计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.20210849
59(12): 2878-2888, 2022

融合语义解析的知识图谱表示方法

胡旭阳 王治政 孙媛媛 徐 博 林鸿飞
(大连理工大学计算机科学与技术学院 辽宁大连 116024)
(huxy912@163.com)

Knowledge Graph Representation Method Combined with Semantic Parsing

Hu Xuyang, Wang Zhizheng, Sun Yuanyuan, Xu Bo, and Lin Hongfei
(*School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning* 116024)

Abstract To solve the problem that the knowledge graph representation learning model only uses triples information, a representation model with semantic analysis is proposed, which is named bidirectional encoder representations from transformers-pruning knowledge embedding (BERT-PKE). It employs bidirectional encoder representations to analyze text, and mines the depth semantic information of entities and relations based on the entities and relations of text description. Since BERT has the heavy consumption in the training time, we propose a pruning strategy with word frequency and k-nearest neighbors to extract the selected text description set. In addition, due to the construction of negative samples has impacts on training model, two strategies are introduced for improving random sampling. One is a negative sampling method based on entity distribution, in which the Bernoulli distribution probability is used to select the replaced entities. It reduces the Pseudo-Labelling problem caused by negative sampling. The other is a negative sampling method based on the similarity of the entities. It mainly uses TransE and k-means to represent the entities as the vectors and classify the entities respectively. High-quality negative triples can be obtained by mutual replacement of entities in the same cluster, which is helpful for feature learning of entities. Experimental results show that the performance of proposed model is significantly improved compared to the SOTA baselines.

Key words knowledge graph representation learning; BERT; semantic analysis; negative sampling; pruning

摘 要 为解决大多数知识图谱表示学习模型仅使用三元组信息的问题,提出融合语义解析的知识图谱表示模型 BERT-PKE.模型利用实体和关系的文本描述,通过 BERT 的双向编码表示进行语义解析,深度挖掘语义信息.由于 BERT 训练代价昂贵,提出一种基于词频和 k 近邻的剪枝策略,提炼选择文本描述集.此外,由于负样本的构造影响了模型的训练,提出 2 种改进随机抽样的策略:一种是基于实体分布的负采样方法,以伯努利分布概率来选择替换的实体,该方法可以减少负采样引起的伪标记问题;另一种是基于实体相似性负采样方法,首先用 TransE 将实体嵌入到向量空间,使用 k-means 聚类算法将实体进行分类.通过同簇实体的相互替换可获得高质量的负三元组,有利于实体的特征学习.实验结果表明,所提出 BERT-PKE 模型与 TransE,KG-BERT,RotatE 等相比,性能有显著提升.

收稿日期:2021-08-20;**修回日期**:2022-01-29
基金项目:国家重点研发计划项目(2018YFC0830603)
This work was supported by the National Key Research and Development Program of China (2018YFC0830603).
通信作者:孙媛媛(syuan@dlut.edu.cn)

关键词 知识图谱表示学习;BERT模型;语义解析;负采样;剪枝

中图法分类号 TP311

伴随进入“大数据”时代,网络中的数据呈指数增长[1].互联网的信息结构多样,多数以HTML格式承载,使用者只能从网页中搜寻自己需要的内容信息,但计算机无法有效地从网页中识别语义信息,数据难被高效利用.于是,“互联网之父”Berners等人[2]提出语义网(semantic Web)的概念,即将万维网中HTML格式链接的网页转化为可被计算机处理的数据链接,将现实世界中的万物联系起来.其中信息以资源描述框架RDF[3](主体-谓词-客体的三元组形式)描述,统一的格式便于计算机处理.随之谷歌提出知识图谱(knowledge graph, KG),其本质是语义网络的知识库,将其用于网页搜索,可从语义层次理解需求,使搜索准确率提高[4].

图谱以图模型可视化地描述了现实世界中信息的关联,继提出概念后,构建和应用知识图谱得到了高速的发展.涌现出大量的开放知识图谱,如WordNet[5], DBpedia[6], NELL[7], YAGO[8], Freebase[9]等.知识图谱揭示了知识的发展规律,并应用于实际任务,如在语义解析[10]、实体消歧[11]、信息提取[12]和问题回答[13]等多个领域均发挥出越来越重要的作用.

尽管知识图谱在结构化表示数据方面很有效,但是这种表示方法由于RDF或类似标准的符号特性需要针对不同的符号设计不同的模型,复杂度高,通用性差、计算效率低.并且知识图谱包含信息极大,符号的表示方法无法缓解数据稀疏性,占用空间大.

近年来,深度学习[14]的迅速发展引起人们广泛的关注,通过表示学习建模在许多方面表现出优越的性能.为解决由知识图谱符号表示所带来的问题,研究人员提出一个新的研究领域——知识表示学习[15],针对知识图谱建模的表示学习也称知识图谱嵌入[16].其核心是在向量空间中建模知识图谱,将符号形式的三元组表示为低维的向量形式,同时保留知识图谱原有的结构.嵌入向量可进一步应用于各种下游任务,如知识图谱补全[17]、关系提取[18]、实体分类[19]和实体解析[20].这种方法具有以下优点[15]:1)便于计算分析;2)融合异质信息[20];3)解决数据稀疏[15,20].

目前,知识图谱表示学习方法大多是仅根据三元组来进行的.即,在向量空间中表示三元组中的实体和关系,并对每个三元组定义一个评分函数衡量其存在的合理性.实体和关系的表示(嵌入)通过最大化三元组的合理性来获得.但这种方法得到的向量表示仅与每个三元组结构有关,而不相连实体之间的隐含关系.因此,得到的向量表示不够准确,对下游任务的预测精度有限[21].为此,研究人员提出融合多源信息进行知识图谱表示学习,如实体类别[22]、关系路径[23]、文本描述[24]、逻辑规则[25]信息等.

由于在给定数据时,不同类型的实体和关系通常均带有文本描述,即一段描述实体或者关系的文字,其文本描述中可能含有复杂的隐藏路径关系.比如给定三元组(中国,首都,北京)、(中国,城市,上海)以及北京的一段描述“北京是中国一座城市,也是中国的首都”,通过这段关于北京的文本描述可以推断出(中国,城市,北京)这样隐含的关系路径.为挖掘更深层次的信息,建模利用的信息更加丰富,更好地学习嵌入,本文旨在将带有复杂语义信息的知识图谱嵌入到低维向量中,以达到知识表示学习的目的,并在具体的下游任务中取得显著效果.

