·人工智能及识别技术· 文章编号:1000-3428(2018)07-0205-07 文献标志码:A 中图分类号:TP18

基于结构与文本联合表示的知识图谱补全方法

鲍开放,顾君忠,杨 静

(华东师范大学 计算机科学技术系,上海 200062)

摘 要: 现有的表示学习算法不能很好地表示知识图谱中的复杂关系,且未能充分利用实体的描述文本。为此,建立一种结合文本表示和结构表示的联合表示学习模型。使用深度卷积神经网络对实体的描述文本进行编码得到文本表示,通过引入非对称映射操作的基于翻译思想的模型生成结构表示,将两者进行联合学习从而得到实体和关系表示,同时使用不同的低秩矩阵分别对头实体和尾实体进行映射,使其能更好地表现知识图谱中的复杂关系。实验结果表明,相对文本表示和结构表示的单独训练模型,该模型具有更好的表示性能。

关键词: 知识图谱补全;表示学习;深度学习;词向量;知识表示

中文引用格式: 鲍开放,顾君忠,杨 静. 基于结构与文本联合表示的知识图谱补全方法[J]. 计算机工程,2018,44(7):205-211.

英文引用格式: BAO Kaifang, GU Junzhong, YANG Jing. Knowledge graph completion method based on jointly representation of structure and text[J]. Computer Engineering,2018,44(7):205-211.

Knowledge Graph Completion Method Based on Jointly Representation of Structure and Text

BAO Kaifang, GU Junzhong, YANG Jing

(Department of Computer Science and Technology, East China Normal University, Shanghai 200062, China)

【Abstract】 The existing representation learning algorithms can not well represent the complex relationship in knowledge graph, and fails to make full use of the description text of entities. To solve this problem, this paper proposes a jointly representation learning model combining text representation and structure representation. The deep Convolution Neural Network(CNN) is used to encode the text of the entity to get the text representation, the structure representation is generated by introducing the translation thought model of asymmetric mapping operation, and the two are jointly studied to get entity and relation representation. At the same time, different low-rank matrices is used to project the head entity and the tail entity separately, so that the proposed model can better express the complex relationship in knowledge graph. Experimental results show that the proposed model has better representation ability than the single training model of text representation and structure representation.

【Key words】 knowledge graph completion; representation learning; deep learning; word vector; knowledge representation

DOI: 10.19678/j.issn.1000-3428.0047598

0 概述

以 Freebase 和 WordNet 为代表的知识图谱因能够提供准确、有效的结构信息,已成为网络检索、推荐系统和自动问答系统等智能应用的重要数据资源[1]。知识图谱往往包含数以百万计的实体和数十亿条的知识,但在实际应用中还不够全面。知识图谱补全旨在解决知识图谱中的数据稀疏问题[2]。例如,由“Queen Elizabeth II,_place_of_birth,London”和“London,_capital_of,United_Kingdom”2 条知识,可以推断“Queen Elizabeth II, nationality, United_Kingdom”很可能是一条潜在知识。因此,无需从外部信息中抽取新的关系数据,仅利用知识图谱中已有的数据就可以挖掘出一些新关系。

知识图谱常以网络形式表示,其中,节点代表实体,边代表 2 个实体间的关系,每一条知识用三元组(h,r,t)形式表示,其中,h 表示头实体,r 表示关系,t 表示尾实体。类似三元组的符号表示方法,要求在知识图谱补全中必须为不同的应用设计不同的图算法。随着知识图谱规模的不断增加,由于

