申请上海交通大学硕士学位论文

SQL 查询语句的自动生成技术研究

论文作者 熊云翔

学　　号 116037910048

导　　师 沈备军

专　　业 软件工程

答辩日期 2018 年 12 月 12 日

Submitted in total fulfillment of the requirements for the degree of Master in Engineering

Research on Automatic Generation Technology of SQL Query Statement

Xiong Yunxiang

Advisor

Prof. Beijun Shen

Software Engineering

Shanghai Jiao Tong University

Shanghai, P.R.China

Dec. 12th, 2018

上海交通大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：_____

日　期：_____年 _____月 _____日

上海交通大学
学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

保　密 □，在 _____ 年解密后适用本授权书。

不保密 □。

（请在以上方框内打 ✓）

学位论文作者签名：_____　　指导教师签名：_____

日　期：_____年 ____月 ____日　　日　期：_____年 ____月 ____日

SQL 查询语句的自动生成技术研究

摘 要

待写 **关键词：**上海交大, 饮水思源, 爱国荣校

RESEARCH ON AUTOMATIC GENERATION TECHNOLOGY OF SQL QUERY STATEMENT

ABSTRACT

write **KEY WORDS:** SJTU, master thesis, XeTeX/LaTeX template

目 录

插图索引

表格索引

算法索引

主要符号对照表

ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数

μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率

第一章 绪论

1.1 研究背景

1.1.1 SQL 查询语句自动生成的重要性和意义

自计算机普及以来，经过数十年的发展，各种系统的数据处理量不断提升。于是，数据库技术应运而生。它是一种建立在计算机存储设备上的仓库，可以将大量数据按照数据结构来组织、存储和管理。SQL（Structured Query Language，结构化查询语句）是一种数据库查询语言，用于存取、查询、更新和管理数据库系统。目前，数据库技术已经广泛的应用于商业和科研领域，通过数据库系统和 SQL，用户可以方便的对大量数据进行各种操作。

数据库系统 DBMS 最早出现在 20 世纪 60 年代，这时随着计算机的普及、数据处理量的增加，计算机系统开始越来越广泛的应用数据管理技术 [1]。然而，传统的文件系统已经逐渐不足以支撑日益增长的需要，于是数据库管理系统 DBMS 应运而生。通过 DBMS，用户可以以不同以往的方式，更统一和集中的管理数据。在数据库系统中，数据的组织，也就是所谓的数据模型是最核心的部分。更详细的说，数据模型是数据库系统中关于数据和联系的逻辑组织的形式表示。而数据库的发展，实际上可以相当程度的体现在数据库应用的数据模型的发展。从时间顺序来看，传统的数据模型包括网状模型和层次模型；接下来关系模型被提出，而且关系型数据库系统也是目前应用最广泛的商业数据库系统；之后随着面向对象思想在计算机技术相关的各大领域的提出与应用，面向对象模型数据库也开始渐渐发展起来 [2]。

SQL（Structured Query Language，结构化查询语言），是一种特定目的程序语言，用于管理关系型数据库管理系统，或在关系流数据管理系统中进行流处理。和关系型数据库一样，SQL 也基于关系代数和元组关系演算。SQL 的分类如下：数据查询语言（DQL）、数据定义语言（DDL）、数据操纵语言（DML）、数据控制语言（DCL）。数据查询语言是 SQL 语言中，负责进行数据查询而不会对数据本身进行修改的语句，这是最基本的 SQL 语句；数据定义语言指 SQL 中负责数据结构定义与数据库对象定义的语言子集；数据操纵语言指 SQL 中对数据库其中的对象和数据运行访问工作的语言子集；数据控制语言，在 SQL 语言中，是一种可对数据访问权进行控制的指令，它可以控制特定用户账户对数据表、查看表、预存程序、用户自定义函数等数据库对象的控制权，也就是权限控制语句。