为得到准确的知识图谱表示,本文提出一种融合语义解析的知识图谱表示学习模型.如图1所示,将BERT用于图谱表示学习中的语义解析,提出表示模型BERT-PKE.将事实三元组的实体和关系的结构和文本描述信息以序列形式输入BERT,通过训练解析语法,将嵌入转化为序列分类问题,通过对下游任务的微调,得到三元组的向量表示并预测三元组和链路的合理性.在多数现有算法的训练中,

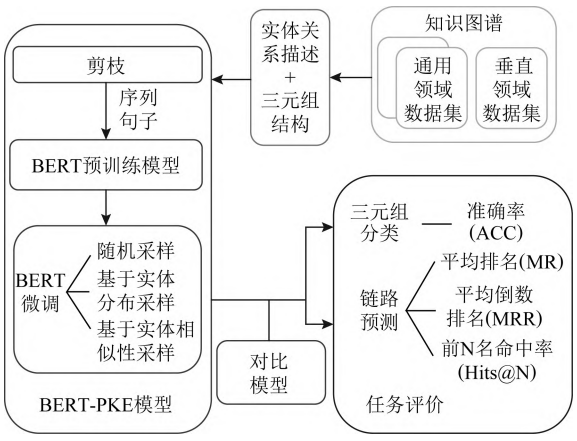


Fig. 1 The research framework

图1 研究内容框架

采用随机负采样方法,生成的负样本是明显的错误样本,很容易通过实体类型区分.因此,本文提出尽量"替换同类实体"的负采样方法——基于实体分布和实体相似度进行采样,生成高质量的负样本用于模型的训练,使模型的训练效果更好.此外,由于BERT计算成本过高,在训练和测试中,解析文本描述微调更新词表花费的时间过长.因此本文提出一种改进策略,将文本描述进行剪枝处理,生成实体和关系的精简文本描述集合,缩短训练时间且性能与原模型基本相同.在构建模型后,将BERT-PKE模型与先进的知识图谱嵌入模型进行调试比较,测试并分析三元组分类和链路预测中的评价结果.经过实验验证,BERT-PKE模型和提出的改进策略在三元组分类和链路预测任务中提升效果显著.

1 相关工作

近年来,知识图谱表示学习研究蓬勃发展,根据研究者利用的信息结构,可分为使用事实三元组建模和融合其他信息建模[26]的知识图谱表示学习模型.

1.1 基于事实三元组的知识图谱表示学习

基于事实三元组的知识图谱表示学习仅根据观察到的事实三元组来执行嵌入任务,将其进行向量表示,向量可用于其他下游任务.模型有3个要点:1)表示形式,实体通常表示为目标空间中的向量,而关系通常表示为目标空间中的操作,如向量、矩阵和高斯分布等;2)得分函数,衡量三元组存在的可能性,其得分越高,三元组在图谱中出现的概率越高;3)优化方法,通常使用梯度下降的方法优化求解.基于事实三元组得分函数定义不同,又可分为基于距离的模型、基于语义匹配的模型[16]等.

1.1.1 距离模型

距离模型学习实体和关系表示,将三元组存在的合理性建模为三元组内部隐含的距离[27].给定一个知识图谱,实体首先被投影至低维向量,然后将关系投影为实体之间的平移或旋转算符,通常表示为向量或矩阵.继而,每个三元组通过2个实体之间的距离评价函数来衡量三元组存在的合理性.合理的三元组往往具有较低的距离值.如TransE[17],TransH[28],TransR[29],TransD[30],RotatE[31].

1.1.2 语义匹配模型

语义匹配模型通过相似性得分函数来学习向量表示的三元组特征,通过张量分解的形式,计算潜在语义相似度并衡量三元组存在的合理性.如RESCAL[21],DistMult[32],HolE[33],ComplEx[34-35],ANALOGY[36],SimplE[37].

1.2 融合多源信息的表示学习

融合多源信息的表示学习除了三元组结构信息外,还通过利用外部资源学习知识图谱的嵌入表示,如实体类别、文本描述、关系路径等.为融合实体类别的信息,语义平滑嵌入模型SSE[22]利用嵌入限制、强正则化约束实体和关系,提出平滑性假设并分别使用2种流形学习算法构建模型.为融合实体和关系的语义信息,融合实体描述的知识表示模型,DKRL[38]利用连续词袋和卷积神经网络学习实体和关系中的语义信息,将语义信息和三元组的结构信息一起进行TransE训练,用改进后的TransE模型学习更深层次的嵌入表示.为融合关系路径的信息,基于关系路径的翻译模型PTransE[39],为特定头实体向量到特定尾实体向量之间途径的所有的实体和关系定义关系路径向量,从而可以利用多个关系中包含的语义信息,建模利用的信息更加丰富,能更好地学习嵌入.

2 融合语义解析的知识图谱表示模型

知识图谱是一种复杂图结构,除三元组之外,还有其他可利用的信息,如上下文、句法和语义信息,这些信息可从更深层次刻画实体和关系的关系,却被以往仅仅依据事实知识的嵌入方法所忽视.其中,实体和关系的文本描述就是一个值得解析利用的信息.

将知识图谱中的三元组视为文本序列,提出了一种融合语义解析的知识图谱表示框架——基于BERT[40]模型的剪枝图谱表示模型BERT-PKE.给定知识图谱,首先将实体和关系的文本描述进行剪枝处理.然后,将三元组和文本描述转化成序列结构输入BERT模型中.最后,利用预训练语言模型BERT对三元组以及实体和关系的描述进行语义解析,得到嵌入模型.在训练过程中,负样本的构造可影响模型的学习.因此,提出2种改进经典方法生成负样本的方法,改变负样本的采集方法来增强模型学习的能力.

2.1 BERT-PKE模型结构

由于BERT[40]可解析深层次的语义信息,因此在融合语义解析的知识图谱表示方法中,本文采用BERT来进行语义解析,输入多层Transformer[41]

结构,使用自注意力机制联合所有层的上下文来训练未标注文本,得到深度双向表示,实现图谱嵌入.由于 BERT 是处理自然语言的模型,只能处理序列结构的句子,图结构无法直接输入.因此 BERT-PKE 模型参考 KG-BERT[42]模型中的输入方法,将三元组结构和文本描述作为文本序列输入预训练语言模型 BERT,将描述实体和关系的词序列作为 BERT 模型的输入句进行微调,然后通过某种训练得到三元组的表示.

