



融合实体类别信息的知识图谱表示学习

金婧, 万怀宇, 林友芳

(北京交通大学 计算机与信息技术学院 交通数据分析与挖掘北京市重点实验室, 北京 100044)

摘要: 知识图谱表示学习通过将实体和关系嵌入连续低维的语义空间中, 获取实体和关系的语义关联信息。设计一种融合实体类别信息的类别增强知识图谱表示学习(CEKGRL)模型, 构建基于结构与基于类别的实体表示, 通过注意力机制捕获实体类别和三元组关系之间的潜在相关性, 结合不同实体类别对于某种特定关系的重要程度及实体类别信息进行知识表示学习。在知识图谱补全和三元组分类任务中的实验结果表明, CEGRL模型在MeanRank和Hit@10评估指标上均取得明显的性能提升, 尤其在实体预测任务的Filter设置下相比TKRL模型约分别提升了23.5%和7.2个百分点, 具有更好的知识表示学习性能。

关键词: 知识图谱; 知识表示学习; 多源信息融合; 注意力机制; 实体消歧

开放科学(资源服务)标志码(OSID):



中文引用格式: 金婧, 万怀宇, 林友芳. 融合实体类别信息的知识图谱表示学习[J]. 计算机工程, 2021, 47(4): 77-83.

英文引用格式: JIN Jing, WAN Huaiyu, LIN Youfang. Knowledge graph representation learning fused with entity category information[J]. Computer Engineering, 2021, 47(4): 77-83.

Knowledge Graph Representation Learning Fused with Entity Category Information

JIN Jing, WAN Huaiyu, LIN Youfang

(Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

[Abstract] Knowledge graph representation learning embeds both entities and relations into a continuous low-dimensional semantic space, so as to obtain the semantic correlation between entities and relations. This paper proposes a Category-Enhanced Knowledge Graph Representation Learning(CEKGRL) model fused with entity category information. The model constructs structure-based and category-based entity representation, and captures the potential correlation between entity categories and triple relations by introducing the attention mechanism. It combines the different importance of different entity categories for a specific relation and entity category information for Knowledge Representation Learning(KRL). The performance of the model in knowledge graph completion and triple classification tasks is tested, and experimental results show that the CEGRL model has made significant improvements in the indicators of MeanRank and Hit@10, which are increased by about 23.5% and 7.2 percentage points than the TKRL model in the Filter setting of the entity prediction task. The results indicate that the model has better KRL performance.

[Key words] knowledge graph; Knowledge Representation Learning(KRL); multi-source information fusion; attention mechanism; entity disambiguation

DOI: 10.19678/j.issn.1000-3428.0057353

0 概述

知识图谱是推动人工智能学科发展和支撑智能信息服务应用的重要技术, 可将人类知识构建成结构化的知识系统。在知识图谱中知识通常以三元组的形式进行表示, 知识图谱以网络图的形式来构建整个知识

系统, 知识表示作为知识图谱中知识获取和应用的基础, 可提升知识图谱的认知和推理能力^[1-2]。随着Freebase、DBpedia等大型知识图谱被提出, 基于网络形式的知识表示在大规模知识图谱下存在计算效率低下和数据稀疏等问题^[3-4]。近年来, 以深度学习为代表的知识图谱表示学习技术得到了广泛关注, 其旨在将