SQL 是对 E．F．Codd 的关系模型的第一个商业化语言实现 [3]，1986 年 10 月，美

国 ANSI 采用 SQL 作为关系数据库管理系统的标准语言，后为国际标准化组织（ISO）采纳为国际标准。目前，所有主要的关系数据库管理系统支持某些形式的 SQL，大部分数据库至少遵守 SQL 的统一标准。SQL 语言相比于在关系型数据库出现之前、应用于网状数据库和层次数据库的查询语言，具有如下特点：1）一体化，SQL 语句的各个组成部分集数据定义、数据操纵、数据控制、数据查询于一体，可以完成关系型数据库管理中的全部工作；2）使用方式灵活，可以直接以命令方式交互使用，也可以嵌入各种程序语言中使用；3）非过程化，只需将需求表达，不必描述操作步骤，也不需要导航式步骤，大大简化了编程者的操作步骤；4）语言简洁，在 ANSI 标准中只包含了 94 个关键词，核心功能只包含 6 个单词，易用性强。

目前，各种计算机系统，尤其是商业领域，应用关系型数据库系统和 SQL 语言来进行数据管理，仍然是最主流和最成熟的方案。需要注意的是，SQL 作为一种数据库操作语言，本质上仍然是一种编程语言，需要操作人员具有一定的专业知识，经过数据库和 SQL 相关知识的培训，才能比较熟练的进行数据库的管理。此外，除了要具备 SQL 和数据库技术的相关知识，具体到真实数据库的操作，数据管理人员还需要对于所使用的关系型数据库的关系模式有所了解，才能将各种操作需求转化为 SQL 语句，来对数据库系统进行管理。然而，随着数据库系统的应用场景越发广泛和复杂，以及数据库数据处理量的不断提升，数据管理人员对数据库的操作逻辑也越来越复杂，数据库查询需求所涉及的数据量也越来越大，相关的关系模式也越来越复杂和多样化。当数据库管理需求达到这样的复杂度，非专业的数据管理人员就越来越无法满足需求。

综上所述，一方面，随着应用于各种领域的数据库系统越来越多，尤其是在商业领域，数据管理人员的需求越来越大；另一方面，越发复杂的应用场景和日益增多的数据量，对数据管理人员的专业能力和知识也提出了越来越高的要求。在实际的应用场景中，数据管理人员很少具备这一方面的专业知识。这样一来，企业就需要雇佣数据专家，或者投入大量的培训成本对数据管理人员进行培训。然而事实上，面向商业领域的数据库管理，需求提出者和操作人员更关心的是数据本身，以及数据从业务层面考虑所具有的意义。一方面，数据管理需求更高专业性的人才，另一方面，数据管理最本质的目的在于获得数据从业务角度考虑所隐含的信息。基于这样的矛盾，一种可以在现代数据操作人员与数据库和 SQL 技术之间架起桥梁的工具，就显得十分必要了。因此，我们需要设计和实现一个 SQL 查询语句自动生成工具，为数据库使用者提供简单、便捷的接口，将数据库信息映射到业务需求，使其无需了解 SQL 语句的使用方式，只需关注数据操作需求对应的业务需求，也能高效率使用数据库操作数据，弥合业务人员与数据操作之间的矛盾。

要使用这个模板撰写学位论文，需要在 *TeX* 系统、*TeX* 技能上有所准备。

1.1.2 SQL 查询语句自动生成的挑战

1.1.2.1 从 xx 自动生成 SQL 查询语句

自然语言接口（NLI，NaturalLanguageInterfaces），是自然语言处理和人机交互的交叉领域，旨在为人类提供通过自然语言与计算机交互的手段。同时，自然语言接口也是人机交互领域研究的终极目标，从各种对话机器人，到今天各种智能穿戴设备装载的语音助手，人机交互领域和自然语言理解领域的专家们一直在朝建立真正智能的自然语言接口这个目标不断前进。

自然语言处理，包含了许多子问题，如分词、词性标注、句法分析、文本分类、信息抽取等等，在许多领域都有着广泛的应用。自然语言的处理，一开始是基于手工查频，获得自然语言的概率模型，通过马尔科夫随机过程、有限状态机等方式，提取自然语言特征；之后，自然语言处理的语料库开始建立，出现了诸如贝叶斯方法、隐马尔科夫、支持向量机等经典算法，来进行自然语言处理；90 年代以来，基于统计的自然语言处理就开始大放异彩了，补充了传统的基于规则的方法；而目前，自然语言的处理则更多的利用了深度学习，神经网络，受图像处理的启发，将自然语言转化为向量、矩阵等形式，通过不同结构的神经网络来进行处理，已经取得了非常优秀的成果，并且在未来还有巨大的进步空间，这也是目前自然语言处理领域的主流研究方向。