对于三元组建模的 BERT-PKE 模型整体框架设计如图 2 所示.首先通过对正三元组负采样得到完整训练集,然后通过匹配训练集中的文本描述得到头实体、关系、尾实体对应的 3 个句子 Sen^h,Sen^r,Sen^t.最后经过剪枝,3 个句子描述表示为一个包含词标记 Tok_1^i,Tok_2^i,…,Tok_j^i,$i\in\{1,2,\cdots,512\}$的集合.在输入序列最前面的是分类词标记[CLS],实体和关系的句子被一个特殊的词标记[SEP]分隔开.对于给定的词标记,输入向量由标记、分段和位置嵌入求和得到.被[SEP]分开的词标记段嵌入不同,头尾实体描述中的词标记共享相同的段嵌入 e_A;而关系描述中的词标记则具有不同的分段嵌入 e_B.不同维度的词标记在同一位置中有相同的位置嵌入.

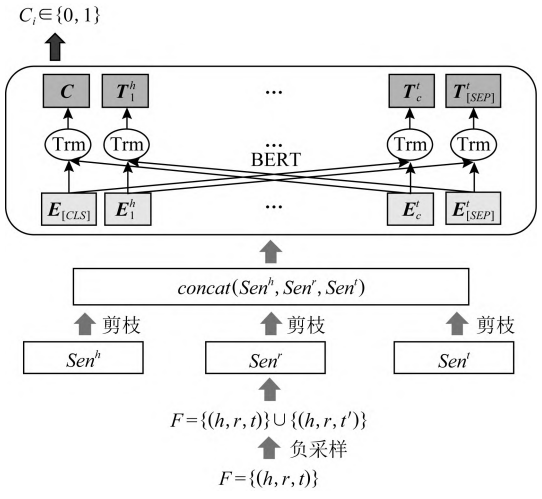


Fig. 2 The overall framework of BERT-PKE

图 2 BERT-PKE 模型整体框架

输入词标记 i 对应的输入向量表示 $\boldsymbol{E}_i$ 输入到 BERT 模型架构中,该架构是基于 Transformer 的双向结构.在隐藏词字机制 MLM 任务中,特殊词标记[CLS]和第 i 个输入词标记中的隐藏向量记为 $\boldsymbol{C}\in\mathbb{R}^H$ 和 $\boldsymbol{T}_i\in\mathbb{R}^H$,其中 H 为预先训练 BERT 中的隐藏块大小.与[CLS]对应的最终隐藏块输出 $\boldsymbol{C}$ 被用于计算三元组的序列表示得分.微调过程中引入的唯一参数 $\boldsymbol{W}\in\mathbb{R}^{2\times H}$,表示输出层的权重.三元组 (h,r,t) 的得分函数为

$$\boldsymbol{s}_\tau=f_r(h,t)=\mathrm{sigmoid}(\boldsymbol{C}\boldsymbol{W}^{\mathrm{T}}),\tag{1}$$

其中权重矩阵 $\boldsymbol{W}$ 与 $\boldsymbol{C}$ 相乘之后可获得三元组是正确的概率 $\boldsymbol{s}_\tau$,$\boldsymbol{s}_\tau\in\mathbb{R}^2$ 是 2 维实向量,且 $s_{\tau0}$,$s_{\tau1}\in[0,1]$ 且 $s_{\tau0}+s_{\tau1}=1$.

在给定正三元组集合 D^+ 和相应构造的负三元组集合 D^-,我们用 $\boldsymbol{s}_\tau$ 和三元组标记计算交叉熵损失:

$$\mathcal{L}=-\sum_{\tau\in D^+\cup D^-}(y_\tau\lg(s_{\tau0})+(1-y_\tau)\lg(s_{\tau1})),\tag{2}$$

其中 $y_\tau\in\{0,1\}$ 是标记该三元组是正例还是负例的标签,即标记是正三元组还是负三元组,而正三元组表示的是正确的三元组,负三元组表示的是错误的三元组,负样本需要我们进行负采样构造.

负采样方法可影响模型的预测能力,在 2.2 节中我们将给出详细介绍.通过梯度下降的方法,可以更新预先训练好的参数权值和新的权值 $\boldsymbol{W}$.

2.2 负采样方法

负采样的目的是帮助模型进行特征学习训练,最终输出正样本.正样本在损失函数学习过程中保留,同时不断更新负样本.通过负采样,在更新隐藏层到输出层的权重时,只需更新负样本而不用更新全部样本,节省计算量.因此负样本的采集质量影响了模型的构建.本文通过负采样的方式降噪,对样本集的正三元组进行负采样,生成的负样本用于计算损失函数.

在现有的知识图谱表示模型中,负采样大多从实体集中随机抽取进行替换,采用这种负采样方法生成的负样本随机且质量较低.这样会带来产生伪标签和模型无法准确地学习训练 2 个问题.针对问题,提出 2 种改进的负采样方法,分别是基于实体分布的负采样方法和基于实体相似度的负采样方法.通过后续试验证明方法的效果.

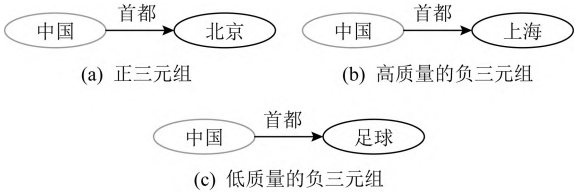
2.2.1 随机抽样的负采样方法

虽然融合语义解析的知识图谱表示方法在实现知识图谱表示学习上有了进一步的突破,但是现有的嵌入模型中普遍存在一个问题,即模型在梯度下降训练中,负三元组集合 D^- 仅仅由实体集合中随机抽取一个实体 h' 或 t',从正三元组 $(h,r,t)\in D^+$ 中替换相应的 h 或 t 得到的,即

$$\begin{aligned}D^-=&\{(h',r,t)\mid(h'\in E)\wedge(h'\neq h)\wedge\\&((h',r,t)\notin D^+)\}\cup\{(h,r,t')\mid\\&(t'\in E)\wedge(t'\neq t)\wedge((h,r,t')\notin D^+)\},\end{aligned}\tag{3}$$

如果三元组已经在正集 D^+ 中,则不会被视为反例.