基金项目: 国家科技支撑计划项目(2015BAH01F02)。

作者简介: 鲍开放(1992—),男,硕士研究生,主研方向为自然语言处理;顾君忠,教授、博士生导师;杨 静,副教授。

收稿日期: 2017-06-14 **修回日期:** 2017-07-25 **E-mail:** 51151201023@stu.ecnu.edu.cn

其扩展性差,导致计算越来越复杂。面对该挑战,知识图谱的表示学习被提出,其将实体和关系训练为分布式表示向量[3-4]。

但是,现有的表示学习算法仍不能很好地表示知识图谱中的复杂关系,且没能充分利用实体的描述文本,针对该问题,本文构建一种结合结构表示和文本表示的联合表示学习模型。针对结构表示,在 TransE 的基础上,分别采用不同的矩阵对头实体和尾实体进行映射以得到新的实体表示,并且规定映射矩阵不为满秩,从而使新的实体能更好地表现知识图谱中的复杂关系;针对文本表示,使用卷积神经网络(Convolutional Neural Network,CNN)来对实体的描述文本进行编码,以提炼出文本中丰富的语义信息。近年来,CNN 在某些自然语言的处理任务(如词性标签、分块、命名实体识别和关系分类)中取得了较好效果[5-6],使用 CNN 来表示文本的优点在于可以考虑到上下文信息和词序,训练速度快、准确率高,同时可以充分表示文本中单词的复杂局部交互,且与循环神经网络相比,CNN 无偏性问题。

1 相关工作

1.1 知识图谱表示学习

知识图谱表示学习将实体和关系投射到连续的低维空间中,并为其求取分布式表示向量。知识图谱表示学习如图 1 所示,其将三元组(h,r,t)训练为分布式向量$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$的形式。通过表示学习得到的实体和关系的低维稠密向量,可以高效地用于与知识图谱补全相关的各任务,如计算实体间的距离,这对于知识的融合和推断具有重要意义。

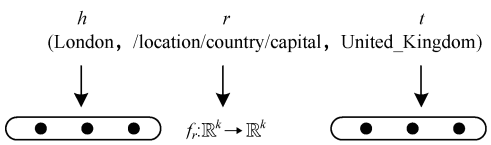

图 1 知识图谱表示学习示意图

近年来,研究者针对知识图谱补全中的表示学习算法建立大量模型,本文按照是否用到文本信息将这些模型分为 2 类:基于翻译思想的模型和融合文本信息的模型。

将一个知识图谱定义为 $KG=(E,R,T)$,E、R 分别表示知识图谱中所有实体、关系的集合,T 表示所有三元组(h,r,t)的集合。表 1 所示为相关模型的简要描述,其中,不同的模型拥有不同的损失函数 $f_r(h,t)$,所有模型都通过最小化对应的损失函数来求得实体和关系的分布式向量。

表 1 相关模型简要描述

模型	评分函数
TransE	$f_r(h,t)=\|\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|\|_2^2$
TransH	$f_r(h,t)=\|\|(\boldsymbol{h}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{h}\boldsymbol{w}_r)+\boldsymbol{r}-(\boldsymbol{t}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{t}\boldsymbol{w}_r)\|\|_2^2$
TransR	$f_r(h,t)=\|\|\boldsymbol{h}\boldsymbol{M}_l+\boldsymbol{r}-\boldsymbol{t}\boldsymbol{M}_t\|\|_2^2$
NTN	$f_r(h,t)=\boldsymbol{u}_r^{\mathrm{T}}g(\boldsymbol{h}\boldsymbol{M}_r\boldsymbol{t}+\boldsymbol{M}_{r,1}\boldsymbol{h}+\boldsymbol{M}_{r,2}\boldsymbol{t}+\boldsymbol{b}_r)$

1.2 基于翻译思想的结构表示模型

受 word2vec 中平移不变现象的启发,结构表示模型中的代表 TransE[7]将关系视为头实体与尾实体间的一种平移,例如,对于三元组(London,_capital_of,United_Kingdom),有 $vec('\text{London}')-vec('\text{United\_Kingdom}')\cong vec('\text{\_capital\_of}')$,即认为 $\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|$ 应尽可能地接近零。考虑到三元组数据中实体和关系除名称外并不能提供其他语义信息,提升空间有限,文献[8]引入一些从维基百科中提取出的文本信息进行联合训练,从而使得实体和关系的表示更准确。

TransE 大幅提升了计算效率,有效缓解了知识图谱中的数据稀疏问题。但其仍存在如下不足:

1)TransE 只适用于 1-to-1 关系(即关系中的一个头/尾实体对应一个尾/头实体),无法表现知识图谱中的复杂关系(如 1-to-n、n-to-1、n-to-n 关系)。图 2 所示的复杂关系中,在求取尾实体“James Cameron”和关系“_directed_by”对应的头实体时,TransE 并不能区别出“Titanic”和“Avatar”。表 2 所示为知识图谱中头实体和尾实体间关系的复杂性表现。

