计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200012

基于多语义提取的知识图谱补全模型研究

李鹏彦　王宝会
北京航空航天大学软件学院　北京 100191
(lipengyan@buaa.edu.cn)

摘　要　在知识图谱补全领域,实体和关系之间丰富的多语义信息对于提升补全任务的准确性具有重要意义。然而,现有模型往往难以充分捕捉和整合这些多语义特征,导致补全效果受限。为此,提出了一种基于多语义提取的知识图谱补全模型(Knowledge Graph Completion Model Based on Multi-Semantic Extraction,MSE)。首先,设计了多语义聚合编码器,对实体和关系嵌入进行维度切分,整合了邻居实体和关系的多语义信息。其次,设计了基于多尺度卷积的解码器,使用不同大小的卷积核提取实体自身的深层语义特征。最后,设计了融合独立性约束的损失函数,引入基于 Pearson 相关系数的正则化项,以提升模型多语义表达能力。实验结果表明,在 FB15k-237 和 WN18RR 数据集上,相较于其他基线的最优模型,MSE 模型的 MRR(Mean Reciprocal Rank)值分别提升 1.7%和 2.3%,验证了其在知识图谱补全任务上的有效性。

关键词:知识图谱;知识补全;多语义信息;独立性约束
中图分类号　TP301

Knowledge Graph Completion Model Based on Multi-semantic Extraction

LI Pengyan and WANG Baohui
School of Software,Beihang University,Beijing 100191,China

Abstract　In the field of knowledge graph completion,the rich multi-semantic information between entities and relationships is of great significance for improving the accuracy of completion tasks. However,existing models often struggle to fully capture and integrate these multi-semantic features,which limits the effectiveness of completion. To address this challenge,this paper proposes a Knowledge Graph Completion Model Based on Multi-Semantic Extraction(MSE). Firstly,a multi-semantic aggregation encoder is designed to dimensionally split entity and relationship embeddings,integrating the multi-semantic information of neighboring entities and relationships. Secondly,a decoder based on multi-scale convolution is proposed,using convolutional kernels of different sizes to extract the deep semantic features of entities. Lastly,a loss function with independence constraints is designed,introducing a regularization term based on Pearson correlation coefficients to enhance the model's multi-semantic expression capability. The experimental results show that on the FB15k-237 and WN18RR datasets,the MRR values of the MSE model are improved by 1.7% and 2.3%,respectively,compared with the optimal models of other baselines,which verifies its effectiveness on the knowledge graph complementation task.

Keywords　Knowledge graph,Knowledge completion,Multi-semantic information,Independence constraint

1　引言

在构建知识图谱时,通常依赖多种数据源,如网络爬虫和数据库集成,这也带来了数据不全和信息稀缺的挑战。数据的不完整性可能导致知识图谱中部分实体关系缺失或实体信息不完整,从而影响知识图谱的完整性与准确性。此外,这些问题还会影响知识图谱在推荐系统、问答系统和智能助手等应用的性能与准确度。因此,知识图谱补全(Knowledge Graph Completion,KGC)成为了一个关键任务,其核心目标是链路预测,主要子任务是实体预测,即在给定头实体和关系的情况下预测尾实体,或在给定尾实体和关系的情况下预测头实体。

知识图谱中的多语义信息包含两个层面。

1)实体自身的多语义信息。实体自身往往承载着复杂而多层次的语义,在不同任务或上下文中呈现出不同的语义信息。例如,实体"苹果公司"不仅仅代表一家科技公司,它还可以在不同维度上被解读为"全球最大公司之一""创新标杆""消费者电子设备制造商"等。这些语义并不是概念上的不同,如"苹果公司"中的"苹果"和水果中的"苹果",而是同一实体概念在不同语义维度下的多重语义视角。然而,传统模型在处理实体时,往往只采用单一的语义表示,如使用平均嵌入或一个固定的特征向量,无法深入挖掘实体自身的多语义结构。

2)邻居关系的多语义信息。实体与其邻居之间的关系不仅限于单一类型的交互,不同的邻居关系为实体带来了丰富的语义层次和信息,同一类型的关系也可能涵盖不同的语义内容。以企业"IBM"为例,它与"微软""谷歌"和"惠普"之间均存在"竞争"关系。这些关系虽统一标注为"竞争",但每一对企业间的竞争关系含义各异。例如,IBM 与微软的竞争主要在云计算和操作系统市场上展开,与谷歌则在人工智能技

通信作者:王宝会(wangbh@buaa.edu.cn)

术领域有更明显的竞争，与惠普则是在硬件销售领域进行竞争。这些关系承载着更丰富的语义层次，不同层次的语义信息对实体表示有着重要作用。然而，现有模型往往采用简单的加权聚合或注意力机制来整合邻居信息，忽略了这些邻居关系的多语义特征。这种处理方式容易导致多语义信息的丢失，从而影响模型对实体的精确表征。

针对以上问题，提出了一种基于多语义提取的知识图谱补全模型。本文贡献如下：

1)设计了一种多语义聚合编码器，通过对实体和关系嵌入进行维度切分，有效整合了邻居实体与关系的多语义信息，丰富了实体的嵌入表示，使模型能适应各种类型和复杂度的关系；