通过梯度下降的方法,负样本更新预先训练好的参数,因此采样的负三元组质量影响了模型的学习和向量的表示.例如,给定三元组(中国,首都,北京)经过随机负采样生成后的三元组可能为(中国,首都,足球),该三元组质量低,对训练过程中参数的更新没有显著帮助.这种采样方法被称为 unif[17] 采样,最初在 TransE 模型中被提出.由于知识图谱数据集中的信息是有限的,通过随机采样产生的负样本可能构造出正三元组,却被当作负样本本来处理,引入伪标签.图 3 是正、负三元组的举例说明.鉴于负采样的基本作用和现有方法的局限性,本文将重点放在负采样上,旨在提高负样本的质量.



(a) 正三元组 (b) 高质量的负三元组

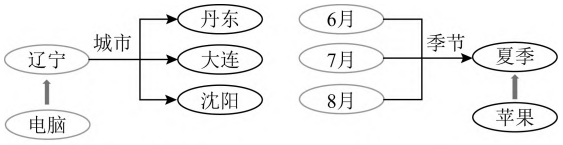
(c) 低质量的负三元组

Fig. 3 Examples of positive and negative triples

图 3 正、负三元组举例

2.2.2 基于实体分布的负采样方法

根据 TransH 中提出的方法,以不同概率按照实体分布来选择替换三元组的头实体或尾实体,可依据伯努利分布提出 bern[28] 采样.本文针对 1_to_N 和 N_to_1 类型的三元组,如果是 1_to_N 三元组,则更大概率破坏头实体 h;如果是 N_to_1 三元组,则更大概率破坏尾实体 t,这样就减少了产生伪标签的机会.图 4 分别展示了不同关系类型下,基于实体分布的 bern 负样本生成过程.



(a) 1_to_N类型 (b) N_to_1类型

Fig. 4 Bern negative sampling

图 4 bern 负采样

对于知识图谱中的每个关系 r,首先得到以下 2 个统计量:1)头实体对应的平均尾实体的数,记为 t_{ph};2)尾实体对应的平均头实体数,记为 h_{pt}.然后为采样定义一个伯努利分布,其参数为

$$p=\frac{t_{ph}}{t_{ph}+h_{pt}}, \tag{4}$$

则替换头或尾实体,服从参数为 p 的伯努利分布,有

$$X=\begin{cases}1, & \text{替换头实体},\\ 0, & \text{替换尾实体},\end{cases} \tag{5}$$

则 X 的分布律为

$$P(X=x)=p^x(1-p)^{1-x},\ x\in[0,1]. \tag{6}$$

对于与关系 r 相关的正三元组(h,r,t),替换 h 构造负三元组的概率为 p,替换 t 构造负三元组的概率为 $1-p$.

2.2.3 基于实体相似性的负采样方法

基于实体分布的负采样方法虽然能够减少了产生伪标签的可能性,但替换实体仍需从整个实体集中选择,生成的三元组质量不佳,对训练过程中的特征学习帮助不大.本文希望替换的实体与原实体语义相似,因此提出一种基于实体相似性[43]的负采样方法,进一步改进 2.2.2 节中基于实体分布的负采样方法.该方法先使用 TransE 将实体表示成 m 维向量将相似性问题简化,然后用 k-means[44] 聚类将实体向量划分为 k 类.在负采样时,正三元组的实体用同类实体进行替换,通过这种负采样方法来提升知识图谱嵌入的质量.图 5 分别展示了 1_to_N 和 N_to_1 类型的 k-means 负样本生成.

(a) 1_to_N类型 (b) N_to_1类型

Fig. 5 k-means negative sampling

图 5 k-means 负采样

首先,本文使用 TransE 将实体和关系表示成 m 维向量,将实体的语义结构相似问题简化为向量距离相近问题.给定三元组(h,r,t),TransE 模型都有 $h+r\approx t$.因此在向量空间中,头实体 h 被 $t-r$ 限制.同理,尾实体 t 和关系 r 分别有 $h+r$ 和 $h-t$ 限制.因此,不同三元组中同一个头实体在不同关系和尾实体的限制是相同的,即三元组(h_1,r_1,t_1)和(h_1,r_2,t_2)中有 $t_1-r_1=t_2-r_2$.因此若 2 个实体相似,则其在空间中的限制也相似,表明在空间中 2 实体的向量坐标越相近,距离越小,则实体越相似.

在得到实体和关系的嵌入向量后,使用 k-means 算法对实体向量进行无监督的分类.首先,在实体向量集合$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\in\mathbb{R}^m$中选择初始化的 k 个样本作为初始聚类中心$\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k\}$;然后,针对实体向量集中每个实体向量 $\boldsymbol{x}_i$ 所属的聚类中的所有点到聚类中心的欧氏距离之和最小,对于每个向量 $\boldsymbol{x}_i$,计算其应该属于的类:即

$$c_i=\arg\min_j\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|_2^2,r_{nk}=\begin{cases}1,\text{若 }\boldsymbol{x}_n\in c_i,\\0,\text{其他情况},\end{cases}\tag{7}$$

其中,c_i 表示样本 $\boldsymbol{x}_i$ 与 k 个距离中心最近的类.arg 是表明样本归于哪个类的运算符.然后,对于每个类中心 $\boldsymbol{\mu}_j$,重新计算该聚类的中心

$$\boldsymbol{\mu}_j:=\frac{\sum_{i=1}^{n}r_{ij}\boldsymbol{x}_i}{\sum_{i=1}^{n}\boldsymbol{\mu}_j}.\tag{8}$$

不断重复划分类 c_i 和更新聚类中心 $\boldsymbol{\mu}_j$ 这 2 个操作,直到达到聚类的中心不变或者变化很小,其损失函数为

$$J=\sum_{i=1}^{C}\sum_{j=1}^{N}r_{ij}\cdot c_i.\tag{9}$$

通过 k-means 聚类算法,本文认为属于同一个类别的实体相似度高,可相互替换生成负样本.基于实体相似性的方法在一定程度上提高了负样本的质量,使表示模型的性能得到了提升.

2.3 剪枝策略

BERT 模型的一个主要局限性是代价太过于昂贵,在学习模型表示的过程中需要将句子的每个词输入多层 Transformer 结构中进行嵌入训练;在测试模型的过程中更是耗费大量时间;在链路预测评估需要遍历所有的实体替换头实体或尾实体,并且所有负三元组序列都被输入到 12 层 Transformer 模型中.由于序列中文本描述通常为一段话,在 50 词以上,过于冗长,包含一些无用信息,如标点、谓词、系动词等.