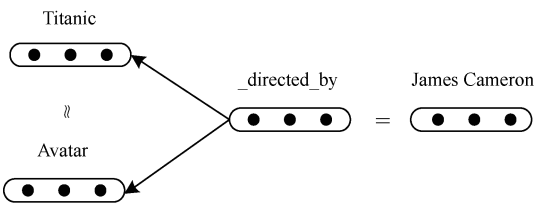

图 2 知识图谱中复杂关系实例

表 2 头实体和尾实体的异质性表现

数据类型	平均值	标准差
头实体对应的尾实体	1.26	0.23
尾实体对应的头实体	2 614.17	9 229.75

2)如图 3 所示,实际知识图谱中大都包含具有丰富语义的实体描述文本,语料质量优于从外部抽取的文本,但其没有得到充分利用。而现有的文本表示模型大都只是简单地用 word2vec 将文本训练成词向量,然后通过求均值、TopK 等方式来得到文本表示,该过程往往损失较多的语义信息。

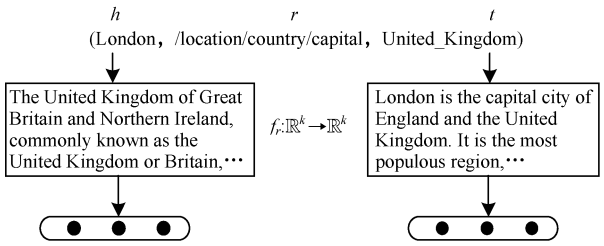

图 3　实体描述文本

针对该问题,TransH[9]提出将实体映射到对应关系的超平面上,允许实体在"多对多"关系下针对特定的关系拥有特定的表示,从而表示知识图谱中的复杂关系。不同于 TransH 在特定关系下对头实体和尾实体使用同一个矩阵进行映射操作,TransR[10]引入非对称映射操作,即使用 2 个不同的矩阵 $\boldsymbol{M}_h$和 $\boldsymbol{M}_t$ 分别对头实体和尾实体进行映射,从而使得实体在不同的关系下得到不同的表示。针对 TransR 参数多、时间复杂度高、投影矩阵没有和实体交互的信息而只和关系相关的问题,TransD[11]使用 2 个投影向量 $\boldsymbol{l}_h$、$\boldsymbol{l}_t$ 来生成投影矩阵 $\boldsymbol{M}_h$、$\boldsymbol{M}_t$,以此达到与实体和关系的交互。

1.3　融入文本信息的模型

基于翻译思想的模型都只关注训练实体间的结构信息,而没有充分利用具有丰富语义的实体描述文本。对此,文献[12]提出 NTN 模型,将每个实体用其名称的词向量的平均值来表示,以此共享相似实体名称中的文本信息。文献[8]提出在 TransE 学习知识图谱的三元组结构信息的基础上,用 word2vec 训练维基百科正文[13],同时通过文本和实体的对应关系,将文本训练得到的与实体对应的词表示与通过三元组结构信息训练得到的实体表示进行对齐。文献[14]提出使用深度学习来对知识图谱中的实体描述文本信息进行编码,但该方法只融合进了 TransE,没有进一步进行探索。文献[15]用实体名称或其描述文本的词向量的平均值来表示实体。文献[16]针对 TransE 的初始输入是随机生成的,提出先用 word2vec 将实体的描述文本训练成词向量,然后使用主成分分析(Principle Component Analysis,PCA)进行降维,以得到最终的实体表示,并将其作为 TransE 的初始输入。

上述文本表示模型均没有充分利用文本中的词序信息,训练过程中会损失较多的语义信息,尤其对自然语言中的复杂词语效果不佳。为此,在充分考虑文本中的复杂局部交互的基础上,本文使用 CNN 来对实体的描述文本进行编码,从而获得文本表示。CNN 简单有效,多层的架构可以较好地表示文本的语义信息。

2　算法设计

2.1　相关定义

为同时利用知识三元组结构信息和实体描述文本信息,本文提出结构表示和文本表示 2 种实体表示类型。对于给定的 $KG=(E,R,T)$,知识三元组 $(h,r,t)\in T$。$\boldsymbol{h}_s$ 和 $\boldsymbol{t}_s$ 分别代表头实体和尾实体基于结构的表示,这种表示一般由现有的基于翻译的模型(如 TransE)训练所得。$\boldsymbol{h}_d$ 和 $\boldsymbol{t}_d$ 分别代表头实体和尾实体关于其描述文本训练所得的表示。结构表示可以较好地捕获各种知识图谱中的知识三元组信息,文本表示则可以较好地表示实体的语义信息。本文将 2 种实体表示进行联合学习,映射到相同的连续向量空间中。模型的损失函数定义为:

$$E=E_s+\alpha E_d+\beta\|\theta\|^2$$

其中,E_s 为结构表示的损失函数,E_d 为文本表示的损失函数,$\|\theta\|^2$ 为正则项,α、β 为超参数,分别衡量文本信息损失和正则项的权重。

算法的伪代码形式如下:

输入　Training set $s=(h,r,t)$, entities and relation sets E and R, margin γ, embeddings dim k, hyper param α,β

Initialize:

$r\leftarrow \text{uniform}\left(-\frac{6}{\sqrt{k}},\frac{6}{\sqrt{k}}\right)$ for each $r\in R$

$r\leftarrow r/\|r\|$ for each $r\in R$

$e\leftarrow \text{uniform}\left(-\frac{6}{\sqrt{k}},\frac{6}{\sqrt{k}}\right)$ for each $e\in E$

left and right project matrix:

$L_r=\text{diag}(1,1,\cdots,1)$

$R_r=\text{diag}(1,1,\cdots,1)$

Loop

$e\leftarrow e/\|e\|$ for each entity $e\in E$

$S_{batch}\leftarrow \text{sample}(S,b)$ //a minibatch of size b

$T_{batch}\leftarrow \varnothing$

for $(h,r,t)\in S_{batch}$ do

$(h',r',t')\leftarrow \text{sample}(S'_{(h,r,t)})$

$T_{batch}\leftarrow T_{batch}\cup\{((h,r,t),(h',r',t'))\}$

end for

get desc-based representations with CNN

update embeddings w. r. t:

$$\sum_{(h,r,t)\in T}\sum_{(h',r',t')\in T_{batch}}\max(0,\gamma+f_r(h,t)-f_{r'}(h',t'))$$

End loop

2.2　结构表示模型 $\boldsymbol{E}_s$

E_s 通过三元组数据来训练,即在 TransE 的基础上,对头实体 h 和尾实体 t 针对不同的关系使用不同的低秩矩阵 $\boldsymbol{L}_r$、$\boldsymbol{R}_r$ 来映射。规定映射矩阵为低秩,可以更好地表示复杂关系,具体证明见文献[17]。E_s 表达式如下:

$$E_s = f_r(h,t) = \| \boldsymbol{h}' + \boldsymbol{r} - \boldsymbol{t}' \|_2^2$$

$$\boldsymbol{h}' = \boldsymbol{L}_r \boldsymbol{h}$$

$$\boldsymbol{t}' = \boldsymbol{R}_r \boldsymbol{t}$$

2.3 文本表示模型 E_d

本文使用 CNN 来对实体的描述文本进行编码，从而获得文本表示 E_d。为使 E_d 的学习过程与 E_s 兼容，定义 E_d 如下：

$$E_d = f_r(h_s,t_d) + f_r(h_d,t_s) + f_r(h_d,t_d)$$

其中，$f_r(h_d,t_s) = \| \boldsymbol{h}_d + \boldsymbol{r} - \boldsymbol{t}_s \|_2^2$、$f_r(h_s,t_d) = \| \boldsymbol{h}_s + \boldsymbol{r} - \boldsymbol{t}_d \|_2^2$ 分别表示在 h 和 t 中一个使用文本表示，另一个使用结构表示，$f_r(h_d,t_d) = \| \boldsymbol{h}_d + \boldsymbol{r} - \boldsymbol{t}_d \|_2^2$ 表示其中的头实体和尾实体都采用文本表示。通过将这 2 种类型的实体表示投射到相同的向量空间中，从而使得 2 种类型的表示互相影响。

2.3.1 CNN 整体结构

图 4 所示为 CNN 模型的整体结构。CNN 共有 5 层，输入层为使用 word2vec 对实体描述文本训练所得的词向量，经过 2 次卷积和 2 次池化，最终输出实体的文本表示向量。