基金项目: 国家重点研发计划(2018YFC0830200)。

作者简介: 金婧(1994—), 女, 硕士研究生, 主研方向为数据挖掘、知识表示学习; 万怀宇(通信作者), 副教授; 林友芳, 教授。

收稿日期: 2020-02-10 修回日期: 2020-03-27 E-mail: hywan@bjtu.edu.cn

研究对象映射到一个连续低维的向量空间中,以便于高效计算实体和关系的语义相似度,同时能有效解决数据稀疏问题。

翻译模型是一种主流的知识表示学习模型,因简单和高效的特点而备受关注,并且许多在翻译模型基础上进行改进的变体模型被陆续提出。这些模型不仅利用了知识图谱所固有的结构信息,而且考虑了实体描述信息、类别信息和图像信息等与实体相关的多源信息,大幅提高了知识表示学习性能。TKRL模型^[5]是一种利用实体类别信息作为外部信息的知识表示学习模型,在该模型中不同类别的实体具有不同的表示,对于实体类别的层次结构,利用两种编码类型对层级结构进行建模,最终证实了实体类别可以在知识表示学习中发挥重要作用。然而,TKRL模型依赖于具有层次结构的类别信息及事先制定好的规则约束,该规则约束具体为当给定一种关系时,约定了该关系的头实体和尾实体的具体类别,但该规则约束对于现实世界的的数据而言不具备灵活性,并且不仅TKRL模型需要利用事先制定好的规则约束,而且很多其他融合实体类别信息的翻译模型也都基于类似的规则。本文建立一种融合实体类别信息的类别增强知识图谱表示学习(Category-Enhanced Knowledge Graph Representation Learning, CEKGRL)模型,引入基于类别的实体表示,通过注意力机制学习实体类别和关系之间的相关性,并结合实体类别信息进行知识表示学习。

1 相关工作

知识表示是对知识进行描述的有效途径,旨在研究如何更准确地表示知识的语义信息以更好地利用知识图谱,从而使得计算机能够接受并运用知识,最终达到智能的目标。知识表示学习是通过机器学习的方式将知识(知识图谱中的实体和关系)表示为稠密低维的实值向量,有效解决了数据稀疏问题,并且学习到的知识表示能够保留知识图谱中的结构和语义关系,从而高效计算实体和关系之间的语义相似度,使其广泛应用于知识图谱补全、自动问答和实体链接等下游任务中。

近年来,随着深度学习技术的发展,知识表示学习方法取得较大进展。以TransE^[2]为代表的翻译模型是知识表示学习中的热门模型,这类模型将关系向量作为头实体向量到尾实体向量之间的平移,即假设尾实体向量 t 近似于头实体向量和关系向量的和 $(h+r)$,并定义能量函数为 $E(h, r, t) = \|h+r-t\|$ 。TransE模型因参数少及计算复杂度低,在1-1简单关系中具有较好的性能表现,但对于1-N、N-1和N-N等复杂关系,由于TransE模型的建模方式过于简单,因此存在一定的局限性。为解决该问题,后续出现了许多以TransE为基础的改进模型,如TransH、TransAH、TransA、TransG、TransR和TransD等。TransH通过将头实体、尾实体向量投影到对应关系

的超平面上,从而令一个实体在不同的关系下具有不同的表示^[6]。TransAH模型在TransH模型的基础上引入了一种自适应的度量方法,通过加入对角权重矩阵将得分函数中的度量由欧氏距离转换为加权欧氏距离^[7]。TransA模型中的自适应度量方法为每一种关系定义一个非负的对称矩阵,从而对表示向量中的每一个维度添加权重,增加了模型的表示能力^[8]。TransG模型使用高斯混合来刻画实体间的多种语义关系,利用最大相似度原理训练数据,解决了多语义问题^[9]。TransR模型假设不同的关系具有不同的语义空间,因此将每个实体投影到对应的关系空间中^[10]。TransD模型通过设置两个关系-实体投影矩阵,并结合头、尾实体位置的属性,解决了TransR模型参数过多的问题^[11]。

除了翻译模型及其改进模型以外,研究人员还提出了一些其他类型的知识表示学习模型,主要包括:1)距离模型,将头、尾实体向量通过投影矩阵投影至对应空间,并通过计算投影向量的距离来反映实体间的语义相似度,如SE模型^[12];2)能量模型,通过定义若干投影矩阵,并利用双线性函数刻画实体与关系的内在联系,如SME模型^[13-14];3)矩阵分解模型,通过矩阵分解的方式得到低维向量表示,如RESCAL模型^[15-16];4)双线性模型,利用基于关系的双线性变换刻画实体和关系之间的二阶联系,如LFM模型^[17]。

以上模型仅利用了知识图谱自身所包含的三元组结构信息,但除了结构信息以外,还有大量与知识相关的其他信息没有得到有效利用,如知识库中所包含的实体和关系的描述信息、类别信息以及知识库以外的海量互联网文本信息等。这些多源信息提供了知识图谱中三元组结构信息以外的额外信息,有助于更准确地学习知识表示。NTN模型^[18]使用实体中单词嵌入的平均值表示实体,从而捕捉实体之间的潜在文本关系。DKRL模型^[19]通过考虑实体的描述信息文本来编码实体描述的语义信息。IKRL模型^[20]引入实体图像信息,并利用神经网络构造实体图像的表示。TKRL模型^[5]通过引入具有层次结构的类别信息以及实体类别与关系之间的约束信息来提高知识表示能力。但并非所有实体类别都具有层次结构,且实体类别与关系的约束方式不具备普适性和灵活性。为解决上述问题,本文提出一种融合实体类别的CEKGRL模型。该模型利用数据集最底层的实体类别,通过注意力机制捕获实体类别和关系之间的相关性,并利用注意力分数对类别表示进行加权以学习知识表示。