2)设计了一种多语义提取解码器，采用多尺度卷积核从嵌入向量中提取实体自身的深层语义特征，使模型更加精确地理解实体的复杂语义，提高了模型在预测任务中的准确性；

3)提出一种基于多语义提取的知识图谱补全模型，通过在两个公共数据集(FB15k-237 和 WN18RR)上进行对比实验，验证了 MSE 模型的有效性，并通过消融实验验证了各个模块的必要性。

2 相关工作

知识图谱补全任务的目的是挖掘和推断出未被明确记录的三元组，以增强知识图谱的完整性和实用性。知识图谱补全技术涉及多种模型，主要可以分为基于翻译的模型、基于张量分解的模型、基于卷积神经网络的模型和基于图神经网络的模型，这些模型通过将实体和关系映射到低维向量空间，或利用网络的结构特性，以提高预测的准确性和效率。

2.1 基于翻译的模型

基于翻译的模型的核心思想是将知识图谱中的关系视为头实体到尾实体的向量平移。

TransE[1]最先将三元组的关系表示为实体向量间的平移，假设头实体 h 加上关系 r 应接近尾实体 t，即 $h+r=t$，通过最小化头实体和尾实体间的距离来优化实体和关系的低维表示。TransE 在处理一对一关系时表现较好，但在多对多关系处理上存在局限。TransH[2]在 TransE 的基础上引入了关系特定的超平面，设定每个关系有一个超平面的范数向量和平移向量，通过在超平面上映射实体，来更好地处理多对多关系。TransR[3]将实体投影到关系空间，关系操作在关系空间内进行，这种分离的空间处理提高了模型的灵活性和表现力，进一步优化了复杂关系的表示。TransD[4]通过构建实体与关系向量间的动态映射矩阵，进一步提高了模型处理实体关系交互时的效率。基于翻译的模型以其简洁直观的思想和高效的计算成为早期的代表性方法。其核心优势在于模型简单，可解释性强。然而，这类方法在处理复杂关系时表征能力有限，本质上是对实体间关系的线性假设，难以捕捉非线性交互和丰富的语义信息。

2.2 基于张量分解的模型

基于张量分解的模型将知识图谱中的实体关系表示为高维张量，通过对这些张量进行分解来学习实体和关系的低维向量表示，并定义评分函数衡量三元组可信度，从而有效地预测和补全知识图谱中的缺失链接。

RESCAL[5]将每个关系表示为完整的矩阵，通过计算头实体向量、关系矩阵和尾实体向量的乘积来得分。而 DistMult[6]模型则使用双线性对角矩阵来简化这一过程，减少参数数量并提高计算效率。ComplEX[7]引入了复数嵌入，将实体和关系映射到复数空间，通过复数的内积操作，进一步提升了模型的表达能力，使得模型能够捕捉反对称性关系。TuckER[8]基于 Tucker 分解，通过核心张量捕捉实体和关系之间的多层次交互，具有较强的表达能力和良好的可解释性，为后续复杂模型提供了坚实的基线。基于张量分解的模型具有坚实的数学理论基础，通过全局矩阵分解能有效捕捉实体和关系间的潜在语义关联，尤其在对称/反对称关系建模上表现出色。但其劣势在于计算复杂度通常较高，且对知识图谱的局部结构信息利用不足，性能在很大程度上依赖于分解算法的选择与优化。

2.3 基于卷积神经网络的模型

卷积神经网络(CNN)如 ConvE[9]首次尝试通过卷积操作提取特征，通过 2D 卷积操作将三元组中的头实体和关系向量折叠成二维矩阵，增强了实体和关系之间的交互信息，在对具有高阶节点数的大规模知识图谱进行建模时十分有效。

InteractE 模型[10]在 ConvE 的基础上增强了实体和关系嵌入之间的交互，通过特征置换和重塑技术，提高了卷积对实体和关系嵌入的特征提取能力。InteractE 在捕捉高阶关系方面更具优势，能够更好地泛化不同的知识图谱，并在较低的参数复杂度下取得良好的性能。ConvKB 模型[11]则直接将头实体、关系和尾实体的向量拼接后，使用一维卷积核进行特征学习，可以捕捉到他们之间的全局交互信息。CapsE[12]采用胶囊网络来建模三元组，首先对实体和关系嵌入进行卷积，然后利用其动态路由机制，能够更好地表示实体和关系的复杂交互信息和空间层次信息，在处理多关系和复杂模式的知识图谱任务中表现出色。CPConvKE[13]采用置信度感知和路径增强的卷积式框架，可同时利用结构、实体类型与规则信息以抑制噪声。基于 CNN 的模型通过卷积操作有效捕捉了实体和关系嵌入中的局部交互模式和复杂特征，非线性表达能力较强。其主要优势在于能够建模高阶交互，并具有良好的扩展性；缺点在于卷积核的感受野有限，难以有效整合远距离的邻域信息，对图谱的整体结构感知能力较弱。