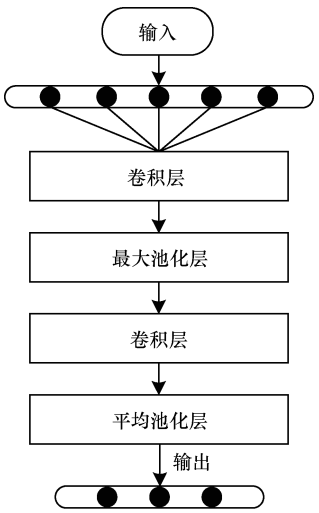

图 4 CNN 整体结构

2.3.2 预处理

预处理阶段，首先去除实体描述文本中所有的停用词；然后将包含多个词的实体进行拼接，形成一个词汇，如将“Molly Shannon”拼接为“Molly_Shannon”；最后使用 word2vec 对文本进行训练，得到所有词向量。

2.3.3 卷积

在卷积层中，用 $\boldsymbol{Z}^{(l)}$ 表示第 l 个卷积层的输出，$\boldsymbol{X}^{(l)}$ 表示第 l 个卷积层的输入。首先，通过一个大小为 k 的窗口对 $\boldsymbol{X}^{(l)}$ 进行滑动处理从而得到 $\boldsymbol{X'}_i^{(l)}$。在第 1 层中，$\boldsymbol{X}^{(l)}$ 是文本经过预处理后通过 word2vec 训练得到的诸如 $(\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)$ 形式的词向量，滑动窗口的过程表示如下：

$$\boldsymbol{x'}_i^{(1)} = \boldsymbol{x}_{i:i+k-1} = [\boldsymbol{x}_i^{\mathrm{T}},\boldsymbol{x}_{i+1}^{\mathrm{T}},\cdots,\boldsymbol{x}_{i+k-1}^{\mathrm{T}}]^{\mathrm{T}}$$

即 $\boldsymbol{x'}_i^{(1)}$ 的第 i 个向量通过将输入语句的第 i 个窗口中的 k 列的向量进行串联获得。由于滑动窗口过程中输入向量的长度可变，本文在输出向量的末尾使用了零填充。卷积层的第 i 个输出向量可表示为：

$$\boldsymbol{z}_i^{(l)} = \sigma(\boldsymbol{W}^{(l)}\boldsymbol{x'}_i^{(l)} + \boldsymbol{b}_i^{(l)})$$

其中，$\boldsymbol{b}_i^{(l)}$ 表示可选的偏置项，$\boldsymbol{W}^{(l)} \in \mathbb{R}^{n_2^{(l)} \times n_1^{(l)}}$ 表示第 l 卷积层的所有经过滑动窗口后的输入向量的卷积核，$n_2^{(l)}$ 表示输出向量的维数，可认为是特征图的大小，$n_1^{(l)} = k \times n_0^{(l)}$，$n_0^{(l)}$ 表示输入向量的维数，σ 是激活函数，如 tanh 函数或 ReLU 函数。需要注意的是，所有零填充的向量在正向传播中不应有贡献，在反向传播中也不能得到更新。通过这种方式，可以对齐输入句子的长度，从而避免零填充的影响。

2.3.4 池化

在每个卷积层后使用池化（pooling）来缩小 CNN 的参数空间并消除噪音。为降低池化过程中的信息损失，对不同的卷积层分别采取不同的池化策略。

为抓取最重要的特征，对第 1 个池化层采用最大池化层（max-pooling）操作。首先将卷积层的输出向量拆分为 n 个窗口，然后选取每个窗口中每个特征的最大值来构成一个新向量，该过程称为 n-max-pooling。通过确定窗口大小 n 来提取出输入向量的每个维度中最显著的特征值：

$$\boldsymbol{x}_i^{(2)} = \max(\boldsymbol{z}_{n\cdot i}^{(1)},\boldsymbol{z}_{n\cdot i+1}^{(1)},\cdots,\boldsymbol{z}_{n\cdot (i+1)-1}^{(1)})$$

n-max-pooling 可以将特征表示缩小 n 倍，从而降低 CNN 编码器的复杂度和参数学习的成本。

然而，有些文本语义较复杂，不同的句子可能含有不同的意思，仅使用 max-pooling 可能造成严重的信息损失。在这种情况下，对于第 2 个池化层，在激励函数之前需要使用 mean-pooling 替代 max-pooling 来生成实体表示：

$$\boldsymbol{x}_i^{(3)} = \sum_{i=1,2,\cdots,m} \frac{\boldsymbol{z}_i^{(2)}}{m}$$