2 CEKGRL模型

知识图谱通常包含实体的类别信息,而类别信息作为实体属性的一部分,能够起到补充实体语义

信息的作用。为有效融合知识图谱中的实体类别信息,同时兼顾翻译模型的高效性,本文提出 CEKGRL 模型,其在 TransE 模型的基础上引入实体的类别表示,旨在通过学习三元组知识的同时,能够通过类别信息得到更加准确的知识表示。该模型无需依赖实体类别与关系之间的固定映射,便于将模型灵活地迁移到其他更加复杂且难以得到该映射关系的场景中。同时,CEKGRL 模型对实体类别的组织形式没有要求,通过将类别的组织结构进行扁平化处理,可适应各种应用场景对类别信息格式的要求,无论是 FB15K 中具有层次结构的类别信息,还是其他形式的类别数据均可以使用。

为更清晰地表述 CEKGRL 模型的基本思想,图 1 通过具体实例说明了实体类别与三元组关系之间的语义相关性。图 1(a)左侧的 George Washington 代表乔治·华盛顿这一实体,其右侧的矩形代表列举出的实体所包含的部分类别属性,包括政治家、美国国会议员、死者、人、名称来源和电影主题。图 1(b)列举了与乔治·华盛顿这一实体有关的两个三元组,括号中的内容从左到右分别是头实体、关系和尾实体,其中,矩形代表乔治·华盛顿的实体类别属性,直线代表类别与关系之间的相关性,直线以及矩形颜色越深代表实体类别与关系之间的相关性越强。以知识图谱中与乔治·华盛顿实体相关的两个三元组为例,乔治·华盛顿的“政治家”和“美国国会议员”这两个类别在(美国大陆会议,官员,乔治·华盛顿)三元组中比其他类别更具相关性,而在(肺炎,死因,乔治·华盛顿)三元组中,“死者”则能表达出更多相关的信息。这说明了同一个实体的不同类别在不同的三元组关系中可以起到提供语义信息的作用,并且不同类别的重要程度与三元组的关系存在一定的关联关系。在此情况下,实体的类别信息可以丰富实体的表示,使知识表示具有更多的语义信息。

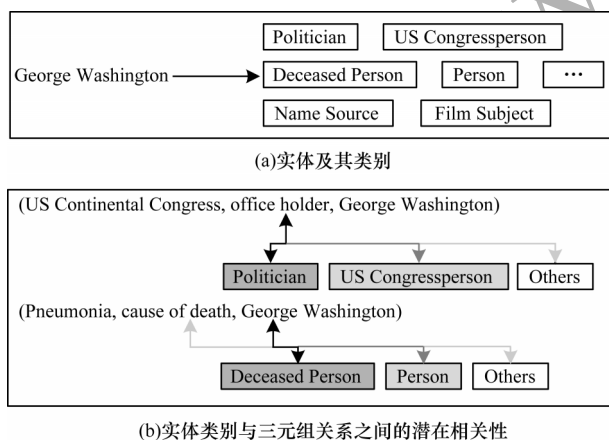


图 1 实体类别与三元组关系之间的语义相关性

Fig.1 Semantic correlation between entity category and triple relationship

为更清晰地描述 CEKGRL 模型,本文给出相关的定义和符号表示,将知识图谱定义为 $G=(E, R, S)$, 其中: E 为实体集; R 为关系集; $S \subseteq E \times R \times E$ 表示三元组集合,三元组集合用 (h, r, t) 进行表示, h 、 r 和 t 分别代表头实体、关系和尾实体。此外,本文引入类别概念,用 C 表示类别集合,并定义基于结构和基于类别的实体表示,分别代表从知识图谱的三元组中学习到的实体表示以及引入类别表示所得到的实体表示。

