分 类 号 TP391　　密　　级
U D C　　编　　号 10486

武漢大學

硕 士 专 业 学 位 论 文

跨域知识图谱的知识表示学习研究

研 究 生 姓 名：杨世杰
学　　号：2021282110124
指导教师姓名、职称：彭敏 教授
专 业 类 别（领 域）：人工智能（自然语言处理）

二〇二三年四月

A Research on Knowledge Representation Learning in Cross-domain Knowledge Graphs

Candidate：Yang Shijie

StudentNumber：2021282110124

Supervisor：Prof.Peng Min

Major：Artificial Intelligence

Speciality：Natural Language Processing

School of Computer Science

Wuhan University

April, 2023

论文原创性声明

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

学位论文作者（签名）：

年　　月　　日

摘　　要

随着知识图谱在工业界的广泛应用，众多基于知识图谱的应用进入了人们的视野，如电商系统、医疗健康管理系统等，其中存储的结构化知识构成了各系统的源域知识图谱。为了充分利用图谱中的知识，知识表示学习将图谱中的实体和关系嵌入到低维连续空间，保留其语义信息的同时有助于各种下游任务的进行。随着知识图谱数据规模的增长以及在个人电脑、手机等移动设备上的应用，知识图谱的处理和存储开始逐渐分布在多个不同的终端上。上述终端中的目标域知识图谱通常是源域知识图谱的一个子集，在用户使用过程中会不断引入其他的实体和关系。然而，这些引入的实体和关系可能包含源域知识图谱实体集合和关系集合中未定义的、新的未见实体和关系，面临着零样本（Zero-Shot）的问题。因此，如何在源域知识图谱上进行训练，并将表示学习能力泛化到包含未知实体和关系的目标域知识图谱中，是非常重要的研究问题。

跨域知识表示学习面临着两个关键问题：（1）当前处理跨域知识图谱的表示学习模型多采用图结构信息来学习未见实体和未见关系的表示，未能充分利用知识图谱的语义信息；（2）对于跨域知识图谱，如何将源域知识图谱上的元知识迁移到目标域图谱，以实现对未见实体和关系的泛化。

本文针对这两个问题，开展面向跨域知识图谱的知识表示学习研究，提出了基于本体信息和元学习的知识表示学习方法，主要包括：对于目标域知识图谱上出现的新关系，构建了一个以关系为节点的视图，同时建模关系拓扑结构和本体语义信息两个方面的特征；对于目标域知识图谱上新出现的实体，通过对实体的邻接关系特征聚合获得其初始化表示，并采用图神经网络，充分利用知识图谱中已知实体和关系的信息，学习和更新整个目标域的实体和关系的向量表示；在模型训练过程中，采用元学习方法划分任务并通过标签模拟跨域场景中的新实体和新关系，从而完成从源域到目标域的知识表示学习任务。

本文基于链接预测任务对提出的模型进行验证，并与多个基准模型进行比较。实验结果证明了本文提出的基于元学习本体增强的跨域知识表示学习模型的有效性。

关键词：知识表示学习；本体嵌入；元学习；图神经网络；归纳推理

ABSTRACT

With the widespread application of knowledge graphs in the industry, numerous knowledge graph-based applications have entered people's vision, such as e-commerce systems, medical and health management systems, etc. The structured knowledge stored in these applications forms the source domain knowledge graph. In order to fully utilize the knowledge in the graph, knowledge representation learning embeds the entities and relationships in the graph into low-dimensional continuous space while preserving their semantic information, which is helpful for various downstream tasks. With the growth of knowledge graph data scale and its application on personal computers, mobile devices, etc., processing and storage of knowledge graphs are gradually distributed across multiple different nodes. The target domain knowledge graph in the above-mentioned nodes is usually a subset of the source domain knowledge graph and constantly introduces other entities and relationships during the user's usage. However, these introduced entities and relationships may contain new and undefined entities and relationships that are not defined in the source domain knowledge graph entity set and relationship set, and we cannot completely cover all the newly added entities and relationships in the dispersed knowledge graph. Therefore, how to train on the source domain knowledge graph and generalize the representation learning ability to the target domain knowledge graph containing unknown entities and relationships has become an inevitable cross-domain knowledge representation problem.

To solve this problem, this paper employs the algorithmic idea of meta-learning and simulates the unseen entities and relationships on the target domain knowledge graph through labels in the training task to help the model acquire the ability to process new entities and relationships on the target domain graph. Meanwhile, ontology information is introduced to provide semantic support for the representation learning of unseen entities and relationships. This paper mainly solves two problems in cross-domain knowledge representation learning: (1) the current representation learning models for cross-domain knowledge graphs mostly use graph structure information to learn the representation of unseen entities and relationships, and do not fully utilize the semantic information of the knowledge graph; (2) for cross-domain knowledge graphs, how to migrate the meta-knowledge on the source domain knowledge graph to the target domain graph to achieve generalization of unseen entities and relationships.

This paper proposes a knowledge representation learning framework based on ontology information and meta-learning for cross-domain knowledge graphs, which includes: for new relationships appearing on the target domain knowledge graph, a view based on relationships

is constructed, and both the topological structure and the semantic information of relationships are modeled as features; for new entities appearing on the target domain knowledge graph, its initial representation is obtained by aggregating the adjacency relationship features of the entity, and graph neural networks are used to fully utilize the known information of the entities and relationships in the knowledge graph to learn and update the vector representations of the entire entities and relationships in the target domain; during the model training process, the meta-learning method is used to divide tasks and simulate new entities and new relationships in the cross-domain scenario through labels, thus completing the knowledge representation learning task from the source domain to the target domain.

This paper verifies the proposed model based on the link prediction task and compares it with multiple benchmark models. The experimental results demonstrate the effectiveness of the proposed meta-learning ontology-enhanced cross-domain knowledge representation learning model.

Key words: knowledge presentation learning; ontology embedding; meta-learning; graph neural networks; inductive reasoning

目　　录

摘　　要 I

ABSTRACT II

1 绪论 1

1.1 研究背景和意义 1

1.2 国内外研究现状及趋势 2

1.2.1 融合辅助信息的跨域表示学习方法 2

1.2.2 基于归纳推理的跨域表示学习方法 3

1.2.3 元学习在知识表示学习中的应用 5

1.3 本文主要研究内容 6

1.4 本文组织结构 7

2 跨域知识图谱的知识表示学习关键技术 8

2.1 跨域知识图谱的知识表示学习定义 8

2.2 本体嵌入关键技术 8

2.3 基于 GNN 的跨域知识表示学习关键技术 10

2.4 元学习训练方法 12

2.5 本章小结 14

3 基于关系拓扑结构及描述文本的本体嵌入 15

3.1 基于关系拓扑结构及描述文本的本体嵌入框架 15

3.2 关系本体图构建 16

3.2.1 关系本体三元组 16

3.2.2 关系位置元关系 17

3.3 基于描述文本的本体表示增强 19

3.3.1 本体结构信息嵌入 20

3.3.2 基于文本的本体嵌入 20

3.4 本章小结 21

4 基于元学习本体增强的跨域知识表示学习模型 22

4.1 模型整体架构 22

4.2 未见关系的特征嵌入 23

4.2.1 关系位置图构建 23

4.2.2 关系相关性系数聚合编码 …… 24
4.3 未见实体的特征嵌入 …… 25
4.3.1 基于关系聚合的实体表示 …… 25
4.3.2 基于 GNN 的实体关系联合嵌入 …… 26
4.4 基于元学习的训练任务设定 …… 26
4.5 基于链接预测任务的模型实现 …… 28
4.6 本章小结 …… 29

5 实验结果及分析 …… 30
5.1 数据集 …… 30
5.1.1 源数据集介绍 …… 30
5.1.2 任务数据集构建 …… 31
5.2 模型参数设置 …… 31
5.3 实验设计及评价指标 …… 32
5.4 实验结果及分析 …… 34
5.5 模型消融实验 …… 36
5.6 未见实体案例分析 …… 37
5.7 未见关系案例分析 …… 38
5.8 本章小结 …… 39

6 总结与展望 …… 40
6.1 总结 …… 40
6.2 未来工作 …… 40

参考文献 …… 42

致谢 …… 46

攻硕期间取得的学术成果和参与的项目 …… 47

1 绪论

1.1 研究背景和意义

近几年，知识图谱（Knowledge Graph，KG）已经成为许多需要访问结构化知识的信息系统的基础[1]。知识图谱将人类知识建模为图的结构进行存储，图中的节点和边代表了现实世界的实体和实体间的关系。典型的知识图谱有 Freebase[2]、NELL-995[3]、DBpedia[4]、YAGO[5] 等。知识图谱中的知识非常庞杂，且往往是隐式或深层次的，难以直接利用或获取到有价值的信息。为了解决这个问题，知识图谱嵌入（Knowledge Graph Embedding，KGE）等知识表示学习（Knowledge Representation Learning，KRL）方法的相关研究迅速起步并获得了广泛的关注。知识图谱嵌入旨在将符号化实体和关系映射到低维稠密的向量空间，以便于计算和应用到下游任务，如知识图谱补全、三元组分类等[6]。随着知识图谱数据规模的增长及在个人电脑、手机等移动设备上的应用，如图1.1所示，知识图谱的处理和存储开始逐渐分布在多个不同的终端上，如移动设备上的电商系统可以利用图谱对用户的商品喜好进行推理和预测。由于资源环境的限制，这些应用上的目标域知识图谱往往是覆盖范围最广、存储知识最多的源域知识图谱中的子集。并且随着图谱的应用，目标域知识图谱会不断引入源域图谱中未定义的、新的未见实体和关系。如何将源域知识图谱的表示学习能力泛化到目标域知识图谱上的跨域知识表示学习问题日益突显。

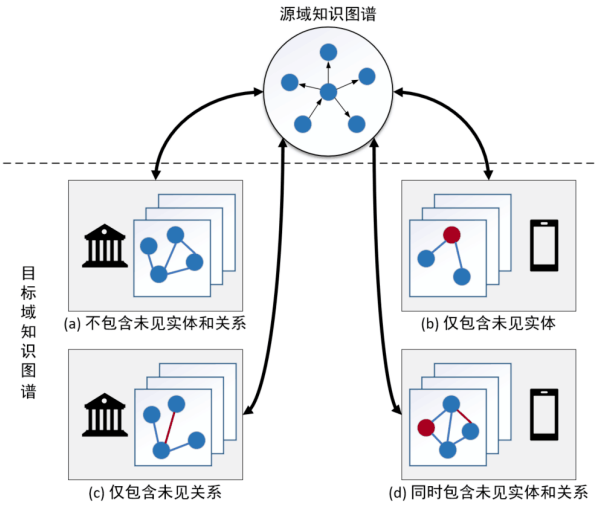

图 1.1 跨域知识图谱的分布

传统的知识表示学习任务在训练集上学习知识图谱的嵌入表示，在测试集上获得实体或关系对应的向量并直接进行计算。而在跨域知识图谱上，目标域知识图谱存在着源域知识图谱中未定义的实体和关系，无法直接从源域知识图谱的嵌入中获取到对应的向

量表示。同时出于用户隐私、成本等方面的考量，无法将新的实体和关系在源域知识图谱中进行知识融合并重新训练。因此跨域知识表示学习任务要求能够从源域知识图谱的训练中学习到对目标域知识图谱中未见实体和未见关系的表示。现有的基于逻辑规则的跨域知识表示学习方法通过自主学习逻辑规则来实现图谱嵌入，但其采用置信度来评估规则的质量，降低了规则的可信度，同时基于规则的模型也存在可扩展性差、大规模知识图谱训练效率低的问题。基于图神经网络的跨域知识表示学习模型聚合邻域特征对目标域知识图谱上新的实体和关系进行表示，但现有模型没有考虑到实体与关系的联系，并且在利用图结构信息进行嵌入的过程中，未能充分利用知识图谱的语义信息。

为了解决上述跨域知识图谱的表示学习上存在的问题，本文提出基于本体信息和元学习的知识表示学习方法。该方法通过对图谱结构信息和本体语义信息的特征学习，能够有效地嵌入表示目标域知识图谱上的新实体和新关系，为跨域知识图谱的表示学习提供一种新的解决方案。同时，本文采用元学习的方法在训练任务上模拟出未见的实体和未见的关系，使得模型能够通过多个训练任务快速收敛，并在测试任务上取得良好的效果。本文所提出的模型也可以应用于现实世界中的真实应用场景，具体流程为：首先在大规模源知识图谱上进行训练，找到最佳的模型参数；随后，利用该模型在各种低资源环境下对目标知识图谱进行小规模的图谱嵌入，以供下游应用如推荐系统、问答系统等使用。本文模型能够有效解决跨域知识图谱的表示学习问题，同时减小整个训练流程的成本消耗，提高源知识图谱的样本利用率。

1.2 国内外研究现状及趋势

跨域知识表示学习一般涉及到融合辅助信息及基于归纳推理的知识表示学习方法。近几年，元学习在知识表示学习中也获得了众多学者的关注，主要用于在少样本、零样本等情境下对表示学习能力的加强。本节将阐述相关方法的研究现状及趋势。

1.2.1 融合辅助信息的跨域表示学习方法

传统的 KGE 方法主要包含基于翻译的表示学习方法和基于语义相似度的方法。基于翻译的方法通过测量低维向量间的距离来计算出三元组事实的可信度。这些方法通常会将关系的嵌入视为从头结点到尾结点的向量平移，代表模型有 TransE[7] 及其变体如 TransH[8]、TransR[9] 等。基于语义相似度的方法通过比较实体和关系在向量空间表示中隐藏的语义特征的相似程度，来评判三元组的合理性，代表模型有 DistMult[10]、ComplEx[11] 等。上述两类方法的本质都是通过设定的打分函数对图谱实体和关系的语义表示向量进行更新，使得表示向量可以更多地贴近源知识图谱中的事实。因此，在传统的、目标域知识图谱不存在源域未见实体和未见关系的知识表示学习上，传统的 KGE 方法能够取得很好的效果。然而，面向跨域知识图谱的知识表示学习，传统的 KGE 方法无法从源知识图谱中学习到新的实体和关系有关的语义信息，不能很好完成对新三元

组的预测任务。

为了对目标域知识图谱中存在的未见实体进行更有效的特征学习，一些学者尝试引入其他辅助信息，以增加表示向量中的隐含特征。知识图谱除了采用事实三元组进行知识存储，还蕴含着其他的丰富信息用于加强知识表示学习，如关系路径信息、图结构信息等。也有一些研究者通过使用图谱外的信息，如文本描述信息等增强传统的 KGE 模型的表示能力。

NTN[12] 模型最早在表示学习中引入实体的描述信息，该模型对实体进行嵌入时并没有采用随机初始化的方法，而是将实体的文本描述信息编码作为实体的表示向量。此后基于描述扩展的知识表示模型 DKRL[13] 试图改进 TransE 模型，使其能够进一步处理实体描述。DKRL 通过将实体结构及实体描述相关的向量联合表示来作为实体嵌入。其中实体描述向量由描述文本经过连续词袋[14] 编码器或卷积神经网络编码器获得。另一种常见的引入辅助信息的方法借助实体的关系信息来进行表示学习，关系信息揭示了实体之间的一个或多个语义联系。PtransE[15] 首次将图谱的多跳关系信息作为知识图谱嵌入的辅助信息使用，在 TransE 的基础上提出了一个以关系路径为基础的表示学习模型。PtransE 的打分函数由两个部分组成，其中一部分是针对头尾实体直接相连关系路径的打分；另一部分是针对头尾实体间其他多跳关系路径的打分。这一改进使得 TransE 的单步推理得以扩展为多步推理，提升了在链接预测任务上的表现。冯俊等人则将知识图谱视为一个大的有向图，提出了基于图感知的表示模型 GAKE[16]，该模型利用知识图的结构信息来学习任意顶点或边的表示。

除了利用图谱自身隐含信息进行表示学习的方法外，一些学者还会引入外部知识以增强表示学习的效果。例如上述的实体描述信息，也可以从新闻稿或者维基百科中获取。图谱外其他模态的信息也同样适用于辅助表示学习，如实体的图像信息、自定义的语义证据等。代表性的 IKRL[17] 模型实现了基于跨模态结构与基于图像的联合表示方法，该模型在遵循平移原则的基础上将图像编码到实体空间。IKRL 提出的跨模态表示可以确保基于结构和基于图像的表示映射在同一个表示空间中。Ren[18] 团队提出的基于语义证据的表示学习方法通过对语义证据的预测及实验验证，研究了 KGE 的外推问题，并通过建模实现了对三种语义证据的加强，在存在有未见实体的知识图谱补全任务上，相比于传统的 KGE 方法取得了更好的表现效果。