2.4 基于图神经网络的模型

图神经网络由于在图结构数据中具有出色表现[14]，已经被广泛应用于知识图谱补全任务中。这些网络主要通过基于谱的方法和基于空间的方法来学习节点的特征。基于谱的方法依赖于基于拉普拉斯特征基的图结构[15]进行变换，而基于空间的方法则通过聚合和传播节点特征来更新节点表示，进而迭代学习每个节点的邻居信息。

R-GCN[16]是首个将图卷积网络应用于知识图谱补全的模型，其通过异构图中的关系进行建模，通过分层传播技术均等地根据关系类型获取每个实体的邻域信息，并区分边的方向性。SACN 模型[17]结合了 WGCN 作为编码器和 Conv-TransE 作为解码器，同时考虑图的结构信息和翻译特性。WGCN 在前向传播过程中控制计算的节点及其邻居节点之间的交互强度，解码器通过使用多个卷积核对实体和关系向量进行卷积，然后利用与备选尾实体向量的内积来计算头实

体和关系与每个尾实体组成的三元组的正确概率。GATFCN[18]模型以图卷积网络为编码器捕获图结构信息，并采用Tucker张量分解作为解码器，通过端到端方式融合局部结构特征与全局语义关联，增强事实表示学习。CompGCN[19]则联合学习多关系图的实体和关系嵌入表示，关系表示向量同样在训练过程中得到学习。KBGAT模型[20]将关系嵌入到计算注意力系数的三元组表示中，允许实体在学习过程中获取更丰富的邻域信息；DMGNN[21]从实体、关系和三元组多视图进行解耦学习，以处理多语言图谱的异构性。基于GNN的模型能够显式地利用知识图谱的拓扑结构，通过消息传递机制聚合多跳邻居信息，从而学习到更丰富的上下文感知的实体表示。这是其最显著的优势。然而，这类模型也存在挑战，例如对图中噪声敏感，深层GNN过平滑，以及计算开销随邻域扩展而增大等。

3 基于多语义提取的知识图谱补全算法

3.1 模型概述

本文提出了基于多语义提取的知识图谱补全模型，模型总体框架如图1所示，分别在编码器和解码器融入多语义提取的思想。首先，设计了多语义聚合编码器，在邻居实体和关系的聚合过程中采用嵌入切分与分别聚合的设计，即将实体和关系嵌入按维度切分，然后分别通过语义分离的邻居映射矩阵进行信息聚合，进而提取出邻居中的多语义信息。其次，设计了多语义提取解码器，采用多尺度卷积核对实体的嵌入向量进行卷积操作，通过不同大小的卷积核提取出不同的特征图张量，即提取实体自身的多语义信息。此外，针对多语义信息设计了融合独立性约束的损失函数，引入Pearson相关系数作为正则化系数。

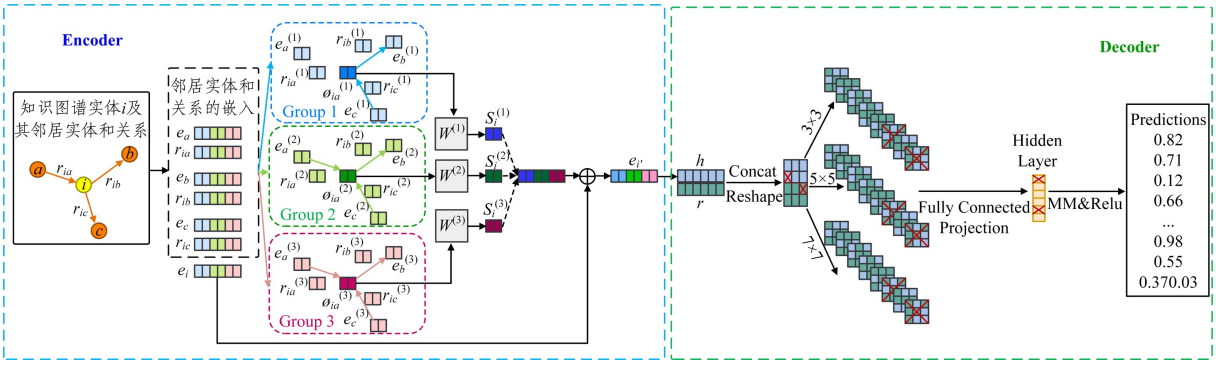

图1 基于多语义提取的知识图谱补全模型框架图

Fig. 1 Framework of knowledge graph completion model based on multi-semantic extraction

多语义聚合编码器中嵌入切分和分别聚合的设计可以有效提取邻居和关系的多语义信息，多语义提取解码器可以通过多尺度卷积核提取实体自身的多语义信息，通过二者的结合，本文提出的基于多语义提取的知识图谱补全模型在多个方面超越了基线模型的性能。这种结合从知识图谱中实体和关系的多语义表示角度出发，分别通过编码器和解码器对实体和关系在多语义层面做了更细粒度的嵌入表示，进而实现了更为准确的补全效果。而融合独立性约束的损失函数则一定程度上减少了引入多语义带来的信息冗余，增强了模型的泛化性能。