CEKGRL 模型的整体架构如图 2 所示,其中,斜线状的圆圈组成的椭圆代表基于结构的向量表示,网格状的圆圈组成的椭圆代表基于类别的向量表示,实心圆圈组成的椭圆代表关系的向量表示,空心的圆圈组成的椭圆代表实体类别的向量表示, a 表示注意力分数。为将两种表示类型进行融合,定义能量函数为:

$$E = E_{ss} + \beta E_{cc} \quad (1)$$

其中: $E_{ss} = \|h_s + r - t_s\|$, 为头实体、尾实体使用基于结构的实体表示得到的能量函数; h_s 、 t_s 分别为基于结构的头实体、尾实体表示; h_c 、 t_c 分别为基于类别的头实体、尾实体表示; 超参数 β 用于调整基于类别的表示在 CEKGRL 模型中的重要程度; $E_{cc} = \|h_c + r - t_c\|$, 为头实体、尾实体使用基于类别的实体表示得到的能量函数。需要说明的是,实体基于结构和基于类别的表示在训练过程中都使用统一的关系表示 r , 保证了两种类型的向量表示空间可通过相同的关系表示达到统一。

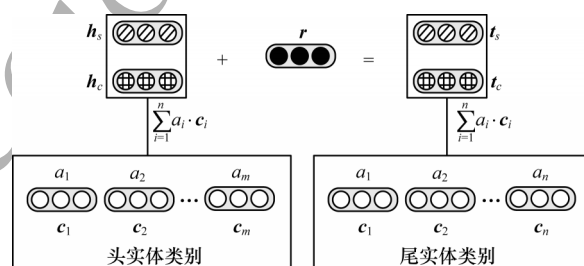


图 2 CEKGRL 模型的整体架构

Fig.2 Overall architecture of CEKGRL model

在训练过程中,首先通过注意力机制得到实体类别表示与三元组关系的相关性,即注意力分数,然后利用该注意力分数对类别表示进行加权求和并将其作为基于类别的实体表示,最后将相同的关系表示作为两种表示空间的联系,将基于结构和基于类别的表示进行联合训练。

2.1 注意力机制

实体的不同类别信息可以从多个角度刻画实体,而同一个实体在不同的关系下会侧重关注其不同的类别信息,具体表现为同一实体的不同类别与不同关系之间的语义相关性不同。为有效利用三元组中关系和

实体类别之间存在的潜在相关性,本文通过以下注意力机制计算并得到两者之间的相似度:

1)基于相似度的注意力(Similarity-based Attention, SA)机制。受STKRL模型^[21]中注意力机制的启发,将实体类别与三元组关系之间的相关性定义为两者向量表示的相似度,并采用余弦相似度进行计算,公式如下:

$$\text{att}(\mathbf{c}, \mathbf{r}) = \frac{\mathbf{c} \cdot \mathbf{r}}{\|\mathbf{c}\| \cdot \|\mathbf{r}\|} \quad (2)$$

其中, $\text{att}()$ 为求解注意力分数 a 的函数, \mathbf{c} 为类别的向量表示。

2)缩放点积注意力(Scaled Dot-Product Attention, SDPA)机制。基于文献[22]中的注意力计算方法,结合CEKGRL模型将关系 \mathbf{r} 作为 query 向量,类别 \mathbf{c} 同时作为 key 向量和 value 向量。在实现过程中,为加快处理效率,通过矩阵的形式计算注意力,因此将多个关系的表示向量及其对应的类别表示向量分别拼接为关系矩阵 \mathbf{R} 和类别矩阵 \mathbf{C} 。然后,引入待训练的权重矩阵 \mathbf{W}^Q 、 \mathbf{W}^K 和 \mathbf{W}^V ,将权重矩阵、关系矩阵和类别矩阵分别做矩阵相乘操作,得到 query、key 和 value 对应的矩阵 \mathbf{Q} 、 \mathbf{K} 、 \mathbf{V} 及注意力分数,如式(3)~式(6)所示:

$$\mathbf{Q} = \mathbf{R} \times \mathbf{W}^Q \quad (3)$$

$$\mathbf{K} = \mathbf{C} \times \mathbf{W}^K \quad (4)$$

$$\mathbf{V} = \mathbf{C} \times \mathbf{W}^V \quad (5)$$

$$\text{att}(\mathbf{C}, \mathbf{R}) = \text{Softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \quad (6)$$

其中, $\mathbf{C} \in \mathbb{R}^{|\mathbf{C}| \times k}$, $\mathbf{R} \in \mathbb{R}^{|\mathbf{R}| \times k}$, $\mathbf{W}^Q \in \mathbb{R}^{k \times d_k}$, $\mathbf{W}^K \in \mathbb{R}^{k \times d_k}$, $\mathbf{W}^V \in \mathbb{R}^{k \times k}$, k 为实体和关系的向量维度, d_k 为权重矩阵维度。