为尽可能正确嵌入训练样本,每个样本句子的词标记的学习过程将不断重复,有时词标记形成的分支过多,这时就有可能把训练集学习得太好,以至于把训练集的某一些特点当成所有数据都具有的性质,这时就发生了过拟合.因此,针对 BERT 模型的局限性,本文将实体和关系的文本描述进行剪枝处理,对冗余的文本描述进行修剪,从而避免嵌入不必要的操作和搜索,更快地获得更好的效果.

本文使用基于词频[45](term frequency)和 k 近邻[46](k-nearest neighbor)的技术.首先,TF 表示的是某个词在文本中出现的次数,即词频,其公式为

$$TF_i=\frac{\text{在句子中词条 i 出现的次数}}{\text{该句子中的词条数}}.\tag{10}$$

对于词频统计的具体做法,本文采用 N 元语法模型(N-gram),N-gram 是一种基于统计语言模型的算法.将描述文本中的单词按字节进行大小为 N 的滑动窗口操作,形成字节片段序列.每个片段称为 gram,对所有 gram 的出现频度进行统计,并且按照阈值过滤,形成文本的向量特征空间.

在 N-gram 中,第 N 个词的出现只与前面 $N-1$ 个词相关,与其他任何词都不相关,整句的概率就是各个词出现概率的乘积.这里只需要获得各个词出现的词频.本文取 $N=1,2,3$.其中,当 $N=1$ 时,称为一元语法模型(unigram model),即当前词的概率分布与给定的历史信息无关,它将文本描述分成单词,统计单词出现的词频;当 $N=2$ 时,称为二元语法模型(bigram model),即当前词的概率分布只与距离最近的词有关,它将文本描述中所有 2 个词组成一个词组,统计词组出现的词频;当 $N=3$ 时,称为三元语法模型(trigram model),即当前词的概率分布与距离最近的 2 个词有关,它将文本描述中的所有相邻 3 个词组成 1 个词组,统计词组出现的词频.

k 近邻表示的是一个样本附近的 k 个最近,即特征空间中最邻近样本,文本是 1 维表示,则其最近邻的度量方式为曼哈顿距离,即

$$L(j,k)=|j-k|.\tag{11}$$

因此,本文在剪枝过程中抽取实体和关系名称的前后 k 跳词语,并抽取除名称、标点、量词、系动词以外出现词频较高出现的词或词组(可以为多个),然后组成一个由逗号分隔、由词语组成的实体和关系的文本描述集合.通常,剪枝后的模型精度稍微有所下降,但相比节省了大量的时间空间,精度基本与原来持平或稍稍下降的误差完全可以忽略.

3 实验设置及结果

本文选用垂直领域数据集 UMLS[47],通用领域数据集 FB14K-237 和 WN18R.其中 FB14K-237 由 FB15K-237[48],WN18R 由 WN18RR[48] 预处理得到,具体信息如表 1 所示:

Table 1 The Information of Data Sets

表 1 数据集信息

知识图谱	实体数	关系数	三元组数		
			训练集	测试集	验证集
FB14k-237	13 986	237	108 846	8 186	7 014
WN18R	30 078	11	34 734	1 253	1 213
UMLS	135	46	5 216	661	652

在完成嵌入后,将嵌入的向量应用于不同的

下游任务中,本文的下游任务为三元组分类和链路预测[49-52].并采用准确率(ACC)作为评价指标用于衡量三元组分类的效果,采用平均排名(MR)、平均倒数排名(MRR)和正确实体排在前 N 名的概率(Hits@N)作为评价指标用于衡量链路预测的效果.

三元组分类的目的是判断三元组(h,r,t)中实体和关系是否正确匹配,本文将各个模型运行 3 次并取其平均值,表 2 给出了 FB14k-237, WN18R, UMLS 在不同模型上的三元组分类任务的准确率.

Table 2 ACC of Triplet Classification

表 2 三元组分类的准确率 %

模型	FB14k-237	WN18R	UMLS	平均值
TransE	60.76	59.18	78.44	66.13
TransH	62.30	58.26	79.27	66.61
TransR	68.11	65.12	83.36	72.20
TransD	60.70	58.46	81.08	66.75
RotatE	62.22	55.63	83.36	67.07
RESCAL				
DistMult	69.41	58.41	83.96	70.59
HolE	68.56	53.47	86.15	69.39
ANALOGY	72.21	57.82	85.47	71.83
ComplEx	76.60	61.61	86.00	74.74
SimplE	61.04	56.30	86.31	67.88
KG-BERT	95.51	96.70	86.05	92.79
BERT-PKE(unif)	95.43	96.41	86.46	92.77
BERT-PKE(bern)	96.46	96.68	88.65	93.93
BERT-PKE(k-means)	96.94	96.89	91.16	95.00

如表 2 可得,所提出 BERT-PKE 模型在三元组分类任务上的准确性显著高于所有基准模型,和 KG-BERT 原型基本相同,证明了本文提出方法的有效性.所提出的剪枝策略改进的 BERT-PKE 模型与原模型 KG-BERT 的准确率相差不多,但训练时间却大大缩短.以 FB14k-237 数据集为例,KG-BERT 算法中词标记有 4 920 563 个,迭代 1 次需要 25 h,而 BERT-PKE 算法中词标记只有 20 409 个,迭代 1 次只需要 1.5 h,而在模型学习过程中需要多次迭代,时间代价成倍数增长.在通用领域数据集 WN18R、FB14k-237 上,其准确率均高于 95%,且采用剪枝策略相差不超过 0.3%,在垂直领域数据集 UMLS 上准确率甚至高于 KG-BERT 模型.因此可以看出,剪枝策略可显著节省模型训练的时间和空间.同时,提出基于实体分布负采样和基于实体相似性负采样负采样改进方法,在 3 个数据集上,这 2 种负采样都能使得 BERT-PKE 模型的性能得到提升.并且基于实体相似性(k-means)的负采样方法在基于实体分布(bern)的方法基础上提出,其准确率也更高.通过实验,证明负采样在模型训练中的重要影响,也证明本文改进采样方法的成效.

链路预测的目的是判断三元组(h,r,t)在已知其中关系和其中一个实体的情况下预测的另一实体是否正确.表 3~5 中给出不同数据集下不同模型的链路预测结果.

