中国科学: 信息科学 2022 年 第 52 卷 第 11 期: 2037–2057
SCIENTIA SINICA Informationis
论文

基于解析图嵌入和加权图卷积网络的知识图谱补全

罗妹秋, 张春霞*, 彭成, 张鑫, 郭贵锁, 牛振东

北京理工大学计算机学院, 北京 100081
* 通信作者. E-mail: cxzhang@bit.edu.cn

收稿日期: 2021–06–28; 修回日期: 2021–10–17; 接受日期: 2021–12–01; 网络出版日期: 2022–11–10

国家重点研发计划 (批准号: 2020AAA0104903) 和国家自然科学基金 (批准号: 62072039) 资助项目

摘要 知识图谱补全是知识图谱构建、自然语言处理和知识工程等领域的重要研究课题. 知识图谱不仅是实现通用领域和专业领域精准知识服务的知识支撑, 而且是信息检索、问答交互和信息推荐等领域取得突破性进展的必要基础. 知识图谱的低质量和小规模是阻碍知识图谱广泛应用的主要瓶颈. 知识图谱补全的目的是构建大规模高质量的知识图谱, 以不断更新和扩充知识图谱. 针对现有知识图谱补全方法难以从非结构化文本等辅助信息中提取深层次语义特征的问题, 本文提出一种基于解析图嵌入和加权图卷积网络的知识图谱补全方法. 一方面, 该方法通过加权图卷积网络, 对实体描述文本的语义依存分析进行建模, 构建语义依存解析图嵌入; 另一方面, 引入了实体描述文本的多粒度句嵌入生成方法, 旨在于构建能够捕获多粒度语义、深层次语义特征的实体表示学习. 通过在两个公开数据集上的实验结果表明了本文知识图谱补全方法优于现有方法, 验证了本文方法的有效性和优越性.

关键词 知识图谱补全, 解析图嵌入, 加权图卷积网络, 语义依存分析, 实体表示学习

1 引言

在大互联网时代, 传统互联网和移动互联网中个体用户和群体用户构成大规模数据和信息的生产者、传播者和接受者. 信息载体的多样性和知识资源的繁杂性和异构性已经成为阻碍精准知识服务的主要瓶颈. 知识图谱是实现人工智能的必不可少的知识基础, 也是实现通用领域和专业领域精准知识服务的必要支撑. 知识图谱是信息检索、信息推荐、自然语言处理等领域取得突破性进展的重要基础 [1~7].

近年来, 知识图谱在学术界和工业界受到了广泛和深入的关注和研究. 目前, 主要的知识图谱包括 Cyc、DBpedia、Freebase、NELL、Wikidata, 以及百度知识图谱等 [5~7]. 例如, 应用知识图谱的系统包括谷歌第二代搜索系统、百度知心搜索系统 [8] 等信息检索系统, 以及苹果 Siri、微软小冰等人工

引用格式: 罗妹秋, 张春霞, 彭成, 等. 基于解析图嵌入和加权图卷积网络的知识图谱补全. 中国科学: 信息科学, 2022, 52: 2037–2057, doi: 10.1360/SSI-2021-0217
Luo M Q, Zhang C X, Peng C, et al. Knowledge graph completion based on parsing graph embedding and a weighted graph convolutional network (in Chinese). Sci Sin Inform, 2022, 52: 2037–2057, doi: 10.1360/SSI-2021-0217

智能机器人. 然而, 在 Freebase 中, 有 71% 的人缺少出生地信息, 75% 的人缺少国籍信息 [9,10]. 一方面, 现有知识图谱需要补全缺失的结构化知识; 另一方面, 不断增长的新知识层出不穷, 需要补全技术来更新和扩充知识图谱. 知识图谱补全是指根据知识图谱中已有的实体和关系, 预测知识图谱中缺失的三元组 (实体、关系、实体), 以提高知识图谱的质量和规模, 进而提高利用知识图谱的众多任务, 包括信息检索、问答系统和意见挖掘等的性能 [9,11,12]. 例如, 对于问答系统, 基于通用领域知识图谱, 可以为用户提供高质量的常见问题解答服务. 基于领域知识图谱, 可以为用户提供更加精准的专业知识解答服务. 引入知识图谱补全的问答系统, 不仅能够利用不断扩充的新知识来进一步拓展回答问题的类型和规模, 而且能够扩大答案覆盖的知识范围和提供更加细粒度的答案.

目前知识图谱补全方法主要包括基于翻译的方法、基于深度神经网络的方法、基于关系路径的方法, 以及基于嵌入辅助信息的方法. 基于翻译的知识图谱补全方法的主要模型包括 TrasnE [13], TransH [14], TransD [15], TransG [16], TranSparse [17], 以及 TransA [18] 等模型. 基于深度神经网络的知识图谱补全模型主要包括 ConvE [19], ConvKB [20] 等卷积神经网络模型, TransGate [21], SENN (shared embedding based neural network) [22] 等循环神经网络模型, 以及 R-GCN (relational graph convolutional network) [23], SACN (structure-aware convolutional network) [24] 等图神经网络模型. 基于关系路径的知识图谱方法的主要模型包括路径排序方法 PRA (path ranking algorithm) [25], PTransE [26] 等模型. 基于辅助信息嵌入的知识图谱补全模型主要包括 NTN (neural tensor network) [27], DKRL (description-embodied knowledge representation learning) [28], ConMask [29], KG-BERT (bidirectional encoder representations from transformers) [30], 以及 MTL-KGC (multi-task learning for knowledge graph completion) [31] 等.

知识图谱补全任务面临的主要困难与挑战包括: 其一, 实体和关系的长尾数据问题. 知识图谱的三元组中很多实体和关系出现频率较高, 即包含这类实体和关系的三元组往往较多. 另外, 存在实体和关系出现频率相对较低, 即包含这类实体和关系的三元组往往较少, 然而这类实体和关系的数量通常较多. 如何提取低频实体和低频关系的特征, 如何获取低频实体和低频关系的三元组知识, 是知识图谱补全任务面临的挑战之一 [32]. 其二, 难以从非结构化文本等辅助信息中捕获深层次语义特征. 现有的知识图谱补全方法主要利用知识图谱的实体结点和关系边的图结构信息, 以及结构化的三元组知识等对知识图谱的实体和关系进行建模. 另外, 已有研究工作利用辅助信息和知识来增强实体和关系的特征, 例如实体类型、实体属性, 以及实体描述文本等 [27~31]. 然而, 这些方法主要根据实体名称、关系名称或实体描述文本的词嵌入进行浅层的特征提取, 难以挖掘和区分不同三元组中同一实体和关系的语义特征.