3.2 多语义聚合编码器

在基于图神经网络的知识图谱补全模型中，编码器将实体邻居的语义信息聚合在实体的嵌入表示中，以丰富实体的嵌入表示承载的信息。然而，传统的编码器无法充分捕捉邻居实体的复杂语义关系，特别是当实体和关系的语义信息多样化时，这种问题尤为显著。此外，简单的聚合策略（如平均或求和）无法有效区分不同邻居的重要性，导致信息表达质量下降，进而影响模型的整体表征能力和最终的预测性能。为了解决这些问题，本文在编码器阶段采用了一种多语义聚合的策略，如图1所示，图中的不同颜色（蓝色、绿色、粉色）代表了嵌入表示中的不同维度分组，在将实体的邻居和关系的嵌入表示聚合到实体的嵌入表示时，通过对实体和关系嵌入进行维度切分，并为每个切分维度分别计算注意力权重，实现了对邻居关系的精细管理。

对于一个待聚合的中心实体 i，其对应的嵌入向量为 e_i，邻居实体 j 对应的嵌入向量为 e_j，而它们之间关系的嵌入为 r_{ij}。聚合过程中，将中心实体、邻居实体以及关系的嵌入向量进行多维度的切分和分组，同一分组代表着同一语义层次，将同一分组的邻居实体和关系向量通过注意力机制进行聚合，然后通过不同的线性矩阵将不同分组聚合后的向量进行聚合，生成如式(1)所示的基于多语义提取的邻居实体和关系信息的嵌入表示 s_i^{tri}。

$$s_i^{\mathrm{tri}}=\bigoplus_{s=1}^{S}\sigma\Big(\sum_{(e_j,r_{ij})\in\mathcal{N}_i}\alpha_{ij}^{(s)}W^{(s)}\phi(e_j^{(s)},r_{ij}^{(s)})\Big) \tag{1}$$

其中，$\oplus$ 表示将所有维度切分聚合后的结果组合在一起；S 是总切分的维度数，对应着不同的语义层面；σ 是一个非线性激活函数；$\mathcal{N}(i)$ 是实体 i 的邻居实体及其关系的集合；$\alpha_{ij}^{(s)}$ 是在第 s 个切片维度上从实体 i 到实体 j 的注意力权重；$W^{(s)}$ 是在第 s 个切片维度上的线性变换矩阵；$\phi(e_j^{(s)},r_{ij}^{(s)})$ 是一个组合函数，用来融合邻居实体和关系在第 s 个切片维度上的语义信息，可以是加法函数 $\phi(e_j^{(s)},r_{ij}^{(s)})=e_j^{(s)}+r_{ij}^{(s)}$，也可以是哈达玛积函数 $\phi(e_j^{(s)},r_{ij}^{(s)})=e_j^{(s)}\odot r_{ij}^{(s)}$，本文使用的是实验效果更好的哈达玛积函数。注意力权重系数 $\alpha_{ij}^{(s)}$ 的计算方式如式(2)所示。

$$\alpha_{ij}^{(s)}=\frac{\exp((e_j^{(s)})^{\mathrm{T}}e_i^{(s)})}{\sum_{k\in\mathcal{N}_i}\exp((e_k^{(s)})^{\mathrm{T}}e_i^{(s)})} \tag{2}$$

最后，将融合邻居实体和关系信息中多语义层面信息的嵌入向量与原实体嵌入向量相加后作为新的实体嵌入向量进行更新，得到如式(3)所示的新的实体嵌入向量 e_i'：

$$e_i'=e_i+s_i^{\mathrm{tri}} \tag{3}$$

使用此方法更新实体的嵌入表示，邻居实体和关系的多语义信息将通过这种方式聚合到实体的嵌入表示中，实体的

嵌入表示在图神经网络的每一层都得到了优化。多语义聚合编码器通过这种多语义聚合方式，能够适应不同类型和复杂度的关系，通过细粒度的控制和动态的权重调整，极大地提升了模型对实体关系的理解深度和预测的准确性。

3.3 基于多尺度卷积的解码器

在知识图谱补全模型中，解码器利用编码器提供的丰富的实体表示来预测实体间的潜在关系。传统解码器通常直接利用简单的嵌入向量进行实体预测或关系预测，这可能导致无法捕捉实体嵌入中的细微特征，尤其是在实体和关系的语义信息较复杂时。为了解决这一问题，本文设计了一种多语义提取解码器，如图 1 所示，在特征提取阶段采用不同尺度的卷积核(图中使用的卷积核大小分别为 3×3，5×5，7×7)来提取实体和关系中不同层次的语义特征。