自然语言接口生成 SQL 也是人们关注的一个领域 [6]，自上个世纪提出以来，人们不断研究从自然语言生成 SQL 语句的可能性，并且的确在研究过程中取得了一些令人振奋的成果。通过自然语言接口生成 SQL 的数据库管理系统，原型已经出现在六十年代和七十年代初期，那时候最著名的自然语言接口数据库是 Lunar，正如其名字，包含月球岩石和化学数据库的自然语言界面。这个原型的实现，是基于特定数据库的，因此无法很容易地修改为和不同的数据库一起使用。之后出现了其他的自然语言接口数据库，用户可以通过对话系统来定制查询，并且这些系统可以配置不同的接口，供不同的底层数据库调用。这时候的自然语言数据库系统使用语义语法，是一种句法和语义处理的综合技术。之后，还有关注于将自然语言输入转化为逻辑语言的技术，以此技术作为自然语言接口数据库的核心技术。在自然语言处理领域，随着语义解析技术的发展，数据库的自然语言技术可以利用基于统计的语义解析，将自然语言和 SQL 结构对应起来，完成自然语言到 SQL 语句的转化 [7]。而近期神经网络、深度学习技术的兴起，更为自然语言接口生成 SQL 的发展提供了新思路和新方向。目前，各种深度神经网络结构可以被应用于 SQL 生成问题，来尝试理解自然语言，利用深度学习的方式生成 SQL 语句。具体的方法有，将从自然语言到 SQL 语句的问题视为机器翻译问题，以从序列到序列的方式将自然语言翻译为 SQL 语句 [5][17]；或者利用深度学习的方式优化语义解析技术，加强对自然语言结构的理解能力，等等。基于深度学习的 SQL 语句生成自然语言

接口，所应用的技术包括词向量的训练和表达 [10]、不同结构的网络模型如卷积神经网络、递归神经网络、神经语法解析器等等。

1.1.2.2 从自然语言自动生成 SQL 查询语句

1.2 研究目标和关键问题

1.2.1 研究目标

1.2.2 关键问题

1.2.3 解决方案

1.3 论文结构

第二章 基于映射的 xx2SQL 生成

2.1 研究问题

2.2 相关技术与研究现状

2.3 解决方案

2.4 实验与分析

2.5 本章小结

一直以来，自然语言接口是人机交互领域的终极追求，也是人机交互、机器学习领域专家孜孜不倦钻研学习的热门问题。对于本文所指出的业务层面与技术层面之间的矛盾，如果能对最终用户提供一个自然语言接口，使得我们的系统能直接从最终用户的自然语言中理解到用户的查询意图，并结合数据库，直接生成符合查询意图的 SQL 语句返回给用户，那么这一矛盾就可以非常自然和优美地得到解决。因此，本章将对从自然语言生成 SQL 语句的技术和模型进行探究，尝试寻找一种解决方案，能提供自然语言接口给非技术用户，让用户通过以自然语言表达的查询意图，得到目标 SQL 语句。

自然语言理解存在许多难题，如歧义、语序，或者存在复杂的依赖结构等等，要完全基于自然语言理解来进行 SQL 生成是很困难的，效果可能会不尽如人意。所以，受到人机交互思想的启发，本文将使用自然语言理解与人机交互相结合的方式，来进行从自然语言到目标 SQL 语句的转化。对于自然语言意图，先对自然语言进行初步解析和理解，并在其中插入人机交互机制，让用户来引导生成的过程，指导自然语言理解，纠正机器理解过程中出现的错误、歧义、含糊不清的问题，从而提升整体的准确性。

图 3-1 是本文从自然语言生成 SQL 语句模型的总体方案，它包含依赖解析树生成、解析树节点映射、解析树优化重构、查询树翻译、交互式对话器、用户接口六个模块：1）用户接口：用户与系统进行交互的接口，包括输入自然语言、返回 SQL 语句、解析过程中的交互等等。2）交互式对话器：管理解析过程与用户的交互，在适当的时候与用户进行交互，让用户对解析过程进行指导。3）依赖解析树生成：将用户输入的，以自然语言表示的查询意图，应用自然语言理解技术，转化为依赖解析树，即将词语进行词性标注，以及识别出词语之间的关系，并将其组织成一个树状结构。详见 3.2 节。4）解析树节点映射：根据解析树节点对应的词语和数据库元数据、数据、SQL 语法等信息，将解析树节点映射为 SQL 语法组件。在这个过程中，如果节点的映射有多个候选