为此, 针对上述困难和挑战, 本文提出一种基于解析图嵌入和加权图卷积网络的知识图谱补全方法. 解析图是指语义依存解析图, 根据非结构化的实体描述文本的语义依存分析生成. 该解析图是由句子中词语之间、短语之间, 以及词语和短语之间语义层面的依存关系构成的有向无环图. 本文提出的知识图谱补全方法的基本思想是: 第一, 采用预训练模型 BERT 和卷积神经网络生成实体描述文本的多粒度句嵌入. 第二, 利用加权图卷积网络, 根据语义依存解析图生成解析图嵌入. 第三, 融合多粒度句嵌入、解析图嵌入、实体嵌入生成混合实体嵌入. 第四, 根据混合实体嵌入和关系嵌入, 通过 Conv-TransE 模型评估三元组成立的概率, 进而获取缺失的三元组集.

本文的主要创新点和贡献具体如下:

(1) 本文提出了一种基于解析图嵌入和加权图卷积网络的知识图谱补全模型框架. 该框架包括实体描述预处理模块、多粒度句嵌入生成模块、解析图嵌入生成模块, 以及知识图谱补全模块. 该框架的特点是: 这 4 个模块具有强独立性和低耦合性. 特别地, 该框架不仅能够动态更新任一模块内部的

处理方法, 而且能够动态地增加或去除进行实体表示学习的模块, 例如多粒度句嵌入生成模块、解析图嵌入生成模块.

(2) 本文通过引入基于语义依存分析的解析图嵌入实现深层次语义的实体建模. 相对于浅层次语义的语义角色标注, 语义依存分析是对自然语言文本的深层次语义分析, 其结果包括句子中语言单位即词语或短语的语义角色, 以及语言单位在语义层面上的依存关系. 基于实体描述文本, 利用加权图卷积网络生成语义依存分析的解析图嵌入对实体建模, 捕获实体描述的深层次语义特征, 有助于解决从非结构化文本的实体描述辅助信息中挖掘深层次语义特征的问题, 从而改进知识图谱的实体表示学习方法, 提高了知识图谱补全的性能.

(3) 本文构建了实体描述文本的多粒度句嵌入, 旨在于实现多粒度语义的实体建模. 现有方法基于预训练模型 BERT 构建句子嵌入时, 通常采用 BERT 最后一层的分类器标记嵌入作为句子嵌入. 然而, 这种方法往往导致应用 BERT 的下游任务的性能偏差较大. 因此, 本文首先将 BERT 中 12 层的分类器标记嵌入进行拼接, 然后通过卷积神经网络生成实体描述文本的多粒度句嵌入, 进而提取 BERT 中 12 种粒度的语义特征.

(4) 本文在两个公开数据集 UMLS 和 FB15k-237 上进行知识图谱补全实验. 实验结果表明本文提出的知识图谱补全方法优于 Conv-TransE, MTL-KGC 和 KG-BERT 等现有方法, 由此表明了本文所提出方法的有效性. 本文研制的基于实体描述、解析图嵌入和多粒度句嵌入的实体表示学习方法, 能够应用于实体识别、关系抽取和信息推荐等领域.

2 国内外相关研究现状

知识图谱补全任务包括头实体预测、尾实体预测、关系预测、链接预测, 以及三元组分类 [9,11,12]. 目前知识图谱补全方法主要包括基于翻译的方法、基于深度神经网络的方法、基于关系路径的方法, 以及基于嵌入辅助信息的方法 [12,32].

基于翻译的知识图谱补全方法的核心思想是对于三元组 (头实体 h, 关系 r, 尾实体 t), 以及头实体、尾实体和关系的向量嵌入, 满足约束关系: 当三元组 (h,r,t) 成立时, 满足 $h+r \approx t$; 否则, $h+r$ 与 t 的距离尽可能增大 [13]. Bordes 等 [13] 提出基于翻译的知识表示学习模型 TrasnE, 将知识图谱中实体和关系都投影在同一个低维向量空间中. 将关系建模为在实体的低维嵌入上进行的翻译操作. 后来, Wang 等 [14] 提出 TransH 模型, 该模型旨在于解决一对多、多对一, 以及多对多的关系. 将实体映射到一个连续向量空间上, 将关系建模为在超平面上的翻译操作. 另外, 在 TransE 的基础上, 研究者构建了 TransD [15], TransG [16], TranSparse [17], TransA [18], HolE [33], DistMult [34], 以及 ANALOGY [35] 等模型.

基于深度神经网络的知识图谱补全方法采用的模型主要包括卷积神经网络 [19,20,36]、循环神经网络 [21,22,37], 以及图神经网络等 [23,24,38]. Dettmers 等 [19] 提出 ConvE 模型, 该模型利用 2D 卷积神经网络来预测知识图谱中缺失的链接关系. 它包括卷积层、投影层, 以及内积层. 另外, Yuan 等 [21] 提出了基于 TransGate 模型的知识图谱嵌入方法. 该模型基于长短期记忆网络的门结构 [39], 设计了共享区分机制, 引入权重向量来重建门结构.

近年来, 随着图卷积网络 [40]、图注意力网络 [41] 等图神经网络模型的发展, 基于图神经网络模型的知识图谱补全受到了越来越多的关注. Nathani 等 [38] 针对知识图谱中的关系预测问题, 采用基于图注意力机制的嵌入表示, 改进了图注意力机制以捕捉实体的多跳邻居中的实体和关系特征.