通过以上两种注意力机制计算得到的注意力分数越高,说明类别 \mathbf{c} 与关系 \mathbf{r} 的相关性越强。因此,本文利用注意力分数对各个类别表示赋予不同权重,再对加权后的所有表示求和得到对应的实体表示,即基于类别的实体表示,其在矩阵形式下的计算公式如下:

$$\mathbf{E}_c = \text{Softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V} \quad (7)$$

其中, \mathbf{E}_c 为基于类别的实体表示向量集合。

2.2 模型训练

CEKGRL模型与TransE模型的训练目标相同,本文采用最大间隔方法增强知识表示的区分能力,定义目标函数为:

$$L = \sum_{(\mathbf{h}, \mathbf{r}, \mathbf{t}) \in T} \sum_{(\mathbf{h}', \mathbf{r}', \mathbf{t}') \in T'} \max(\gamma + E(\mathbf{h}, \mathbf{r}, \mathbf{t}) - E(\mathbf{h}', \mathbf{r}', \mathbf{t}'), 0) \quad (8)$$

其中, $E(\mathbf{h}, \mathbf{r}, \mathbf{t})$ 为正例三元组的能量函数, $E(\mathbf{h}', \mathbf{r}', \mathbf{t}')$ 为负例三元组的能量函数, γ 为间隔的超参数且 $\gamma > 0$, T 为训练集, T' 为利用 T 进行负采样得到的集合,定义为:

$$T' = \{(\mathbf{h}', \mathbf{r}, \mathbf{t}) | \mathbf{h}' \in E\} \cup \{(\mathbf{h}, \mathbf{r}, \mathbf{t}') | \mathbf{t}' \in E\} \cup \{(\mathbf{h}, \mathbf{r}', \mathbf{t}) | \mathbf{r}' \in R\} \quad (9)$$

其中,头实体、尾实体或者关系被随机替换为其他实体或者关系,另外,如果替换后的新三元组仍在 T 中,则不会被当作负样本。

在模型训练过程中,实体、关系和类别的表示均可以随机初始化,实体和关系的表示也可以采用简单的翻译模型预训练得到。在具体的模型实现过程中,为使初始的类别表示具有一定的语义信息,本文借助预训练得到的实体表示对其进行初始化,相较于随机初始化能够缩短模型收敛时间。初始化类别的具体方法为利用所有包含类别 c_i 的实体表示的平均值作为该类别的表示,形式化为:

$$\mathbf{c}_i = \frac{1}{|e_{c_i}|} \sum_{j=1}^{|e_{c_i}|} \mathbf{e}_j \quad (10)$$

其中, $|e_{c_i}|$ 表示具有 c_i 类别的实体数量且 i 满足 $i \in [1, |\mathbf{C}|]$ 。

CEKGRL模型在最小化目标函数的同时,可学习基于结构和基于类别的两种表示,并在模型训练过程中,采用Adam^[23]优化算法提升学习效果。

3 实验与结果分析

3.1 实验数据集

实验使用FB15K数据集,通过知识图谱补全和三元组分类任务对模型进行性能评估。FB15K是从Freebase中抽取出的数据集,在实验中将其划分为训练集、验证集和测试集,具体的统计信息如表1所示,其中, #Rel表示关系, #Ent表示实体, #Train表示训练集, #Valid表示验证集, #Test表示测试集。所有的事实三元组即实验中的训练集、验证集和测试集的并集,在下文中称为黄金三元组。

表1 FB15K数据集统计信息
Table 1 Statistics of FB15K dataset

对比项目	项目数量
#Rel	1 345
#Ent	14 951
#Train	483 142
#Valid	50 000
#Test	59 071

对于类别数据,本文采用文献[5]公开的数据集,该数据集包含Freebase知识库所涉及的type/instance字段,即类别信息,通过匹配Freebase中FB15K所包含的实体,并为这些实体添加知识库中实体对应的类别信息得到。在数据处理过程中,发现有10个实体出现在原始FB15K数据集中,但没有与之对应的实体类别信息。在处理数据缺失问题时,为保证原始FB15K数据

