文章编号：1003-0077(2021)10-0048-08

一种基于 IC 参数的知识图谱嵌入方法

赵晓函，周子力，李天宇，陈丹华，王凯莉

（曲阜师范大学 网络空间安全学院，山东 曲阜 273100）

摘 要：TransC 是一种高效的知识图谱嵌入方法，通过区分概念和实例来建立概念、实例及关系的嵌入。TransC 将概念编码为球体，球体半径被随机初始化并在训练中迭代更新。由此导致模型出现两个问题：一是训练得到的部分球体半径与模型训练目标不符；二是忽略了概念本身提供的语义信息。针对上述两个问题，该文提出了 TransIC 模型，首先，基于 IC 参数给出新的概念球体半径求解方法，使求得的半径满足 TransC 目标，并且丰富了概念嵌入向量的语义信息。其次，该模型以 TransC 为基础，在概念编码阶段引入基于 IC 参数的概念球体半径。最后，在公开的数据集 YAGO39K 上完成链接预测和三元组分类两个任务，并将该文方法实验所得性能与 TransC 及其他模型的性能进行对比。结果表明，TransIC 在多数指标上均取得显著提升。

关键词：知识图谱嵌入；TransC；信息量

中图分类号：TP391 **文献标识码**：A

Knowledge Graph Embedding Based on IC Parameters

ZHAO Xiaohan, ZHOU Zili, LI Tianyu, CHEN Danhua, WANG Kaili

(School of Cyber Science and Security, Qufu Normal University, Qufu, Shandong 273100, China)

Abstract: TransC is an efficient method for embedding knowledge graphs. It establishes the embedding of concepts, instances, and relations by distinguishing concepts and instances. TransC encodes the concept as a sphere, and the radius of the sphere is randomly initialized and updated iteratively during training.This leads to two problems in the model. First, part of the sphere radius obtained from training does not match the model training target. Second, the semantic information provided by the concept itself is ignored. This paper proposes a model named TransIC to deal with the two issues above. TransIC adopts a novel concept sphere radius solution method based on IC parameters, so that the obtained radius meets the TransC goal, and enriches the semantic information of the concept embedding vector. Then it is based on TransC and introduces a concept sphere radius based on IC parameters during the concept coding phase.Finally, the two tasks of link prediction and triple classification are completed on the public data set YAGO39K, and the experimental performance of the method in this paper is compared with the performance of TransC and other models. The results show that TransIC has achieved a significant improvement in most indicators.

Keywords: knowledge graph embedding; TransC; information content

0 引言

近年来，许多知识图谱，如 WordNet[1]、Freebase[2]、NELL[3]以及 YAGO[4]等，已经成为智能问答、新闻推荐等很多实际应用的重要资源。知识图谱是由节点和不同类型的边组成的语义网，其中节点表示真实世界中的实体，不同类型的边表示实体之间的语义关系[5]。知识图谱是结构化的知识库，通常将知识表示为（头实体，关系，尾实体）形式，即三元组 (h,r,t)[6]。但是元组的底层符号特性使其不能有效表示实体之间的语义关联，并且难以实现知识图谱的智能化应用，智能化应用通常以深度学习等机器学习算法为支撑，而这些算法需要数值形式的输入[7]。

为了解决上述问题，知识图谱嵌入应运而生，并涌现了多种嵌入方法。其中，以深度学习为代表的嵌入学习技术在人工智能领域广受关注，其主要思

收稿日期：2020-03-02 **定稿日期**：2020-04-07

基金项目：国家自然科学基金(61871185)；山东省自然科学基金(ZR2017MD019)；教育部高教司产学合作协同育人项目(201701020098)；赛尔网络下一代互联网技术创新项目(NGII20190516)

想是通过机器学习的方法将知识图谱中的实体和关系表示成低维稠密向量，使得实体和关系分别嵌入表示[8]。目前，基于距离的翻译模型简单有效，如 TransE[9] 适用于一对一关系，将关系看作头实体和尾实体之间的一种平移。之后出现的拓展模型 TransH[10]，TransR[11]，TransD[12]，TransA[13] 等，主要解决了 TransE 无法处理复杂关系的问题。TransC[14] 在知识图谱嵌入时区分了概念和实例，将概念和实例分别建模为空间中的球体和点。然而，模型训练结束后的部分概念球体半径不符合其目标函数；同时学习三元组时并未考虑概念本身具有的信息量。