Table 3 Link Prediction Results of UMLS

表 3 UMLS 数据集的链路预测结果

模型	MR		Hits@N/%		
	MR	MRR	N=1	N=3	N=10
TransE	3.42	0.59	36.46	79.20	93.87
TransH	3.41	0.58	34.64	78.80	93.57
TransR	2.30	0.60	35.40	81.09	94.48
TransD	3.26	0.59	35.85	79.58	95.01
RotatE	3.06	0.72	58.32	83.13	94.40
RESCAL	16.29	0.30	20.65	30.94	50.00
DistMult	3.14	0.71	56.96	81.24	93.42
HolE	2.28	0.81	71.86	89.03	97.88
ANALOGY	2.63	0.76	64.52	85.02	96.82
ComplEx	2.57	0.77	63.54	86.84	98.00
SimplE	2.60	0.75	60.06	86.84	95.84
BERT-PKE (unif)	2.10	0.81	71.67	89.09	97.87
BERT-PKE (bern)	1.88	0.82	72.35	92.73	98.48
BERT-PKE(k-means)	1.81	0.85	78.03	93.65	98.80

Table 4 Link Prediction Results of WN18R

表 4 WN18R 数据集的链路预测结果

模型	MR		Hits@N/%		
	MR	MRR	N=1	N=3	N=10
TransE	3 716	0.013	0.56	1.16	2.51
TransH	3 822	0.014	0.60	1.36	2.63
TransR	3 365	0.037	0.36	5.02	9.70
TransD	3 821	0.014	0.52	1.14	2.39
RotatE	4 529	0.017	0.80	1.72	2.95
RESCAL	4 450	0.004	0.07	0.24	0.56
DistMult	3 275	0.029	1.56	2.59	5.15
HolE	3 489	0.021	1.32	2.19	3.19
ANALOGY	3 199	0.089	7.10	9.38	12.10
ComplEx	2 974	0.099	11.25	15.28	18.36
SimplE	3 444	0.007	0.20	0.44	1.08
BERT-PKE (unif)	234.7	0.104	3.55	8.66	14.95
BERT-PKE(bern)	224.5	0.110	6.02	16.44	38.99
BERT-PKE(k-means)	196.8	0.148	6.76	18.44	43.46

Table 5 Link Prediction Results of FB14k-237

表 5 FB14k-237 数据集的链路预测结果

模型	MR		Hits@N/%		
	MR	MRR	N=1	N=3	N=10
TransE	277.6	0.102	6.07	9.89	17.00
TransH	276.7	0.116	7.42	11.15	18.66
TransR	269.5	0.151	10.27	15.34	23.89
TransD	275.5	0.101	5.90	9.88	17.21
RotatE	260.4	0.165	11.23	16.62	26.80
RESCAL	333.0	0.077	4.44	7.07	13.02
DistMult	146.2	0.133	8.01	13.17	22.97
HolE	244.9	0.129	7.64	13.00	22.25
ANALOGY	233.9	0.147	9.09	14.60	25.58
ComplEx	235.7	0.185	12.63	19.21	29.90
SimplE	266.8	0.164	11.31	16.19	26.41
BERT-PKE (unif)	190.7	0.221	6.93	12.14	21.82
BERT-PKE (bern)	203.7	0.230	7.95	13.81	22.59
BERT-PKE(k-means)	144.1	0.259	10.41	16.01	25.54

通过表 3～5 以及可视化图 6～8 上的结果,可以总结出:1)3 种 BERT-PKE 模型的 MR 值均比基准模型的 MR 值更低,MRR 值更接近于 1,并且提升较为明显.且采用基于实体相似性和实体分布的负采样策略也有明显的提升效果.2)3 种 BERT-PKE 模型中有一小部分负采样方法中的 Hits@N 值低于一些最先进的方法,如 ANALOGY,ComplEx;但采用改进负采样策略的 BERT-PKE 模型的 Hits@N 值较随机负采样方法相比有明显提升.这是由于 BERT-PKE 模型没有对知识图谱的整体图结构信息进行准确建模,从而无法使得实体和关系描述的

