问题与表名列名之间的隐式关系

[7]仅在 Graphix-T5[**8**] 模型论文中进行比较，并未指明出处

2.2 多轮任务: MIGA + PICARD

该模型参考 T5 模型统一多个 NLP 任务的思想，模型将 Text-to-SQL 问题分解为三个子任务，并将其统一到一个模型，同时训练使用的三个辅助任务以添加特殊 token 即插即用的方式进行训练，这使得框架很容易容纳更多的辅助任务。MIGA 模型 [**11**] 整体使用 T5 模型的框架以及参数，基于 Text-to-SQL 领域提出三个预训练任务[8]进行参数更新，最后在 T5 模型 decoder 后添加 PICARD 模型约束生成的 SQL。

3 关于 Text-to-SQL 的思考

当今 Text-to-SQL 领域许多工作都倾向于使用大型预训练模型 T5 以及约束生成 SQL 语句的 PICARD 模型，并取得了优秀的效果。但对于单轮和多轮任务 sota 模型仍然存在可以改进的方面。如都可以从训练数据集的方面改进，可以采用论文 [**9**] 提出的合成带有噪音的数据，以及 Dr.Spider[**10**] 中生成鲁棒性评估数据的方式通过原训练集生成新的训练数据，对模型进行再次训练，增强模型稳健型和鲁棒性。除此之外，该领域内针对单轮以及多轮任务的评价指标依然有优化的空间。

对于多轮任务的 sota 模型 MIGA[**11**]，我认为还可以在其编码器上学习 Graphix-T5[**8**] 的编码层，加入图注意机制。并且，由于 MIGA 中对上下文的存储通过拼接直接作为输入，该方法很容易使输入达到 512 个 tokne 导致出现截断，使得高轮次的回答效果降低，该部分可以借鉴 CQR-SQL 模型 [**7**] 中对前面轮次进行总结的方法。

References

[1] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. Liu, Peter J..Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.. arXiv:1910.10683 [cs.LG]

[2] Scholak, T., Schucher, N. Bahdanau, D.PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models.. arXiv:2109.05093 [cs.CL]

[3] Wang, B., Shin, R., Liu, X., Polozov, O. Richardson, M.RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers.. arXiv:1911.04942 [cs.CL]

[4] Li, J., Hui, B., Cheng, R., Qin, B., Ma, C., Huo, N., Huang, F., Du, W., Si, L. Li, Y.Graphix-T5: Mixing Pre-Trained Transformers with Graph-Aware Layers for Text-to-SQL Parsing.. arXiv:2301.07507 [cs.CL]

[5] Zhao, Y., Jiang, J., Hu, Y., Lan, W., Zhu, H., Chauhan, A., Li, A., Pan, L., Wang, J., Hang, C., Zhang, S., Dong, M., Lilien, J., Ng, P., Wang, Z., Castelli, V. Xiang, B.Importance of Synthesizing High-quality Data for Text-to-SQL Parsing.. arXiv:2212.08785 [cs.CL]

[6] Chang, S., Wang, J., Dong, M., Pan, L., Zhu, H., Li, A., Lan, W., Zhang, S., Jiang, J., Lilien, J., Ash, S., Wang, W., Wang, Z., Castelli, V., Ng, P. Xiang, B.Dr.Spider: A Diagnostic Evaluation Benchmark towards Text-to-SQL Robustness.. arXiv:2301.08881 [cs.CL]

[7] Xiao, D., Chai, L., Zhang, Q., Yan, Z., Li, Z. Cao, Y.CQR-SQL: Conversational Question Reformulation Enhanced Context-Dependent Text-to-SQL Parsers.. arXiv:2205.07686 [cs.CL]

[8] Li, J., Hui, B., Cheng, R., Qin, B., Ma, C., Huo, N., Huang, F., Du, W., Si, L. Li, Y.Graphix-T5: Mixing Pre-Trained Transformers with Graph-Aware Layers for Text-to-SQL Parsing.. arXiv:2301.07507 [cs.CL]

[8]预测生成 SQL 需要使用的列和表：Related Schema Prediction、预测当前 SQL 相对上一个 SQL 做的改变：Turn Switch Prediction 与类似于 CQR，生成整个交互的总结性问题：Final Utterance Prediction

[9] Zhao, Y., Jiang, J., Hu, Y., Lan, W., Zhu, H., Chauhan, A., Li, A., Pan, L., Wang, J., Hang, C., Zhang, S., Dong, M., Lilien, J., Ng, P., Wang, Z., Castelli, V. Xiang, B.Importance of Synthesizing High-quality Data for Text-to-SQL Parsing.. arXiv:2212.08785 [cs.CL]

[10] Chang, S., Wang, J., Dong, M., Pan, L., Zhu, H., Li, A., Lan, W., Zhang, S., Jiang, J., Lilien, J., Ash, S., Wang, W., Wang, Z., Castelli, V., Ng, P. Xiang, B.Dr.Spider: A Diagnostic Evaluation Benchmark towards Text-to-SQL Robustness.. arXiv:2301.08881 [cs.CL]

[11] Fu, Y., Ou, W., Yu, Z. Lin, Y.MIGA: A Unified Multi-task Generation Framework for Conversational Text-to-SQL.. arXiv:2212.09278 [cs.CL]