1.2.2 基于归纳推理的跨域表示学习方法

为了解决传统 KGE 方法无法处理目标域知识图谱中存在未见实体和未见关系的问题，学者们开始探索基于归纳推理的知识表示学习方法。这类方法通过对结构或一般规则的学习尽量减小对实体和关系的依赖，将表示学习能力泛化到对目标域知识图谱的新三元组的预测任务上。基于归纳推理的知识表示学习方法主要可分为基于逻辑规则和基于 GNN 的表示学习方法，本节将分别介绍两种方法的研究现状。

（1）基于逻辑规则的表示学习方法

基于逻辑规则的表示学习方法通过学习关系的共现模式来挖掘逻辑规则。从形式语言表达能力的角度可将规则分为“命题规则”和“一阶逻辑规则”[19]。命题规则限定了具体的关系及实体，不包含变量，因此具有很强的限制性，无法应用到其他实体的推理任务上。一阶逻辑规则包含了关系及变量，可看作对命题规则的抽象，与具体的实体无关。这些规则从本质上获得了归纳推理的能力，摆脱了实体依赖的限制，可以对存在新实体和关系的三元组上的任务进行推理指导。

在对知识图谱事实三元组中隐含的规则进行挖掘上，传统的规则挖掘方法如 AMIE[20] 模型基于路径遍历的思想实现挖掘算法。这些模型将知识图的关系路径近似视为规则进行提取，通过统计度量或者固定的人工设定模式进行规则的学习。但是由于知识图谱的复杂关系，遍历路径会使得规则提取的成本大大提高，无法应用于大型知识图谱，并且这些传统的规则提取方法存在无法扩展的问题。

随着知识表示学习方法的广泛应用，基于表示学习的规则挖掘方法通过对实体和关系的向量打分来挖掘规则，典型的模型有 RLvLR[21]、HARL[22]、IterE[23] 等。RLvLR 模型通过联合表示学习和子图采样方法从图谱中学习规则，能够在大型的知识图谱如 Freebase、YAGO 上进行有效的规则提取。HARL 在 RLvLR 的基础上考虑了实体的属性信息，添加了对实体属性的规则学习。IterE 模型为了获得对知识图谱中存在的大量稀疏实体和关系更准确的表示，提出了表示学习和规则学习同时进行的迭代算法。该模型在表示学习的基础上添加了公理归纳和公理注入的模块，通过对稀疏的实体和关系的公理注入来加强表示学习的效果，获得了规则和表示学习的共同提升。国内刘藤[24] 等人在 IterE 模型的基础上，专注于规则学习和规则融合模块的改进。他们基于三元组的打分函数，对规则置信度计算方法进行了改进，扩大了模型的适用性。此外，他们还实现了利用表示更新规则及利用规则增强表示的迭代算法，进一步提高了该方法的性能。

在专注于跨域知识图谱的链接预测任务上，一些可微规则学习方法如 Neural-LP[25] 模型和 DRUM[26] 模型，采用端到端的方法学习规则逻辑。此类方法通过设计可微的逻辑规则学习模型，采用基于梯度的方法进行优化求解。然而这些方法不能解决知识图谱中缺边的问题，同时在处理候选规则中的不合理部分仍存在不足。不同于直接从事实三元组中学习逻辑规则的方法，一些方法利用子图隐式地表示逻辑规则。GraIL[27] 和 TACT[28] 通过学习未见实体周围的封闭子图结构，并将其用作逻辑规则。但是随着实体邻居数目的指数增长，这些子图的规模可能会很庞大从而导致效率低下的问题。

（2）基于 GNN 的表示学习方法

基于 GNN 的表示学习方法在传统的知识图谱嵌入方法上进行延伸，利用图神经网络作为编码器学习图结构信息，并采用 TransE、ComplEx 等传统知识图谱嵌入方法作为解码器。

Hamaguchi[29] 等人将图神经网络应用到知识图谱上。不同于先前学者利用 GNN 将子图整体进行编码的方法，Hamaguchi 利用图神经网络，将图中的实体和关系分别映射到单独的向量中。对于三元组中待预测的新实体，该方法通过汇总所有已知实体的信息来生成新实体的嵌入向量。该方法通过 GNN 编码节点信息和图的拓扑结构，能够处理跨域知识图谱中的未见实体，展现了 GNN 在知识图谱上的有效应用，催生了许多基于 GNN 的知识图谱嵌入方法的实现。然而，该方法在聚合邻居节点特征时，仅采用简单的池化操作，没有对邻居节点进行区分或引入关系特征。

图卷积网络[30]（Graph Convolutional Network，GCN）是在图神经网络和卷积神经网络的基础上发展而来的。GCN 通过在节点特征和邻居信息上进行卷积操作来更新节点表示，充分地进行邻居节点之间的信息传递。然而，GCN 只适用于同构图，无法处理知识图谱中复杂的关系结构。而关系图卷积网络[31]（Relational Graph Convolutional Network，R-GCN）则为不同类型的关系设置不同的权重矩阵，将 GCN 模型扩展到了多关系图中。在此基础上，VR-GCN[32] 在对实体和关系的嵌入过程中，考虑了知识图谱中不同关系方向和不同实体类型可能具有不同特征空间的情况。而 TransGCN[33] 则在 VR-GCN 的基础上，，利用 RotatE 思想对聚合函数进行扩展，提出了一种新的处理异构关系的方法。CompGCN[34] 基于 TransE、DistMult 和 HoIE 的思想设置了三种不同的聚合操作，对图谱中的实体和关系信息同时进行建模和学习，并通过共享各层关系嵌入解决了过参数化的缺陷。

不同于上述模型对关系信息的关注和加强，受序列任务中注意力机制的启发，图注意力网络[35]Graph Attention Networks，GAT）将注意力机制和图卷积网络结合，增加对邻接点关注度的区分，并应用于图数据中的节点分类。相关实验结果表明，该机制的引入可以高效地作用于跨域知识表示问题。此外，r-GAT[36]（Relational Graph Attention Network）采用多通道的方法学习实体的嵌入表示，其中每个通道对应不同的实体语义信息，同时利用关系特征聚合邻居信息，从而能够有效处理复杂的多关系图。

1.2.3 元学习在知识表示学习中的应用

元学习的目的是帮助模型掌握如何学习的能力。其主要应用场景为数据稀少的零样本学习和小样本学习等，能够更快、更好地学习新的知识表示，并泛化至新的任务。

MetaR[37] 模型首次将元学习应用到少样本的链接预测任务上。该模型尝试通过元学习的方法捕捉源域知识图谱中事实三元组隐含的、与关系相关的元信息。这类元信息通常是一类任务中常见且通用的信息，因此可迁移到新的三元组中。并且，通过使用小批量的三元组进行元学习，可以加速整个学习过程。MetaR 结合元学习的方法解决少样本场景下的知识图谱补全任务，证明了元学习在跨域图谱的知识表示学习上的有效性。Meta-KGR[38] 模型提出了一种将强化学习与元学习结合的方法，以解决少样本的关系推理问题。该模型通过采用强化学习，训练了一个代理，用于搜索目标实体和多跳推理

路径，从而提高了模型的可解释性。为了解决知识图谱补全过程中由于邻域稀疏导致的噪音过大问题，GANA[39] 模型在 MetaR 模型的基础上引入了图注意力机制和门控网络对噪音进行过滤，同时采用 TransH 作为评分函数，提高了模型对复杂关系的表示能力。

随着图神经网络在知识表示学习中的良好应用，结合 GNN 和元学习方法的研究也受到了学者们的关注。Meta-iKG[40] 包含基于子图的元学习器。它将链接预测任务转换为子图建模问题，并采用局部子图传递子图特定信息。然而，该方法在对子图进行编码的过程中，仅提取了子图的结构语义，无法很好地处理反对称关系的三元组。最近提出的 MorsE[41] 同样结合了 GNN 和元学习方法，将元知识学习分为实体初始化器和图神经网络调制器两个模块。实体初始化器通过两个与实体无关的嵌入（关系域嵌入和关系范围嵌入）初始化每个实体嵌入。而 GNN 调制器则利用实体的邻居结构信息增强实体嵌入。通过对与实体无关的元知识进行建模和学习，MorsE 可以为新实体生成高质量的嵌入。

整体而言，元学习方法侧重于捕获关系信息。模型整体结构分为两个部分：第一部分负责融合信息，获取任务相关的表示；第二部分利用元学习训练模型参数，从而将学习能力快速泛化到新关系上。为了满足具体任务的需要，这些方法往往会通过结合其他方法进行加强。

1.3 本文主要研究内容

对于跨域知识图谱的知识表示学习，传统的知识图谱嵌入方法由于学习固定三元组的实体嵌入，无法很好地处理目标域知识图谱中的新实体和新关系。现有的基于归纳推理的模型尽管取得了一些效果，但利用逻辑规则的方法无法挖掘实体和关系间复杂的语义相关性；利用图结构信息对实体和关系进行嵌入的方法没有充分利用到未见实体与关系的相关性，且未考虑知识图谱语义信息（如本体）对表示学习的补充。一些学者尝试利用实体类型等辅助信息来增强表示学习，但这些信息往往集中于局部的特征。考虑到知识图谱本体可以提供更加丰富且完整的语义信息，包括实体类型、层次信息和关系信息，本文提出了一个本体信息增强的跨域知识表示学习模型，能够结合图的拓扑结构信息和本体语义信息对未见实体和未见关系进行建模，同时从实例关系位置结构中获得了关系本体三元组，结合本体描述文本对本体嵌入中的关系信息进行加强。

为了能更好的将源域的知识迁移到目标域，实现更好的泛化能力，本文提出了一种基于元学习的方法，将训练集划分成多个训练任务，并在每个任务中通过标签模拟未见实体和关系，以提高对目标域知识图谱上新实体和新关系的泛化学习能力。最后，本文进行了充分的实验和分析，验证了模型的有效性。

综上，本文的主要工作包括：

1) 提出了一种嵌入学习模型框架，采用元学习的模型训练方法，分别训练多个单任务，并在各个训练任务中通过标签模拟未见实体和关系。在源域知识图谱上训练

模型，并将训练的参数用于目标域的图谱嵌入上，实现了跨域知识表示学习任务。

2) 针对目标域知识图谱上的未见关系，构建了一个以关系为结点的视图，融合本体信息和结构信息对关系进行建模，并基于关系拓扑结构和本体描述文本对本体嵌入加强。
3) 在多个数据集上进行了充分的实验并与其他基准模型进行对比分析，实验结果表明了模型的有效性。同时通过一系列消融实验，证明了模型各部分的重要性。

1.4 本文组织结构

本文的内容分为六章，以下主要概括各章的内容：

第一章，绪论部分：主要介绍跨域知识图谱的知识表示学习相关研究，探讨其现实意义，阐述了国内外学者在融合辅助信息、基于归纳推理的知识表示学习和元学习等方面的研究进展。最后概括了当前研究所面临的问题，并提出了解决思路。

第二章，跨域知识图谱的知识表示学习关键技术研究：分三个小节介绍本文模型主要涉及到的三种技术的发展状况及原理分析。其中，本体嵌入作为模型主要的语义信息补充，介绍了现有的本体嵌入方法。本文模型作为基于 GNN 的表示学习方法的实现，介绍了基于 GNN 的跨域知识表示学习模型所涉及的关键技术。最后简要介绍了元学习的主要技术及相关思想。

第三章，基于关系拓扑结构及描述文本的本体信息嵌入：介绍了如何捕捉本体的语义信息，并进行本体三元组的嵌入。为了补充本体中关系相关的三元组，本文采用了两种方法进行关系相关本体三元组的提取。在本体三元组结构嵌入的基础上，本章还详细说明了如何使用本体描述文本来增强本体嵌入。

第四章，基于元学习本体增强的跨域知识表示模型：详细介绍了本文提出的跨域知识表示学习模型的各个组成部分，主要包括未见关系嵌入、未见实体嵌入以及基于元学习的训练任务设定。为了验证知识表示学习模型的有效性，介绍了基于链接预测任务的模型实现流程。

第五章，实验结果及分析：介绍了本文使用的跨域表示学习效果评测所需的两个数据集的构建方法。在目标数据集上，对本文模型和相关基准模型进行了充分的实验，并分析实验结果，验证了模型的有效性。同时进行消融实验，对本文模型的各组成模块进行了验证，证明了各个模块的重要性。

第六章，总结部分：综合全文的研究内容及实验结果，总结了本文的主要工作，并对未来的工作进行规划。

2 跨域知识图谱的知识表示学习关键技术

传统的知识图谱嵌入模型通过提取三元组中的事实特征来对实体和关系进行编码。这类模型设计了评分函数如 TransE、RESCAL 等对三元组进行打分。然而在跨域知识图谱的场景下，目标域知识图谱存在新的实体和关系，传统的嵌入模型不能很好地学习到这些实体和关系的表示。基于规则和归纳推理的方法从子图或关系结构中学习向量表示，未能充分利用知识图谱的语义信息。为了解决跨域知识表示问题，本文设计了一个基于本体信息和元学习的跨域知识表示学习模型。本章将介绍模型涉及到的本体嵌入方法、基于 GNN 的表示学习方法以及元学习相关的关键技术。

2.1 跨域知识图谱的知识表示学习定义

知识图谱由众多的事实三元组组成，通常可以被定义为 $\mathcal{G}=(\mathcal{E},\mathcal{R},\mathcal{T})$，其中 $\mathcal{E}$，$\mathcal{R}$，$\mathcal{T}$ 分别指代图谱实体的集合、关系的集合和实体三元组的集合。事实三元组中头尾实体和关系分别来自于实体集 $\mathcal{E}$ 和关系集 $\mathcal{R}$，即 $\mathcal{T}=\{(h,r,t)\subseteq\mathcal{E}\times\mathcal{R}\times\mathcal{E}\}$。传统的知识图谱表示学习旨在将实体和关系映射到连续的低维向量空间，同时保留知识图谱的结构特征。为了验证知识表示学习的有效性，一般会评价其在下游任务的性能表现，并将三元组划分为用于模型参数学习的三元组 $\mathcal{T}_{support}$ 及用于测试的三元组 $\mathcal{T}_{query}$。如在尾结链接预测任务中，给定测试三元组中的一个事实 $(h,r,t)\in\mathcal{T}_{query}$，通过模型计算其所有可能的候选三元组 $\{(h,r,e)|e\in\mathcal{E},(h,r,e)\notin\mathcal{T}_{support}\cup\mathcal{T}_{query}\}$ 的排名，在所有预测三元组中，事实三元组 (h,r,t) 的排名越靠前，则表明该表示学习模型的效果越好。

在跨域知识图谱的现实场景中，需要将源知识图谱上训练的模型应用于目标知识图谱上。由于成本、用户数据隐私等限制，无法将目标知识图谱与源知识图谱合并后重新训练，因此划分出跨域知识图谱。其中 **源域知识图谱** 包含大量已知事实三元组用于训练，**目标域知识图谱** 作为源域知识图谱子集可能包含未定义的实体和关系用于测试。现给定一个用于训练的源域知识图谱 $\mathcal{G}^{train}=(\mathcal{E}^{train})$，并以在源域知识图谱上进行模型参数的学习为目标任务，从而能够将该模型应用在包含未见实体和未见关系的目标域知识图谱上，即 $\mathcal{G}^{test}=(\mathcal{E}^{test},\mathcal{R}^{test},\mathcal{T}^{test}_{support},\mathcal{T}^{test}_{query})$。跨域知识图谱中的实体集和关系集遵循 $(\mathcal{E}^{train}\neq\mathcal{E}^{test},\mathcal{E}^{train}\cap\mathcal{E}^{test}\neq\emptyset)$ 及 $(\mathcal{R}^{train}\neq\mathcal{R}^{test},\mathcal{R}^{train}\cap\mathcal{R}^{test}\neq\emptyset)$。其中 $\mathcal{T}^{test}_{support}$ 只用于标记测试集中实体与关系的结构，不用于对模型的训练。

2.2 本体嵌入关键技术

本体图是一种特殊的知识图谱，本体规定了一系列基本概念之间的语义关系。通常本体以层次概念为主干，通过属性来描述概念的语义关系，以表示通用或特定领域的知

识。随着知识表示学习的不断发展，传统基于三元组结构信息进行表示学习的方法在面对跨域知识图谱等场景下，存在一定局限性。因此，越来越多的方法尝试引入本体信息，以提高表示学习的效果。如图2.1所示，这些本体信息不仅描述实体或关系的限制条件，如属性域和取值范围等，也是对知识图谱语义信息的抽象和集中表示。