的完整性,需要保留这10个实体及其所涉及的所有三元组,使这10个实体也具有类别信息。经过数据统计发现,99%的实体都包含 common/topic 类别,因此在实验中采用众数规则对这10个缺失类别的实体人为添加 common/topic 类别。经过处理的数据集具有3 852个类别,每个实体平均约有12个类别。

3.2 实验设置

为验证 CEKGRL 模型的学习效果,将其与 TransE、TransR 和 TKRL 等模型进行对比,在训练阶段对 TransE 模型增加关系负采样操作,提升关系预测性能。对于 TransR 模型,本文采用文献[10]的开源代码进行实验,并与 TransE 模型在负采样过程中的操作相同,在生成负样本时也对关系进行替换操作。对于 RESCAL、SE、SME、LFM 及 TKRL 模型,本文直接引用文献[5,10]中的实验结果。

关于模型的参数选择问题,实验设置初始学习率 α 为 0.000 5、0.001 0、0.002 0,批量大小 B 为 20、240、1 200、4 800,实体和关系的向量维度 k 为 50、100、200,阈值 γ 为 0.5、1.0、1.5、2.0。对于缩放点积注意力机制,设置权重矩阵中的 d_k 为 49、64、100。考虑到基于结构和基于类别的表示所起作用不同,因此本文为基于类别的表示设置权重,用超参数 β 进行表示,超参数 β 用于调整其在 CEKGRL 模型中的重要程度。实验得到的模型最优参数设置为 $\alpha=0.001 0$ 、 $B=4 800$ 、 $k=200$ 、 $\gamma=1.0$ 、 $d_k=100$ 和 $\beta=0.5$ 。

3.3 知识图谱补全实验与结果分析

知识图谱补全任务是在给定事实三元组 (h, r, t) 中两项的前提下预测缺失的一项,即给定 (h, r) 预测 t ,给定 (r, t) 预测 h 或给定 (h, t) 预测 r ,因此知识图谱补全包括实体预测和关系预测这两个子任务。

本文采用 MeanRank 和 Hit@ n 两种评估指标,其分别表示正确的实体和关系在预测结果中的平均排名以及正确的实体和关系排在预测结果前 n 名的比例。针对每个指标给定 Raw 和 Filter 两种不同设置,Raw 设置只要预测结果不是当前三元组所期待的结果,就将其视作错误的预测结果,即使该预测结果属于黄金三元组,Filter 设置则是剔除属于黄金三元组的预测结果后所得的预测结果。

对于这两种设置,Raw 设置会忽略黄金三元组的存在,如果预测出的结果属于黄金三元组,但并非当前所关注的特定三元组,则认为预测结果错误,从而导致预测性能变差,但这部分由于黄金三元组而造成预测错误的结果,实际上的预测结果为正确,不应影响模型预测性能,因此本文认为 Filter 设置的预测结果更具说服力。

3.3.1 实体预测

CEKGRL 模型在实体预测任务中的评估结果如表 2 所示,结果表明 CEKGRL 模型除了 MeanRank 的 Raw 指标较 TKRL 和 TransR 模型略低以外,其他指标均得到提升。在 Filter 设置下,与 TKRL 模型相比,CEKGRL(SA)模型的 Hit@10 指标约提升了 7.2 个百分点,MeanRank 指标提升了约 23.5%。

表 2 实体预测的评估结果

Table 2 Evaluation results on entity prediction

模型	MeanRank		Hit@10/%	
	Raw	Filter	Raw	Filter
RESCAL	828	683	28.4	44.1
SE	273	162	28.8	39.8
SME(linear)	274	154	30.7	40.8
SME(bilinear)	284	158	31.3	41.3
LFM	283	164	26.0	33.1
TransE	250	102	46.1	69.6
TransR	199	77	47.2	67.2
TKRL	184	68	49.2	69.4
CEKGRL(SDPA)	205	53	49.2	76.1
CEKGRL(SA)	205	52	49.3	76.6

3.3.2 关系预测

关系预测任务的评估结果如表 3 所示,结果表明 CEKGRL(SA)模型的 MeanRank 指标优于其他模型,这说明 CEKGRL 模型具有较好的关系预测性能。同时可以看出,TKRL 模型的 Hit@1 指标略优于 CEKGRL 模型,主要原因为 TKRL 模型利用关系与类别之间的约束关系来对层次结构信息进行编码,相较 CEKGRL 模型额外引入了约束关系信息来提升模型性能。若要获得该约束关系,则需要对数据集有一定的要求或者对一些不容易提取的约束关系数据集进行人工构造,这样会导致 TKRL 模型的通用性和灵活性变差。本文提出的 CEKGRL 模型获取约束信息的方式更具普适性和灵活性,适用于基于多源信息融合的知识表示学习。