这样，所有包含不同信息的 m 个输入向量都会对最终的实体表示有所贡献，并且在反向传播中也可以得到更新。由于采用了 2 种不同的池化策略，因此面对不同长度的输入向量时，本文模型都可以为每个实体获得固定长度的表示向量，且不会有太多的信息丢失。

2.4 训练过程

本文模型可表示为参数集 $\phi = (\boldsymbol{X},\boldsymbol{W}^{(1)},\boldsymbol{W}^{(2)},\boldsymbol{E},\boldsymbol{R})$，其中，$\boldsymbol{X}$、$\boldsymbol{E}$、$\boldsymbol{R}$ 分别代表单词、实体和关系的向量表示，$\boldsymbol{W}^{(1)}$、$\boldsymbol{W}^{(2)}$ 分别代表不同层的卷积核。与文

献[7]类似,本文将最小化基于间隔的损失函数作为最终的训练目标。

$$L=\sum_{(h,r,t)\in T}\sum_{(h',r',t')\in T'}\min(0,\gamma+f_r(h,t)-f_{r'}(h',t'))$$

其中,$\gamma>0$ 是间隔超参数,T' 是 T 的负样本集,其构造方式如下:

$$T'=\{(h',r,t)\mid h'\in E\}\cup\{(h,r,t')\mid t'\in E\}\cup\{(h,r',t)\mid r'\in R\}\}$$

T'中头实体、尾实体和关系被另一个随机三元组中的实体和关系代替。特别地,如果替换后的三元组已经在 T 中,则不会被加入到负样本中。由于 h 和 t 都有 2 种类型的表示,因此在基于间隔的损失函数中也有基于结构的表示和基于文本的表示,本文使用随机梯度下降(Stochastic Gradient Descent, SGD)来最小化上述损失函数。

3　实验

3.1　数据集

本文使用 FreeBase 的一个子集 FB15k[18] 来进行链接预测和三元组分类实验,实验前删除其中没有描述文本或者描述文本经过预处理后少于 3 个单词的实体。预处理后实体描述文本的平均长度为 69 个单词。表 3 所示为 FB15k 数据集的统计情况。

表 3　FB15k 数据集统计情况

数据集	关系	实体	训练集	验证集	测试集
FB15k	1 341	14 904	472 860	48 991	57 803

3.2　参数设置

设定 SGD 的学习速率 $\lambda\in\{0.1,0.01,0.001\}$,间隔超参数 $\gamma\in\{0.5,1,2\}$,实体和关系表示向量的维度 $k\in\{50,80,100\}$,实体描述文本的词向量维度 $n\in\{50,80,100\}$,特征图维度 $f\in\{50,100,150\}$,batch 大小 $B\in\{20,120,480,600,960,1\,440,4\,800\}$,$\alpha\in\{0,0.1,0.2,0.25,0.5,0.75,0.8,1\}$,卷积层窗口大小 $w\in\{1,2,3\}$,第一池化层使用 4-max-pooling,β 设定为 1。所有实验进行 1 000 轮的迭代训练,最优参数根据验证集中的平均排名确定。

3.3　链接预测

3.3.1　实验设计

链接预测即预测知识三元组中缺失的实体或关系。具体实验中,对测试集中的每个三元组,分别去掉其中的头实体、关系、尾实体,然后依次用数据集中所有存在的实体、关系来替换,形成待评分的三元组。对构造出来的待评分三元组使用损失函数计算相似度,以此对所有替换进去的实体、关系进行排序,相似度越高排名越靠前,从而得到正确的实体、关系排名。

3.3.2　评价指标

与文献[3]类似,本文选择如下 2 个指标作为评估依据:所有正确实体与关系的平均排名(MeanRank)和正确实体与关系排名进前 10% 的比例(Hits@10)。MeanRank 越低或 Hits@10 越高,代表模型的效果越好,表示学习能力越强。但是,在该度量方式中,可能存在一些三元组被“污染”,即构造的待评分三元组中可能会有一些包含于训练集中,这些待评分三元组的得分会比测试集中的得分好。对此,本文在训练集、测试集和验证集中删除被“污染”的三元组。将原来的三元组称为“Raw”,删除掉的、0 被“污染”的三元组称为“Filter”。因为各模型使用的数据集都一样,所以直接引用各模型的实验结果作为对比。