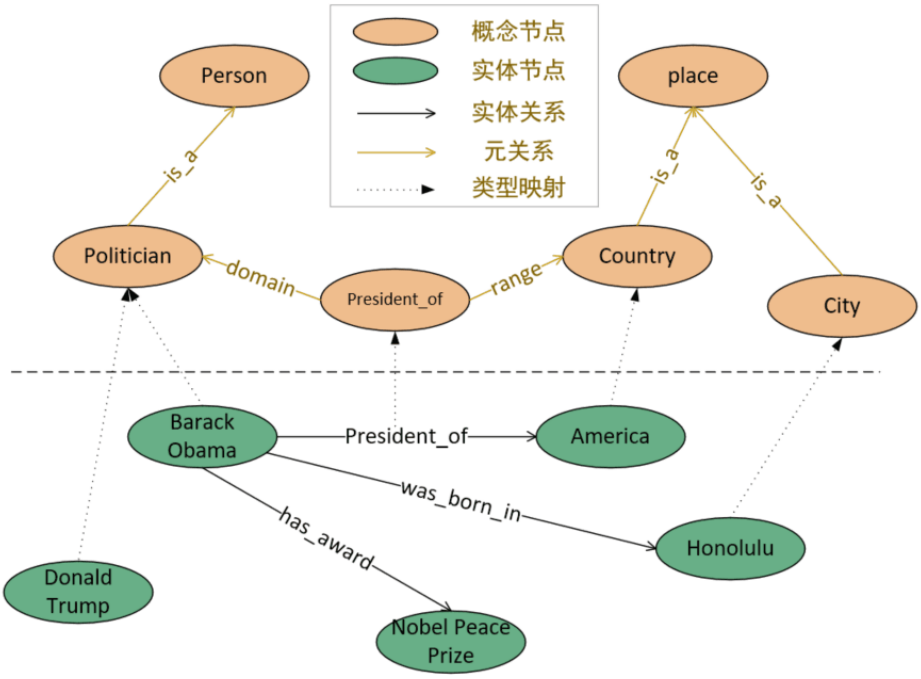

图 2.1　知识图谱的本体视角和实例视角

本体图包含概念和概念间的元关系，通常被定义为 $\mathcal{G}=(\mathcal{C},\mathcal{P})$，其中 $\mathcal{C}$ 是概念的集合，$\mathcal{P}$ 是元关系的集合。类似于实体三元组，一个本体三元组（s，r，t）表示概念 s，t 通过元关系 r 进行联系。然而，与实例知识图谱复杂多样的关系不同，元关系可进一步分类为传递关系、对称关系、层次关系和其他简单关系[42]。受 KGE 方法的启发，On2Vec[42]尝试将本体概念和元关系映射为低维向量。但该模型认为本体关系大多具有传递、对称等特性，不能直接将 KGE 方法应用在本体图上。例如对称关系 r 的两个三元组（c_1，r，c_1）和（c_1，r，c_1），当采用 TransE 等方法学习嵌入时，无法同时兼顾三元组对应的向量满足 $\mathbf{c}_{1,r}+\mathbf{r}\approx\mathbf{c}_{2,r}$ 和 $\mathbf{c}_{2,r}+\mathbf{r}\approx\mathbf{c}_{1,r}$。为了解决上述问题，On2Vec 通过设置两个关系特定的投影函数，来区分同一概念在特定关系头尾位置的不同编码，如公式2.1所示：

$$S_d(T)=||f_{1,r}(\mathbf{s})+\mathbf{r}-f_{2,r}(\mathbf{t})||, \tag{2.1}$$

其中 $f_{1,r}$ 和 $f_{2,r}$ 分别代表了对特定关系 r 的三元组作为头本体和尾本体不同的投影操作。通过对头部本体和尾部本体分别进行不同的投影，可以解决上述传递元关系和对称元关系引起的矛盾问题。对于头尾本体投影操作的选择，On2Vec 采用了简单的线性变

换处理，如公式2.2所示：

$$\begin{aligned} f_{1,r}(\mathbf{s}) &= \mathbf{M}_{1,r}\mathbf{s}, \quad \mathbf{M}_{1,r} \in \mathbb{R}^{k\times k}, \\ f_{2,r}(\mathbf{s}) &= \mathbf{M}_{2,r}\mathbf{s}, \quad \mathbf{M}_{2,r} \in \mathbb{R}^{k\times k}, \end{aligned} \tag{2.2}$$

特别对于层次关系，On2Vec 将层次关系进一步划分为 R_r 和 R_c，R_r 表示粗略概念被划分为更细致概念的细化关系，而 R_c 表示将更细致概念分组为更粗略概念的简略关系。该模型为使细致概念的嵌入汇聚在一个更紧密的邻域内，采用层次模型对层次关系进行单独的处理。对层次关系嵌入的评分函数设置如公式2.3所示：

$$\begin{aligned} S_{hm}(G) = &\sum_{r\in R_r}\sum_{s\in C}\sum_{r\in\sigma(s,r)} \omega(f_{1,r}(\mathbf{s})+\mathbf{r}, f_{2,r}(\mathbf{t})) \\ &+\sum_{r\in R_c}\sum_{t\in C}\sum_{r\in\sigma(t,r)} \omega(f_{2,r}(\mathbf{t})-\mathbf{r}, f_{1,r}(\mathbf{s})), \end{aligned} \tag{2.3}$$

其中，ω 为两个参数向量的角度或距离单调递增的函数，On2Vec 采用余弦距离函数。σ 为对相应层次关系的本体节点的搜索操作，包括对细化关系寻找所有该关系下的所有尾本体、对简略关系寻找该关系下的所有头本体。On2Vec 扩展了 TransE 方法，通过捕获本体关系的关系属性和层次结构，实现了对本体概念和元关系的表示，同时证明了在本体图上应用知识图谱嵌入方法的有效性。

现有的大多数本体模型在知识嵌入过程中都只包含单一的本体信息，对表示学习的补充作用有限，如 SSE[43]、TKRL[44] 都仅采用了本体中实体类型信息，无法实现对所有可用本体信息的嵌入。本文能够在知识嵌入过程中整合所有可用的本体信息，充分补充图谱，提高复杂场景下的决策能力。同时本文使用了一种简单的本体形式，即 RDF 模式 (RDF Schema，RDFS)，而那些更复杂的 OWL 本体可以按照一定的标准转换为 RDFS 本体。

2.3 基于 GNN 的跨域知识表示学习关键技术

图神经网络模型是专门用于处理图数据的模型，对图结构进行编码，并进行节点级别、边级别及图级别的预测任务。CNN 模型只能作用在具有相同结构的图像或者特定序列的语音和文字上，而图数据没有固定的形式且邻居节点也都是无序的，因此 CNN 模型无法作用在复杂的知识图谱上。相比之下，GNN 通过聚合和更新操作，能够学习到图谱结构和节点特征的有效信息。

为学习到图的结构信息和节点特征，GNN 主要包含了两个部分：聚合函数和更新函数。聚合函数可以将相邻节点的特征进行聚合，可以使用诸如 sum、mean 和 max 等聚合操作。一次聚合操作可以提取邻接节点的信息，这些邻接节点距离当前节点为一跳。GNN 通常包含多层，每一层都会用上一层的信息进行聚合和传递。因此，n 层聚合后传递的信息包含了 n 层邻接节点的结构信息和节点本身的特征信息。每层的特征聚合

函数，如公式2.4所示：

$$h_v^k = \sigma(W_k \sum \frac{h_u^{k-1}}{|N(v)|} + B_k h_v^{k-1}) \quad where \quad k = 1, ..., k-1 \tag{2.4}$$

其中本层输出特征 h_v^k 包含两个组成部分。$W_k \sum \frac{h_u^{k-1}}{|N(v)|}$ 表示邻接点特征的聚合操作，后一部分是上一层聚合的特征与本层权重参数的乘积。这两部分特征通过激活函数更新，输出本层节点的特征表示。在 GNN 模型中，聚合函数没有区分邻接节点的重要性，仅采用简单的池化操作。

基于 GNN 的一种改进方向为集中于对图中关系表示的补充。图神经网络关注节点特征的聚合和更新操作，但在信息传递的过程中，图的关系结构仅用于指明邻接点，关系特征未参与节点的更新过程。为了增加关系信息对节点的影响，R-GCN 在每层节点特征计算中引入了邻接点间的对应关系，更新函数如公式2.5所示：

$$h_i^{l+1} = \sigma\left(\sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \frac{1}{c_{i,r}} W_r^l h_j^l + W_0^l h_i^l\right), \tag{2.5}$$

与 GNN 对邻接点特征聚合的操作不同，R-GCN 引入了关系特定的转换。这种转换取决于边的类型和方向。为确保第 l 层的节点表示可以受到自身层次表示的影响，R-GCN 在基础的图关系上为每个节点添加了一个自连接的特殊关系。如图2.2所示，在 R-GCN 中每层对一个实体节点（红色块表示）进行特征生成的过程中，首先从邻接点获取特征（蓝色块表示）并根据该节点与邻接节点的关系类型进行特征转换得到该种关系对应的表示（绿色块表示），其中关系类型分别由入关系、出关系以及自循环关系组成。然后将所有根据关系类型转换后的邻接节点信息累加求和，并通过一个如 ReLU 的激活函数更新，即可获得该节点本层的输出表示。相比于 R-GCN，近期提出的 CompGCN 在 R-GCN 模型的基础上进一步引入了注意力机制，针对每种边类型和方向分别进行了注意力计算以加强对重要信息的关注。而且在计算效率方面，它减少了每个节点的嵌入大小及依赖于固定卷积核的计算量，因此更适合于大规模图数据。

图神经网络通过聚合实体的邻域信息进行表示学习，一定程度上已经可以处理跨域知识表示问题，许多方法将 GNN 与其他方法结合进一步加强跨域知识表示学习效果。例如，INDIGO[45] 模型使用实体三元组与 GNN 的内层和外层的特征向量元素之间的一一对应关系对知识图谱进行编码，并避免了额外的打分函数，充分利用了 GNN 的特征聚合能力。其他方法如 Zhao[46] 等人通过基于注意力的图网络聚合未见关系的邻接结构特征，来作为未见关系的表示。但这些方法要么专注于传统的知识表示学习领域，要么通过结构信息对未见的实体或关系进行嵌入，忽略了图谱的语义信息。本文通过全局本体信息的嵌入表示，在加强知识引入的同时，借助关系的位置结构信息来训练一个兼顾未见实体和关系的基于 GNN 的表示学习模型。

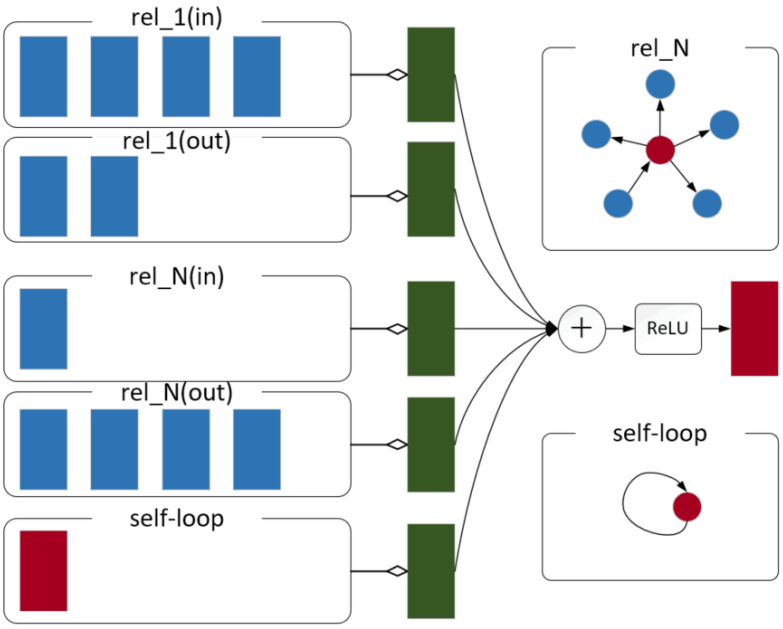

图 2.2　R-GCN 的特征传递

2.4　元学习训练方法

元学习最普适性的算法思想可以理解为“learning to learn”，它通过多个学习任务的训练来改进学习算法，而传统的机器学习算法则是在多个数据实例上进行模型的学习。

传统的机器学习方法会设置一个训练集 $\mathcal{D}=\{(x_1,y_1),...,(x_N,y_N)\}$，如样本对 (输入的图片，图片的标签)。而元学习的目的是训练出一个模型函数 $y=f_\theta(x)$，通过训练来获得其中的参数 θ，求解公式如公式2.6所示：

$$\theta^*=\arg\min_\theta \mathcal{L}(\mathcal{D};\theta,\omega) \tag{2.6}$$

其中的 $\mathcal{L}$ 是一个用于计算真实标签与模型预测标签之间的误差的损失函数，ω 指代了模型如何学习的假设，例如如何为参数 θ 选择合适的优化器或者为 f 选择函数类型等。传统的机器学习方法实现过程中，该部分由研究者手动设置。模型的泛化性能则通过评估模型在已知标签测试集上的任务性能来衡量。传统的机器学习假设模型的优化在每个训练集 $\mathcal{D}$ 上由初始参数开始执行，模型如何学习的设定是预先指定的，而这些设定将极大地影响模型的准确性和数据效率等性能指标。元学习试图从任务中通过学习学习算法本身来改进这些指标，而不是假设学习算法是预先指定或者固定的。

元学习可以视为一个双层优化问题，其中包含内外两层。双层优化[47] 是指一类层次性优化问题，其中一个优化问题包含另一个优化问题作为约束[48][49]。经典的内外双层模型的算法如 MAML，其算法流程如算法1所示：

Data: $p(\mathcal{T})$:distribution over tasks
Data: α, β step size hyperparameters
randomly initialize θ
while *not done* **do**
  Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$ **for** $\mathcal{T}_i$ **do**
    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ wtih respect to K examples
    Compute adapted parameters with gradient descent:$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
  end
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'})$
end

算法 1　MAML 模型算法流程

在该视角下，元学习任务可以通过公式2.7所示来规范化：

$$\omega^* = \arg\min_{\omega} \sum_{i=1}^{M} \mathcal{L}^{\text{meta}}(\mathcal{D}_{\text{source}}^{\text{val}\,(i)}; \theta^{*(i)}, \omega), \tag{2.7}$$

$$s.t. \qquad \theta^{*(i)}(\omega) = \arg\min_{\theta} \mathcal{L}^{\text{task}}(\mathcal{D}_{\text{source}}^{\text{train}\,(i)}; \theta, \omega), \tag{2.8}$$

其中 $\mathcal{L}^{\text{meta}}$ 和 $\mathcal{L}^{\text{task}}$ 分别指代外层的优化目标和内层的优化目标，如在分类任务下的交叉熵。但是这两层的优化级别并不对称，内层优化在基于外层参数 ω 的优化过程中不能对 ω 进行修改。公式中 ω 可以指代如非凸优化[50] 的内层模型的初始化参数或其他可学习的超参数。因此，元学习的整个模型训练分为了两层：内层模型首先接收外层模型的参数 ω，然后根据自己的任务在该任务的训练集上进行训练，并在任务的测试集上计算出损失；外层模型接收内层模型计算出的损失，并对参数 ω 进行更新，使得内层函数的损失趋向最优。元学习的思想即通过外层模型的训练，学习到内层模型一个较优的设定，可以让内层模型更好的完成其他任务。

如前所述，内层模型在训练的时候需要针对面向的问题提供相应的训练集和测试集，这里以任务为训练单位的设定也是元学习方法区别于传统机器学习方法的一大特点。从训练任务的角度而言，元学习的目标即学习一种通用的、能够作用在各任务上的学习算法，这些学习到的算法能够在新的任务上获得更好的表现效果。内层模型可以视为带有外层模型参数 ω 的传统机器学习算法，其数据集为 $\mathcal{D} = (\mathcal{D}^{train}, \mathcal{D}^{val})$，针对单个任务的损失函数为 $\mathcal{L}(\mathcal{D}; \omega) = \mathcal{L}(\mathcal{D}^{val}; \omega^*(\mathcal{D}^{train}, \omega), \omega)$。在实际应用中，通常只有一个训练集和测试集。因此，一般会从源训练集中抽样出一组任务用于训练。这些任务的训练集和测试集被称为 support 集和 query 集，以避免与最终模型训练后进行评估的测试集混淆。

本文旨在进行跨域知识图谱上进行知识表示的相关研究，并在含有未见实体和关系的目标域知识图谱上的链接预测任务上进行模型效果评估。传统的知识图谱表示学习的

链接预测任务采用的数据集通常包含一个训练集和测试集，测试集中不包含新的实体和关系。因此，本文从现有数据集中构建符合跨域场景的测试集，并借鉴元学习“learning to learn”的思想，从训练数据中抽取多个任务用于训练，从任务上学习到对未见实体和关系的表示能力。

2.5 本章小结

本章首先介绍了跨域知识图谱的知识表示学习的定义，强调了跨域知识表示学习中新的实体和关系对传统知识表示学习的影响。对于如何将本体信息用于到知识表示中，介绍了使用本体嵌入的基本方法和代表性的一些应用模型。本体的向量表示，能够为实例图谱的实体和关系提供较为完整的语义信息。同时，本文结合 GNN 模型对邻接实体和关系的特征进行学习，在第三部分介绍了 GNN 在知识图谱的知识表示学习上的应用及基于 GNN 改进的一些模型方法。最后介绍了元学习相关的思想、原理及方法，为后续模型的任务划分及训练流程提供理论支撑和参考。