概念 c 的信息量[15]（Information Content，IC）是指概念 c 在某给定语料库中出现的概率 $p(c)$ 的负对数。一个概念在树形结构中所处的位置越高，所提供的有效信息就越少。IC 值主要用于语义相似度计算，而对于一些涉及概念的知识图谱嵌入模型（如 TransC），引入 IC 参数能够使模型对概念进行表示时捕获更多的语义特征，从而更有效地完成链接预测、三元组分类等任务。因此本文利用树形结构 IC 计算模型计算出每一个概念的 IC 值，进而求出对应的球体半径，使概念拥有丰富的语义信息，得以进一步增强知识图谱的嵌入效果。

1　相关工作

知识图谱嵌入方法可分为平移距离模型和语义匹配模型两大类[16]。前者又包括基于距离的翻译模型和其他距离模型，使用基于距离的损失函数；后者主要包括矩阵分解模型和神经网络模型，使用基于相似度的损失函数。其中，基于距离的翻译模型是当前知识图谱表示方法的研究热点，主要包括 Trans 系列模型。2013 年 Bordes 等人提出 TransE[9]，其主要思想是将三元组 (h,r,t) 中的实体 h，t 和关系 r 表示在同一低维语义空间，使得 $h+r\approx t$。然而 TransE 在处理一对多，多对一及多对多复杂关系时仍然存在一些缺陷。

为解决 TransE 中的问题，一系列拓展模型随之产生。当涉及不同关系时，TransH[10] 使得一个实体具有不同嵌入表示。首先将一个关系 r 建模在一个法向量为 W_r 的超平面，然后将三元组 (h,r,t) 中的实体嵌入 h，t 投影到关系超平面。虽然可以有效地处理复杂关系，但假设实体、关系在同一语义空间不利于 TransH 的充分表示。

TransR[11] 致力于解决 TransE 和 TransH 的表示问题，为每个关系设置一个传递矩阵 M_r，将三元组 (h,r,t) 中的 h，t 从实体空间嵌入到关系空间。虽然效果显著提升，但这一过程属于实体和关系的交互，TransR 中的传递矩阵不应该仅考虑实体之间的关系。

TransD[12] 同时考虑到不同类型的实体和关系，每一关系实体对 (r,e) 都会具有一个投影矩阵 M_{re}，用于将实体信息嵌入投影到关系向量空间。如此改进了 TransR 的不足，使性能得到提高。

另外，矩阵分解模型通过矩阵分解进行知识图谱表示学习。以 RESCAL[17] 为例，将每个实体关联向量从而捕获其潜在语义，每个关系都表示为一个矩阵，该矩阵对潜在因子间的成对交互关系进行建模。许多 RESCAL 的扩展如 DistMult[18]，将矩阵约束为对角矩阵以简化 RESCAL；HolE[19] 结合了 RESCAL 的表达能力及 DistMult 的简单性。还有 DistMult 的扩展如 ComplEx[20]，通过引入复值嵌入来更好地建模不对称关系。

2　TransIC 知识嵌入方法

针对 TransC 中部分球体半径不满足模型训练目标和概念语义信息表示不足这两个问题，TransIC 结合 IC 参数提出了新的计算概念球体半径的方法来加以解决。

2.1　TransC 问题分析

TransC[14] 提出之前，Trans 系列模型都将实体和关系以相同的方式编码为低维语义空间中的向量。TransC 尝试在同一语义空间中按照不同方式表示概念、实例和关系。具体而言，TransC 将每个概念编码为一个球体 $s(p,m)$，其中，球心 $p\in R^k$ 表示概念向量；R^k 表示维度为 k 的向量空间，m 为球体半径，用于判断实例与概念和子概念与概念的相对位置关系。对于每个实例 $i\in I$ 和实例关系 $r\in R_I$，分别学习一个低维向量 $i\in R^k$ 和 $r\in R^k$，其中实例用点表示。然后使用球体与点的相对位置关系表示概念与实例间的关系（instanceOf），球体与球体的相对位置关系表示子概念与概念间的关系（subClassOf），点与点的相对位置关系表示实例与实例间的关系（relational）。对于 subClassOf 关系，两个概念之间可能的相对位置如图 1 所示，其中，符号“∧”表示“并且”。