基于关系路径的知识图谱补全方法基本思想是, 将预测缺失三元组问题转换为对由实体结点和

关系边构成的图上关系路径的搜索和排序问题 [25, 26, 42]. 该方法的特点是能够解决大规模复杂关系路径的建模问题 [25, 26, 42]. 例如, Lao 等 [25] 提出一种路径约束随机游走的加权组合学习方法. 另外, 在 TransE 模型的基础上, Lin 等 [26] 提出 PTransE 模型, 引入路径约束资源分配方法来度量关系路径的可靠性, 通过关系嵌入的相加、相乘和循环神经网络 3 种语义组合来对关系路径建模.

对于知识图谱补全任务, 利用辅助信息嵌入的目的是为实体和关系建模提供更加丰富和具有上下文语境的知识表示特征. 目前利用的辅助信息包括实体属性、文本描述和类型约束等 [43~45]. 例如, 实体的非结构化文本描述通常包括知识图谱中的实体名称、实体类型、实体描述等. 实体名称是指表达实体的词语, 实体类型为实体所属的概念层次的语义类别. 又如, 实体 “The Last King of Scotland” 的实体类型为 “film”. 基于辅助信息嵌入的知识图谱补全方法的主要模型包括 NTN [26], DKRL [28], SSP (semantic space projection) [46], ConMask [29], KG-BERT [30], 以及 MTL-KGC [31] 等. 例如, Yao 等 [30] 研制一种预训练语言模型 KG-BERT, 其思想是将上下文表示引入到预训练模型 BERT 中, 从而实现知识图谱补全. 三元组中的头实体和尾实体的输入可以为实体名称、或描述实体的句子. Kim 等 [31] 指出 KG-BERT 模型存在如下两个问题, 即难以充分学习知识图谱中的关系信息, 以及难以从词汇层面相似的候选实体中识别正确的实体. 为此, 提出一种多任务学习模型 MTL-KGC, 通过在三元组分类、链接预测和关系预测任务的实验结果表明了 MTL-KGC 的有效性.

3 知识图谱补全方法

知识图谱补全的目的是发现知识图谱中缺失的三元组, 补充或更新到现有知识图谱中. 根据 Shi 等 [29] 的工作, 首先给出知识图谱补全任务的定义.

定义1 (知识图谱补全任务) 给定一个不完整知识图谱 $G = (E, R, T)$, 其中 E 表示实体集合, R 表示关系集合, T 表示三元组集合. 知识图谱补全任务是指根据知识图谱 G, 获取缺失的三元组集合 $T' = \{(h, r, t)|h \in E, r \in R, t \in E, (h, r, t) \notin T\}$, 其中 h 表示三元组中的头实体, r 表示关系, t 表示尾实体.

图 1 给出了本文提出的基于解析图嵌入和加权图卷积神经网络 (parsing graph embedding and weighted graph convolutional network, PGE-WGCN) 的知识图谱补全模型整体框架. PGE-WGCN 模型包括实体描述预处理模块、多粒度句嵌入生成模块、解析图嵌入生成模块, 以及知识图谱补全模块. PGE-WGCN 的输入为三元组集、实体集、关系集和实体描述集, 输出为缺失的三元组集合. 其中, 每个实体的实体描述为非结构化文本. 实体描述预处理模块包括构建实体词典 entityDict、关系词典 relationDict 和三元组词典 tripleDict, 以及对实体描述文本进行数据清洗. 实体词典、关系词典和三元组词典分别由数据集中的实体、关系和三元组构成. 数据清洗是指将实体描述文本进行分词, 同时删除乱码字符.

多粒度句嵌入生成模块采用预训练模型 BERT 和卷积神经网络, 生成实体描述文本的多粒度句嵌入. 解析图嵌入生成模块中, 根据实体描述文本的语义依存解析图, 利用加权图卷积网络生成语义依存解析图嵌入. 知识图谱补全模块中, 首先, 融合多粒度句嵌入、语义依存解析图嵌入, 以及实体嵌入, 构建混合实体嵌入. 然后, 通过卷积神经网络对混合实体嵌入和关系嵌入进行卷积, 生成组合嵌入. 最后, 将组合嵌入与实体嵌入矩阵点乘, 通过 Sigmoid 函数计算候选实体作为头实体或尾实体的概率, 并根据概率选择缺失的三元组集合.