3.3.3　结果与分析

该实验中,最优参数 $\lambda=0.001,\gamma=1,k=n=f=100,B=1\,440,\alpha=0.8$。实验结果如表 4、表 5 所示。

表 4　各模型链接预测实验结果

目标	模型	MeanRank(Raw)	MeanRank(Filter)	Hits@10(Raw)	Hits@10(Filter)
实体预测	TransE	243.00	125.00	34.9	47.1
	TransR	198.00	77.00	48.2	68.7
	CNN	200.00	113.00	44.3	57.6
	word2vec(SG) + TransE	195.00	101.00	40.1	60.1
	CNN + TransE	191.00	91.00	49.6	67.4
	本文模型	185.00	72.00	56.6	77.2
关系预测	TransE	2.91	2.53	69.5	90.2
	TransR	2.68	2.33	69.6	90.5
	CNN	2.91	2.55	69.8	89.0
	word2vec(SG) + TransE	2.71	2.41	69.7	90.1
	CNN + TransE	2.41	2.23	69.8	90.8
	本文模型	2.37	1.98	69.9	90.9

表 5 各模型在不同类型关系下链接预测实验结果

目标	模型	1-to-1	1-to-*n*	*n*-to-1	*n*-to-*n*
预测头实体（Hits@10）	TransE	43.7	65.7	18.2	47.2
	TransR	78.8	89.2	34.1	69.2
	CNN	51.2	71.4	22.6	61.3
	word2vec(SG) + TransE	55.6	78.2	20.1	70.2
	CNN + TransE	79.1	85.5	37.8	72.5
	本文模型	85.1	93.2	45.8	78.1
预测尾实体（Hits@10）	TransE	28.2	13.1	76.0	41.8
	TransR	79.2	37.4	90.4	72.1
	CNN	40.2	27.6	81.2	56.9
	word2vec(SG) + TransE	50.2	35.2	85.6	61.7
	CNN + TransE	78.4	41.6	82.4	75.2
	本文模型	86.1	51.3	90.5	81.4

从表 4 可以看出：1）文本表示和结构表示联合训练（CNN + TransE、本文模型）的预测结果好于文本表示（CNN）和结构表示（TransE）的单独训练，说明联合训练可以更好地表示知识图谱中的数据；2）CNN + TransE 的预测结果好于 word2vec(SG) + TransE，说明对词向量使用 CNN 处理可以更好地表示文本信息；3）Filter 下的结果好于 Raw，说明删除掉被“污染”的三元组可以提升预测结果。

从表 5 可以看出：1）TransR、CNN + TransE 的预测结果好于 TransE，说明对头实体和尾实体使用不同矩阵进行映射可以更好地表示不同类型的关系；2）预测头实体和预测尾实体的结果有所差别，验证了知识图谱中的异质性和不均衡性。

3.4 三元组分类

3.4.1 实验设计

三元组分类是一个二分类任务，旨在判断一个给定三元组(h,r,t)的正确性。文献[8,10,12]均对该分类进行过介绍。本文按照文献[12]中的方法构造一个负样本集来进行分类实验。

在三元组分类实验中，首先设定一个阈值σ_r，对于三元组(h,r,t)，如果由损失函数计算得到的值低于σ_r，三元组则被分类为正；否则，分类为负。σ_r的值通过最大化验证集上的分类准确度得到。

3.4.2 结果与分析

该实验中，最优参数$\lambda=0.001$，$\gamma=2$，$k=n=f=100$，$B=120$，$\alpha=0.8$。实验结果如表 6 所示。从表 6 中可以看出：1）本文模型取得了 89.9% 的准确率，优于所有其他模型，说明在三元组分类中，文本表示和结构表示的联合训练同样有效；2）从单模型角度看，CNN 单独表示实体比结构模型好，体现出文本表示在知识图谱表示学习中的潜力，即优秀的文本表示模型可以在一定程度上提升表示效果。而基于翻译思想的模型效果都好于 NTN，说明翻译思想更能表示知识图谱中的数据。

表 6 不同模型三元组分类实验结果 %

模型	准确率
NTN	68.5
TransE	79.2
TransR	83.9
CNN	85.2
word2vec(SG) + TransE	83.5
CNN + TransE	86.4
本文模型	89.9

4 结束语

针对知识图谱补全，本文提出一种联合学习三元组结构信息和实体描述文本的表示学习算法。通过引入非对称映射操作的基于翻译思想模型生成结构表示，运用 CNN 对实体描述文本进行编码得到文本表示，将两者进行联合学习从而得到实体和关系的表示。链接预测实验与三元组分类实验结果均表明，相对 TransE、TransR、CNN 等模型，本文模型可以更好地表示知识图谱中的数据。在实际中，除实体的描述信息提供的文本外，互联网中还有大量的文本信息，下一步考虑将不同来源的文本融入到知识表示学习中。此外，今后还将考虑使用一些较新的 CNN 模型来表示文本。

参考文献

[1] 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.