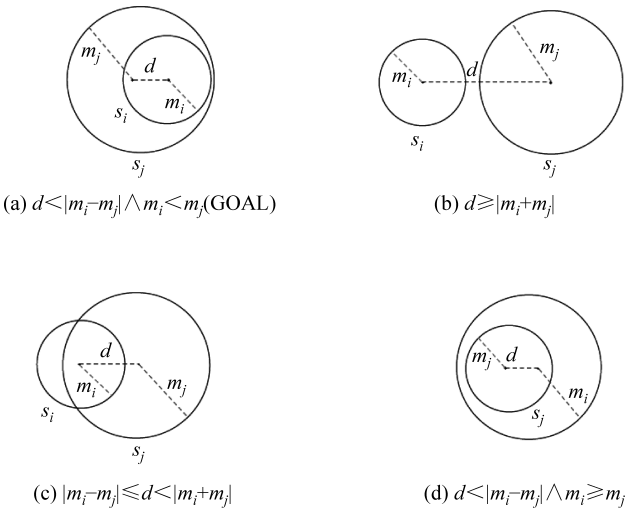

图 1 球体 s_i 和 s_j 之间的四个相对位置

对于给定正确三元组 (c_i, r_c, c_j)，其中，c_i 和 c_j 分别表示子概念和概念，r_c 表示 subClassOf 关系，如图 1(a)所示，TransC 的训练目标为概念球体 s_j 包含子概念球体 s_i，故表示子概念大小的球体半径 m_i 应小于表示概念大小的球体半径 m_j。然而，表 1 中列举出了 3 对经 TransC 训练得到的子概念球体半径大于概念球体半径的情况，如子概念“Artists_from_California”的球体半径(0.597 961)大于概念“artist”的球体半径(0.406 852)，表明 TransC 在训练中迭代更新得到的部分球体半径仍存在不合理之处，导致学习得到的概念嵌入向量不够准确。

2.2 TransIC

2.2.1 概念球体半径求解方法

表 1 TransC 部分子概念与概念球体半径结果分析

	subClassOf 1		subClassOf 2		subClassOf 3	
	子概念 c_i	概念 c_j	子概念 c_i	概念 c_j	子概念 c_i	概念 c_j
concept	Artists_from_California	artist	Olympic_footballers_of_Switzerland	Regional_capitals_in_Ghana	Political_parties_established_in_1956	party
TransC-m	0.597 961>0.406 852		0.687 164>0.494 344		0.777 434>0.515 779	

概念 IC 值是概念本身所包含的信息量，具体含义为：概念 c 出现的概率越大，其包含的自信息量就越小[21]。IC 计算主要分为基于统计和基于树形结构两种方法[22]，前者通过求解概念 c 在给定语料库中出现的概率得到 IC 值；后者最早由 Nuno[22] 提出，Nuno 认为，树形结构中叶子节点比非叶子节点包含更大的概念信息量。该方法的 IC 计算模型如式(1)所示。

$$\mathrm{IC}(c)=-\log_{10}p(c)=1-\frac{\log_{10}(\mathrm{hypo}(c)+1)}{\log_{10}(\mathrm{Node}_{\max})} \tag{1}$$

其中，hype(c)是概念 c 在树形结构中所有子节点的总数，$\mathrm{Node}_{\max}$是概念节点的总个数。

根据信息论，同一个概念出现的次数与其包含的信息量(IC)成反比。对于子概念－概念结构，子概念出现的次数小于概念出现次数，则子概念 IC 值大于概念 IC 值。然而 TransC 的训练目标为概念球体包含子概念球体，即概念球体半径大于子概念半径，因此根据 Nuno 提出的 IC 计算模型给出概念球体半径求解方法，半径计算方法如式(2)所示。

$$m=1-\mathrm{IC}(c)=\frac{\log_{10}(\mathrm{hypo}(c)+1)}{\log_{10}(\mathrm{Node}_{\max})} \tag{2}$$