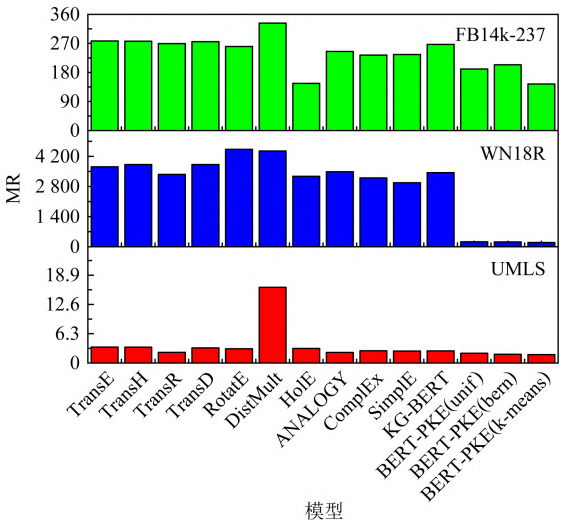


Fig. 6 MR of link prediction

图 6 链路预测的 MR

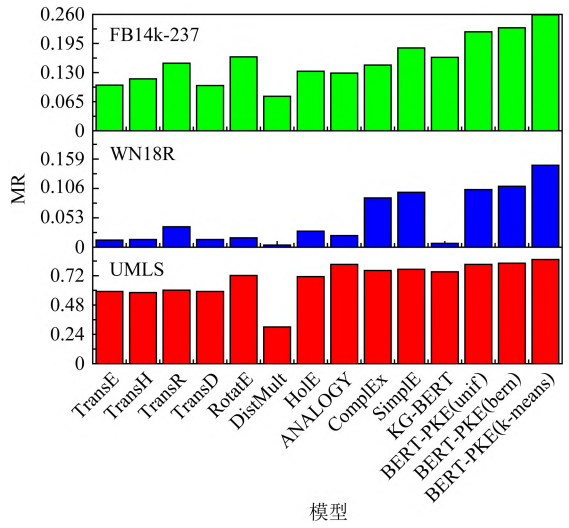


Fig. 7 MRR of link prediction

图 7 链路预测的 MRR

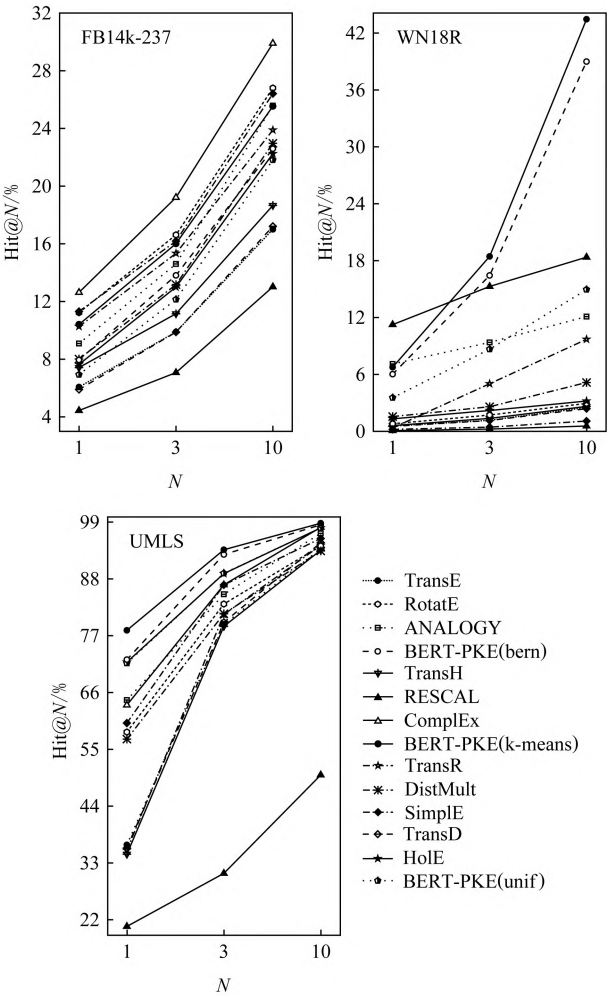


Fig. 8 Hits@N for top 1,3,10 of link prediction

图 8 链路预测的前 1,3,10 名命中率

语义相关度很高,因此不能将给定实体的某些邻居实体排在前 10 位.通过基于实体分布和实体相似度

的负采样改进方法可提高 Hits@N 值、判断实体关系的种类、并通过 TransE 预先得到实体相似度分布、然后进行归类，该方法都对图结构有一些整体把握，因此可提升模型的性能.由此可得，负采样策略可提升知识图谱表示学习的能力，并且通过剪枝策略，可大大缩短模型训练和测试的时间，如 FB14k-237 数据集，剪枝前迭代一次需要 25 h，剪枝后只需要 1.5 h；剪枝前测试匹配一个实体需要 8 min，而剪枝后只需要 50 s.

4 结 论

本文提出一种融合语义解析的知识图谱表示模型——BERT-PKE，该模型将 BERT 用于语义解析，提出基于词频和 k 近邻的剪枝策略以缩短训练时间.此外，提出 2 种负采样策略，基于实体分布的负采样方法可减少伪标签产生；基于实体相似性的负采样可通过同簇实体的替换提高负三元组质量，帮助特征训练.本文填补了已有表示模型中挖掘文本描述深度关联的空白.此外，本文还将 BERT 模型应用于知识图谱补全任务.未来的研究方向包括通过图结构联合建模等.将 BERT-PKE 模型作为一种知识增强语言模型应用于语言理解任务是我们未来要探索的一项工作.

作者贡献声明：胡旭阳设计了算法思路和实验方案，完成了所有实验以及文章撰写；王治政参与设计了算法实验、论文架构并完成了实验分析；孙媛媛指导了论文思路，对实验提出指导意见并修改论文；徐博参与了论文想法的讨论，对于实验方案提出指导意见并完善论文内容；林鸿飞负责提出选题并确定论文框架.

参 考 文 献

[1] Xia Chunguan. Knowledge representation learning with semantic interactions and semantic constrains [D]. Zhengzhou: Zhengzhou University, 2020 (in Chinese)
(夏春管. 融合语义交互和语义限制的知识表示学习[D]. 郑州：郑州大学，2020)

[2] Berners L, Hendler J, Lassila O. The Semantic Web [J]. Scientific American, 2001, 284(5): 34-43

[3] Miller E. An introduction to the resource description framework [J]. Bulletin of the American Society for Information Science and Technology, 1998, 25(1): 15-19

[4] Su Yu, Yang Shengqi, Sun Huan, et al. Exploiting relevance feedback in knowledge graph search [C] //Proc of the 21st ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2015: 1135-1144

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

[6] Lehmann J, Isele R, Jakob M, et al. DBpedia—A large-scale, multilingual knowledge base extracted from wikipedia [J]. Semantic Web, 2015, 6(2): 167-195

[7] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning [C] //Proc of the 24th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2010: 1306-1313

[8] Fabian M, Gjergji K, Gerhard W. Yago: A core of semantic knowledge [C] //Proc of the 16th Int Conf on World Wide Web. New York: ACM, 2007: 697-706

[9] Bollacker K D, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge [C] //Proc of the 2008 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2008: 1247-1250

[10] Berant J, Chou A, Frostig R, et al. Semantic parsing on Freebase from question-answer Pairs [C] //Proc of the 2013 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2013: 1533-1544

[11] Alhelbawy A, Gaizauskas R. Graph ranking for collective named entity disambiguation [C] //Proc of the 52nd Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2014: 75-80

[12] Hoffmann R, Zhang C, Xiao Ling, et al. Knowledge-based weak supervision for information extraction of overlapping relations [C] //Proc of the 49th Annual Meeting of the ACL: Human Language Technologies ACL. Stroudsburg, PA: ACL, 2011: 541-550

[13] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models [C] //Proc of the 2014 Machine Learning and Knowledge Discovery in Databases-European Conf. Berlin: Springer, 2014: 165-180

[14] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444

[15] Liu Zhiyuan, Sun Maosong, Lin Yankai, et al. Knowledge representation learning: A review [J]. Journal of Computer Research and Development, 2016, 53(2): 247-261 (in Chinese)
(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)

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

[17] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of the 27th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2787-2795

[18] Weston J, Bordes A, Yakhnenko O, et al. Connecting language and knowledge bases with embedding models for relation extraction [C] //Proc of the 2013 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2013: 1366-1371

[19] Nickel M, Tresp V, Kriegel H P. Factorizing YAGO: Scalable machine learning for linked data [C] //Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 271-280

[20] Tang Huilin. Research on knowledge graph completion by combining structural and semantic information [D]. Beijing: Beijing University of Posts and Telecommunications, 2017 (in Chinese)
(唐慧琳. 融合结构和语义信息的知识图谱补全算法研究[D]. 北京: 北京邮电大学, 2017)

[21] Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data [C] //Proc of the 28th Int Conf on Machine Learning. New York: ACM, 2011: 809-816

[22] Guo Shu, Wang Quan, Wang Bin, et al. SSE: Semantically smooth embedding for knowledge graphs [J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(4): 884-897