3　基于关系拓扑结构及描述文本的本体嵌入

本体是领域内公认的概念的集合，是对一系列关系和实体的抽象描述，能够对目标域知识图谱中的未见关系和未见实体提供语义信息，实现零样本学习。为了能够从本体中学习到关系和实体的语义信息，以便将本体信息进行高效的表示学习并用于后续的知识图谱嵌入，需要在本体中补充与关系相关的三元组。本文通过提取实例知识图谱三元组中的关系头尾节点信息和关系间的拓扑信息，得到了关系的定义域、值域三元组以及关系位置元关系三元组，以此构建了关系加强的本体图。之后利用本体概念的描述文本信息对结构层面的本体嵌入进行增强，学习到了融合本体三元组结构信息和描述文本信息的本体嵌入。

3.1　基于关系拓扑结构及描述文本的本体嵌入框架

如图3.1所示，基于关系拓扑结构及描述文本的本体信息嵌入框架主要包含关系本体构建和基于描述文本的本体表示两个部分。作为本体嵌入的基础，本文了解到大多数引入本体信息的模型仅通过实体的类型和层次信息作为本体信息的体现，忽略了关系在本体中的体现。因此，本文首先对实例知识图谱中的关系进行处理，将关系的首尾实体抽象为实体类型，构建了关系的 domain、range 相关的本体三元组。考虑到知识图谱中关系的语义相关性，本文提出四种关系的位置元关系进一步对本体中的关系信息进行加强。而后在本体三元组的基础上学习对本体信息的嵌入。在通过传统 KGE 方法对本体三元组结构信息嵌入的基础上，本文又引入了本体节点的描述文本对本体嵌入的语义进行加强，获得了最终的本体向量表示。

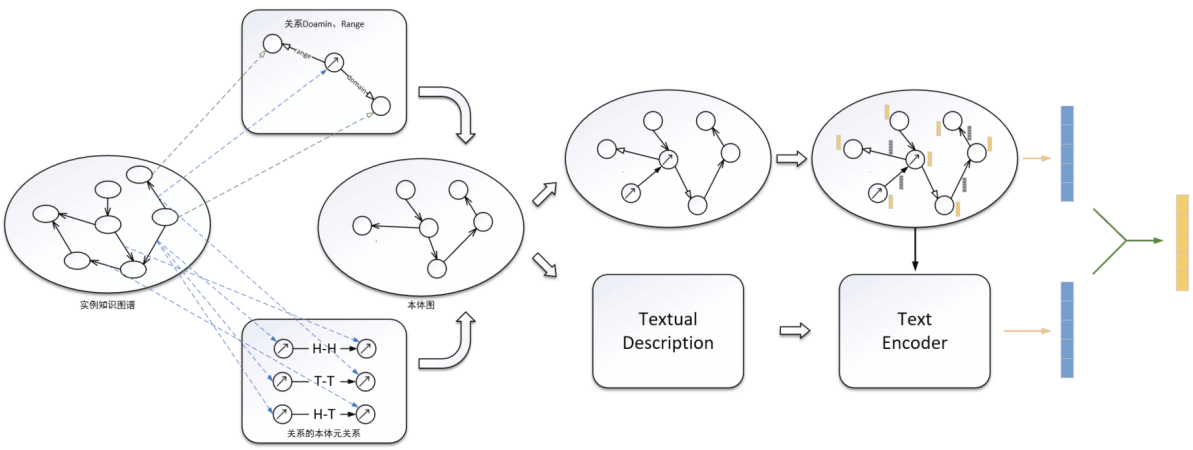

图 3.1　本体嵌入整体框架

3.2 关系本体图构建

结合本体的知识表示模型往往仅使用了实体类型、层次等信息，忽略了关系在本体中的重要体现。本节将介绍本文如何在现有的本体三元组基础上，对关系相关的本体三元组进行补充。同时，根据关系的相对位置定义了四种关系的位置元关系，对本体中的关系信息进行加强。

3.2.1 关系本体三元组

在资源描述框架（Resource Description Framework，RDF）的设置下，知识图谱中的知识总是以三元组的形式出现，通过 RDF 的主语、谓语和宾语来描述事实。RDF 通过类和属性描述个体之间的关系。这些类和属性由模式定义。RDF 模式（RDF Schema，RDFS）提供了最基本的对类和属性的定义。其中，类有 type 和 subClassOf 两种定义，分别用于指定个体所属的类以及子类和父类之间的关系。对于属性，有三种核心属性，分别是 subPropertyOf、domain 和 range，用于指定子属性与父属性之间的关系，以及属性适用范围和取值范围。

RDFS 通过定义的方式来描述元数据之间的关系。同时，通过定义可以将知识分为两类：一类是数据层面的知识，如（Obama，type，Person）说明 Obama 是 Person 的一个实例；另一类是模式层面的知识，如（speaker，domain，Person）说明 speaker 属性的定义域是 Person 类。从简单意义上讲，数据层面的知识更多作用于实体，而模式层面的知识更多作用于关系。但当下将本体信息引入知识图谱嵌入的方法大多数仅采用数据层面的知识，忽略了模式层面知识对实例关系的知识补充。例如 TransT[51] 模型根据头尾实体的类型计算相似度，并作为知识图谱的先验知识改进三元组评分；JOIE[52] 模型通过实体的本体类型信息将本体图和实例图进行跨视图的表示学习。这些模型主要通过本体的数据层面的知识，如实体类型对表示学习进行加强，更多偏向于实体的知识补充，忽略了对关系的补充。而跨域知识图谱中存在未见关系，本体中对关系模式的补充是非常必要的。

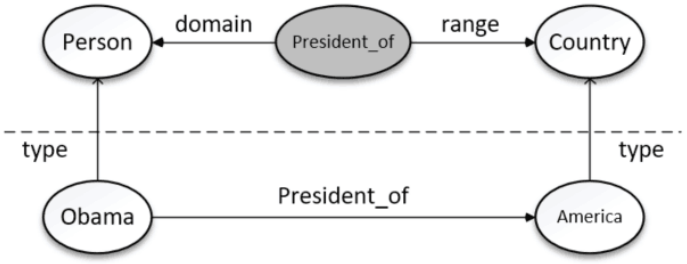

图 3.2　关系本体三元组

同样借助于实体的类型，本文通过对实例图谱中事实三元组的处理提取关系的本体三元组。如图3.2所示，对一个事实三元组（Obama，President_of，America），其中头尾实

体的本体类型三元组为（Obama，type，Person）和（America，type，Country），可知对于关系 President_of 在本体中的定义域即头本体应该是类型 Person，值域即尾本体，因此构建出相应的关系本体三元组（President_of，domain，Person）和（President_of，range，country）。

但从源知识图谱直接抽取会产生大量的 domain 和 range 相关的关系本体三元组，其中包含了所有关系可能的值域和定义域。为了能够将抽取的关系的值域和定义域的本体三元组尽可能地表示最普适的元信息，对于抽取所有的关系本体三元组，本文通过统计关系本体三元组在实例知识图谱中出现的频率，设置阈值进行筛选，以出现频率来代表关系本体三元组的普适程度。除此之外，除了关系的 domain 和 range 的模式三元组，本文认为关系与关系之间存在相似性的关联，因此通过对关系描述文本的相似度匹配，计算关系与关系间的相似程度。本文将关系与关系的相似性关联及现有本体三元组中的 isa、synonym 元关系统一归类为 generalizations 元关系加入到本体三元组数据中。其中对关系描述文本的相似度匹配，本文采用 word2vec 进行文本嵌入和相似度计算。

3.2.2 关系位置元关系

关系本体三元组从值域和定义域的层面对本体中的关系信息进行了补充。除此之外，本文希望通过知识图谱中关系的语义相关性进一步对关系相关的语义信息进行捕捉。知识图谱中的实体具有隐含的语义相关性，例如北京和武汉在作为类型城市的实例中具有相似的语义。而关系的语义的相关性也非常常见，例如关系“/people/person/nationality”和“/people/ethnicity/languages_spoken”具有很强的语义相关性，由于一个人说的语言很大程度上与他的国籍有关，在知识图谱上则可能直接表示为上述两个关系与同一个实体相关联。相反的，上述关系与语义差别很大的其他关系，如“/film/film/country”，则没有很强的语义相关性。并且这种语义相关性与关系的方向密切相关，如图3.3的“sister_of”和通过 e1 节点连接的“has_gender”，及通过 e2 节点连接的“sister_gender”因为拓扑关系的不同，包含了完全不同的隐含信息。

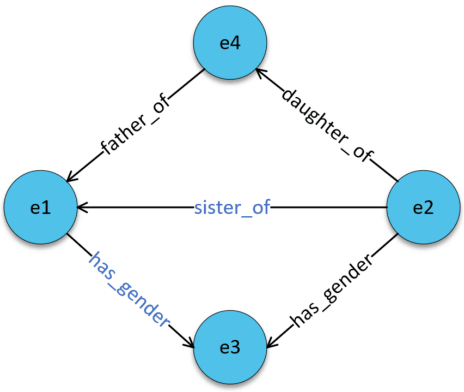

图 3.3 关系位置实例

为了对两个关系之间的相关性进行建模，本文将关系与关系之间的拓扑关系建模为四种关系的位置元关系，如图3.4所示，分别为 tail-head、head-tail、tail-tail、head-head。位置元关系的头结点和尾结点都代表了两个相邻关系的指向，比如（relation1，tail-head，relation2）代表同一个实体连接的两个相邻关系 1 和关系 2，且关系 1 指向该实体而关系 2 则从该实体指向其他实体。对于训练三元组中的两个关系，如果它们符合其中一种的相对位置关系，则提取一个位置元关系补充到本体三元组中。

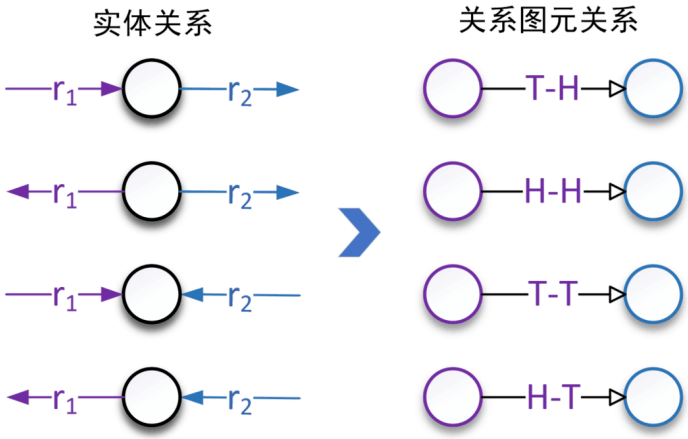

图 3.4　实体关系与关系图元关系的映射

将位置元关系补充到本体三元组中后，本文希望通过传统的知识图谱嵌入方法对本体概念学习向量表示。但对位置元关系分析可知，“head-head”元关系自身、“tail-tail”元关系自身以及“head-tail”和“tail-head”元关系对都具有对称性。如图3.3中通过 e1 连接的“sister_of”和“has_gender”关系存在有（sister_of，head-tail，has_gender）与（has_gender，tail-head，sister_of），通过本文第 2 章对 On2Vec 模型的分析可知无法采用传统的 KGE 方法对有对称的元关系直接进行嵌入。因此，本文在补充位置元关系时仅保留了“head-tail”、“head-head”和“tail-tail”元关系，并对“head-head”、“tail-tail”进行去重，避免了具有对称性位置元关系的存在。

通过对上述两种关系本体信息的补充，结合实体类型相关的本体三元组，本文在本体三元组中保留了 domain、range、generalizations 以及三个位置元关系总计 6 种元关系。本体三元组统计数据如表3.1所示，最终构建出的本体图局部如图3.5所示。

表 3.1　本体三元组统计信息

	关系三元组数	元关系三元组数	其他三元组	总计
NELL_Ext	1816	2135	332	4326
DB_Ext	1727	1104	464	3295

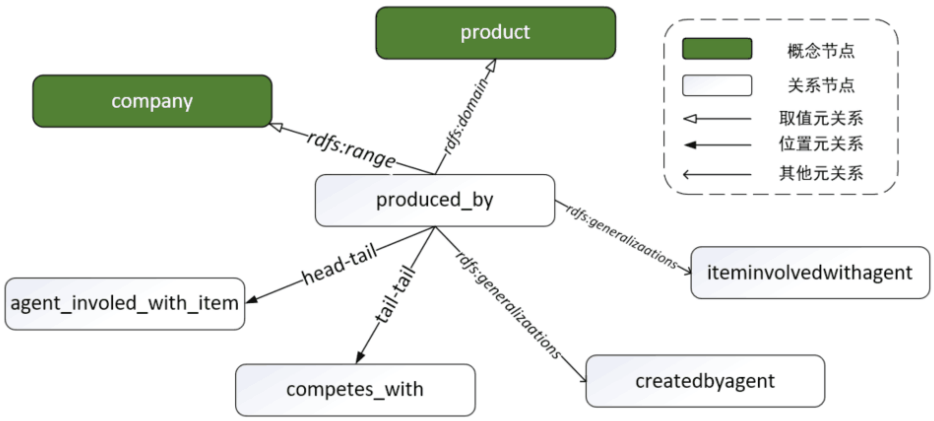

图 3.5　关系本体图

3.3　基于描述文本的本体表示增强

为了将符号化的本体三元组用于后续对未见实体和未见关系的语义补充，需要将本体三元组转化为低维向量表示。本节先从基础的本体三元组数据中学习到结构化的嵌入表示，然后从三元组的概念节点的描述信息中使用词嵌入学习到概念节点的描述文本嵌入。为了将描述文本的嵌入补充到本体信息的结构化表示中去，本文使用一个共享参数的线性层将结构化嵌入表示和描述文本嵌入表示映射到同一个表示空间中。在映射后的线性层中，参照 TransE 的评分方法，本文采用三个距离打分函数将映射后的两种嵌入表示联合更新，学习到兼顾结构信息和描述文本信息的嵌入表示。最后将两种表示拼接

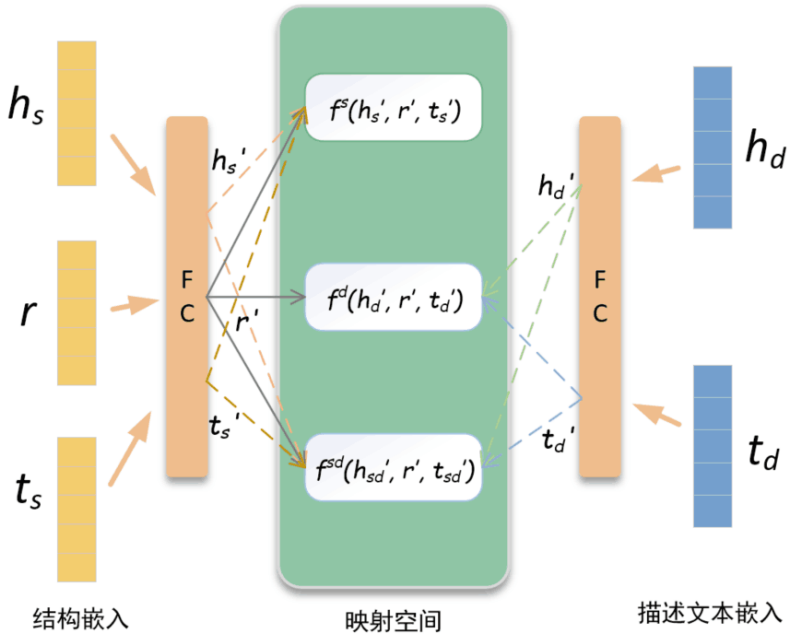

图 3.6　本体嵌入架构

后作为本体信息进行后续操作。基于描述文本的本体表示增强结构图3.6所示。

3.3.1 本体结构信息嵌入

如前文所述，本文在构建本体三元组时降低了对称关系对嵌入效果的影响，因此可以应用多种传统的知识图谱嵌入方法学习本体表示。与传统的知识图谱嵌入表示类似，对于一个本体三元组 (c_i, p, c_j)，本体语义信息编码的目的就是设计出一个打分函数 $f(c_i, p, c_j)$ 作为编码模型的激活函数，本文采用 RotatE 对本体三元组进行编码。按照 RotatE 模型的设定，本体三元组的元关系属性是头尾两个实体节点的转移量，打分函数如公式3.1所示：

$$f_{RotatE}(c_i, p, c_j) = -\|\mathbf{c}_i \circ \mathbf{p} + \mathbf{c}_j\|, \tag{3.1}$$

其中 $\mathbf{c}_i$，$\mathbf{p}$，$\mathbf{c}_j$ 是一个本体三元组相应的概念编码和元关系编码，$\circ$ 指代了向量的旋转操作。同时为了提高所有本体三元组的嵌入效果，本文采用自对抗负抽样损失函数来计算损失并更新模型，如公式3.2所示：

$$\mathcal{L}_{\mathcal{O}} = \frac{1}{|\mathcal{T}_{\mathcal{O}}|} \sum_{(c_i,p,c_j)\in\mathcal{T}_{\mathcal{O}}} [\gamma_o + f(c_i', p, c_j') + f(c_i, p, c_j)], \tag{3.2}$$