概念球体半径具体求解步骤如下：

(1) 统计出训练数据中所有概念的总数 N；

(2) 分别统计出每个概念出现的次数，记为 N_{c_i}。如图 2 所示，左子树自底向上递归过程为：

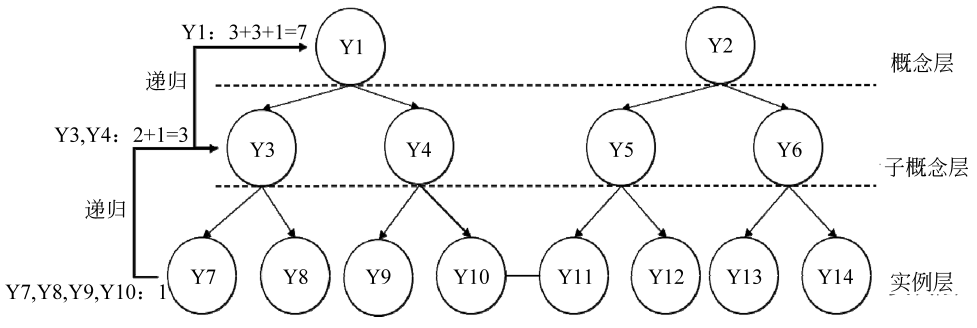

图 2 概念次数统计示意图

a. 实例层节点 Y7～Y10 各出现 1 次；

b. 子概念层 Y3＝Y7＋Y8＋1，Y4＝Y9＋Y10＋1；

c. 概念层 Y1＝Y3＋Y4＋1，右子树同理。

(3) 计算得到概念球体半径如式(3)所示。

$$m_i = 1 - \mathrm{IC}_i(c) = \frac{\log_{10}(N_{c_i})}{\log_{10}(N)} \tag{3}$$

根据式(3)求出的概念球体半径满足 TransC 模型训练目标，即子概念球体半径小于概念球体半径。具体对比结果如表 2 所示。

表 2　TransC 与 TransIC 的部分概念与子概念球体半径对比结果

	subClassOf 1		subClassOf 2		subClassOf 3	
	子概念 c_i	概念 c_j	子概念 c_i	概念 c_j	子概念 c_i	概念 c_j
concept	Artists_from_California	artist	Olympic_footballers_of_Switzerland	Regional_capitals_in_Ghana	Political_parti-es_established_in_1956	party
TransC-m	0.597 961＞0.406 852		0.687 164＞0.494 344		0.777 434＞0.515 779	
TransIC-m	0.343 511＜0.706 582		0.166 850＜0.681 815		0.102 303＜0.613 565	

2.2.2　模型构建

TransC 采用随机值的方式初始化实例向量 $\boldsymbol{i}$、概念向量 $\boldsymbol{p}$ 和概念半径 m。其中，m 在(0，1)内随机取值初始化，并在随机梯度下降 SGD(stochastic gradient descent)中不断迭代更新，导致 TransC 中的部分球体半径在训练结束后仍存在与优化目标相反的情况，并且忽略了概念中应包含的语义信息内容，使得概念的嵌入向量存在偏差。

TransIC 模型中，概念球体的半径通过新的半径计算方法生成，不再由随机值初始化，且随机梯度下降中不再迭代更新，所有的球体半径一直保持初始值，如式(4)所示。

$$m_i = 1 - \mathrm{IC}_i(c) \tag{4}$$

其中，m_i 表示第 i 个球体的半径大小，$\mathrm{IC}_i(c)$ 表示第 i 个概念的 IC 值。

TransIC 可建模的关系分为三类：实例与概念间的关系(instanceOf)、子概念与概念间的关系(subClassOf)以及实例与实例间的关系(relational)。对此，模型分别定义不同的损失函数来测量嵌入空间中的实例与概念、子概念与概念及实例与实例之间的相对位置，并且同时学习概念、实例和关系的表示。

(1) **instanceOf**：给定一个实例 i，概念 c，若它是正确三元组 (i, r_e, c)，则 i 应该在球体 s 内部，r_e 为 instanceOf 关系。损失函数如式(5)所示。

$$f_e(i,c) = \| \boldsymbol{i} - \boldsymbol{p} \|_2 - m \tag{5}$$

(2) **subClassOf**：对于一个正确的子概念与概念三元组 (c_i, r_c, c_j)，子概念 c_i，概念 c_j 分别被编码为球体 $s_i(\boldsymbol{p}_i, m_i)$ 与 $s_j(\boldsymbol{p}_j, m_j)$，且 s_j 应包含 s_i，r_c 为 subClassOf 关系。两个球心之间的距离定义为 $d = \| \boldsymbol{p}_i - \boldsymbol{p}_j \|_2$。损失函数如式(6)～式(8)所示。

$$f_c(c_i,c_j) = \| \boldsymbol{p}_i - \boldsymbol{p}_j \|_2 + m_i - m_j \quad (d \geqslant | m_i + m_j |) \tag{6}$$

$$f_c(c_i,c_j) = \| \boldsymbol{p}_i - \boldsymbol{p}_j \|_2 + m_i - m_j \quad (| m_i + m_j | \leqslant d < | m_i + m_j |) \tag{7}$$

$$f_c(c_i,c_j) = m_i - m_j \quad (d < | m_i + m_j | \wedge m_i \geqslant m_j) \tag{8}$$

其中，式(5)中 $m = 1 - \mathrm{IC}(c)$，式(6)～式(8)中 $m_i = 1 - \mathrm{IC}_i(c)$，$m_j = 1 - \mathrm{IC}_j(c)$。

(3) **relational**：给定一个实例关系三元组 $(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$，采用与 TransE[5] 相同的学习方式。损失函数如式(9)所示。

$$f_r(h,t) = \| \boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t} \|_2^2 \tag{9}$$