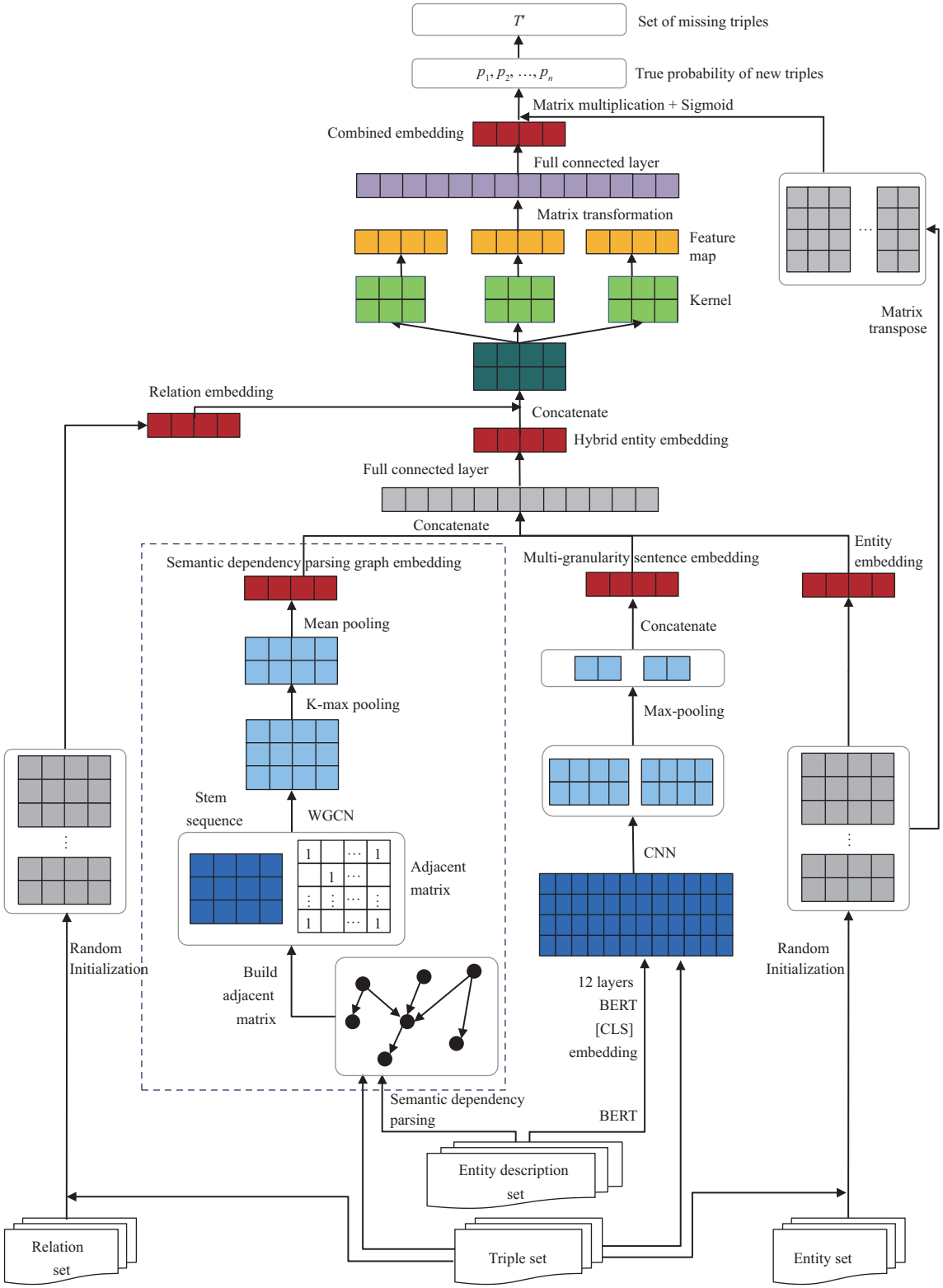

图 1 (网络版彩图) 基于解析图嵌入和加权图卷积神经网络的知识图谱补全模型图

Figure 1 (Color online) Knowledge graph completion model based on parsing graph embedding and a weighted graph convolutional network

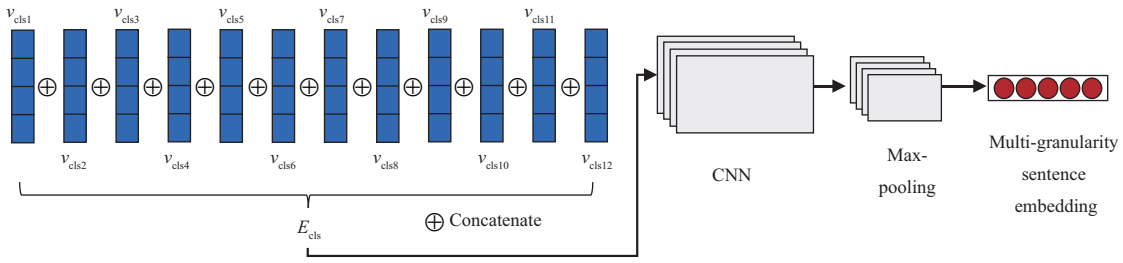

图 2 (网络版彩图) 基于 BERT12 层分类器标记的句嵌入生成模型图

Figure 2 (Color online) Sentence embedding generation model based on the classifier tags of 12 layers of BERT

3.1 多粒度句嵌入生成模块: 基于实体描述的实体表示学习

对于实体 e, e 的实体描述文本 d, 多粒度句嵌入生成模块用于生成实体描述文本 d 的多粒度句嵌入. 每个实体的实体描述文本蕴含丰富的上下文语义信息. 该模块的目的是利用非结构化文本形式的实体描述, 训练生成实体描述文本的多粒度句嵌入来对实体建模, 因此也称为基于实体描述的实体表示学习. 现有的文本句嵌入生成方法是, 对句子中所有词语的词嵌入进行平均池化、最大池化、加权求和等方式生成句嵌入. 然而, 这些方法难以有效捕获句子的上下文信息.

目前, 利用预训练模型 BERT 生成句嵌入时, 通常方法是直接采用最后一层的分类器标记 [CLS] 嵌入. 然而, BERT 中不同 Transformer 层编码的语义信息粒度不同, 不同应用任务关注的语义信息粒度也不同. 因此, BERT 在下游应用任务中性能偏差较大, 需要根据不同的任务来选择不同层的嵌入 [47,48]. 由此, 本文提出一种基于卷积神经网络的 BERT 多粒度语义融合的句嵌入生成方法. 在多粒度句嵌入生成方法中, 受知识图谱表示学习方法 DKRL 的启发 [28], 利用卷积神经网络对实体描述建模. 本文与 DKRL 方法的不同点是: DKRL 方法是基于实体描述的词嵌入, 利用卷积神经网络生成实体描述的句嵌入; 本文的多粒度句嵌入生成方法是基于 BERT 生成句嵌入, 然后采用卷积神经网络生成实体描述的包含多粒度语义的句嵌入. 多粒度句嵌入生成的主要步骤包括: 首先, 通过预训练模型 BERT 获取实体描述的所有 12 层分类器标记 [CLS] 嵌入. 图 2 给出了利用预训练模型 BERT 按层逐级输出这 12 层分类器标记嵌入的网络结构. 然后, 对 12 层分类器标记 [CLS] 嵌入, 按层级顺序拼接成一个新的特征 (记为 E_{cls}), 并将其作为后接卷积神经网络 CNN (convolutional neural network) 的输入. 由于每层分类器标记捕捉每一种粒度的语义信息, 因此利用 12 层分类器标记则可捕捉 12 种粒度的语义信息. 最后, 通过最大池化生成实体描述的多粒度句嵌入.