表 3 关系预测的评估结果

Table 3 Evaluation results on relation prediction

模型	MeanRank		Hit@1/%	
	Raw	Filter	Raw	Filter
TransE	2.79	2.43	68.4	87.2
TransR	2.49	2.09	70.2	91.6
TKRL	2.12	1.73	71.1	92.8
CEKGRL(SDPA)	2.36	1.95	70.3	91.8
CEKGRL(SA)	2.11	1.69	69.4	92.2

3.4 三元组分类实验与结果分析

三元组分类是一个二分类任务,用于判断给定的三元组是否准确。在生成负样本三元组时,本文采取与文献[18]相同的策略,对生成负样本时所需替换的实体或者关系进行一定的限制,使得负样本难以区分,从而提升模型在三元组分类任务中的性能。在分类过程中,对于给定的三元组 (h, r, t) ,如果其注意力得分低于给定的阈值 γ ,则预测其为正确的三元组,反之为错误的三元组。每种关系的阈值设置不同,具体通过最大化验证集中对应关系下的分类准确率进行设置。三元组分类的评估结果如表 4 所示,结果表明 CEKGRL 模型具有较优的分类性能。

表4 三元组分类的评估结果

Table 4 Evaluation results on triple classification %

模型	准确率
TransE	83.1
TransR	83.7
CEKGRL(SDPA)	84.9
CEKGRL(SA)	83.4

3.5 案例分析

为进一步验证CEKGRL模型可以学习到特定关系下不同类别的相关性,并更清晰地表示模型的作用效果,本文通过具体案例进行分析与说明。图3给出了在(Gangs of New York, film_festivals, 2010 Berlin Film Festival)三元组中,2010 Berlin Film Festival作为尾实体所具有的类别在实验中的注意力分数排名,其中,“Head:Gangs of New York”表示头实体为Gangs of New York(电影名称),“Relation:film_festivals”表示关系为film_festivals,“Tail(interest):2010 Berlin Film Festival”表示尾实体为2010 Berlin Film Festival,并且是本文所关注的类别排名的实体。根据类别注意力分数得到的排名结果可以看出,排在最靠前的类别与三元组中的关系相关性最强,排在靠后位置的类别一般覆盖范围更广。

Head: Gangs of New York Relation: film_festivals Tail(interest): 2010 Berlin Film Festival	
Rank	Categories of tail
1	/film/film_festival_event
2	/film/film_screening_venue
3	/time/event
4	/common/topic

图3 2010 Berlin Film Festival实体类别根据注意力分数的排名情况

Fig.3 Rank of 2010 Berlin Film Festival entity category according to attention score

由于CEKGRL模型可以区分出不同关系中实体类别的重要程度,因此当具有一词多义的实体处于不同语境时,可以通过最相关的类别来判断其具体含义,用于辅助实体消歧任务。实体消歧是由于同一实体指称在不同上下文可以指代不同实体,为能够明确实体指称所指代的实体而提出的任务,在语义分析、搜索和问答等自然语言处理相关应用中都是需要解决的关键性问题。本文将实体类别中最高的注意力分数作为不同语义环境下区分实体的参考依据,并基于此设计实体消歧实验。在实验中,从测试集中随机选取100个三元组,通过模型得到头、尾实体类别的注意力分数,并检验最高分数的类别是否能够直接体现出对应实体在该三元组中的语义。实验结果显示,其中有61个三元组符合上述实验假设,因此证明了CEKGRL模型在实体消歧任务中也具有一定的指导意义。由于篇幅限制,在此对该实验过程不再赘述。

4 结束语

现有融合实体类别信息知识表示学习模型中的类别与关系间通常需要设置约束条件。为高效利用实体类别与三元组关系之间的潜在相关性,本文提出一种采用注意力机制学习类别与关系间相关性的CEKGRL模型。在具有实体类别信息的FB15K数据集上,利用知识图谱补全和三元组分类任务对CEKGRL模型进行性能评估,结果表明其相比现有知识表示学习模型在MeanRank和Hit@n评估指标上均取得一定的性能提升,并通过案例分析验证了注意力机制的有效性。由于CEKGRL模型仅利用了知识图谱中的类别信息,因此后续可在该模型中融入更多具有丰富语义的多源信息进行联合训练,拓宽其在自然语言处理领域的应用范围,进一步提升适用性与实用性。