2.2.3　训练目标及算法

对于 instanceOf 三元组，S_e 和 S'_e 分别表示正例三元组集合和负例三元组集合。将下面基于边际的损失函数定义为训练目标，如式(10)所示。

$$L_e = \sum_{(i,c) \in S_e} \sum_{(i',c') \in S'_e} [\gamma_e + f_e(i,c) - f_e(i',c')]_+ \tag{10}$$

其中，$[x]_+ \triangleq \max(0, x)$ 表示模型的输出值为 0 或 x，γ 是边际参数。同样地，对于 subClassOf 三元组，S_c 和 S'_c 分别表示正概念三元组集合和负概念三元组集合，训练目标定义如式(11)所示。

$$L_c = \sum_{(c_i,c_j) \in S_c} \sum_{(c'_i,c'_j) \in S'_c} [\gamma_c + f_c(c_i,c_j) - f_c(c'_i,c'_j)]_+ \tag{11}$$

对于 relational 三元组，S_l 和 S'_l 分别表示正关系三元组集合和负关系三元组集合，训练目标定义如式(12)所示。

$$L_l = \sum_{(h,r,t) \in S_l} \sum_{(h',r,t') \in S'_l} [\gamma + f_r(h,t) - f_r(h',t')]_+ \tag{12}$$

最后，整体训练目标定义为上述三个函数的线性组合如式(13)所示。

$$L = L_e + L_c + L_l \tag{13}$$

TransIC 采用 TransH[10] 中的抽样策略替换头实体或尾实体，并使用随机梯度下降的方法对目标函数进行优化，具体训练过程如表 3 所示。

表 3 TransIC 训练过程

Algorithm 1 Learning TransIC

Input : training set $S_e=(i,c)$, $S_c=(c_i,c_j)$, $S_l=(h,r,t)$, margin γ_e, γ_c, γ_l, embeddings dim n, learn rate λ, spheres $s(\boldsymbol{p},m)$, concept information content IC

Output : vector $\boldsymbol{i},\boldsymbol{c},\boldsymbol{c}_i,\boldsymbol{c}_j,\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}\in R^k$

Initialization i,c,c_i,c_j,h,r,t, $m=1-IC(c)$, $m_i=1-\mathrm{IC}_i(c)$, $m_j=1-IC_j(c)$

for each epoch ∈ epochs do

  for each $(h,r,t)\in S_l$ do

    for each $(h',r,t')\in S_l'$ do

      $f_r(h,t)=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_2^2$

      $\sum_{(h,r,t)\in Sl}\sum_{(h',r,t')\in Sl'}[\gamma+f_r(h,t)-f_r(h',t')]_+$

    end for

    Normalize h,r,t

  end for

  for each $(i,c)\in S_e$ do

    for each $(i',c')\in S_e'$ do

      $f_e(i,c)=\|\boldsymbol{i}-\boldsymbol{p}\|_2-m$

      $\sum_{(i,c)\in Se}\sum_{(i',c')\in Se'}[\gamma_e+f_e(i,c)-f_e(i',c')]_+$

    end for

    Normalize i,c

  end for

  for each $(c_i,c_j)\in S_c$ do

    for each $(c_i',c_j')\in S_c'$ do

      $d=\|\boldsymbol{p_i}-\boldsymbol{p_j}\|_2$

      if $d\geqslant|m_i+m_j|$

        $f_c(c_i,c_j)=\|\boldsymbol{p_i}-\boldsymbol{p_j}\|_2+m_i-m_j$

      if $|m_i-m_j|\leqslant d<|m_i+m_j|$

        $f_c(c_i,c_j)=\|\boldsymbol{p_i}-\boldsymbol{p_j}\|_2+m_i-m_j$

      if $d<|m_i-m_j|\wedge m_i\geqslant m_j$

        $f_c(c_i,c_j)=m_i-m_j$

        $\sum_{(ci,cj)\in Sc}\sum_{(ci',cj')\in Sc'}[\gamma_c+f_c(c_i,c_j)-f_c(c_i',c_j')]_+$

      end if

      end if

      end if

    end for

    Normalize c_i, c_j

  end for

end for

3 实验验证

本文采用 YAGO 数据集的一个子集 YAGO39K 验证与评估 TransIC 模型的性能。完成知识图谱嵌入中的两个典型任务——链接预测和三元组分类，并与其他模型进行对比分析。YAGO39K 的统计信息如表 4 所示。

表 4 YAGO39K 数据集统计信息 （单位：个）

Data	YAGO39K
# instance	39 374
# concept	46 109
# relation	37
# relational Triple	354 996
# instanceOf Triple	437 836
# subClassOf Triple	29 181
# Valid (Relational Triple)	9 341
# Test (Relational Triple)	9 364
# Valid (instanceOf Triple)	5 000
# Test (instanceOf Triple)	5 000
# Valid (subClassOf Triple)	1 000
# Test (subClassOf Triple)	1 000