[23] Lin Yankai, Liu Zhiyuan, Sun Maosong, et al. Modeling relation paths for representation learning of knowledge bases [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language. Stroudsburg, PA: ACL, 2015: 705-714

[24] Wang Zhen, Zhang Jianwen, Feng Jianlin, et al. Knowledge graph and text jointly embedding [C] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1591-1601

[25] Guo Shu, Wang Quan, Wang Lihong, et al. Jointly embedding knowledge graphs and logical rules [C] //Proc of the 2016 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 192-202

[26] Nie Binling. Graph structure information for knowledge representation Learning [D]. Hangzhou: Zhejiang University, 2019 (in Chinese)
(聂斌玲. 基于图结构信息的知识表示学习方法研究[D]. 杭州: 浙江大学, 2019)

[27] Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs [J]. Proceedings of the IEEE, 2015, 104(1): 11-33

[28] Wang Zhen, Zhang Jiawen, Feng Jianlin, et al. Knowledge graph embedding by translating on hyperplanes [C] //Proc of the 28th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2014: 1112-1119

[29] Lin Yankai, Liu Zhiyuan, Sun Maoping, et al. Learning entity and relation embeddings for knowledge graph completion [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 2181-2187

[30] Ji Guoliang, He Shizhu, Xu Liheng, et al. Knowledge graph embedding via dynamic mapping matrix [C] //Proc of the 53rd Annual Meeting of the ACL and the 7th Int Joint Conf on Natural Language Processing of the Asian Federation of Natural Language Processing. Stroudsburg, PA: ACL, 2015: 687-696

[31] Sun Zhiqing, Deng Zhihong, Nie Jianyun, et al. RotatE: Knowledge graph embedding by relational rotation in complex space [C/OL] //Proc of the 7th Int Conf on Learning Representations. 2019 [2021-05-22]. https://openreview.net/attachment?id=HkgEQnRqYQ&-name=pdf

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

[33] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs [C] //Proc of the 20th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 1955-1961

[34] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 2071-2080

[35] Hayashi K, Shimbo M. On the equivalence of holographic and complex embeddings for link prediction [C] //Proc of the 55th Annual Meeting of the ACM. Stroudsburg, PA: ACL, 2017: 554-559

[36] Liu Hanxiao, Wu Yuexin, Yang Yiming. Analogical inference for multi-relational embeddings [C] //Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 2168-2178

[37] Kazemi S M, Poole D. SimplE embedding for link prediction in knowledge graphs [C] //Proc of the 31st Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018: 4289-4300

[38] Xie Ruobing, Liu Zhiyuan, Jia Jia, et al. Representation learning of knowledge graphs with entity descriptions [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 2659-2665

[39] Lin Yankai, Liu Zhiyuan, Sun Maoping, et al. Modeling relation paths for representation learning of knowledge bases [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 705-714

[40] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint, arXiv: 1810.04805, 2018

[41] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Proc of the 30th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 5998-6008

[42] Yao Liang, Mao Chengsheng, Luo Yuan. KG-BERT: BERT for knowledge graph completion [J]. arXiv preprint, arXiv: 1909.03193, 2019

[43] Rao Guanjun, Gu Tianlong, Chang Liang, et al. Knowledge graph embedding based on similarity negative sampling [J]. CAAI Transactions on Intelligent Systems, 2020, 15(2): 218-226 (in Chinese)
(饶官军, 古天龙, 常亮, 等. 基于相似性负采样的知识图谱嵌入[J]. 智能系统学报, 2020, 15(2): 218-226)

[44] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm [J]. Journal of the Royal Statistical Society, 1979, 28(1): 100-108

[45] Jones S. A statistical interpretation of term specificity and its application in retrieval [J]. Journal of Documentation, 1972, 28(1): 11-21

[46] Wang Yu. Research on text classification based on decision tree and K-nearest neighbor algorithm [D]. Tianjin: Tianjin University, 2006 (in Chinese)
(王煜. 基于决策树和 K 最近邻算法的文本分类研究[D]. 天津: 天津大学, 2006)

[47] Olivier B. The unified medical language system (UMLS): Integrating biomedical terminology [J]. Nucleic Acids Research: Database-Issue, 2004, 32: 267-270

[48] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D knowledge graph embeddings [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 1811-1818

[49] Vrandecic D, Krtoetzsch M. Wikidata: A free collaborateve knowledgebase [J]. Communications of the ACM, 2014, 57(10): 78-85

[50] Ji Shaoxiong, Pan Shirui, Cambria E, et al. A survey on knowledge graphs: Representation, acquisition and applications [J]. arXiv preprint, arXiv: 2002.00388, 2020

[51] Wang Zhen. Embedding model based knowledge graph completion [D]. Guangzhou: Sun Yat-sen University, 2017 (in Chinese)
(王桢. 基于嵌入模型的知识图谱补全[D]. 广州: 中山大学, 2017)

[52] Ding Jianhui, Jia Weijia. A review of knowledge graph completion algorithms [J]. Information and Communication Technology, 2018(1): 56-62 (in Chinese)
(丁建辉, 贾维嘉. 知识图谱补全算法综述[J]. 信息通信技术, 2018, 2018(1): 56-62)

Hu Xuyang, born in 1999. Master candidate. Her main research interest includes knowledge graph representation learning.
胡旭阳,1999 年生.硕士研究生.主要研究方向为知识图谱表示学习.

Wang Zhizheng, born in 1993. PhD candidate. His main research interests include representation learning and knowledge graph reasoning.
王治政,1993 年生.博士研究生.主要研究方向为表示学习和知识图推理.

Sun Yuanyuan, born in 1978. PhD, professor, PhD supervisor. Her main research interests include natural language processing, nonlinear theory and applications, machine learning algorithm. (syuan@dlut.edu.cn)
孙媛媛,1978 年生.博士,教授,博士生导师.主要研究方向为自然语言处理、非线性理论及应用、机器学习算法.

Xu Bo, born in 1988. PhD, associate professor, master supervisor. His main research interests include understanding user intention and learning to rank models.
徐　博,1988 年生.博士,副教授,硕士生导师.主要研究方向为用户意图理解和排序学习模型.

Lin Hongfei, born in 1962. PhD, professor, PhD supervisor. His main research interests include sentiment analysis and opinion mining, information retrieval and recommendation, and knowledge mining.
林鸿飞,1962 年生.博士,教授,博士生导师.主要研究方向为情感分析与观点挖掘、信息检索与信息推荐、知识挖掘.