其中 γ_o 是控制正负样本得分的参数，c_i，c_j 是不存在于本体三元组中的负样本。为了生成这些负样本，本文在所有的本体概念中分别遮盖住已存在的本体三元组的头节点和尾结点，然后从所有本体节点中随机筛选出其他节点组成负样本。

3.3.2 基于文本的本体嵌入

除了结构化的本体三元组外，本体信息还有许多对本体进行详细描述的文本，如本体概念“companyceo”的描述文本“specifies that a particular CEO is the CEO of a particular company”。这些描述文本可以为本体信息提取提供额外的语义信息，因此本文通过使用文本描述来加强对本体三元组的语义嵌入。然而，描述文本的建模与一般的三元组建模因模型的差异而无法直接融合。因此对于给定的本体三元组 (c_i, p, c_j)，本文首先获得了三元组的结构嵌入 $h_s/r/t_s \in \mathbb{R}^{d_1}$ 和每个本体节点描述文本的向量表示 $h_d/t_d \in \mathbb{R}^{d_2}$，表示文本描述信息。为了融合这两个不同层面的嵌入，本文引入了一个全连接层，将两个不同的嵌入映射到同一表示维度上。映射后的结构嵌入和文本嵌入分别表示为 h_s' 和 h_d'，在统一表示空间中，使用 TransE 对三元组结构的嵌入进行打分，打分函数如公式3.3所示：

$$f^s = -\| h_s' + r' - t_s' \|, \tag{3.3}$$

对描述文本的嵌入进行打分，打分函数如公式3.4所示：

$$f^d = -\| h_d' + r' - t_d' \|, \tag{3.4}$$

同时为了使这两种类型的表示相互兼容和互补，本文遵循 DKRL 模型的设定来定义交叉和相加得分函数，打分函数如公式3.5所示：

$$f^{sd} = - \| h'_s + r' - t'_d \| - \| h'_d + r' - t'_s \|, \tag{3.5}$$

三个得分函数的综合可以保证两个层面的嵌入表示可以在相同空间里进行学习和更新，最后本体嵌入的得分函数如公式3.6所示：

$$f'(c_i, p, c_j) = f^s + f^d + f^{sd}, \tag{3.6}$$

从而在本体嵌入的损失函数也相应的转化为公式3.7：

$$\mathcal{L}_{\mathcal{O}}^{ont} = \frac{1}{|\mathcal{T}_{\mathcal{O}}|} \sum_{(c_i,p,c_j)\in\mathcal{T}_{\mathcal{O}}} [\gamma_o + f'(c'_i, p, c'_j) + f'(c_i, p, c_j)], \tag{3.7}$$

经过训练后每个本体节点都有两个层面的嵌入：三元组结构嵌入和描述文本嵌入。本文将这两种映射后的嵌入进行拼接作为本体节点的最终向量表示。上述描述文本的向量表示，本文采用了词袋模型进行生成。

3.4 本章小结

本章通过对本体三元组的关系部分构建，强调了关系在本体信息中的体现。尤其是在处理跨域知识图谱中存在的新的关系，本体中的信息能够对关系表示进行有效的语义补充。通过本体三元组的结构信息和本体的描述文本信息，能够学习到好的本体嵌入，为下一章未见关系的建模提供了良好的先验知识。

4　基于元学习本体增强的跨域知识表示学习模型

跨域知识图谱的知识表示学习关键在于如何对目标域知识图谱中的未见关系和未见实体进行嵌入，以解决零样本问题。受到元学习算法思想的启发，本文通过设置多个训练任务，在任务中模拟存在新的实体和新的关系的跨域场景，并基于训练任务进行模型训练，使得模型能够学习到对未见实体和未见关系的嵌入能力。在每一个训练任务上，模型通过对关系的拓扑信息和本体信息进行学习获得对未见关系的嵌入表示，通过对未见实体邻接的关系信息聚合获得未见实体的嵌入表示。为充分利用到已知关系和已知实体的特征信息，本文采用 CompGCN 在实例知识图谱上对所有实体和关系的初始化嵌入再进行一次更新，获得所有实体和关系最终的向量表示。最后通过多个 KGE 方法计算损失并进行模型优化。该模型的主要创新点如下：1）关系和实体的表示能够同时学习到事实三元组结构上的拓扑信息和本体层面上的语义信息。2）能够通过元学习的训练流程模拟出目标问题的学习任务并对模型参数进行更新。3）能够同时对测试集中未见的关系和未见的实体嵌入编码。

4.1　模型整体架构

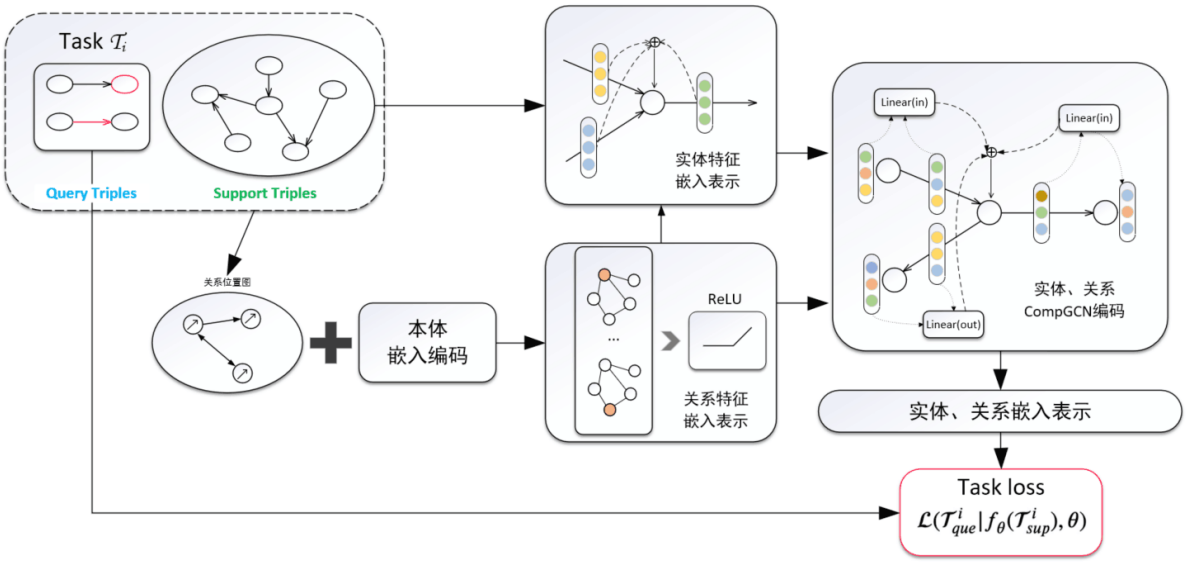

图 4.1　模型架构图

本文提出的模型整体结构如图4.1所示，包含三个主要的部分：未见关系嵌入、未见实体嵌入以及元学习训练设定。对于跨域知识图谱上未见关系的特征提取，本文首先根据实例图谱中关系的相对位置构建了关系位置图，其中节点为实例图谱中的关系。本体信息和关系位置图结合后，通过两层 GCN 对关系的语义信息和结构信息进行更新和

学习，获得关系特征的嵌入表示。对于未见实体的表示学习，本文对未见实体的邻接关系进行聚合，并使用关系方向特定的调整矩阵，提取所有关系的特征，以获得实体特征的嵌入表示。为了利用实例图谱中的已知关系和已知实体，本文在实例图谱上采用两层 CompGCN 聚合所有实体和关系及其邻域特征，并对它们的嵌入表示进行更新。模型修改了 CompGCN 的输出层，使得关系和实体的维度不要求一致，从而可以采用多种 KGE 模型作为打分函数来计算损失，以对模型进行调优。而为了能够让模型在训练过程中获得对目标域知识图谱未见实体和关系表示学习的能力，本文设置了多个训练任务对跨域场景进行模拟，在训练任务上获得最优参数，并将训练的参数用于目标域的图谱嵌入上。

对于训练集中可见的关系和实体，本文采用了传统的知识图谱表示学习的学习流程，即分别设置关系特征矩阵和实体特征矩阵，在初始化阶段对这两个矩阵按照各自目标维度随机向量初始化。在模型训练过程中通过 TransE 等打分函数对嵌入结果打分后进行更新，可以得到可见部分的特征嵌入。

4.2 未见关系的特征嵌入

4.2.1 关系位置图构建

为了能够对目标域知识图谱中的未见关系进行有效的特征学习，由于在源知识图谱中不存在相关的事实三元组可以提供知识，必须要从其他层面获取到对未见关系编码有效的语义信息。本文在第三章中为本体添加了关系的本体三元组，学习到了关系在本体层面的语义信息。同样，基于第三章的关系位置元关系的设定，源域知识图谱中所有的关系可以在关系位置图中进行表示，关系与关系之间通过已定义的四种关系位置元关系（tail-head，tail-tail，head-head，head-tail）进行连接。在关系位置图中，关系与关系间的位置联系可以作为关系的一种拓扑特征信息。这些结构性的信息会减少对具体实体和关系的依赖，在对关系特征进行学习的时候，通过聚合相连其他关系的信息来对关系进行特征表示，能够作用在未见的关系上。

与第三章抽取关系的位置元关系不同，在位置关系图中，本节没有对图中具有对称性的位置元关系进行筛选和去除。为了防止对图中对称的元关系重复的进行信息传递，在对关系节点进行特征聚合时，本节仅考虑每个关系节点的入向关系，以此避免了位置元关系的对称性的影响。

对于输入的实体三元组，可以将原始的图结构转变为关系相关图，如图4.2所示。同时在构建图的过程中没有使用到任何额外的实体属性或关系属性，可以作用到任何未见或已知的关系上。关系相关图中的节点代表原始三元组中的关系，边表示在原始三元组任意两个关系对应的元关系。

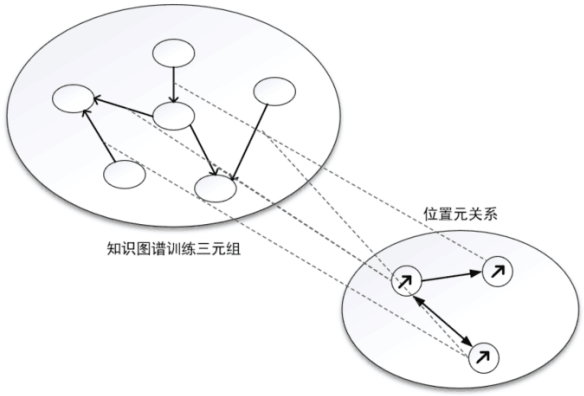

图 4.2　实例图构建关系图

4.2.2　关系相关性系数聚合编码

在关系位置图中，结点为原三元组中的关系，可以通过聚合关系结点的其他相邻关系结点的特征作为该关系的特征。为了能够将本体的嵌入作为对关系嵌入的语义补充，在构建好的关系位置图上，本文通过关系节点到本体概念的映射获取到各关系节点的初始化特征表示。然后通过多层图卷积网络聚合邻接节点的特征来对关系节点的特征学习，对于节点的更新公式如公式4.1所示：

$$h_v^{l+1} = f\left(\sum_{u\in\mathcal{N}(v)} \frac{1}{c_u} h_u^l W^l\right) \tag{4.1}$$

其中 h_u^l 是关系节点 v 相邻节点的特征。每层图卷积会有两步的操作，首先左乘归一化因子（如节点度的倒数）聚合邻接点的特征，然后右乘一个可学习的线性转换矩阵 W，再使用一个非线性的激活函数来获取本层的特征输出。

为了在聚合关系本体信息和拓扑信息时区分不同元关系的重要性，本文在对不同元关系连接的关系特征进行聚合时设置一个可学习的权重参数，根据任务的表现来学习不同元关系对应的重要程度，计算公式转化如公式4.2所示：

$$h_v^{l+1} = f\left(\sum_{(u,r)\in\mathcal{N}(v)} \frac{1}{c_{u,r}} h_u^l W_{dir(r)}^l\right) \tag{4.2}$$

其中 $W_{dir(r)}$ 是两个关系节点相连的元关系类型相对应的参数，根据本文设定的四种不同的元关系，该系数由四个不同的参数控制，如公式4.3所示：

$$W_{dir(r)} = \left\{ \begin{array}{ll} W_{t-h}, & r \in R_{tail-head} \\ W_{h-t}, & r \in R_{head-tail} \\ W_{t-t}, & r \in R_{tail-tail} \\ W_{h-h}, & r \in R_{head-head} \end{array} \right\} \tag{4.3}$$

4.3 未见实体的特征嵌入

4.3.1 基于关系聚合的实体表示

传统的知识表示学习通过对知识图谱三元组的结构信息学习，能够使表示向量尽可能贴近事实。但在目标域知识图谱中存在未见实体的情况下，由于缺乏事实三元组的支撑，传统的 KGE 方法无法学习到未见实体的特征信息。

为了解决上述问题，本文借鉴人类对未见实体的推理过程来对未见实体进行编码。如图4.3所示，传统的 KGE 方法能够通过对事实三元组的学习，对已知的 Tom 节点进行有效的学习。对于存在未见实体的目标域知识图谱，常人虽然无法获得节点 X、Y、Z、A 的具体内容，但是通过对左右两个图谱的推理可知，X 具有和 Tom 类似的邻域结构信息，如 student_of、advisor_of 及 lives_in 等，因此，节点 X 应该是一个类似于 Tom 的一个学生类型的节点。这些结构性的信息可以帮助人去理解一个新的实体，同时这些实体的邻域结构信息与实体本身具体信息是无关且通用的。

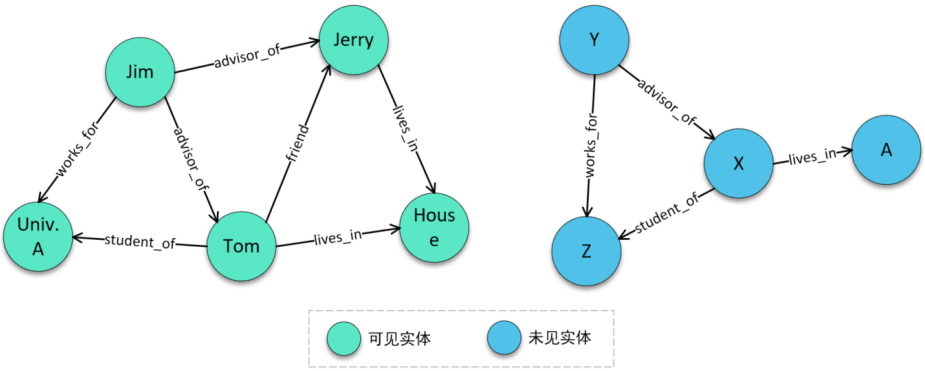

图 4.3 未见实体的关系特征

本节通过模拟人类对新实体的推理过程，对实体邻域的结构信息进行建模来获得实体的特征编码。实体邻域结构信息最直接的体现便是与实体直接联系的各个关系，因此本节通过使用实体相关的关系信息获得未见实体的嵌入。同时考虑到实体关系的方向性，本文通过公式4.4来进行实体关系的聚合：

$$\mathbf{h}_e = \frac{1}{|\mathcal{I}(e)| + |\mathcal{O}(e)|}\left(\sum_{r\in\mathcal{I}(e)} \mathbf{W}_{in}^{ent}\mathbf{h}_r + \sum_{r\in\mathcal{O}(e)} \mathbf{W}_{out}^{ent}\mathbf{h}_r\right) \tag{4.4}$$

其中 $\mathcal{I}(e)$ 是实体 e 所有连接关系中指向实体 e 的集合，$\mathcal{O}(e)$ 是实体 e 所有连接关系中由实体 e 出发指向其他实体的关系集合，$\mathbf{W}_{in}^{ent}$ 和 $\mathbf{W}_{out}^{ent}$ 分别是用于在聚合实体邻接关系中入关系和出关系是对关系嵌入的权重矩阵。

然而，通过对实体的关系特征进行聚合只能获得实体的初始化嵌入，因为实体的邻域结构只传递类型级别的信息，而不是实例级别的。例如，对于节点 X，我们只能推断

出该节点是一个类似 Tom 的学生类型节点，但无法得知节点 X 具体是谁。为了解决这个问题，本文在下一节引入了基于 GNN 的实体关系联合嵌入模块，根据实体的多跳邻域结构调制每个实体和关系的初始化嵌入。