其中，subClassOf 关系中包含概念-子概念(2 层树结构)、概念(1 层树结构)及子概念(1 层树结构)三种层级；instanceOf 关系分为子概念-实例(2 层树结构)和概念-实例(2 层树结构)两种层级；relational 关系无层级。

3.1 链接预测任务定义及实验分析

链接预测的目的是预测关系三元组 (h,r,t) 中缺少的头实体 h、尾实体 t 或关系 r。具体实验如下：

(1) **评估标准**。对于测试集中的每个关系三元组 (h,r,t)，将头实例 h 或尾实例 t 随机替换为给定知识图谱中任一实例得到负例关系三元组 (h',r,t) 或 (h,r,t')；然后利用损失函数 f_r 计算关系三元组和负例关系三元组的距离并对其升序排列。此任务使用两个评估指标：所有正确实例排名倒数的平均值 MRR 和排名不大于 N 的正确实体的比例 Hits@N。MRR 和 Hits@N 越高说明实验结果越好。然而，若负例关系三元组仍然存在于知识图谱中，则会对关系三元组的排序结果产生干扰。所以排名前需要从训练、验证、测试集中过滤掉干扰三元组，此过程称为 Filter，用“Filt”表示。未过滤

的评价设置表示为“Raw”。

（2）**实验设置**。TransIC 进行训练时，随机梯度下降学习率 λ 从{0.1,0.01,0.001}中选择，边际参数 γ_l，γ_e 和 γ_c 从{0.1,0.3,0.4,1,2}中选择，实例向量和关系向量维度 n 从{20,50,100}中选择。多次实验得出最优参数配置为：$\gamma_l=1$，$\gamma_e=0.4$，$\gamma_c=0.3$，$n=100$，以 L_1 范式作为相似性度量。构造负例关系三元组的方法为：使用“unif”表示分别以 50% 的概率替换头实体或尾实体的传统方法，使用 “bern”表示伯努利抽样策略，即对于一对多关系三元组，以更高概率（大于 50%）替换头实体；对于多对一关系三元组，以更高概率替换尾实体。模型训练时迭代 1 000 次。TransIC 与其他模型链接预测对比结果如表 5 所示。

表 5　关系三元组链接预测对比结果

Model	MRR		Hits@N/%		
	Raw	Filt	1	3	10
TransE	**0.114**	0.248	12.3	28.7	51.1
TransH	0.102	0.215	10.4	24.0	45.1
TransR	0.112	0.289	15.8	33.8	56.7
TransD	0.113	0.176	8.9	19.0	35.4
TransC	0.112	0.420	29.8	50.2	69.8
HolE	0.063	0.198	11.0	23.0	38.4
ComplEx	0.058	0.362	29.2	40.7	48.1
TransIC(unif)	0.087	**0.445**	**30.7**	**53.2**	**70.7**
TransIC(bern)	0.113	0.436	30.0	51.7	70.5

从表 5 中得出：(1)TransIC 在 Filt 设置下的平均排名倒数 MRR 及排名前 1、3、10 的比例 Hits@1、Hits@3、Hits@10 指标均高于 TransC 及之前的其他模型。尤其是 unif 采样下的 Hits@3，与模型中最好的 TransC 相比提高了 3 个百分点，bern 采样下的 Hits@3 提高了 1.5%；其次是 unif 采样下的 Hits@1 和 Hits@10，均比 TransC 提高了 0.9 个百分点，而 bern 采样下分别提高了 0.2% 和 0.7%，这进一步表明，TransC 和 IC 参数结合求出概念球体半径的方法是合理有效的，TransIC 模型包含丰富的概念语义信息，使得链接预测指标结果得到提高。(2)与其他模型相比，TransIC 在 Raw 设置下的 MRR 指标稍低。其中 unif 采样下的 MRR(Raw) 比模型中最好的 TransE 降低了 0.027，不过 bern 采样下的 MRR(Raw) 仅降低了 0.001，差别相对不明显。分析其原因，在选取最优实验参数配置时，TransIC 根据整体指标效果设置参数，没有单独针对 MRR 选取实验参数的最优配置。(3)另外，TransIC 实验中 unif 采样方法整体优于 bern 采样方法，Filt 实验结果优于 Raw 实验结果。

3.2　三元组分类任务定义及实验分析

三元组分类用来判断给定关系三元组是否正确，也就是对关系三元组进行二分类。在 YAGO39K 数据集上，采用与链接预测相同的方式得到负例关系三元组。根据损失函数 f_r 计算给定关系三元组 (h,r,t) 的相异性得分，若得分低于一个阈值 δ，则预测为给定关系三元组，否则为负例关系三元组。