答案，交互式对话器会发起与用户的交互，将多个候选映射展示给用户，让用户来进行选择。详见 3.3 节。5）解析树优化重构：节点映射完毕后，解析树节点经过系统匹配和用户指导后，得到了比较准确的结果，但解析树的结构仍是最初由依赖解析树生成得到的结构，这一原始结构的准确度并不高，受限于自然语言的复杂性和省略性，可能会有错误关系、缺失关系、缺失节点等等。为了使解析树能得到比较准确的结构，将设计算法和规则，修正错误关系，补全缺失节点和缺失关系，得到较为准确的解析树，称为查询树。详见 3.4 节。6）查询树翻译：解析树的结构已经符合 SQL 语法，节点映射结果也对应于真实的 SQL 语法、数据库 schema 或数据，可以很自然地将树状结构的翻译为 SQL 语句，最后将 SQL 语句通过用户接口返回给用户。详见 3.5 节。接下来，将详细阐述依赖解析树模块、解析树节点映射模块、解析树优化重构模块、查询树翻译器的设计思路和实现细节。

依赖解析树生成本模块将用户输入的自然语言查询意图解析为解析树，包含各词语的词性标注、关系提取等等信息。在具体实现过程中，这一模块基于 StanfordCoreNLP[8] 实现。StanfordCoreNLP 是斯坦福大学推出的自然语言处理工具集，支持多种语言，还提供了 C++、Python、Java 等多种程序语言的编程接口，提供依赖解析、命名实体识别、词性标注、情感分析、机器翻译等多种功能。本模块调用该工具，将自然语言解析为依赖解析树。

解析树节点映射本模块将依赖解析树的节点对应的词语，映射到 SQL 组件上。解析树节点的类型定义如表 3-1 所示。

表 3-1 解析树节点的类型节点类型对应的 SQL 组件选择节点（SN）SELECT 操作符节点（ON）一个操作符，如等于（“=”）、小于（“<”）聚合函数节点（FN）一个聚合函数，如 AVG、MAX 名字节点（NN）业务数据库中的一个数据表的名字，或数据表的一个字段的名字值节点（VN）业务数据库中某字段的一个值度量节点（QN）ALL，ANY，EACH 逻辑节点（LN）AND，OR，NOT

其中名字节点和值节点是与当前应用的业务数据库有关，其余五种节点都与业务数据库无关，仅与 SQL 语法规则相关。所以，本系统建立了一个五种与业务数据库无关的节点类型与自然语言单词的词典映射。映射过程的实现如下：对每一个解析树节点对应的单词 n，分别计算其与业务数据库元数据、存储数据、词典映射中词语 v 的相似度 Sim(n,v)，Sim(n,v) 的定义如下：

其中为 n 和 v 的 WUP 相似度 [12]，(n,v) 为 n 和 v 的 q-gram 的 Jaccard 相似度的平方根 [13]。经公式计算，可以得到节点单词 n 与所有 SQL 组件的相似度；对相似度进行排序，可以得到前五相似的 SQL 组件，如果前五相似的 SQL 组件的相似度 Sim(n,v) 的值差别较大，则直接以相似度最高的 SQL 组件作为当前单词 n 的映射，并赋予该组件对

应的节点类型；若前五相似的 SQL 组件的相似度差别较小，则视作歧义，将候选的 SQL 组件返回给用户，让用户来进行选择，最后用户选择的结果会作为当前节点的映射。解析树优化重构节点映射完成后，需要对解析树的结构进行重构，保证解析树的结构能有较高的准确性。由于解析树可能会存在关系解析错误和节点关系缺失，所以这一模块对于解析树的优化重构会分为两个步骤进行，分别为结构调整和隐藏节点插入。结构调整在进行结构调整之前，首先定义什么样的树结构是好的、合法的。我们从两个角度来考虑这个问题：第一点是树结构与经依赖解析器解析后的原始结构的差别有多大；第二点是树结构是否符合 SQL 的语法，这一点的评估可以根据表 3-2 的定义来确定 [9]。表 3-2 根据 SQL 语句的语法，结合了树状结构，定义了能合理的转化为 SQL 语句的语法树应该满足什么样的规则，这样的语法树我们称之为查询树。在表 3-2 中，“+”代表父子节点的关系，“*”代表兄弟节点的关系，上标“*”代表可重复出现的兄弟节点，“|”代表“或者关系”，表示当前节点可能存在的情况。