4.3.2 基于 GNN 的实体关系联合嵌入

在上述两个章节中，本文通过关系的本体信息和拓扑信息学习到了未见关系的初始化嵌入，通过实体的邻域结构信息的聚合学习到了未见实体的初始化嵌入。但是这些初始化的嵌入仅传递了部分信息，如实体偏向于类型的初始化嵌入，没有充分利用到已知实体和已知关系的信息。因此，本文在实例知识图谱上，通过借鉴 CompGCN 的模型结构，设计了一个能够同时对实体和关系进行邻域信息聚合的模块。该模块通过对每一个实体和关系聚合多跳邻域结构信息，充分利用到所有已知关系和已知实体的信息，学习到未见实体和未见关系更充分的语义表示。

该层 GNN 网络虽然以 CompGCN 模型为基础，但是 CompGCN 原模型在对关系和实体进行聚合操作的时候采用了 TransE、HoLE 和 DistMul 模型的评分函数，限制了关系和实体的维度必须一致。为了使得该 GNN 网络能够输出维度不同的实体和关系的编码，本文借鉴 MaKEr[41] 模型的设置，将 CompGCN 模型原有的实体关系聚合器修改为一个线性转化层，可以让模型更好适应以多种传统 KGE 方法如 RotatE 来作为解码器进行下游任务。各层的实体嵌入更新公式如公式4.5所示：

$$\mathbf{h}_e^{l+1} = f\left(\frac{\sum_{(r,e)\in\mathcal{N}(e)} \mathbf{W}^l_{dir(r)}[\mathbf{h}^l_r;\mathbf{h}^l_e]}{|\mathcal{N}(e)|} + \mathbf{W}^l_{self}\mathbf{h}^l_e\right) \tag{4.5}$$

其中 $\mathcal{N}(e)$ 是实体 e 的所有相连的关系集合，$\mathbf{W}^l_{dir(r)}$ 是对集合中关系不同方向特定的权重参数，入方向和出方向时分别记为 $\mathbf{W}^l_{in}$ 和 $\mathbf{W}^l_{out}$，$[;]$ 指代两个向量的连接操作。$\mathbf{W}^l_{self}$ 是针对实体 e 本身特征自循环更新的模型学习参数，f 指代模型 GNN 模型的激活函数。同时，关系也在该层网络中也进行了更新操作，如公式4.6所示：

$$\mathbf{h}_r^{l+1} = \mathbf{W}^l_{rel}\mathbf{h}^l_r \tag{4.6}$$

经过两层 GNN 网络，关系嵌入和实体嵌入得到了更新，该模块输出的是该训练任务下所有实体和关系的知识表示嵌入，用于本次任务的损失计算及模型参数更新。

4.4 基于元学习的训练任务设定

为了能够对知识图谱中的未见关系和未见实体进行表示学习，借鉴于元学习“learning to learn”的思想，本文设置了训练元任务和测试元任务。训练元任务用于训练元学习算法，即在训练过程中让算法从元任务中获取经验并调整参数，以便于在测试任务中能够快速地适应。测试元任务则是在训练完成后用于测试元学习算法性能。不同于元学习一般划分多个不同元任务的设定，本文采用单任务设定，即训练元任务和测试元任务

均视为单一的任务。单任务设定可以使得元学习算法更加适应目标任务，并提高模型的泛化能力。同时，模型可以更加专注于学习适应固定任务的策略，还可以减少模型的计算和存储成本，并使得模型更加容易调整和解释。

本文从训练集中抽取了一系列包含未见实体和关系的训练任务来模拟跨域环境，并在该训练环境中对模型进行训练。每个训练任务 $T^i = (\mathcal{E}^i, \mathcal{R}^i, \mathcal{T}^i_{sup}, \mathcal{T}^i_{que})$ 包含训练实体集、关系集、support 训练三元组集以及 query 测试三元组集。为了模拟未见的组件，将部分实体和关系标记为未见，每个训练任务被重新定义为如4.7所示：

$$T^i = (\mathcal{E}^i = (\hat{\mathcal{E}}^i, \tilde{\mathcal{E}}^i), \mathcal{R}^i = (\hat{\mathcal{R}}^i, \tilde{\mathcal{R}}^i), \mathcal{T}^i_{sup}, \mathcal{T}^i_{que}) \tag{4.7}$$

其中 $\hat{\mathcal{E}}^i \in \mathcal{E}^{train}$ 指代实体集中的可见实体而 $\tilde{\mathcal{E}}^i \notin \mathcal{E}^{train}$ 指代未见的实体；$\hat{\mathcal{R}}^i \in \mathcal{R}^{train}$ 指代关系集中的可见关系 $\tilde{\mathcal{R}}^i \notin \mathcal{R}^{train}$ 指代未见的关系。对训练集中三元组抽取训练任务和测试任务的流程如算法2所示：

Data: 读取所有三元组，构建大图 $\mathcal{G}_{train}$
for *i* < *num_subgraph* **do**
  选取 num_root 个根节点
  while *子图节点* < *50* **do**
    从根节点随机游走 num_step 步，构建子图 $\mathcal{G}_i$
  end
  从子图三元组中抽取 10% 作为 query 三元组
  对 query 三元组中的关系和实体随机标记为未见
  构建 support 三元组中的实体和关系联系矩阵，构建关系位置图
  将 support 三元组、query 三元组、关系位置图存储
end

算法 2　任务子图构建流程

所有训练的总目标是：在每个任务的 support 集上对未见实体和关系的学习能力训练，使得在 query 集上的评估得分最高，计算函数如公式4.8所示：

$$\max_{\theta} \mathbb{E}_{T^i \sim p(T)} \left[\sum_{(h,r,t) \in \mathcal{T}^i_{\text{que}}} \frac{1}{|\mathcal{T}^i_{\text{que}}|} \mathcal{M}(h,r,t|\mathcal{T}^i_{\text{sup}}) \right] \tag{4.8}$$

其中 $\mathcal{M}$ 通过在 support 集上学习到实体和关系的嵌入表示，并在 query 集上对三元组计算得分。本文采用自对抗负抽样损失函数来对计算损失更新模型，如公式4.9所示：

$$\begin{aligned} \mathcal{L}(T^i) = & \frac{1}{|\mathcal{T}^i_{que}|} \sum_{(h,r,t) \in \mathcal{T}^i_{que}} -log\sigma(\gamma + s(h,r,t)) \\ & - \sum_{i=1}^{n} p(h'_i, r, t'_i) log\sigma(-\gamma - s(h,r,t)) \end{aligned} \tag{4.9}$$

n 指代负采样的数量，$p(h_i', r, t_i')$ 指代负采样的权重系数，$s(h, r, t)$ 是用于计算嵌入向量在 query 三元组上的得分函数。

4.5 基于链接预测任务的模型实现

知识图谱的知识表示学习通过学习实体和关系的向量表示，来捕捉隐含的语义关系和规律。跨域知识图谱的知识表示学习在传统知识表示学习的基础上，能够学习到目标域知识图谱上新的实体和关系的特征表示。为了验证知识表示学习的有效性，本文结合下游链接预测任务对提出的知识表示学习方法进行评判。

链接预测任务通常涉及知识图谱中实体对之间的关系推理。该任务基于现有事实来推断出未知的三元组，以帮助补全知识图谱中的缺失信息、推断未知实体之间的关系及预测新实体属性等。链接预测任务可分为头实体预测和尾实体预测两种类型。为了评估模型在链接预测方面的能力，常用的指标有 MRR 和 Hits@k 等。其中 MRR 被定义为测试数据集中每个三元组的预测排序的平均排名的倒数，而 Hits@k 则是指在测试数据集中每个三元组的预测概率排名中前 k 名的正确率，即命中率。这些指标可以帮助确定预测结果的准确性和可靠性，同时有助于对不同模型进行比较和评估。

本文在训练任务的 support 上学习到实体和关系的嵌入，并在 query 集上进行损失计算。在进行每次任务损失的过程中，通过正负样本的联合计算出损失，并以此进行梯度下降求解。负样本损失计算公式如公式4.10所示：

$$\mathcal{L}_{neg} = \sum_{i=1}^{n} \left(softmax(W[s_{tail}; s_{head}]) log\sigma(-[s_{tail}; s_{head}]) \right) \tag{4.10}$$

其中 s_{tail} 和 s_{head} 是对一个测试三元组对应的尾结点负采样得分和头结点负采样得分，n 代表负采样的数量，[;] 操作将正负样本的得分矩阵进行简单的连接操作，W 是对得分进行微调的权重参数，本文默认设置为 1。正负样本的得分求得算术平均即为该测试三元组的单个损失。

对于模型的训练流程，本文按照元学习单任务训练方法进行，每个任务的目标都是使该任务下的连接预测效果最优，单个任务的训练流程包含有 5 步：

1) 初始化模型参数，主要包含了训练数据集中可见实体和关系的嵌入初始化、元关系的权重参数以及两个不同阶段的 GCN 层和 CompGCN 层的传递参数。
2) 获得任务的支持集和查询集，主要从原训练集中抽取单个任务的子图，并划分为支持集和查询集，同时对关系和节点的标签进行设置以模拟出未见组件。
3) 单次任务训练获得实体关系的嵌入，通过 KGE 模型计算查询集损失，利用梯度下降求解。
4) 更新模型参数，重复下一个任务的训练直至效果不再更优。
5) 对测试集进行模型测试，输出任务得分。

4.6 本章小结

本章从本文模型的总体架构开始，详细介绍了模型的各个组成部分，包括主要的未见关系嵌入、未见实体嵌入以及基于元学习的训练任务设定和基于链接预测任务的模型实现的说明。对未见关系嵌入，说明了本文模型如何对关系的拓扑结构进行提取以及如何通过图卷积层对关系的本体信息和拓扑信息进行联合学习，从而获得未见关系的有效表示；对未见实体嵌入，通过未见实体的邻接关系的特征聚合，并且在实例知识图谱上通过 GNN 来聚合其他关系和实体信息，完成最后的实体和关系的特征更新。之后的章节中将对本文模型在测试数据集上进行充分的实验，检验本文模型的有效性。

5 实验结果及分析

在前文中，本文针对跨域知识图谱中存在未见实体和未见关系的问题，提出了一个基于本体信息和元学习的知识表示学习模型。同时针对跨域知识表示学习中知识图谱语义信息的不足，本文提出了一种基于关系拓扑结构和描述文本的本体嵌入框架，通过获取本体信息并与图结构信息联合学习，实现了对未见实体和未见关系的特征聚合。结合元学习和本体嵌入技术，能够有效地处理跨域知识图谱中新实体和新关系的表示。为了验证模型的有效性，本章主要验证和评估了本文模型（NAMER）的实验效果。实验主要面向跨域情景的测试数据集进行，与现有的处理类似问题的归纳知识图谱表示模型进行对比。结果表明，引入了多层次的特征信息后，本文模型在实验数据上明显优于其他现有模型。最后的多个消融实验也证明了模型各个组成部分的重要性。

5.1 数据集

由于传统的知识图谱数据集通常基于“closed-world”设定，测试三元组中的实体和关系在训练数据集中已知，不存在跨域场景中的未见实体和未见关系。因此，为测试模型在包含未知实体和关系的跨域场景下的有效性，本文通过抽取包含本体层次三元组的DBpedia 和 NELL-995 知识图谱的子集构建了新的数据集 DB_Ext 和 NELL_Ext。

5.1.1 源数据集介绍

DBpedia：DBpedia 是一个基于维基百科的语义知识库。该数据集由开源社区维护并使用维基百科文章和其他在线网络资源进行扩展。DBpedia 从 Wikipedia 中提取结构化的数据，转化为事实三元组进行存储，包括图像、标签、描述文本等结构化属性，最新的数据集快照含有 8.5 亿条事实三元组数据。此外，DBpedia 注重本体论的构建，本体类型数共 768 个，主要包括人物、地点、工作、组织等概念。这些实体、属性和关系通常以 RDF 图的形式存储，标准查询语言为 SPARQL。DBpedia 开源社区提供网页公布数据集的最新版本和统计数据，同时支持各数据集版本的检索和下载，为研究者使用提供了便利。

NELL-995：NELL-995 数据集是根据 NELL 知识库抽取用于知识图谱补全任务的基准数据集。NELL (Never Ending Language Learning) 是由卡内基梅隆大学领导的自动化学习系统，旨在从互联网的非结构化文本中自动提取知识。基于 NELL 知识库，NELL-995 数据集包含 995 个实体和 129 个关系，覆盖地理、医学、体育、音乐、文化等广泛领域。NELL-995 数据集的实体和关系被细粒度分类，每个实体被分配到一组多层次的关系类型中。因此，NELL-995 数据集是学习细粒度知识表示和关系推理模型的理想数据集。

5.1.2 任务数据集构建

为了抽取出包含未见实体和关系的数据子集，本文首先构建一个包含所有三元组的实例知识图谱，在其中随机选取 100 个根实体节点并将每个节点的 10 个相邻接点组成测试子图，将子图中 1/10 的实体和关系标记为测试集，并从训练集中剔除这些实体和关系相关的三元组作为测试三元组；最终每个数据集中都包含了两个基本的子知识图谱，训练知识图谱 $\mathcal{G}^{train}$ 和测试图谱 $\mathcal{G}^{test}$，后者包含了训练图谱中不存在的实体和关系。此外，为了测试模型对未见关系及未见实体的单独学习性能，在测试集中的 query 集设置中，本文将测试三元组分为三类：1、所有测试三元组中只包含了未见的实体（unseen_ent）；2、所有测试三元组中只包含了未见的关系（unseen_rel）；3、所有测试三元组中同时包含了未见的实体和关系（unseen_both）。两个数据集中的统计数据如表5.1所示。其中，DB_Ext 数据集包含 243 个仅含未见实体的测试三元组、10 个仅含未见关系的三元组和 243 个同时包含两个未见组件的三元组；NELL_Ext 数据集包含 565 个仅含未见实体的测试三元组、12 个仅含未见关系的三元组和 115 个同时包含两个未见组件的三元组。各数据集包含三元组的统计数量如表5.1所示。

表 5.1 数据集统计数据 (括号中为未见组件数量)

	训练图谱			测试图谱			
	实体数	关系数	三元组数	实体数	关系数	support 三元组数	query 三元组数
NELL_Ext	1583	153	5269	851(753)	140(30)	2160	692
DB_Ext	795	115	1508	913(884)	128(46)	1930	496

5.2 模型参数设置

对于用于对比的基准模型，本文采用了相关论文给出的最优超参数设置，本文模型采用的相关参数设置如表5.2所示。

其中主要包含以下三个方面的参数设置：

1) 本体嵌入相关参数：模型学习率 lr、本体三元组概念节点结构嵌入维度 ent_str_dim、本体概念节点文本嵌入维度 ent_text_dim、线性隐藏层维度 mapping_size 以及训练的 epoch 数量和 batch 的大小。
2) 元学习相关参数：任务学习率 lr、单任务支持集的 batch 数量 train_bs、单任务查询集的 batch 数量 eval_bs、元训练总任务数 num_step、元训练提前结束无效训练任务计数 early_stop_patience 以及单个 batch 采样的根节点数。
3) 图谱嵌入相关参数：基本维度的设置 dim、关系位置图对关系进行 GCN 的层数

num_gcn、GCN 中间传递维度 gcn_dim 以及 GCN 层的丢弃率 hid_drop、对关系和实体进行联合学习的 CompGCN 的层数 num_compgcn。

表 5.2　模型超参数设置

本体嵌入参数	设置值	元学习训练参数	设置值	嵌入参数	设置值
lr	0.00005	lr	0.001	dim	300
ent_str_dim	150	train_bs	64	num_gcn	2
ent_text_dim	300	eval_bs	16	num_compgcn	2
mapping_size	300	num_step	100000	gcn_dim	300
training_epochs	1000	early_stop_patience	20	hid_drop	0.3
batch_size	100	num_sample_size	10	-	-

在对三元组进行打分评估时，本文通过对 CompGCN 第二层输出的改进，可支持多种 KGE 模型作为评分函数。实际采用的 KGE 模型包含 TransE、DistMult、ComplEx 及 RotatE。实体和关系的维度根据采用的 KGE 模型在基础嵌入维度上进行调整，具体如下表5.3所示：

表 5.3　评分函数

模型名	实体维度	关系维度	评分函数
TranE	dim	dim	$F = -\| h + r - t \|$
DistMult	dim	dim	$F = h^T \, diag(r) \, t$
ComplEx	2 * dim	2 * dim	$F = Re(h^T \, diag(r) \, t)$
RotatE	2 * dim	dim	$F = -\| h \circ r - t \|$

其中评分函数中的 h、r、t 分别指代头实体、关系和尾实体的嵌入表示，Re 表示复向量的实部分量，$\circ$ 操作表示旋转操作。

5.3　实验设计及评价指标

本实验包含本体嵌入表示学习和图谱表示学习两个阶段，第一个阶段主要学习到融合描述文本信息的本体嵌入表示，第二阶段使用本体信息进行跨域知识图谱的表示学习。

第一阶段首先采用传统的表示学习方法对本体三元组数据进行初步嵌入表示。其次，从预训练词嵌入 glove 中获取本体描述信息中各描述单词的初始化词向量。使用 TF-IDF 统计方法识别单词的重要程度，并对单词的词向量进行加权聚合计算，以获得本体节点描述文本的初始化向量嵌入。最后，将本体三元组的结构化表示嵌入和本体描