首先，将实体和关系的嵌入向量连接在一起；然后，为了让不同尺度的卷积核都能够在嵌入的不同部分捕获丰富的特征，并满足后续的多尺度卷积操作的输入要求，通过重塑操作将连接后的实体和关系的嵌入向量转换成二维形式，如式(4)所示。

$$\boldsymbol{X}=reshape([\boldsymbol{h};\boldsymbol{r}]) \tag{4}$$

其中，h 代表实体的嵌入向量，是多语义聚合编码器聚合后的实体嵌入向量；r 代表关系的嵌入向量，通过重塑操作，将它们转换成一个二维矩阵 X。接着，在特征提取阶段，使用多尺度卷积滤波器对这些重塑后的嵌入向量进行操作，不同的卷积核会提取不同的特征图张量组，它们分别代表着不同语义层面的信息，使得模型能够从原始的嵌入中提取出多语义的特征表示，如式(5)所示。

$$T_i=f(\omega_i * X) \tag{5}$$

其中，ω_i 表示第 i 个卷积核；$*$ 表示卷积操作；f 是激活函数，如 ReLU，用于引入非线性来增强模型的表达能力。

多个特征图 $\boldsymbol{T}_i$ 生成后，会被向量化并拼接在一起，通过一个全连接层进行进一步融合和变换，如式(6)所示。

$$z=\boldsymbol{W}_f\cdot \mathrm{vec}([\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n]) \tag{6}$$

全连接层的权重矩阵 $\boldsymbol{W}_f$ 将这些特征映射到目标空间，整合和转换卷积层提取的局部特征，将卷积层的高维特征转换为与原实体嵌入表示一致的维度，最终通过与实体矩阵相乘后得到各实体最终的相似度计算结果和链接预测结果。最终的评分函数如式(7)所示。

$$\psi(h,r,t)=z\cdot e_0^{\mathrm{T}} \tag{7}$$

通过这种方法，多语义提取解码器可以有效捕捉到多语义聚合编码器处理后的嵌入向量中的多语义层面的信息，更准确地理解知识图谱中实体本身的复杂语义，进一步提升了预测的准确性。

3.4 融合独立性约束的损失函数

现有的知识图谱补全模型通常仅考虑了简单的嵌入聚合方式，忽略了多语义特征可能存在的相似性问题。具体而言，当对多语义特征进行聚合时，不同语义的嵌入在特征空间中可能过于相似，导致模型在表达多样化语义时出现信息冗余。为此，本文设计了一种基于 Pearson 相关系数的正则化方法，用于降低不同语义的共线性，进而增强模型在多语义信息上的区分能力。一些其他正则化方法(如 L1 正则化或基于模型复杂度的惩罚项)并非专门设计用于共线性问题，而是更广义地作用于特征选择或模型稀疏性。相比之下，Pearson 相关系数作为正则化项，更加聚焦于解决特征间的冗余和相关性问题。

设定模型中嵌入的切分组数为 S，每个切分组对应的线性层权重分别为 $W^{(i)}$ 和 $W^{(j)}$，其中 $i,j\in\{1,2,\cdots,S\}$ 且 $i\neq j$，为减少多语义特征之间的相关性，引入 Pearson 相关系数作为正则化项。

首先，如式(8)所示，为了计算多语义特征的相关性，将每个线性层的权重展平为一维向量。

$$w_i=Flatten(W^{(i)}),w_j=Flatten(W^{(j)}) \tag{8}$$

然后，对展平的权重向量进行去均值处理，以消除权重偏移的影响，如式(9)所示。

$$\hat{w}_i=w_i-\bar{w}_i,\hat{w}_j=w_j-\bar{w}_j \tag{9}$$

其中，$\bar{w}_i$ 和 $\bar{w}_j$ 是权重向量的均值。

定义标准化权重向量的 Pearson 相关系数为：

$$\rho_{i,j}=\frac{\sum_{k=1}^{n}\hat{w}_{i,k}\cdot\hat{w}_{j,k}}{\sqrt{\sum_{k=1}^{n}\hat{w}_{i,k}^2}\cdot\sqrt{\sum_{k=1}^{n}\hat{w}_{j,k}^2}} \tag{10}$$

其中，n 是展平后的权重向量的长度。式(10)衡量了不同线性层权重向量之间的相似性。

最后，通过将所有线性层权重对之间的 Pearson 相关系数累加，得到如式(11)的正则化损失函数 $\mathcal{L}_{\mathrm{orth}}$。

$$\mathcal{L}_{\mathrm{orth}}=\frac{1}{\left(\frac{S}{2}\right)}\sum_{i=1}^{S}\sum_{j=i+1}^{S}|\rho_{i,j}| \tag{11}$$

其中，$\binom{S}{2}$ 表示总共的矩阵对数。正则化项的目标是最小化不同语义特征之间的相似性，从而促进多语义特征的去相关化。

在链路预测任务中，已知一个三元组中的头实体和关系，目标是预测该头实体在给定关系下的可能尾实体。如式(12)所示，在多语义提取解码器输出三元组得分 $\psi(h,r,t)$ 之后，使用 sigmoid 函数将其映射到(0,1)区间。

$$y=\sigma(\psi(h,r,t)) \tag{12}$$

其中，y 是三元组被解释为真的概率，当 y 越接近于 1 时，三元组越倾向为正样本。

为了衡量预测的准确性并进行优化，本文使用交叉熵损失函数来计算模型在链路预测任务上的损失，如式(13)所示。

$$\mathcal{L}_{\mathrm{task}}=-\frac{1}{N}\sum_i(\hat{y}_i\cdot\log(y_i)+(1-\hat{y}_i)\cdot\log(1-y_i)) \tag{13}$$