表 3-3 展示了结构调整的算法，算法的基本思想是，建立一个优先级队列，对于当前的解析树，调用 adjust() 函数（第 8 行），通过一次移动子树操作，生成在这一次操作后所有可能的结构；然后记录当前树的哈希值（第 12 行），防止之后出现重复的树结构；若当前树结构没有出现过（第 10 行），且 edit 值小于一个阈值，则对此树进行下一步操作；由于移动了一次子树，返回的树的 edit 属性加一（第 11 行），这一属性将用来评估生成的树结构与原始结构的差异；调用 evaluate() 函数，记录当前结构有多少节点不满足表 3-2 设定的规则，不满足规则的节点数将用来评估树结构在语法上的合法性；综合这两方面评估标准，为树打分，如果分数比之前的树结构要高，则加入优先级队列；若该树完全符合语法规则，则视为一颗查询树，加入 result 集合；之后对优先级队列内的树结构重复以上操作，直到优先级队列为空；最后根据评估分数对 result 集合排序，将结果返回。

结构调整之后的解析树结果集，将会通过交互式对话器，与用户进行交互，因为结果集中的解析树虽然都符合 SQL 语法规则，但仍有可能存在与用户意图不同的情况，如 SELECT 子句中的名字节点和 WHERE 子句中的名字节点，位置可能会互换，虽然仍然符合 SQL 语法规则，但与原始查询意图已经有比较大的差别了。所以，在这里需要再一次应用人机交互机制，让用户来选择结果集中与自己的查询意图比较相似的解析树。交互完成后，筛选出的结果集，会进入下一步——隐藏节点插入。隐藏节点插入结构调整完成后，对经过排序和用户交互筛选后的结果集合，进行隐藏节点插入。在给出隐藏节点插入的方法之前，先给出需要用到的概念定义，即“核心节点”。核心节点指的是在节点类型为 leftSubtree 和 rightSubtree 的情况下（表 3-2），leftSubtree（rightSubtree）的所有子节点中，高度最高的名称节点被称为核心节点。经过研究，需要进行隐藏节点插入的

情况有以下几种：1）左子树（leftSubtree）与右子树（rightSubtree）的核心节点对应了不同的 SQL 组件，即认为右子树真正的核心节点在自然语言表达时被省略了 [14]；这是十分常见的现象，因为人在进行自然语言表达时，对两个值进行比较时，会很自然的省略掉后者的一部分，如“Ihavemorebooksthanyours”这句话，就将最后的“yourbooks”给省略成了“yours”；2）左右子树的约束条件应该一致，如果不一致，则认为右子树一部分约束条件被省略了，如“returnauthorswhosepaperspublishedin2018morethanJack’s”这句话，过滤条件的左子树有“in2018”这一约束，而右子树在解析之后没有这一约束，事实上右子树的这一约束被隐藏了，需要作为隐藏节点插入进去；3）某些函数会被省略，如聚合函数“COUNT”，在自然语言表达中经常会被省略，如“Ihavemorebooksthanyours”这句话，“thenumberof”就被省略了。在树结构中，如果过滤条件的操作符为“more”、“less”等词语，而左右子树的核心节点对应的是非数字类型的 SQL 组件，那么就认为“COUNT”被省略了，需要作为隐藏节点插入解析树。进行隐藏节点插入之后，查询树的结构就比较完整了，将会输入下一模块进行翻译。查询树翻译这一阶段的查询树已经在节点映射、树结构、完整性方面都比较可信、合法了，翻译步骤如下：1）根据树结构，在 SClause 子树下的结构为 SELECT 子句，读取 SClause 下的名称节点，根据对应的 SQL 组件（如果 SQL 组件对应某数据表，该数据表会预定义一个核心字段，如用户的名字、城市的名称，该数据表的核心字段将作为结果返回），填充入 SELECT 子句，并记录 SQL 组件对应的数据表，以备 FROM 子句的生成 2）在 ComplexCondition 子树下的结构为 WHERE 子句，分别读取左子树和右子树核心节点对应的 SQL 组件，记录下 SQL 组件对应的数据表，并查看左右子树中所有的节点，根据其节点类型，将其翻译为对应的 SQL 组件，并应用于核心节点；左右子树解析完后，以操作符连接左右子树，将其填充入 WHERE 子句 3）根据之前记录的相关数据表，生成 SQL 语句的 FROM 子句 4）将三部分按照语法连接起来，作为合法的 SQL 语句返回给用户。