述文本的表示嵌入映射到同一个空间，以进行评分和更新，从而获得拼接后的最终本体嵌入。

第二阶段对跨域知识图谱进行表示学习，为了在元学习中模拟出跨域场景，在每个元学习任务的设置中都人为抹除了一些实体和关系的标签，使得这些实体和关系必须通过本文模型的未见关系和未见实体嵌入模块学习得到向量表示，而不是从传统嵌入方法的嵌入矩阵中取得。

对于模型在链接预测任务上的评价指标，本文选取了 MRR 和 Hit@10 作为评判的标准。其中 MRR 通过预测三元组排名的倒数来进行计算，即对测试三元组中的所有事实三元组，如果该三元组在预测排名靠前，对应的倒数也会比较大，因此链接预测性能与 MRR 评价指标的数值大小成正相关。而 Hits@n 描述的是在所有预测三元组排名中前 n 的三元组所占的平均比例，其计算公式如公式5.1所示：

$$HITS@n = \frac{1}{|S|}\sum_{i=1}^{|S|} ||(rank_i \leqslant n) \tag{5.1}$$

假设 n 设置为 10，那么统计事实三元组在预测三元组中前 n 名的个数，最后再除以总个数就得到了 Hits@10 的结果，其中 $||(\cdot)$ 为 indicator 函数（若条件真则函数值为 1，否则为 0）。

参与比较的模型如下：

Neural-LP[25]（2017）：该模型基于知识库构建了一个可学习逻辑规则的可微模型。逻辑规则是独立于实体和关系的，因此理论上该模型可在任何未见的实体上应用，并在归纳图谱补全任务中相比传统方法（即对实体进行结构信息表示学习）有明显提升。

DURM[26]（2019）：提出了一种可微的、可同时学习规则逻辑及其置信度得分的方法。该方法可以使用梯度优化来处理归纳逻辑编程任务，并可用于处理含有未知实体的链接预测任务。

GraIL[27]（2020）：该模型不直接学习实体节点嵌入，也没有使用任何节点的属性。相反，它在测试三元组候选关系的周围构建子图，并利用子图的结构和结构化的节点特征来预测三元组。这使得该模型能够很好地应用于未知的实体三元组预测任务。

CoMPILE[53]（2021）：该模型对 GraIL 模型子图归纳模型进行了改进，包括加强对子图关系方向性的限制，并在未知节点特征聚合的信息传递过程中增加了先前模型中忽略的关系特征。

MaKEr[37]（2022）：该模型通过学习关系结构的特征来聚合邻接关系的特征，进而对关系进行表示，并聚合关系特征对实体进行编码。模型利用拓扑结构的信息，在一定程度上实现对未见实体和未见关系的表示。

5.4 实验结果及分析

各模型在 NELL_Ext 上的链接预测任务实验结果如下表5.4所示，各模型在 DB_Ext 上的链接预测任务实验结果如表5.5所示。根据测试数据集的三元组对未知实体和关系的包含情况，将结果分为了只包含未见实体的结果 (u_ent)、只包含未见关系的结果 (u_rel) 以及同时包含未见实体和未见关系的结果 (u_both)。表格中加粗部分为最优的实验效果，带有下划线的则是该类基准模型中表现最优的得分，模型括号中指代的是在评分阶段采用的 KGE 评分函数。

表 5.4 NELL_Ext 数据集结果

NELL_Ext						
	u_ent		u_rel		u_both	
	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
Neural-LP	30.48	47.96	-	-	-	-
DRUM	31.82	48.32	-	-	-	-
GraIL	71.62	92.92	-	-	-	-
CoMPILE	<u>75.94</u>	<u>93.62</u>	-	-	-	-
MaKEr(TransE)	70.82	92.00	24.56	54.17	21.53	51.74
MaKEr(DistMult)	70.63	91.33	27.02	60.00	**41.39**	57.65
MaKEr(ComplEx)	72.24	91.91	18.27	34.17	29.39	59.65
MaKEr(RotatE)	<u>77.09</u>	<u>94.64</u>	<u>31.53</u>	<u>55.00</u>	31.45	<u>62.35</u>
NAMER(TransE)	78.28	94.69	20.72	53.34	27.11	55.85
NAMER(DistMult)	75.98	92.46	19.30	22.50	31.37	55.65
NAMER(ComplEx)	73.61	90.60	24.44	38.33	29.70	54.96
NAMER(RotatE)	**79.92**	**94.73**	**45.07**	**75.63**	<u>40.33</u>	**67.06**

表5.4和表5.5展示了各模型在 NELL_Ext 和 DB_Ext 上的链接预测结果。计算 Hit@10 分数时，本文选取了 50 个候选进行评估，并对于不同的补全任务（u_ent、u_rel 和 u_both）显示了不同模型的得分。GraIL、Neural-LP、DRUM 和 CoMPILE 仅针对未见实体补全任务设计了实验，因此未列出其在未见关系上的实验结果。上述数据均为模型运行 4 次后取平均值的结果。

结果表明，本文提出的 NAMER 模型相比其他基准模型有所改进，并在不同的 KGE 评分模型上有不同程度的提升。此外，基于 RotatE 的 NAMER 模型总体取得了最好的成绩。我们认为 RotatE 模型采用了更加复杂的关系和实体的映射关系，因此能表现更多没有重叠的特征信息。这也证明了本文模型的有效性。

表 5.5 DB_Ext 数据集结果

	DB_Ext					
	u_ent		u_rel		u_both	
	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
Neural-LP	57.15	73.46	-	-	-	-
DRUM	59.88	73.25	-	-	-	-
GraIL	59.44	80.86	-	-	-	-
CoMPILE	60.66	79.93	-	-	-	-
MaKEr(TransE)	54.4	83.7	31.13	54.00	38.66	66.50
MaKEr(DistMult)	46.24	81.07	16.43	11.00	32.16	56.71
MaKEr(ComplEx)	53.79	82.47	19.95	29.00	36.88	59.26
MaKEr(RotatE)	59.55	86.09	32.93	55.00	**41.27**	66.54
NAMER(TransE)	64.63	89.60	**44.77**	70.25	34.92	**69.42**
NAMER(DistMult)	56.73	80.25	13.68	11.00	33.94	61.74
NAMER(ComplEx)	52.22	77.34	14.40	15.00	31.72	59.49
NAMER(RotatE)	**66.43**	**89.67**	41.80	**74.00**	35.11	63.81

本文在处理未见实体的测试集上，首先比较了依据规则学习来处理未见实体的 Neural-LP 模型和 DRUM 模型以及基于子图推理的 GraIL 模型和 CoMPILE 模型。结果表明，基于规则的模型的效果不如基于子图的模型，在 NELL_Ext 数据集上，Neural-LP 和 DRUM 的实验得分远低于 GraIL 和 CoMPILE 模型。这是由于基于规则的模型依赖于针对数据集学习出的规则，需要大量的数据集或对样本均衡性有严格的要求，因此模型效果受数据集影响较大。CoMPILE 模型在 GraIL 模型的基础上强调了关系的重要性，总体效果要比 GraIL 表现更好。然而，上述四个模型都无法处理未见关系，而基于子图归纳推理的模型强调测试三元组中头尾实体间的局部子图信息，没有完全利用到实体周围的结构特征信息以及关系信息，整体效果都比本文的模型差。比较本文的模型与 MaKEr 模型，MaKEr 模型在 MRR 的评价指标上平均低了 4.42%，而在对应 RotatE 评分函数下的 Hits@10 评价指标上，本文模型表现领先了 1.85%。尽管 MaKEr 模型考虑到了关系对实体的重要性，但在关系表示的特征学习方面，没有充分利用图谱的语义知识，所以效果比本文模型差。这进一步说明了本文模型在未见实体表示方面的有效性。

本文在仅包含未见关系和同时包含未见实体和关系的测试集上，与通过结构信息对关系和实体进行编码的 MaKEr 模型进行了比较。针对未知关系，本文引入了额外的本体知识作为关系语义信息的补充，并利用关系图卷积对关系的表示进行了更新，以学习未见关系周围的结构拓扑信息。两个测试集上的实验结果表明，相比于仅使用结构信息

编码关系的 MaKEr 模型，本文引入本体信息能够有效补充关系表示的语义信息。在处理未知关系时，本文采用 RotatE 作为评分函数的模型表现出明显的优势，在 NELL_Ext 数据集上的 MRR 得分比 MaKEr 高出 14.54%，而在 Hits@10 的得分方面，70.63 的得分比 MaKEr 的得分高出约 20%。这表明本文模型在捕捉关系语义和结构信息方面具有更好的性能，在考虑局部关系和全局关系时都表现良好。在同时包含两种未知组件的测试集上，本文模型相比 MaKEr 模型在 NELL_Ext 数据集的 Hits@10 评分上平均提升了 1.82%，在 DB_Ext 数据集的 MRR 和 Hits@10 分别平均提升了 1.18% 和 0.53%。

此外，本文实验发现，相比于 DistMult 和 ComplEx 模型，将 TransE 和 RotatE 用作解码器时模型效果更好，尤其在处理关系方面。在本体嵌入实验中，本文采用 RotatE 作为评分函数来对本体信息进行表示学习。DistMult 和 ComplEx 模型因其复杂性与 RotatE 的本体嵌入方法不兼容而效果下降。相比之下，TransE 模型简单易操作，在进行特征提取时更为适合。与本文模型搭配整体表现最好的 RotatE 模型不仅与本体嵌入方法相匹配，还能够提供更全面的对实体和关系低维度嵌入表示，因此在本文中表现为最优的模型之一。

5.5 模型消融实验

本节将介绍模型中几个重要模块的多项消融实验，以展示本文模型各部分的重要性，主要设置了 5 项不同的消融设置实验：(1) 去除元学习的设置；(2) 去除本体的设置；(3) 去除实体聚合表示的设置；(4) 去除关系 GCN 聚合的设置；(5) 同时去除本体和元学习的设置。获得的实验结果如表5.6所示：

表 5.6 在 NELL_Ext 上的消融实验结果

NELL_Ext						
	u_ent		u_rel		u_both	
	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
NAMER(TransE)	**78.28**	**94.86**	20.72	53.34	**27.11**	**55.85**
no_meta_TransE	29.31	43.82	12.65	30.83	10.18	22.26
no_ont_TransE	76.89	94.07	14.96	27.50	14.06	28.78
no_ent_TransE	76.85	94.07	18.00	55.83	19.43	50.7
no_gcn_TransE	74.68	92.27	**23.88**	**61.67**	19.85	46.26
no_meta_ont_TransE	26.57	38.69	11.15	25.00	12.62	27.74

其中每个消融实验的具体设置如下：

1) 去除元学习的设置：在模型训练阶段，本文引入了元学习，在训练集上提取任务子图并通过随机标签来模拟出未见的组件从而训练出模型处理未见组件的能力。在

该消融实验下，取消掉了对子图标签的模拟效果即按照传统 KGE 模型的训练方法，在训练集上对所有三元组进行嵌入和损失计算。

2) 去除本体的设置：本文模型的一个创新点即在对关系嵌入表示的时候通过关系位置图加入本体信息进行学习，在该消融实验设置下，将本体信息替换为随机表示对未见关系进行特征学习。
3) 去除实体聚合表示的设置：本文模型在处理未见实体时，采用该实体周围的关系信息聚合表示，在该消融实验设置下对未见实体进行随机化嵌入设置，考察未见实体的表示模块。
4) 去除关系 GCN 聚合的设置：在构建完关系的位置图并引入本体嵌入后，本文模型通过 GCN 来加强关系对邻接结构的特征学习，该消融实验下去除两层 GCN 层，直接使用本体信息。
5) 同时去除本体和元学习设置：同时结合消融实验 (1) 和 (2) 的设置。

将消融实验结果与 NAMER(TransE) 模型在 NELL_Ext 上的实验结果进行比较，可以看出，基本所有的消融设置都会导致性能下降，表明上述各模块的重要性。但是在去除关系 GCN 聚合的设置上，在仅含有未见关系的测试集上效果反而提升，分析可知该层 GCN 的作用是在关系本体嵌入的基础上学习关系邻接关系的结构信息。这些结构信息的引入一定程度上会影响仅对关系的表示效果，导致在仅包含未见关系的测试集上的效果下降。但是这些结构信息的引入在未见实体的表示中，因为实体需要聚合邻接关系进行表示，因此引入的结构信息会对实体的表示效果进行提升，可以发现去掉 GCN 在包含未见实体和同时包含未见实体和未见关系的测试集上效果都有明显的下降，因此也侧面证明了 GCN 模型对实验效果提升的必要性。此外，本文观察到元学习设置对模型性能至关重要，表明在推广到测试知识图谱的任务上对模型进行元训练的有效性。

5.6 未见实体案例分析

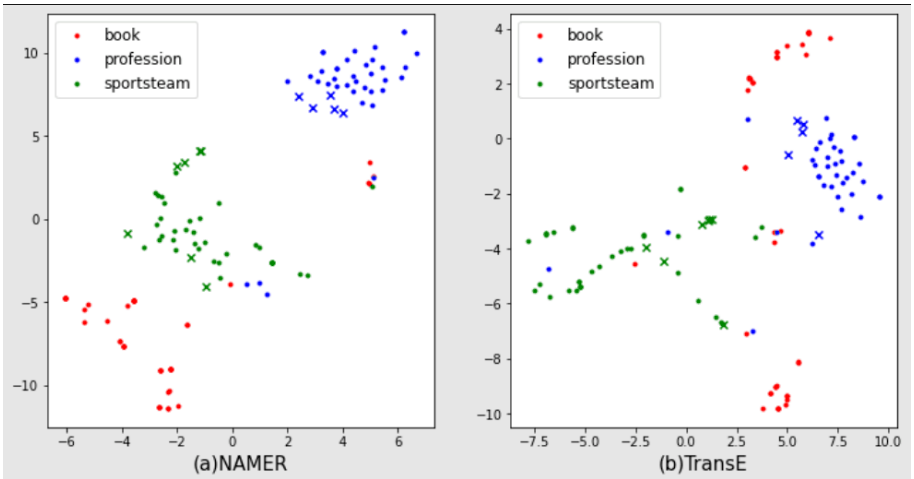

图 5.1 实体嵌入可视化分析

图5.1中展示了本文提出的 NAMER 模型和去除元学习和本体嵌入的传统 TransE 模型在 NELL-Ext 数据集的测试集上实体嵌入的可视化。图中不同颜色展示了不同类型的实体，圆点代表未见的实体，叉号则代表了该类型下的已知实体。NAMER 在实体获得初始化的向量表示后，通过两层 GCN 来学习邻域节点和相连关系的特征，使得同一类型的实体在表示空间上尽可能聚集在较紧密的邻域中。从可视化中也可以看出 NAMER 产生的嵌入分布与对应类型更加一致，在嵌入映射的距离上更近紧凑，而 TransE-KGE 产生的嵌入则混合了不同实体类型。NAMER 将嵌入映射到了不同的聚类中，而 TransE 中不同实体类型的嵌入则混合在一起。此外，本文的模型在对未见的实体进行表示学习的过程中，采用聚合实体关系的特征来初始化实体向量，相邻关系的结构特征可以表现出实体的类型信息，因此模型可以将未见的实体的嵌入与同一类型的已知实体聚类。不同类型实体的聚类表明，NAMER 能够用包含合理语义和信息性知识的嵌入来表示未见的实体。