(1) 基于预训练模型 BERT 的实体描述建模. 对于实体 e 的实体描述文本 $d=(w_1,w_2,\ldots,w_n)$, 获取 BERT 中所有 12 层的分类器标记嵌入并进行拼接, 生成句嵌入 $E_{\text{cls}}=[v_{\text{cls}1};v_{\text{cls}2};\ldots;v_{\text{cls}12}]\in\mathbb{R}^{12\times 768}$, 如图 2 所示. 其中, n 为实体描述的长度, w_i $(i=1,2,\ldots,n)$ 为实体描述中的词语, $v_{\text{cls}j}$ $(j=1,2,\ldots,12)$ 为 BERT 的第 j 层分类器标记 [CLS] 嵌入, 符号 ";" 表示拼接操作. BERT 的层数为 12、隐藏尺寸为 768、自注意力头数为 12. 实体描述的最大长度设置为 512, 即实体描述包含的词语数最大为 512. 对于长度大于 512 的实体描述进行截断处理, 长度小于 512 词语的实体描述采取补零操作. 另外, Sahoo 等 [49] 研制了一种在线深度学习框架和 HBP (Herge backpropagation) 方法用于从训练数据序列中学习深度神经网络的自适应深度. Yang 等 [50] 提出了一种增量自适应深度模型 (incremental adaptive deep model), 能够从流数据中学习具有自适应深度的网络模型, 并实现容量可扩展性. 本文中实体描述的多粒度句嵌入生成方法的特点是利用卷积神经网络对句嵌入编码, 以及易于实现.

(2) 基于卷积神经网络生成实体描述的多粒度句嵌入. 本文采用卷积神经网络对句嵌入进行特征抽取 [51,52]. 卷积神经网络能够捕获 12 层卷积中每层隐含的复杂语义关系, 从而利用 12 种不同粒度的语义信息. 首先, 对 E_{cls} 进行卷积, 生成特征图 $v_{\text{bertCnn}} = (v_1, v_2, \ldots, v_{\text{ch}}) \in \mathbb{R}^{\text{ch}}$, 其中 ch 为卷积核个数. 其次, 通过 BatchNorm 方法对 v_{bertCnn} 进行批归一化, 如式 (1) 所示, 其中 γ 和 β 为可学习的参数, ϵ 为平均移动的动量, $\text{mean}(v_{\text{bertCnn}})$ 为 v_{bertCnn} 的均值, $\sqrt{\text{Var}(v_{\text{bertCnn}})}$ 为 v_{bertCnn} 的方差. 然后, 利用激活函数 ReLU 对 v_{bertCnn} 进行非线性转换获得特征图, 如式 (2) 所示.

$$v_{\text{bertCnn}} = \gamma \frac{v_{\text{bertCnn}} - \text{mean}(v_{\text{bertCnn}})}{\sqrt{\text{Var}(v_{\text{bertCnn}}) + \epsilon}} + \beta, \tag{1}$$

$$v_{\text{bertCnn}} = \max(0, v_{\text{bertCnn}}). \tag{2}$$

最后, 对特征图进行自适应最大池化, 生成实体描述的多粒度句子嵌入 $v_{\text{desc}} \in \mathbb{R}^{\text{EmbSize}}$, 其中 EmbSize 为嵌入维度.

3.2 解析图嵌入生成模块: 基于解析图的实体表示学习

现有基于辅助信息嵌入的知识图谱补全方法中, 难以从非结构化的实体描述文本中挖掘深层次的语义特征 [27~31]. 为了捕获实体描述文本中深层次的语义特征来对实体建模, 本文采用一种基于加权图卷积网络和语义依存分析的语义依存解析图嵌入生成方法. 语义依存解析图是指对实体描述文本进行语义依存分析获得的解析图, 实现了从非结构化文本到结构化图的转换. 语义依存分析剖析如下类型关系, 包括谓词与谓词、谓词与论元、论元与论元, 以及论元构成成分之间的语义关系 [53]. 另外, 加权图卷积神经网络 WGCN [24] 的特点是, 将包含多种关系的知识图谱分解为多个子图, 每个子图对应一种关系. WGCN 为表示不同类型关系的边训练一个可自学习的参数权重, 根据关系边来为邻居节点赋予不同权重. 语义依存解析图嵌入的构建方法包括: 首先, 对实体描述文本进行语义依存分析. 然后, 基于语义依存分析结果, 采用加权图卷积网络生成语义依存解析图嵌入, 从而挖掘实体描述文本的深层次语义特征.

(1) 语义依存分析. 在自然语言处理领域, 句子分析主要包括句法层面和语义层面. 句子的句法信息通常利用句法依存分析工具获得. 句子的语义信息主要通过语义角色标注工具、或语义依存分析工具获得. 句法依存分析以句子为处理对象, 以句子的谓词为中心, 分析句子中语言单位即词语之间、短语之间, 以及词语和短语之间的句法层面的结构关系, 例如, 主谓关系、动宾关系、定中关系和动补关系等.

语义依存分析是一种深层次的语义分析, 分析句子中语言单位即词语和短语之间语义层面的依存关系. 语义依存分析结果包括句子中语言单位的语义角色, 以及语义依存关系 [53]. 例如, 语义角色包括施事者、受事者、方式角色, 以及空间角色等. 语义依存关系包括当事关系、施事关系、客事关系, 以及领事关系等. 特别地, 句法依存分析与语义依存分析的区别具体如下: (a) 前者关注句子句法层面的结构关系, 后者强调句子语义层面的结构关系; (b) 对于同一种含义的不同表达方式的句子集合 S, S 中句子往往呈现不同的句法结构. 然而, S 中不同句子的同一语言单位的语义特征或语义角色是相同的. 因此, 语义依存分析突破了句法结构的约束, 分析谓词与谓词、论元与论元、论元构成成分与论元构成成分之间的关系, 以及谓词与论元之间的关系.

(2) 语义依存解析图嵌入生成. 解析图嵌入生成模块用于生成实体描述文本的语义依存解析图嵌入. 语义依存解析图嵌入生成主要包括如下步骤: 语义依存分析、语义依存解析图嵌入生成.