数据集本次实验所使用的业务数据库为 MySQL 的样例数据库 Classicmodels，图 3-2 为 Classicmodels 数据库的数据库模式图

实验所使用的自然语言查询数据集由作者根据 Classicmodels 数据库的模式建立，根据查询意图的复杂度，分为简单、中等、困难三类。每种类别，提出了 20 条自然语言查询，共计 60 条，用来测试模型的准确性。表 3-4 给出了三种类别的自然语言查询示例。

结果分析 1）经过两次实验，分别检验完整模型和无交互机制模型生成 SQL 语句的准确性。可以看出，交互机制对于模型的准确性有很大的提升；而加入了交互机制后，模型可以比较准确的处理简单和中等复杂度的查询意图，即便是复杂度较高的查询意图，本文提出的模型仍然可以正确生成一部分困难复杂度的 SQL 语句。在实验过程中，经常出现节点歧义需要映射，如“price”一词可以映射到“products”表中的“buyPrice”，也

有可能映射到“orderdetails”表中的“priceEach”。那么对于测试语句“return order details whose price is higher than 50”，如果没有交互机制，对于“price”这个节点的映射就会优先映射到“products”中的“buyPrice”，而实际上应该映射到“orderdetails”表中的“priceEach”。可以看出，交互机制在节点映射这一部分准确率的提升，大大影响了整体模型的准确率。2）在作者构造的自然语句查询意图数据集中，简单复杂度的查询意图大体上是返回某数据表中某一字段，中等复杂度的查询意图会添加一些聚合函数和简单过滤条件，困难复杂度的意图会增加更多的聚合函数和跨表查询，总体上来说结构都比较简单。在表 3-5 所示的结果中，困难复杂度意图的错误生成情况，一部分来源于查询意图对应的 SQL 语句需要子查询，如“return the customer who has the most orders”。目前的模型还不能很好的处理这一种情况，在解析树语法和翻译过程中都还没考虑子查询的情况，这可能是接下来需要进一步进行的工作，即提升模型的处理能力，使其能处理更复杂的查询意图。3）目前模型基于相似度的节点映射机制仍然有不足之处，如“return the mobile number of custormer whose name is Australian Gift Network”这一查询意图，对于顾客名称“Australian Gift Network”，系统会将其映射为三个不同的节点，导致生成出错。尽管模型对于数据库模式中的一些连接词或短语，如“thenumberof”、“customername”，进行了特殊处理，但对于上文所示的这些特殊短语或词语，目前没有较好的方法来进行处理。4）本文自行建立了一个解析树节点类型与自然语言单词的映射词典（3.3 节），能处理一些常见的单词与 SQL 语法的对应关系，如“return”、“in”、“have”、“thenumberof”，但映射关系仍然不足，所以本实验所使用的测试查询意图都需要使用这些比较固定的词语来构建。事实上，自然语言的表达非常多样，如果要增加模型的处理能力，扩充这个映射词典也是非常必要的。5）尽管人机交互机制对于节点映射的准确率有较大提升，但根据观察，立足于目前数据量较小、数据库模式简单的前提下，这一机制能保证映射相似度排名前五的 SQL 组件包含正确的映射关系；但随着数据库数据量的增加、数据库模式的复杂化，目前节点映射的相似度计算机制，可能无法确保正确的 SQL 组件能有较高的相似度。

第三章 基于深度强化学习的 NL2SQL 生成

第四章 基于多任务学习的 NL2SQL 生成

4.1 研究问题

4.2 相关技术与研究现状

4.3 解决方案

4.4 实验与分析

4.5 本章小结

Machine ComprehensionQuestion answering (QA) mod-els receive a question and a context that contains informationnecessary to output the desired answer. We use the StanfordQuestion Answering Dataset (SQuAD) [] for this task. Con-texts are paragraphs taken from the English Wikipedia, andanswers are sequences of words copied from the context.SQuAD uses a normalized F1 (nF1) metric that strip outarticles and punctuation.Machine Translation.Machine translation models receivean input document in a source language that must be trans-lated into a target language. We use the 2016 English to