5.7 未见关系案例分析

对于未见关系，本文同样选取了部分关系映射到了二维空间上进行可视化分析，如图5.2所示，其中圆点代表训练集中可见的关系，叉号代表测试集中的未见关系。

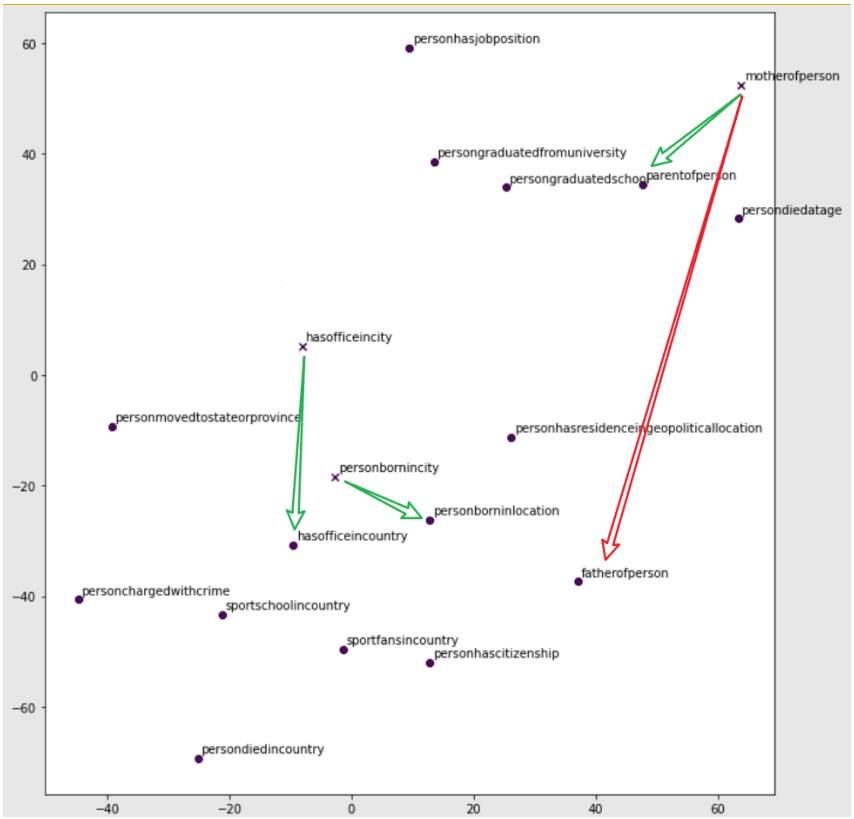

图 5.2 实体嵌入可视化分析

本文模型通过采用关系位置图、根据关系的相对位置对关系局部的邻接关系进行

建模学习到了关系的拓扑结构信息，同时引入本体信息作为语义补充，因此同一实体的具有相近语义的相邻关系在距离上应该表现为更加相近，从图上可以看出，对于未见关系 has_office_in_city 贴近于具有类似语义的关系 has_office_in_coutry，未见关系 person_born_in_city 更贴近于已知关系 person_born_in_location。而对于已知关系 parent_of_person、father_of_person 和未见关系 mother_of_person，本文可知一个实体如果存在 parent_of_person 的关系那么该实体节点的邻接关系中只能存在其中一个 father 或者 mother 的关系，因此在图中本文可以观察到 mother_of_person 在距离上更接近于 parent_of_person 关系，而远离 father_of_person 关系。由此可见，本文模型通过在关系的位置图上联合本体语义信息有效的学习到了对应关系的语义关系，且其中相似的关系在向量空间中靠近，证明了本文提出的 NAMER 在嵌入未知关系方面的有效性。

5.8 本章小结

本章将本文提出的模型在测试数据集上 NELL_Ext 和 DB_Ext 上进行了链接预测任务的相关实验。与多个基准模型相比，本文提出的模型在任务得分上均有不同程度的提升，并通过对实验结果的分析验证了模型对于表示学习效果增强的有效性。同时通过对各个模型组件的消融实验发现明显的效果下降，证明了模型模块的重要性；最后对未见实体和未见关系的案例分析可知得到的嵌入表示符合模型原理设定，再次证明了该模型对未见组件表示上的突出能力。

6 总结与展望

6.1 总结

当前，知识图谱技术的应用已经深入到了人们的日常生活中，从搜索引擎到推荐系统，不断挖掘知识图谱知识的应用是人工智能发展的基础。然而，出于数据隐私和成本等多方面的考虑，大到公司多个图谱存储服务器，小到人们每日使用的移动终端，我们无法将所有分散知识图谱新添加的实体和关系完全覆盖。因此，面向跨领域知识表示学习问题的研究已成为不可避免的需求和研究方向。

本文针对跨域知识图谱的知识表示学习问题，采用元学习的方法，在训练任务中模拟跨域场景下的未见关系和未见实体，从而获得了跨域知识表示的能力。元学习方法具有训练效率高的优点，可以在算力珍贵的时代大幅降低成本消耗。对于未见关系的表示，本文依据关系间的相对位置关系构建了一个以关系为结点的图，并通过预定义的元关系将其连接起来，以学习关系的拓扑信息。此外，本文将本体信息嵌入作为图谱语义信息的补充，通过图卷积网络对关系节点进行拓扑信息和语义信息的联合学习，获得了对未见关系的表示。对于未见实体的表示，本文认为相似类型的实体具有相似的邻接结构，并采用了实体的邻接关系特征聚合作为未见实体的初始化表示。为了能够充分利用到已知实体和关系的特征信息，本文使用图神经网络对实体和关系进行邻接信息的学习和更新。通过实验、基准模型比较和案例分析，本文证明了所提出模型的有效性。总的来说，本文的工作主要包括：

1) 提出了一个基于本体信息和元学习的跨域知识表示学习框架，采用元学习的任务设定在模型训练中对跨域场景进行模拟，并结合图的拓扑结构信息和本体语义信息对未见实体和未见关系进行建模，使得模型具备处理未见实体和关系的能力。
2) 对于未见关系和实体，本文创新性的同时考虑关系的拓扑结构和本体语义两个方面的特征，通过在本体视图中引入拓扑元关系，并利用图网络实现拓扑关系和语义信息的联合表示。
3) 在两个测试数据集上进行了模型实验，并将结果与多个基准模型进行了比较。实验结果表明本文模型相比于其他基准模型均有不同程度的提升，证明了本文模型的有效性。

6.2 未来工作

虽然本文在测试数据集上的效果相比于其他基准模型已经获得了明显的提升，但是在本文的整个研究过程中仍旧发现了当下模型在应用方面值得继续探究的几个研究方向：

1) 本文引入了本体信息，以增强跨域知识表示学习。其中，本体三元组是不可或缺的。然而，目前一些基准知识图谱缺少相应的本体三元组。例如，广泛使用的 FB15K-237 数据集，其源数据集已经停止维护，因此本体类型信息较为杂乱难以使用。如何更充分地获取数据集的本体信息仍然是待解决的难题。
2) 在 DB_Ext 数据集上的实验结果表明，基于规则提取的模型在处理未见的实体时也能表现出不错的效果。此外，规则可以同时作用于实体和关系，并对其表示学习进行约束，如何将规则信息融入到本文模型也是未来工作的一个值得期待的方向。

参考文献

[1] Zou X. A survey on application of knowledge graph[A]. Journal of Physics: Conference Series: Vol 1487[C], 2020: 012016.

[2] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[A]. Proceedings of the 2008 ACM SIGMOD international conference on Management of data[C], 2008: 1247–1250.

[3] Xiong W, Hoang T, Wang W Y. Deeppath: A reinforcement learning method for knowledge graph reasoning[J]. arXiv preprint arXiv:1707.06690, 2017.

[4] Bizer C, Lehmann J, Kobilarov G, et al. Dbpedia-a crystallization point for the web of data[J]. Journal of web semantics, 2009, 7(3): 154–165.

[5] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[A]. Proceedings of the 16th international conference on World Wide Web[C], 2007: 697–706.

[6] 张正航, 钱育蓉, 行艳妮, et al. 基于 TransE 的表示学习方法研究综述 [J]. 计算机应用研究, 2021, 38(03): 656–663.

[7] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[J]. Advances in neural information processing systems, 2013, 26.

[8] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[A]. Proceedings of the AAAI conference on artificial intelligence: Vol 28[C], 2014.

[9] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[A]. Proceedings of the AAAI conference on artificial intelligence: Vol 29[C], 2015.

[10] Yang B, Yih W-t, He X, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint arXiv:1412.6575, 2014.

[11] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[A]. International conference on machine learning[C], 2016: 2071–2080.

[12] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion[J]. Advances in neural information processing systems, 2013, 26.

[13] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions[A]. Proceedings of the AAAI Conference on Artificial Intelligence: Vol 30[C], 2016.

[14] Valverde-Rebaza J C, de Andrade Lopes A. Link prediction in complex networks based

on cluster information[A]. Advances in Artificial Intelligence-SBIA 2012: 21th Brazilian Symposium on Artificial Intelligence, Curitiba, Brazil, October 20-25, 2012. Proceedings[C], 2012 : 92 – 101.

[15] Lin Y, Liu Z, Luan H, et al. Modeling relation paths for representation learning of knowledge bases[J]. arXiv preprint arXiv:1506.00379, 2015.

[16] Feng J, Huang M, Yang Y, et al. GAKE: Graph aware knowledge embedding[A]. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers[C], 2016 : 641 – 651.

[17] Xie R, Liu Z, Luan H, et al. Image-embodied knowledge representation learning[J]. arXiv preprint arXiv:1609.07028, 2016.

[18] Li R, Cao Y, Zhu Q, et al. How does knowledge graph embedding extrapolate to unseen data: a semantic evidence view[A]. Proceedings of the AAAI Conference on Artificial Intelligence : Vol 36[C], 2022 : 5781 – 5791.

[19] 刘洪波, 陈越, 卢记仓, et al. 面向知识图谱的规则挖掘研究综述 [J]. 计算机工程与应用, : 1 – 11.

[20] Galárraga L A, Teflioudi C, Hose K, et al. AMIE: association rule mining under incomplete evidence in ontological knowledge bases[A]. Proceedings of the 22nd international conference on World Wide Web[C], 2013 : 413 – 422.

[21] Omran P G, Wang K, Wang Z. Scalable Rule Learning via Learning Representation.[A]. IJCAI[C], 2018 : 2149 – 2155.

[22] Omran P G, Wang Z, Wang K. Learning Rules With Attributes and Relations in Knowledge Graphs.[A]. AAAI Spring Symposium: MAKE[C], 2022.

[23] Zhang W, Paudel B, Wang L, et al. Iteratively learning embeddings and rules for knowledge graph reasoning[A]. The world wide web conference[C], 2019 : 2366 – 2377.

[24] 刘藤, 陈恒, 李冠宇. 联合 FOL 规则的知识图谱表示学习方法 [J]. 计算机工程与应用, 2021, 57(04) : 100 – 107.

[25] Yang F, Yang Z, Cohen W W. Differentiable learning of logical rules for knowledge base reasoning[J]. Advances in neural information processing systems, 2017, 30.

[26] Sadeghian A, Armandpour M, Ding P, et al. Drum: End-to-end differentiable rule mining on knowledge graphs[J]. Advances in Neural Information Processing Systems, 2019, 32.

[27] Teru K, Denis E, Hamilton W. Inductive relation prediction by subgraph reasoning[A]. International Conference on Machine Learning[C], 2020 : 9448 – 9457.

[28] Chen J, He H, Wu F, et al. Topology-aware correlations between relations for inductive link prediction in knowledge graphs[A]. Proceedings of the AAAI Conference on Artificial Intelligence : Vol 35[C], 2021 : 6271 – 6278.

[29] Hamaguchi T, Oiwa H, Shimbo M, et al. Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach[J]. arXiv preprint arXiv:1706.05674, 2017.

[30] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[31] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[A]. The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15[C], 2018: 593–607.

[32] Ye R, Li X, Fang Y, et al. A Vectorized Relational Graph Convolutional Network for Multi-Relational Network Alignment.[A]. IJCAI[C], 2019: 4135–4141.

[33] Cai L, Yan B, Mai G, et al. TransGCN: Coupling transformation assumptions with graph convolutional networks for link prediction[A]. Proceedings of the 10th international conference on knowledge capture[C], 2019: 131–138.

[34] Vashishth S, Sanyal S, Nitin V, et al. Composition-based multi-relational graph convolutional networks[J]. arXiv preprint arXiv:1911.03082, 2019.

[35] Nathani D, Chauhan J, Sharma C, et al. Learning attention-based embeddings for relation prediction in knowledge graphs[J]. arXiv preprint arXiv:1906.01195, 2019.

[36] Chen M, Zhang Y, Kou X, et al. r-GAT: Relational Graph Attention Network for Multi-Relational Graphs[J]. arXiv preprint arXiv:2109.05922, 2021.

[37] Chen M, Zhang W, Zhang W, et al. Meta relational learning for few-shot link prediction in knowledge graphs[J]. arXiv preprint arXiv:1909.01515, 2019.

[38] Lv X, Gu Y, Han X, et al. Adapting meta knowledge graph information for multi-hop reasoning over few-shot relations[J]. arXiv preprint arXiv:1908.11513, 2019.

[39] Niu G, Li Y, Tang C, et al. Relational learning with gated and attentive neighbor aggregator for few-shot knowledge graph completion[A]. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval[C], 2021: 213–222.

[40] Zheng S, Mai S, Sun Y, et al. Subgraph-aware few-shot inductive link prediction via meta-learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2022.

[41] Chen M, Zhang W, Zhu Y, et al. Meta-knowledge transfer for inductive knowledge graph embedding[A]. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval[C], 2022: 927–937.

[42] Chen M, Tian Y, Chen X, et al. On2vec: Embedding-based relation prediction for ontology population[A]. Proceedings of the 2018 SIAM International Conference on Data Mining[C], 2018: 315–323.

[43] Guo S, Wang Q, Wang B, et al. SSE: Semantically smooth embedding for knowledge

graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(4): 884–897.

[44] Xie R, Liu Z, Sun M, et al. Representation learning of knowledge graphs with hierarchical types.[A]. IJCAI: Vol 2016[C], 2016: 2965–2971.

[45] Liu S, Grau B, Horrocks I, et al. Indigo: Gnn-based inductive knowledge graph completion using pair-wise encoding[J]. Advances in Neural Information Processing Systems, 2021, 34: 2034–2045.

[46] Zhao M, Jia W, Huang Y. Attention-based aggregation graph networks for knowledge graph information transfer[A]. Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24[C], 2020: 542–554.

[47] Stackelberg H v, others. Theory of the market economy[J], 1952.

[48] Franceschi L, Frasconi P, Salzo S, et al. Bilevel programming for hyperparameter optimization and meta-learning[A]. International Conference on Machine Learning[C], 2018: 1568–1577.

[49] Sinha A, Malo P, Deb K. A review on bilevel optimization: From classical to evolutionary approaches and applications[J]. IEEE Transactions on Evolutionary Computation, 2017, 22(2): 276–295.

[50] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[A]. International conference on machine learning[C], 2017: 1126–1135.

[51] Ma S, Ding J, Jia W, et al. Transt: Type-based multiple embedding representations for knowledge graph completion[A]. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10[C], 2017: 717–733.

[52] Hao J, Chen M, Yu W, et al. Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts[A]. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining[C], 2019: 1709–1719.

[53] Mai S, Zheng S, Yang Y, et al. Communicative message passing for inductive relation reasoning[A]. Proceedings of the AAAI Conference on Artificial Intelligence: Vol 35[C], 2021: 4294–4302.

致谢

岁月流转，转眼已经陪伴武汉大学六个春秋。随着毕业论文的完结，我的研究生生涯也即将告一段落。

在度过两年平平无奇的大学生涯后，我很庆幸能够鼓起勇气敲开了彭敏教授办公室的门，向她表达了希望能够通过实验室的实习来提升自己的意愿。更让我高兴的是彭老师愿意给我尝试的机会，最终我也如愿留在了实验室从事科研和项目的工作。我从心底里非常感谢我的指导老师彭敏教授，是您一直精益求精的教学精神和耐心细致的指导，让我能够在探索学术领域的路上有所收获。也是在实验室项目的不断磨砺中，让我从一个什么都不太懂的项目小白成功收获了不错的工作 offer，您的教诲和启发，对我而言永久难忘，必将深深烙印在我的心中。

同时，我要感谢我所在的实验室团队。在他们的帮助下，我体会到科研和工作的融洽和乐趣。尤其是刘奔博士在我毕业论文撰写的过程中，不辞辛苦地为我解忧答惑，对我遇到的每一个问题都能够给我指明一个正确的解决方向。也是在刘奔博士的悉心帮助下，我能够对论文的结构进行优化和调整，完成了整个论文的修改工作。除此之外其他的实验室成员也都是我求学路上重要的伙伴，正是在他们的支持下，我能够勇往直前并应对每一个挑战。

最后，我要感谢我的家人，是他们一直支撑着我继续求学，不断给予我鼓励和支持。我还要感谢我的女友月月，是她在我心情最失落、最困难的时候陪伴着我，让我能够有继续走下去的勇气。

山水一程，愿我们各自未来的旅程都继续迸发出新的精彩！

攻硕期间取得的学术成果和参与的项目

[1] 湖北省科技厅重点研发计划：智慧农业·草莓大棚智能化生产

[2] 武汉市智慧城管信息化建设项目：武汉市城管应急管理平台

[3] 岩土工程监测管理及预测预警系统

[4] 东湖核心区污水传输系统工程主隧结构健康监测平台

武汉大学学位论文使用授权协议书

本学位论文作者愿意遵守武汉大学关于保存、使用学位论文的管理办法及规定，即：学校有权保存学位论文的印刷本和电子版，并提供文献检索与阅览服务；学校可以采用影印、缩印、数字化或其它复制手段保存论文；在以教学与科研服务为目的前提下，学校可以在校园网内公布部分或全部内容。

1、 在本论文提交当年，同意在校园网内以及中国高等教育文献保障系统（CALIS）、高校学位论文系统提供查询及前十六页浏览服务。

2、 在本论文提交 □当年/ □一年/□两年/ □三年以后，同意在校园网内允许读者在线浏览并下载全文，学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。（保密论文解密后遵守此规定）

论文作者（签名）：_____

学　　号：_____

学　　院：_____

日期：　　年　　月　　日