本文利用 HanLP [54] 工具对实体描述文本中的句子进行语义依存分析. 给定实体 e 及其实体描述 $d = (w_1, w_2, \ldots, w_n)$, 通过语义依存分析工具获得词干序列 $\text{stem} = (w'_1, w'_2, \ldots, w'_n)$, 语义依存边序

列 semantic = $[[h_{\mathrm{se}_1}, t_{\mathrm{se}_1}, \mathrm{se}_1], \ldots, [h_{\mathrm{se}_{\mathrm{ne}}}, t_{\mathrm{se}_{\mathrm{ne}}}, \mathrm{se}_{\mathrm{ne}}]]$, 其中, n 为实体描述的长度, ne 为语义依存边的个数. 进一步, 构建词干词典 wordDict、语义依存标签字典 depDict, 以及语义依存邻接矩阵 A. 例如, 对于 “实体/m/07l450 (The Last King of Scotland)”, 其实体描述为 “The Last King of Scotland is a 2006 British drama film ...”, 通过 HanLP 工具获得词干序列 “(the, Last, King, of, Scotland, be, a, 2006, british, drama, film, ...)”, 语义依存边序列 “[[1, 2, orphan], [1, 3, BV], [2, 3, compound], [4, 3, ARG1], [6, 3, ARG1], [1, 4, orphan],...]”.

通过 Glove 对词干词典 wordDict 进行初始化, 采用随机初始化方式对 depDict 进行初始化, 生成实体描述 d 的词嵌入矩阵 $V_{\mathrm{word}} \in \mathbb{R}^{n\times \mathrm{dim}}$、可学习的语义依存边权重矩阵 $W_{\mathrm{dep}} \in \mathbb{R}^{n_{\mathrm{dep}}\times \mathrm{dim}}$. 并且使用 xavier normal 使得嵌入服从正态分布. 其中, n 为实体描述词的个数, n_{dep} 为语义依存标签的个数, dim 为嵌入维度. 因为加权图卷积神经网络每次训练需要使用图中的所有节点和边, 因此训练过程将实体描述 d 的词嵌入矩阵 V_{word}、语义依存邻接矩阵 A、语义依存边权重矩阵 W_{dep} 输入加权图卷积神经网络, 获取基于解析图邻域的词嵌入矩阵 V_{wgcn}.

加权图卷积神经网络 [24] 分为聚合邻域特征、更新当前节点两个步骤. 在第 1 步, 首先根据语义依存边权重矩阵 W_{dep}, 获取语义依存邻接矩阵 A 中依存边的权重, 生成注意力权重矩阵 A_α. 然后, 对实体描述 d 的词嵌入矩阵 V_{word} 进行维度转换生成 V'_{word}, 利用注意力权重矩阵 A_α 对 V'_{word} 加权求和生成聚合邻域特征的词嵌入矩阵 V_{wgcn}, 如式 (3) 所示. 其中, W 为进行维度转换的权重矩阵.

$$V_{\mathrm{wgcn}} = A_\alpha(V_{\mathrm{word}}W) = A_\alpha V'_{\mathrm{word}}. \tag{3}$$

第 2 步是更新当前节点. 首先利用 BatchNorm 方法对聚合邻域特征的词嵌入矩阵 V_{wgcn} 进行批归一化, 然后通过 tanh 函数进行非线性转换, 生成基于语义依存解析图的词嵌入矩阵 V_{wgcn}, 如式 (4) 所示. 需要指出的是, V_{wgcn} 给出了词典 wordDict 中每个词语的基于语义依存解析图的词嵌入.

$$V_{\mathrm{wgcn}} = \tanh(\mathrm{BatchNorm}(V_{\mathrm{wgcn}})). \tag{4}$$

然后, 对于头实体 h, 以及 h 实体描述中的所有词语, 在词嵌入矩阵 V_{wgcn} 中查找这些词语对应的词嵌入, 进而构建头实体的实体描述的词嵌入矩阵 $v_{\mathrm{dep}} \in \mathbb{R}^{n\times \mathrm{dim}}$, 如式 (5) 所示. 其中 n 为实体描述的长度, dim 为嵌入维度.

$$v_{\mathrm{dep}} = [v_{\mathrm{wgcn}_1}, v_{\mathrm{wgcn}_2}, \ldots, v_{\mathrm{wgcn}_n}]. \tag{5}$$

最后, 对实体描述的所有词嵌入 v_{dep} 先进行 k 最大池化, 再进行均值池化, 生成实体描述的语义依存解析图嵌入 v_{dep}. 与均值池化方法相比, 该方法能够保留显著的特征值. 与最大池化方法相比, 该方法能够在一定程度上避免信息损失.

3.3 知识图谱补全

对于实体 e 及其实体描述 d, 首先, 根据实体描述 d 生成多粒度句嵌入 v_{desc}、语义依存解析图嵌入 v_{dep}. 然后, 将 v_{desc}, v_{dep} 以及实体嵌入 v_{entity} 拼接后输入线性层, 生成混合实体嵌入. 最后, 根据混合实体嵌入和关系嵌入, 通过 Conv-TransE 模型 [24] 评估三元组成立的概率, 进而获取缺失的三元组集合.

(1) 生成混合实体嵌入. 对实体词典 entityDict、关系词典 relationDict 进行随机初始化, 生成实体嵌入矩阵 $V_{\mathrm{entity}} \in \mathbb{R}^{n_{\mathrm{entity}}\times \mathrm{dim}}$、关系嵌入矩阵 $V_{\mathrm{realtion}} \in \mathbb{R}^{n_{\mathrm{realtion}}\times \mathrm{dim}}$, 并且使用 Xavier Normal [55] 使得嵌入服从正态分布. 其中 n_{entity} 为实体个数, n_{realtion} 为关系个数. 首先, 对于头实体 h, 通过索引的

方式, 从实体嵌入矩阵 V_{entity} 中获取 h 的初始实体嵌入 v_{entity}. 对于关系 r, 通过索引的方式, 从关系嵌入矩阵 V_{relation} 中获取 r 的关系嵌入 v_{relation}. 其次, 将 h 初始实体嵌入 v_{entity}, h 实体描述的多粒度句嵌入 v_{desc}, 及其语义依存解析图嵌入 v_{dep} 拼接后通过线性层生成混合实体嵌入 v_h, 如式 (6) 所示, 其中 W_e 和 b_e 分别为线性层的权重矩阵和偏置值.

$$v_h = W_e[v_{\text{desc}}, v_{\text{dep}}, v_{\text{entity}}] + b_e. \tag{6}$$

(2) 预测缺失三元组. 本文采用 Conv-TransE 模型 [24] 预测缺失三元组. 将关系嵌入 v_{realtion} 和混合实体嵌入 v_h 拼接后输入卷积神经网络生成特征图 v_{hr}, 如式 (7) 所示. 其中 channels 是卷积核的个数, ccorr$(*)$ 表示卷积操作, ω 表示卷积核矩阵, b_c 表示偏置.