German training data prepared for the International Workshopon Spoken Language Translation (IWSLT) []. Examples arefrom transcribed TED presentations that cover a wide varietyof topics with conversational language. We evaluate with acorpus-level BLEU score [] on the 2013 and 2014 test sets asvalidation and test sets, respectively.Natural Language Inference.Natural Language Infer-ence (NLI) models receive two input sentences: a premiseand a hypothesis. Models must then classify the inferencerelationship between the two as one of entailment, neutrality,or contradiction. We use the Multi-Genre Natural LanguageInference Corpus (MNLI) [] which provides training examplesfrom multiple domains (transcribed speech, popular fiction,government reports) and test pairs from seen and unseen

domains. MNLI uses an exact match (EM) score. SentimentAnalysis. Sentiment analysis models are trained to classifythe sentiment expressed by input text. The Stanford Senti-ment Treebank (SST) [] consists of movie reviews with thecorresponding sentiment (positive, neutral, negative). We usethe unparsed, binary version []. SST also uses an EM score.Semantic Parsing.SQL query generation is related tosemantic parsing. Models based on the WikiSQL dataset[] translate natural language questions into structured SQLqueries so that users can

interact with a database in naturallanguage. WikiSQL is evaluated by a logical form exact match(lfEM) to ensure that models do not obtain correct answersfrom incorrectly generated queries.Summarization.Summarization models take in a documentand output a summary of that document. Most important torecent progress in summarization was the transformation ofthe CNN/DailyMail (CNN/DM) corpus [Hermann et al., 2015]into a summarization dataset [Nallapati et al., 2016]. We in-clude the non-anonymized version of this dataset in decaNLP.On average, these examples contain the longest documentsin decaNLP and force models to balance extracting from thecontext with generation of novel, abstractive sequences ofwords. CNN/DM uses ROUGE-1, ROUGE-2, and ROUGE-L scores [Lin, 2004]. We average these three measures tocompute an overall ROUGE score.Sentiment Analysis.Sentiment analysis models are trainedto classify the sentiment expressed by input text. The StanfordSentiment Treebank (SST) [Socher et al., 2013] consists ofmovie reviews with the corresponding sentiment (positive,neutral, negative). We use the unparsed, binary version [Rad-ford et al., 2017]. SST also uses an EM score.Semantic Role Labeling.Semantic role labeling (SRL)models are given a sentence and predicate (typically a verb)and must determine who did what to whom, when, and where[Johansson and Nugues, 2008]. We use an SRL dataset thattreats the task as question answering, QA-SRL [He et al.,2015]. This dataset covers both news and Wikipedia domains,but we only use the latter in order to ensure that all datafor decaNLP can be freely downloaded. We evaluate QA-SRLwith the nF1 metric used for SQuAD.Relation Extraction.Relation extraction systems take in apiece of unstructured text and the kind of relation that is tobe extracted from that text. In this setting, it is important thatmodels can report that the relation is not present and cannot beextracted. As with SRL, we use a dataset that maps relations toa set of questions so that relation extraction can be treated asquestion answering: QA-ZRE [Levy et al., 2017]. Evaluationof the dataset is designed to measure zero shot performanceon new kinds of relations the dataset is split so that relationsseen at test time are unseen at train time. This kind of zero-shot relation extraction, framed as question answering, makesit possible to generalize to new relations. QA-ZRE uses acorpus-level F1 metric (cF1) in order to accurately accountfor unanswerable questions. This F1 metric defines precisionas the true positive count divided by the number of timesthe system returned a non-null answer and recall as the true

positive count divided by the number of instances that havean answer.Goal-Oriented Dialogue.Goal-Oriented Dialogue. Dia-logue state tracking is a key component of goal-orienteddialogue systems. Based on user utterances, actions takenalready, and conversation history, dialogue state trackers keeptrack of which predefined goals the user has for the dia-