其中，$\hat{y}_i$ 表示训练集中的标签，当三元组为正样本时为 1，否则为 0；y_i 是模型预测的概率；N 是训练样本的总数。

总损失函数由任务损失和正则化损失组成，如式(14)所示。

$$\mathcal{L}_{\mathrm{total}}=\mathcal{L}_{\mathrm{task}}+\lambda\cdot\mathcal{L}_{\mathrm{orth}} \tag{14}$$

其中，λ 是正则化项的权重超参数，用于控制正则化项在总损失中的影响。

通过引入基于 Pearson 相关系数的正则化项，模型不仅能够在任务性能上保持高效，还能显著降低多语义特征的共线性，从而更准确地捕捉多样化的语义信息。这种设计能够有效提升知识图谱补全任务中的多语义表达能力，为下游任务提供更丰富和精确的语义支持。

4 实验与分析

为验证提出的 MSE 性能，在 RTX4090(24G)，Python 版本 3.9.19、PyTorch 版本 2.1.0、dgl 版本 2.4.0 和 CUDA 版本 11.8 的环境下进行对比实验分析。

4.1 数据集与预处理

本文使用了 2 个公开的数据集来验证方法的有效性。

1)FB15k-237[22]。FB15k-237 数据集是从原始的 FB15k 数据集中衍生出来的，后者基于涵盖多个领域信息的 Freebase 知识库，广泛用作多种知识图谱嵌入模型的基准测试集。该数据集包括大约 14541 个实体和 237 种关系类型，共有 310116 个三元组。

2)WN18RR[9]。WN18RR 数据集是从广泛使用的 WN18 数据集衍生的，后者基于 WordNet。此数据集包含 40943 个实体和 11 种关系类型，共 93003 个三元组。

各数据集处理后的详细数据如表 1 所列。

表 1 数据集详细信息

Table 1 Detailed statistics of datasets

	FB15k-237	WN18RR
实体数量	14541	40943
关系数量	237	11
训练集三元组数	27115	86835
验证集三元组数	17535	3034
测试集三元组数	20466	3134

4.2 评价指标

为了全面评估本文多语义提取知识图谱补全模型的性能，采用平均倒数排名(Mean Reciprocal Ranking，MRR)和命中率(Hit@N)作为评价指标。这些指标能够从不同角度反映模型在知识图谱补全任务上的效果。

平均倒数排名(MRR)是一个常用来衡量排名任务性能的指标。它考虑了模型预测正确答案的排名的倒数的平均值，更加注重排名靠前的预测。

$$MRR=\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{15}$$

其中，$rank_i$ 是模型在第 i 个问题中预测正确答案的排名。

命中率(Hit@N)是评估模型能否将正确答案排在前 N 位的频率。这是一个直观的指标，用于衡量模型在预测最相关答案时的效果。

$$Hit@N=\frac{|\{i|rank_i\leqslant N\}|}{|Q|} \tag{16}$$

其中，$rank_i\leqslant N$ 表示模型预测的正确答案排在前 N 位的情况，$|Q|$ 是问题总数。

4.3 实验设置

本文在训练过程中采用了 Adam 优化器，学习率为 3.5×10^{-4}，并设置标签平滑为 0.1，以提升网络优化的稳定性。

在 FB15k-237 数据集上，模型的训练设置为 600 个 epoch，batch size 为 512。实体和关系的嵌入维度为 600，卷积核采用 3 种不同的尺度 3，5，7，输出通道数设置为 200。在关系表示更新方面，模型采用了 2 层图神经网络(GNN)，并通过 20 和 30 的 k_h 和 k_w 参数重塑实体-关系输入矩阵。此外，模型采用了一对多的评分方式，即将头实体与关系(或尾实体与关系)组合后，与所有其他实体进行评分。

在 WN18RR 数据集上，训练设置为：学习率为 1.5×10^{-3}，实体和关系的嵌入维度为 300，卷积核尺寸同样为 5，7，9，输出通道数为 250，图神经网络层数为 1。同时，重塑后的输入矩阵高和宽分别为 10 和 30。训练过程共进行 800 个 epoch，batch size 为 256。

模型的最终参数根据验证集上评估的 MRR 指标进行确定，即参数选取在 MRR 指标上表现最佳的一组。

4.4 链路预测实验

本文将提出的 MSE 模型与主流知识图谱补全模型在链路预测任务上开展了对比实验，包括平移类模型(TransE)，基于张量分解的模型(DistMult，ComplEx，TuckER)，基于卷积的模型(ConvE，InteractE)，以及图神经网络模型(SACN，CompGCN)等，结果如表 2 所列。