$$v_{hr} = \sum_{k=0}^{\text{channels}} \text{ccorr}(([v_h, v_{\text{realtion}}]), \omega) + b_c. \tag{7}$$

首先, 采用 BatchNorm 方法对特征图进行归一化, 利用 ReLU 函数对特征图进行非线性转换, 通过全连接层对 v_{hr} 进行维度变换. 其次, 通过 BtahcNorm 对 v_{hr} 进行归一化, 采用 ReLU 函数对特征图进行非线性转换. 然后, 预测缺失的三元组集合, 即已知头实体 h (或尾实体 t) 和关系 r, 预测尾实体 t (或头实体 h). 计算三元组成立的概率 $f(h,r,t)$, 如式 (8) 所示, 其中 V_{entity} 为实体嵌入矩阵.

$$f(h,r,t) = \text{sigmoid}(\text{dot\_product}(v_{hr}, V_{\text{entity}}^{\text{T}})). \tag{8}$$

最后, 将训练样本中三元组 (h,r,t') 的尾实体 t' (或头实体) 转换为 one-hot 向量, 再根据式 (9) 来计算交叉损失值 $l(h,r,t)$. 其中 target 表示训练样本中三元组 (h,r,t') 是否为正确三元组, 若为正确三元组, 则值为 1, 否则为 0.

$$\ell(h,r,t) = -(\text{target} \cdot \log f(h,r,t) + (1 - \text{target}) \cdot \log(1 - f(h,r,t))). \tag{9}$$

本文方法的核心是进行三元组概率预测. 技术上, 三元组概率为零的样本, 即可以认为是负样本. 因此, 本文方法无需执行一个明晰的负样本构建过程. 网络训练的主要驱动力在于最小化损失函数. 本文知识图谱补全算法训练过程如算法 1 所示. 首先, 对实体描述文本进行数据清洗, 并构建实体词典、关系词典和三元组词典. 采用预训练模型 BERT、卷积神经网络 CNN 生成实体描述文本的多粒度句嵌入. 其次, 使用加权图卷积网络 WGCN 对语义依存解析图生成语义依存解析图嵌入. 然后, 对多粒度句嵌入、语义依存解析图嵌入、实体嵌入进行融合, 生成混合实体嵌入. 进一步, 利用卷积神经网络对混合实体嵌入和关系嵌入进行卷积, 生成组合嵌入. 最后, 将组合嵌入与实体嵌入矩阵点乘, 通过 Sigmoid 函数计算候选实体作为尾实体 (头实体) 的概率, 并根据概率构建缺失三元组集合. 计算出损失值之后, 并对参数进行更新. 算法上, 参数用 Xavier 初始化器初始化, 损失函数由 AdaGrad 优化器进行优化.

4 实验结果

本节阐述本文研制的知识图谱补全方法在两个公开数据集 UMLS 和 FB15k-237 上的实验结果比较与分析. 本节给出了与相关研究工作的比较实验、消融实验, 以及参数敏感性分析实验.

UMLS 是一种医学语义网络 [56]. UMLS 数据集包含 135 个实体、46 种关系. 训练集、验证集和测试集分别包括 5126, 652 和 661 个三元组, 如表 1 所示. UMLS 中的实体是高层次概念, 每个实体都含

表 1　数据集统计信息

Table 1　Statistics of datasets

Dataset	Entities	Relations	Train	Validation	Test
UMLS	135	46	5216	652	661
FB15k-237	14541	237	272115	17535	20466

Algorithm 1 Knowledge graph completion method based on parsing graph embedding and a weighted graph convolutional network

Input: Entity set E, relation set R, triple set T, entity description set D.

Output: Missed triples T'.

for e_i in E **do**: entityDict $\leftarrow e_i$; **end for** //Build entity dictionary

for r_i in R **do**: relationDict $\leftarrow r_i$; **end for** //Construct relation dictionary

for triple$_i$ in T **do**: tripleDict $\leftarrow$ triple$_i$; **end for** //Build triple dictionary

for d_i in D **do**:

  desc$_i \leftarrow$ Preprocessing(d_i); //Data cleaning

  $D' \leftarrow$ desc$_i$; //Build input of multi-granularity sentence embedding

  if useSemantic: //Semantic dependency analysis

    wordDict, wordList, depDict, $A \leftarrow$ HanLP(desc$_i$);

  end if

end for

$V_{\text{entity}} \xleftarrow{\text{Embedding}}$ entityDict; //Obtain entity embedding

$V_{\text{relation}} \xleftarrow{\text{Embedding}}$ relationDict; //Build relation embedding

$V_{\text{word}} \xleftarrow{\text{Glove}}$ wordDict; //Construct entity descriptive word embedding

$W_{\text{dep}} \xleftarrow{\text{Embedding}}$ depDict; //Obtain learned weight

for $q = 1, 2, \ldots, Q$ in Epoch **do**

  for i in batch **do**:

    $h_{\text{batch}}, r_{\text{batch}}, t_{\text{batch}} \leftarrow$ entityDict, relationDict, tripleDict;

    $v_{\text{desc}} \xleftarrow{\text{BERT-CNN}} h_{\text{batch}}, D'$; //Construct multi-granularity sentence embedding of head entity

    $V_{\text{wgcn}} \xleftarrow{\text{WGCN}} V_{\text{word}}, A, W_{\text{dep}}$; //Obtain entity descriptive word embeeding based on parsing graph

    $v_{\text{dep}} \leftarrow V_{\text{wgcn}}, h_{\text{batch}}$; //Generate parsing graph embedding of head entity

    $v_{\text{entity}} \leftarrow V_{\text{entity}}, h_{\text{batch}}$; //Obtain entity embedding

    $v_h \xleftarrow{\text{FC}} [v_{\text{desc}}, v_{\text{dep}}, v_{\text{entity}}]$; //Build hybrid entity embedding

    $v_r \leftarrow V_{\text{relation}}, r_{\text{batch}}$; //Obtain relation embedding