此任务使用四个评估指标：准确率(Accuracy)，查准率(Precision)，查全率(Recall)，查准率与查全率的调和平均(F_1-Score)。四个指标越高说明模型的效果越好。实验中采用与链接预测相同的参数配置以及 L_1 度量，采样策略为 unif 方法和 bern 方法，所有训练元组迭代 1 000 次。TransIC 与其他模型的三元组分类实验对比结果如表 6 所示。

表 6　三元组分类实验对比结果　（单位：%）

Model	Accuracy	Precision	Recall	F_1-Score
TransE	92.1	92.8	91.2	92.0
TransH	90.8	91.2	90.3	90.8
TransR	91.7	91.6	91.9	91.7
TransD	89.3	88.1	91.0	89.5
TransC	**93.8**	**94.8**	92.7	**93.7**
HolE	92.3	92.6	91.9	92.3
ComplEx	92.8	92.6	**93.1**	92.9
TransIC(unif)	92.4	93.2	91.3	92.3
TransIC(bern)	92.9	**94.8**	90.8	92.7

从表 6 可以得出：(1)在 Precision 指标上，bern 采样下的 TransIC 与模型中最好的 TransC 取得了相同的实验结果，并对比于其他模型效果提升。(2)在 Accuracy、Recall 及 F_1-Score 指标上，TransIC 均相较于其他模型欠佳。其中，在 unif 采样下比各自最好的模型分别降低了 1.4%、1.8% 和 1.4%，而在 bern 采样下比各自最好的模型分别降低了 0.9%、2.3% 和 1%。这说明 TransIC 在三元组分类中表现的不是很好，仅在 Precision 指标上与最好结果持平。分析主要原因有两个：一是实验数据集中具有 subClassOf 关系的三元组数量较少，根据表 4 可知仅有 29 181 个，使得基于 IC 计算概念球体半径未能在该实验中表现出优势。更重要的是，sub-

ClassOf 三元组的训练往往存在较高的出错率，也就是说，若该类三元组数量增多，其训练的出错率相应增高，TransIC 模型测试效果会更加明显，三元组分类实验相应指标也会得到一定提升；二是本实验直接采用了链接预测的实验参数配置，该组参数并非三元组分类实验的最佳配置。(3)本实验中，TransIC 在 bern 采样下的测试结果整体优于 unif 采样下的结果。

3.3 球体半径定性分析

经过统计，TransC 与 TransIC 训练结束后均得到 29 181 个 subClassOf 三元组，即(子概念，概念)结构中子概念球体半径－概念球体半径为 29 181 对。然而，TransC 中存在 6 789 对球体半径与其训练目标完全相反。TransIC 弥补了其不足，训练得到的球体半径均满足其目标函数。图 3 对比了 TransC 及 TransIC 的 subClassOf 三元组训练出错率。

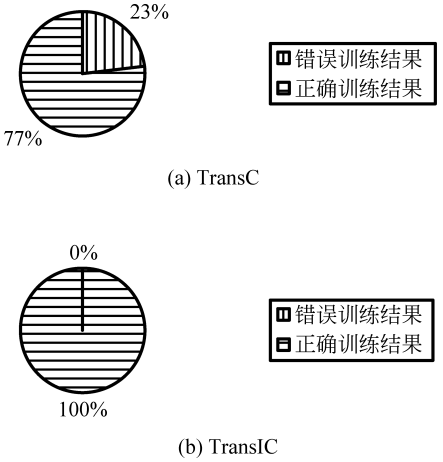

图 3 TransC 与 TransIC 的 subClassOf 三元组训练出错率对比

根据表 4 中 YAGO39K 数据集统计信息可知，具有 subClassOf 关系的三元组仅占总三元组的 3.5%，又由图 3 得出，对于数量较少的此类三元组，TransC 的训练出错率达到 23%，而 TransIC 的训练出错率为 0%，进一步说明 TransIC 中求解概念球体半径的方法是有效的。

4 结论

为了解决 TransC 中部分球体半径与优化目标相反以及概念语义信息表示不足这两个问题，本文提出了 TransIC 知识图谱嵌入模型。TransIC 利用 IC 参数计算出每个概念的 IC 值，再通过概念 IC 值计算出对应概念球体的半径，以保证子概念与概念之间的半径关系满足模型目标。其中，IC 参数提供了概念本身的语义信息内容，使其在嵌入学习中获取更丰富的语义信息。

在下一步工作中，计划扩充数据集，从 YAGO 中提取更多的 subClassOf 关系三元组加入到训练数据集中，使模型得到更充分地训练；同时继续优化 TransIC，寻找更准确的 IC 计算方法以得到更精确的概念 IC 值，从而进一步提升模型性能。此外，尝试将 IC 参数用于其他知识图谱嵌入模型中。

参考文献

[1] Miller G A. WordNet: A lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[2] Bollacker K, Evans C, Paritosh P, et al. Freebase: Acollaboratively created graph database for structuring human knowledge[C]//Proceedings of the ACM SIGMOD International Conference on Management of data, 2008: 1247-1250.