表 2 链路预测结果对比

Table 2 Comparison of link prediction results

模型	FB15k-237				WN18RR			
	MRR	*Hits*@1	*Hits*@3	*Hits*@10	*MRR*	*Hits*@1	*Hits*@3	*Hits*@10
TransE	0.287	0.192	0.325	0.475	0.193	0.003	0.37	0.445
DistMult	0.178	0.092	0.204	0.352	0.332	0.260	0.380	0.456
ComplEx	0.234	0.146	0.265	0.407	0.394	0.353	0.419	0.461
TuckER	0.358	0.266	0.394	0.544	0.470	0.443	0.482	0.526
ConvE	0.325	0.237	0.356	0.501	0.430	0.400	0.440	0.520
InteractE	0.354	0.263	—	0.535	0.463	0.430	—	0.528
SACN	0.350	0.260	0.390	0.540	0.470	0.430	0.480	0.540
GATFCN	0.348	0.258	0.381	0.531	—	—	—	—
CompGCN	0.355	0.264	0.390	0.535	0.479	0.443	0.494	0.546
MSE	**0.364**	**0.272**	**0.399**	**0.549**	**0.490**	**0.453**	**0.504**	**0.562**

从表 2 中可以看出，在 FB15k-237 数据集上，MSE 模型相较于基线模型中表现最好的 TuckER 模型，*MRR* 提升 1.7%，*Hits*@1 提升 2.3%，*Hit*@3 提升 1.3%，*Hit*@10 提升 0.9%；在 WN18RR 数据集上，MSE 模型对比基线模型中表现最好的 CompGCN 模型，*MRR* 提升 2.3%，*Hits*@1 提升 2.3%，*Hit*@3 提升 2%，*Hit*@10 提升 2.9%。模型性能的稳定提升，主要归功于其在多语义信息提取方面的创新设计。首先，模型采用了嵌入切分的多语义聚合方法。通过将嵌入向量划分为多个子空间，并分别学习不同语义下的特征表示，模型能更全面地捕捉实体邻居的多语义信息。同时，通过引入基于 Pearson 相关系数的正则化项，降低了聚合过程中特征向量的线性相关性，使模型能够更准确地表达多语义。此外，模

型引入了多尺度卷积的特征提取机制，使其能够同时关注局部和全局的结构信息，进一步深入理解实体自身的多语义信息。

4.5 消融实验

为了评估基于多语义提取的知识图谱补全模型中各个关键组件的有效性和贡献，在 WN18RR 数据集上进行了一系列消融实验。通过从完整模型中逐一移除或替换特定组件，来分析每个组件对模型性能的影响。如表 3 所列，考虑了 4 种模型配置：完整的基于多语义提取的知识图谱补全模型，移除图信息，移除多尺度卷积，以及移除正则化损失。

表 3 消融实验结果

Table 3 Results of ablation experiments

模型配置	*MRR*	*Hits*@1	*Hits*@3	*Hits*@10
本文模型	**0.490**	**0.453**	**0.504**	**0.562**
移除图信息	0.472	0.439	0.485	0.538
移除多尺度卷积	0.482	0.446	0.494	0.555
移除正则化损失	0.486	0.448	0.501	0.559

实验结果表明，去除任一组件都会导致模型性能的下降。其中，去除图信息后的性能下降最为显著，这表明图结构在多语义信息提取中的重要性，图结构不仅提供了实体间的复杂关系，而且这些关系对于理解和预测实体间的潜在链接至关重要。分别去除多尺度卷积和正则化损失后，模型的评估指标均出现了一定程度的下降，这进一步证明了多尺度卷积在自身多语义提取中的作用，以及正则化项在减少特征共线性及提升模型泛化性方面的贡献。

4.6 参数分析实验

本文中主要关注嵌入维度和解码器的多尺度卷积核大小组合两个角度的参数。通过在 WN18RR 数据集上进行实验，分析了不同参数设置对模型性能的影响。

为了分析不同嵌入维度对模型性能的影响，在不同维度(100，200，300，400，500)下进行了实验，并综合评估了模型在各个指标下的表现，实验结果如表 4 所列。

表 4 嵌入维度参数实验结果

Table 4 Results of embedding dimension experiments

模型配置	*MRR*	*Hits*@1	*Hits*@3	*Hits*@10
100	0.369	0.288	0.409	0.524
200	0.484	0.443	0.500	0.565
300	**0.490**	**0.453**	**0.504**	**0.562**
400	0.482	0.444	0.496	0.554
500	0.485	0.448	0.500	0.557

随着嵌入维度的增加，模型性能呈现出先升后降的趋势。当嵌入维度从 100 增加到 300 时，模型的性能显著提升，这表明适当增加维度有助于模型捕捉更丰富的特征表示，从而增强其表达能力。当维度设置为 400 时，指标略有下降，可能是模型参数增加导致过拟合或学习不到更有效的特征。当维度设置为 500 时，指标较 400 维略微提升，但仍不及 300 维性能，这可能是维度过高增加了模型复杂度，未显著提高泛化能力。

在模型的解码器设计中，本文采用了多尺度卷积核来提取实体的多语义信息。为了评估不同卷积核大小组合对模型性能的具体影响，进行了多组实验，选取了不同的卷积核大小组合进行测试(如 2，3，5；3，5，7；5，7，9 等)，实验结果如表 5 所列。

表 5 卷积核大小组合参数实验结果

Table 5 Results of convolution kernel size combination experiments

卷积核组合	*MRR*	*Hits*@1	*Hits*@3	*Hits*@10
2，3，5	0.484	0.447	0.497	0.556
3，5，7	0.459	0.451	0.504	0.563
5，7，9	**0.490**	**0.453**	**0.504**	**0.562**
7，9，11	0.486	0.446	0.504	0.561