参考文献

- [1] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a Web-scale approach to probabilistic knowledge fusion [C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2014: 601-610.
- [2] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of Annual Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2013: 2787-2795.
- [3] YANG B S, YIH W T, HE X D, et al. Embedding entities and relations for learning and inference in knowledge bases [EB/OL]. [2020-01-09]. <https://arxiv.org/abs/1412.6575>.
- [4] NEELAKANTAN A, ROTH B, MCCALLUM A. Compositional vector space models for knowledge base completion [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Philadelphia, USA: ACL Press, 2015: 156-166.
- [5] XIE Ruobing, LIU Zhiyuan, SUN Maosong. Representation learning of knowledge graphs with hierarchical types [C]//Proceedings of International Joint Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2016: 2965-2971.
- [6] WANG Zhen, ZHANG Jianwen, FENG Jianlin, et al. Knowledge graph embedding by translating on hyperplanes [C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2014: 1112-1119.
- [7] FANG Yang, ZHAO Xiang, TAN Zhen, et al. A revised translation-based method for knowledge graph representation [J]. Journal of Computer Research and Development, 2018, 55(1): 139-150. (in Chinese)
方阳, 赵翔, 谭真, 等. 一种改进的基于翻译的知识图谱表示方法 [J]. 计算机研究与发展, 2018, 55(1): 139-150.
- [8] XIAO Han, HUANG Minlie, HAO Yu, et al. TransA: an adaptive approach for knowledge graph embedding [EB/OL]. [2020-02-09]. <http://arxiv.org/abs/1509.05490>.
- [9] XIAO Han, HUANG Minlie, HAO Yu, et al. TransG: a generative mixture model for knowledge graph embedd-

- ing[EB/OL]. [2020-02-09]. <https://arxiv.org/pdf/1509.05488v2.pdf>.
- [10] LIN Yankai, LIU Zhiyuan, SUN Maosong, et al. Learning entity and relation embeddings for knowledge graph completion [C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2015; 2181-2187.
- [11] JI Guoliang, HE Shizhu, XU Liheng, et al. Knowledge graph embedding via dynamic mapping matrix [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Philadelphia, USA: ACL Press, 2015; 687-696.
- [12] BORDES A, WESTON J, COLLOBERT R, et al. Learning structured embeddings of knowledge bases [C]//Proceedings of the 25th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2011; 301-306.
- [13] BORDES A, GLOROT X, WESTON J, et al. Joint learning of words and meaning representations for open-text semantic parsing [C]//Proceedings of Artificial Intelligence and Statistics Conference. Cambridge, USA: MIT Press, 2012; 127-135.
- [14] BORDES A, GLOROT X, WESTON J, et al. A semantic matching energy function for learning with multi-relational data [J]. Machine Learning, 2014, 94(2): 233-259.
- [15] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data [C]//Proceedings of the 28th International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2011; 809-816.
- [16] NICKEL M, TRESP V, KRIEGEL H P. Factorizing YAGO: scalable machine learning for linked data [C]//Proceedings of the 21st International Conference on World Wide Web. New York, USA: ACM Press, 2012; 271-280.
- [17] JENATTON R, ROUX N L, BORDES A, et al. A latent factor model for highly multi-relational data [C]//Proceedings of Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2012; 3167-3175.
- [18] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion [C]//Proceedings of Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2013; 926-934.
- [19] XIE Ruobing, LIU Zhiyuan, JIA Jia, et al. Representation learning of knowledge graphs with entity descriptions [C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2016; 2659-2665.
- [20] XIE Ruobing, LIU Zhiyuan, LUAN Huanbo, et al. Image-embodied knowledge representation learning [C]//Proceedings of International Joint Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2017; 3140-3146.
- [21] WU Jiawei, XIE Ruobing, LIU Zhiyuan, et al. Knowledge representation via joint learning of sequential text and knowledge graphs [EB/OL]. [2020-02-09]. <https://arxiv.org/pdf/1609.07075.pdf>.
- [22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2017; 5998-6008.
- [23] KINGMA D, BA J. Adam: a method for stochastic optimization [EB/OL]. [2020-02-09]. <https://arxiv.org/abs/1412.6980>.