[2] 袁书寒，向 阳. 词汇语义表示研究综述[J]. 中文信息学报，2016，30(5)：1-8.

[3] 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.

[4] 安　波,韩先培,孙　乐,等. 基于分布式表示和多特征融合的知识库三元组分类[J]. 中文信息学报, 2016, 30(6): 84-89.

[5] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[6] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[EB/OL]. [2017-06-10]. http://www.aclweb.org/anthology/C/C14/C14-1220.pdf.

[7] 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. [S.l.]: Curran Associates., Inc., 2013: 2787-2795.

[8] WANG Z, ZHANG J, FENG J, et al. Knowledge graph and text jointly embedding[EB/OL]. [2017-06-10]. http://www.aclweb.org/anthology/attachments/D/D14/D14-1167.Attachment.pdf.

[9] 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. [S.l.]: AAAI Press, 2014: 1112-1119.

[10] 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. [S.l.]: AAAI Press, 2015: 2181-2187.

[11] JI G, HE S, XU L, et al. Knowledge graph embedding via dynamic mapping matrix[EB/OL]. [2017-06-10]. http://or.nsfc.gov.cn/bitstream/00001903-5/149814/1/1000014952718.pdf.

[12] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates., Inc., 2013: 926-934.

[13] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2017-06-10]. http://www.surdeanu.info/mihai/teaching/ista555-spring15/readings/mikolov2013.pdf.

[14] XIE R, LIU Z, JIA J, et al. Representation learning of knowledge graphs with entity descriptions[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2016: 2659-2665.

[15] ZHANG D, YUAN B, WANG D, et al. Joint semantic relevance learning with text data and graph knowledge[EB/OL]. [2017-06-10]. http://wing.comp.nus.edu.sg/~antho/W/W15/W15-4004.pdf.

[16] LONG T, LOWE R, CHEUNG J C K, et al. Leveraging lexical resources for learning entity embeddings in multi-relational data[EB/OL]. [2017-06-10]. http://aclweb.org/anthology/P16-2019.

[17] TIAN F, GAO B, CHEN E H, et al. Learning better word embedding by asymmetric low-rank projection of knowledge graph[J]. Journal of Computer Science and Technology, 2016, 31(3): 624-634.

[18] 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.

编辑　吴云芳

(上接第 204 页)

[18] LIU P, JOTY S, MENG H. Fine-grained opinion mining with recurrent neural networks and word embeddings[C]// Proceedings of ACM Conference on Empirical Methods in Natural Language Processing. New York, USA: ACM Press, 2015: 1433-1443.

[19] WANG B, LIU M. Deep learning for aspect-based sentiment analysis[EB/OL]. [2017-05-21]. http://web.stanford.edu/class.

[20] TANG D, QIN B, FENG X, et al. Effective LSTMs for target-dependent sentiment classification[EB/OL]. [2017-05-21]. http://www.aclweb.org/.

[21] TANG D, QIN B, LIU T. Aspect level sentiment classification with deep memory network[EB/OL]. [2017-05-21]. http://wing.comp.nus.edu.sg/.

[22] SUKHBAATAR S, WESTON J, FERGUS R. End-to-end memory networks[J]. Neural Information Processing Systems, 2015(1): 2440-2448.

[23] WANG Y, HUANG M, ZHU X, et al. Attention-based LSTM for aspect-level sentiment classification[C]// Proceedings of IEEE EMNLP'06. Washington D. C., USA: IEEE Press, 2016: 606-615.

[24] ZHANG Y, ZHANG H, ZHANG M, et al. Do users rate or review?: boost phrase-level sentiment labeling with review-level sentiment classification[C]//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. New York, USA: ACM Press, 2014: 1027-1030.

编辑　索书志