实验结果表明，卷积核大小组合为 5，7，9 时模型性能最佳，这种组合能够在不同尺度上有效地捕捉实体的多语义信息，从而提升模型的性能。相比之下，组合为 3，5，7 和 7，9，11 时性能略低，尽管 *Hits*@3 和 *Hits*@10 接近，但 *Hits*@1 和 *MRR* 稍有下降，这说明增大或减小最佳组合中的卷积核大小都会影响模型对细粒度语义的捕捉能力。组合为 2，3，5 时性能最低，各项指标均低于其他组合，这可能是卷积核范围过小，无法充分捕捉语义特征。因此，5，7，9 是推荐的卷积核大小组合，以获得最佳性能。

结束语 本文提出的基于多语义提取的知识图谱补全模型通过创新的多语义聚合编码器和多尺度卷积解码器，有效地整合了实体和关系的多语义信息，稳定提升了知识图谱补全的准确性。在 FB15k-237 和 WN18RR 数据集上的实验，验证了基于多语义提取的知识图谱补全模型在链路预测任务上相较于现有基线模型取得了稳定的性能提升。此外，融合独立性约束的损失函数进一步增强了模型在多语义信息上的区分能力，减少了信息冗余。未来研究可致力于提升高维度的模型泛化能力，以及如何将图像、文本等多模态数据与知识图谱相结合，增强模型对多源信息的理解和推理能力。

参 考 文 献

[1] 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.

[2] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2014.

[3] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2015.

[4] JI G, HE S, XU L, 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(volume 1: Long papers). 2015: 687-696.

[5] NICKEL M, TRESP V, KRIEGELH P. A three-way model for collective learning on multi-relational data[C]// ICML. 2011: 3104482-3104584.

[6] YANG B, YIH W, HE X, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv: 1412.6575, 2014.

[7] TROUILLON T, WELBL J, RIEDEL S, et al. Complex embeddings for simple link prediction[C]// International Conference on Machine Learning. PMLR, 2016: 2071-2080.

[8] BALAŽEVIĆ I, ALLEN C, HOSPEDALES T M. Tucker: Tensor factorization for knowledge graph completion[J]. arXiv: 1901.09590, 2019.

[9] DETTMERS T，MINERVINI P，STENETORP P，et al. Convolutional 2d knowledge graph embeddings[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[10] VASHISHTH S，SANYAL S，NITIN V，et al. Interacte：Improving convolution-based knowledge graph embeddings by increasing feature interactions[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：3009-3016.

[11] NGUYEN D Q，NGUYEN T D，NGUYEN D Q，et al. A novel embedding model for knowledge base completion based on convolutional neural network[J]. arXiv：1712. 02121，2017.

[12] NGUYEN D Q，VU T，NGUYEN T D，et al. A capsule network-based embedding model for knowledge graph completion and search personalization[J]. arXiv：1808. 04122，2018.

[13] YANG X，WANG N. A confidence-aware and path-enhanced convolutional neural network embedding framework on noisy knowledge graph[J]. Neurocomputing，2023，545：126261.

[14] GORI M，MONFARDINI G，SCARSELLI F. A new model for learning in graph domains[C]// 2005 IEEE International Joint Conference on Neural Networks. IEEE，2005，729-734.

[15] WU B，LIANG X，ZHANG S S，et al. Advances and Applications in Graph Neural Network[J]. Chinese Journal of Computers，2022，45(1)：35-68.

[16] SCHLICHTKRULL M，KIPF T N，BLOEM P，et al. Modeling relational data with graph convolutional networks[C]// The Semantic Web：15th International Conference(ESWC 2018). Heraklion，Crete，Greece，Springer International Publishing，2018：593-607.

[17] SHANG C，TANG Y，HUANG J，et al. End-to-end structure-aware convolutional networks for knowledge base completion [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019：3060-3067.

[18] JIN Y，YANG L. Graph-aware tensor factorization convolutional network for knowledge graph completion[J]. International Journal of Machine Learning and Cybernetics，2024，15(5)：1755-1766.

[19] VASHISHTH S，SANYAL S，NITIN V，et al. Composition-based multi-relational graph convolutional networks[J]. arXiv：1911. 03082，2019.

[20] NATHANI D，CHAUHAN J，SHARMA C，et al. Learning attention-based embeddings for relation prediction in knowledge graphs[J]. arXiv：1906. 01195，2019.

[21] DONG B，BU C，WANG Y，et al. Disentangled Multi-view Graph Neural Network for multilingual knowledge graph completion[J]. Applied Soft Computing，2025，183：113605.

[22] TOUTANOVA K，CHEN D. Observed versus latent features for knowledge base and text inference[C]// Proceedings of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality. 2015：57-66.

LI Pengyan，born in 1993，postgraduate，intermediate engineer. His main research interests include knowledge graph and recommender system.

WANG Baohui，born in 1973，senior engineer，master supervisor. His main research interests include software architecture，big data，artificial intelligence.