[3] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning[C]//Proceedings of the 24th AAAI Conference on Artificial Intelligence, 2010: 1306-1313.

[4] Suchanek F M, Kasneci G, Weikum G. Yago: A core of semantic knowledge[C]//Proceedings of the 16th International Conference on World Wide Web, 2007: 697-706.

[5] 刘知远，孙茂松，林衍凯，等. 知识表示学习研究进展[J]. 计算机研究与发展，2016，53(2)：247-261.

[6] 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.

[7] Han X, Zhang C, Guo C, et al. A generalization of recurrent neural networks for graph embedding[C]//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2018: 247-259.

[8] 朱艳丽，杨小平，王良，等. TransRD：一种不对等特征的知识图谱嵌入表示模型[J]. 中文信息学报，2019，33(11)：73-82.

[9] Bordes A, Usunier N, Garcia Duran A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013: 2787-2795.

[10] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014.

[11] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015.

[12] 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 Com-

putational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015：687-696.

[13] Xiao H，Huang M，Hao Y，et al. TransA：An adaptive approach for knowledge graph embedding[J]. arXiv preprint arXiv：1509.05490，2015.

[14] Lv X，Hou L，Li J，et al. Differentiating concepts and instances for knowledge graph embedding[J]. arXiv preprint arXiv：1811.04588，2018.

[15] 游彬，严岳松，孙英阁，等. 基于 HowNet 的信息量计算语义相似度算法[J]. 计算机系统应用，2013(1)：129-133.

[16] Wang Q，Mao Z，Wang B，et al. Knowledge graph embedding：A survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering，2017，29(12)：2724-2743.

[17] Nickel M，Tresp V，Kriegel H P. A three-way model for collective learning on multi-relational data[C]//Proceedings of the ICML，2011，11：809-816.

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

[19] Nickel M，Rosasco L，Poggio T. Holographic embeddings of knowledge graphs[C]//Proceedings of the 13th AAAI Conference on Artificial Intelligence，2016.

[20] Trouillon T，Welbl J，Sebastian Riedel E A，et al. Complex embeddings for simple link prediction[C]//Proceedings of the International Conference on Machine Learning，2071-2080.

[21] Zhou Z，Wang Y，Gu J. New model of semantic similarity measuring in wordnet[C]//Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering. IEEE，2008，1：256-261.

[22] Seco N，Veale T，Hayes J. An intrinsic information content metric for semantic similarity in WordNet[C]//Proceedings of the ECAI，2004，16：1089.

赵晓函(1994—)，硕士研究生，主要研究领域为知识图谱表示学习。
E-mail：xhzhao0917@163.com

李天宇(1996—)，硕士研究生，主要研究领域为知识图谱表示学习。
E-mail：tyli0214@163.com

周子力(1973—)，通信作者，博士，副教授，主要研究领域为本体理论及应用、知识图谱。
E-mail：zlzhou999@163.com

(上接第 47 页)

[23] 柳位平，朱艳辉，栗春亮，等.中文基础情感词词典构建方法研究[J]. 计算机应用，2009，29(10)：2875-2877.

[24] 饶洋辉，李青，刘文印，等.公众文本之情感词典研究进展[J]. 中国科学(信息科学)，2014，44(07)：825-835.

[25] 赵妍妍，秦兵，刘挺，等.文本情感分析[J]. 软件学报，2010，21(8)：1834-1848.

[26] Bo Yuan，Ying Liu，Hui Li. Sentiment classification in Chinese microblogs：Lexicon-based and learning-based approaches[C]//Proceedings of Economics Development and Research，2013，68：1-6.

[27] Han He. HanLP：Han language processing[OL]. http://github.com/hankcs/HanLP，2020.

[28] 计峰，邱锡鹏.基于序列标注的中文依存句法分析方法[J]. 计算机应用与软件，2009，44(10)：133-135.

王弘睿(1996—)，硕士，主要研究领域为计算语言学。
E-mail：whongrui18@163.com

于东(1982—)，通信作者，博士，副教授，主要研究领域为自然语言处理。
E-mail：yudong_blcu@126.com

刘畅(1995—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：liuchang2014@gmail.com