loguesystem and which kinds of requests the user makes as thesystem and user interact turn-by-turn. We use the EnglishWizard of Oz (WOZ) restaurant reservation task [Wen et al.,2016], which comes with a predefined ontology of foods,dates, times, addresses, and other information that would helpan agent make a reservation for a customer. WOZ is evaluatedby turn-based dialogue state EM (dsEM) over the goals of thecustomers.Semantic Parsing.SQL query generation is related tosemantic parsing. Models based on the WikiSQL dataset[Zhong et al., 2017] translate natural language questions intostructured SQL queries so that users can interact with adatabase in natural language. WikiSQL is evaluated by alogical form exact match (lfEM) to ensure that models donot obtain correct answers from incorrectly generated queries.Pronoun Resolution.Our final task is based on Winogradschemas [Winograd, 1972], which require pronoun resolu-tion: ”Joan made sure to thank Susan for the help she had[given/received]. Who had [given/received] help? Susan orJoan?”. We started with examples taken from the WinogradSchema Challenge [Levesque et al., 2011] and modified themto ensure that answers were a single word from the context.This modified Winograd Schema Challenge (MWSC) ensuresthat scores are neither inflated nor deflated by oddities inphrasing or inconsistencies between context, question, andanswer. We evaluate with an EM score.The Decathlon Score (decaScore).Models competing ondecaNLP are evaluated using an additive combination of eachtask-specific metric. All metrics fall between 0 and 100, so thatthe decaScore naturally falls between 0 and 1000 for ten tasks.Using an additive combination avoids issues that arise fromweighing different metrics. All metrics are case insensitive.As shown in Table II. All metrics are case insensitive. nF1is the normalized F1 metric used by SQuAD that strips outarticles and punctuation. EM is an exact match comparison:for text classification, this amounts to accuracy; for WOZ itis equivalent to turn-based dialogue state exact match (dsEM)and for WikiSQL it is equivalent to exact match of logicalforms (lfEM). F1 for QA-ZRE is a corpus level metric (cF1)that takes into account that some question are unanswerable.Precision is the true positive count divided by the number oftimes the system returned a non-null answer. Recall is the truepositive count divided by the number of instances that havean answer.

Because every task is framed as question answering andtrained jointly, we call our model a multitask question answer-

ing network (MQAN). Each example consists of a context,question, and answer as shown in Fig. 1. Many recent QAmodels for question answering typically assume the answercan be copied from the context [Wang and Jiang, 2017, Seoet al., 2017, Xiong et al., 2018], but this assumption doesnot hold for general question answering. The question oftencontains key

information that constrains the answer space.Noting this, we extend the coattention of [Xiong et al., 2017]to enrich the representation of not only the input but also thequestion. Also, the pointer-mechanism of [See et al., 2017]is generalized into a hierarchical, multi-pointer-generator thatenables the capacity to copy directly from the question andthe context.During training, the MQAN takes as input three sequences:a context c with l tokens, a question q with m tokens, andan answer a with n tokens. Each of these is represented by amatrix where the ith row of the matrix corresponds to a demb-dimensional embedding (such as word or character vectors) forthe ith token in the sequence:

Encoder: xxxxx Decoder: xxxxx

第五章 总结与展望

5.1 本文工作小结

5.2 展望

致 谢

感谢所有测试和使用交大学位论文 LaTeX 模板的同学！
感谢那位最先制作出博士学位论文 LaTeX 模板的交大物理系同学！
感谢 William Wang 同学对模板移植做出的巨大贡献！
感谢 @weijianwen 学长一直以来的开发和维护工作！
感谢 @sjtug 以及 @dyweb 对 0.9.5 之后版本的开发和维护工作！
感谢所有为模板贡献过代码的同学们, 以及所有测试和使用模板的各位同学！

攻读学位期间发表的学术论文

[1] CHEN H, CHAN C T. Acoustic cloaking in three dimensions using acoustic metamaterials[J]. Applied Physics Letters, 2007, 91:183518.

[2] CHEN H, WU B I, ZHANG B, et al. Electromagnetic Wave Interactions with a Metamaterial Cloak[J]. Physical Review Letters, 2007, 99(6):63903.

攻读学位期间参与的项目

[1] 973 项目“XXX”

[2] 自然基金项目“XXX”

[3] 国防项目“XXX”

简 历

基本情况

某某，yyyy 年 mm 月生于 xxxx。

教育背景

yyyy 年 mm 月至今，上海交通大学，博士研究生，xx 专业

yyyy 年 mm 月至 yyyy 年 mm 月，上海交通大学，硕士研究生，xx 专业

yyyy 年 mm 月至 yyyy 年 mm 月，上海交通大学，本科，xx 专业

研究兴趣

LaTeX 排版

联系方式

地址：上海市闵行区东川路 800 号，200240

E-mail: xxx@sjtu.edu.cn