    $v_{hr} \xleftarrow{\text{CNN}} [v_h, v_r]$; //Build combined embedding

    $v_{hr} \xleftarrow{\text{FC}} v_{hr}$; //Use the full-connected layer for dimension transformation

    $f(h,r,t) \xleftarrow{\text{Sigmoid}} \text{dot\_product}(v_{hr}, V_{\text{entity}}^{\text{T}})$; //Calculate the probability of the new triple

    loss = BCEloss($t, f(h,r,t)$); //Compute loss

    Minimize loss;

    Update parameters;

  end for

end for

有对应的实体描述. 例如, UMLS 数据集包括实体 “disease or syndrome (疾病或综合症)”、“biologically active substance (生物活性物质)” 等. 实体 “disease or syndrome” 的实体描述为 “Restless legs syndrome (RLS) is generally a long term disorder that causes a strong urge to move one’s legs ...”. 关系是二元谓词, 例如, UMLS 数据集包括关系 “causes”, “affects”. 头实体、尾实体, 以及关系构成三元组. 例如, 三元组 (nucleic acid nucleoside or nucleotide, causes, disease or syndrome), 即 (核酸核苷或核苷酸, 导致,

表 2 数据集 UMLS 上相关工作的对比实验结果

Table 2 Comparsion of experimental results with related studies on the UMLS dataset

Model	Hits@10 (%)	Hits@3 (%)	MR	MRR
TransE [13]	98.9	–	1.84	–
TransH [14]	99.5	–	1.80	–
TransR [59]	99.4	–	1.81	–
TransD [15]	99.3	–	1.71	–
DistMult [34]	84.6	–	5.52	–
ComplEx [60]	96.7	–	2.59	–
Conv-TransE [24]	99.29	95.62	1.4851	0.9089
SACN [24]	99.37	95.70	1.5125	0.8798
KG-BERT [30]	99.0	–	1.47	–
Our method	**99.92**	**98.43**	**1.2429**	**0.9142**

疾病或综合症).

数据集 FB15k-237 是由数据集 FB15k 去除逆序三元组后构建而成 [57], 数据集 FB15k 来源于数据集 Freebase [11, 58]. FB15k-237 以三元组的形式存储, 且每个实体均有对应的实体描述. 例如, 三元组 ($/m/07l450$, /film/film/genre, $/m/082gq$) 中, 实体$/m/07l450$ (the last king of scotland)的实体描述为 "The Last King of Scotland is a 2006 British drama film ...". 数据集 FB15k-237 包括实体 14541 个, 关系类型 237 种. 训练集、验证集和测试集分别包括 272, 115, 17535 和 20466 个三元组, 如表 1 所示.

本文采用在知识图谱补全任务常用的评估指标 Hits@10, Hits@3, MRR (mean reciprocal rank) 和 MR (mean rank) 来评估知识补全方法的性能. 其中, Hits@k 表示在测试集中正确三元组排名在前 k 的概率, 该评估值越高表明知识补全方法的性能越高. 平均倒数排名 (MRR) 表示测试集中正确三元组排名倒数的平均值, 该评估值越高表明方法性能越高. 平均排名 (MR) 表示测试集中正确三元组排名的平均值, 该评估值越低表明方法性能越高.

4.1 相关工作对比实验

表 2 和 3 给出了本文提出的基于解析图嵌入和加权图卷积网络的知识图谱补全方法与其他相关研究工作的对比实验结果. 其他相关工作包括 TransE [13], TransH [14], TransR [59], TransD [15], DistMult [34], ComplEx [60], Conv-TransE [24], KG-BERT [30], SACN [24], DSKG [37], MTL-KGC [31] 和 ParamE-MLP [61] 方法. 知识图谱补全实验的主要参数包括 Conv-TransE 中卷积神经网络的卷积核个数 (number of channels)、Conv-TransE 中卷积神经网络的卷积核大小 (kernel size)、多粒度句嵌入中卷积神经网络的卷积核个数 (number of BERT channels)、多粒度句嵌入中卷积神经网络的卷积核大小 (BERT kernel size)、加权图卷积网络的随机失活率 (WGCN dropout rate)、学习率 (learning rate)、实体和关系的嵌入维度 (embedding size), 以及隐含层随机失活率 (hidden dropout rate). 对于数据集 UMLS 和 FB15k-237, 这些参数分别取值为 (50, 3, 100, 2, 0.5, 0.01, 50, 0.5) 和 (300, 5, 100, 2, 0.2, 0.001, 100, 0.5).

根据表 2 的实验结果可以看出, 本文的基于解析图嵌入和加权图卷积网络的知识图谱补全方法在数据集 UMLS 上 Hits@10, Hits@3, MR 和 MRR 评估值分别为 99.92%, 98.43%, 1.2429 和 0.9142. 相比于 KG-BERT 方法 [30], 本文方法的 Hits@10, MR 分别提高了 0.92% 和 0.2271. 对于 FB15k-237 数

表 3　数据集 FB15k-237 上相关工作的对比实验结果

Table 3　Comparsion of experimental results with related stuides on FB15k-237

Model	Hits@10 (%)	Hits@3 (%)	MR	MRR
TransE [13]	47.4	–	223	–
TransH [14]	48.6	–	255	–
TransR [59]	51.1	–	237	–
TransD [15]	48.4	–	246	–
DistMult [34]	42.0	26.0	–	0.24
ComplEx [60]	43.0	28.0	–	0.25
Conv-TransE [24]	51.0	37.0	–	0.33
KG-BERT [30]	42.0	–	153	–
DSKG [37]	52.1	–	175	0.339
MTL-KGC [31]	45.8	29.8	**132**	0.267
ParamE-MLP [61]	45.9	33.9	–	0.314
Our method	**52.80**	**37.80**	216.70	**0.3441**

据集, 表 3 中本文方法的 Hits@10, Hits@3, MR 和 MRR 分别为 52.80%, 37.80%, 216.70 和 0.3441. 相比于 ParamE-MLP 方法 [61], 本文方法的 Hits@10, Hits@3 和 MRR 分别提高了 6.9%, 3.9% 和 0.0301. 表 2 和 3 的实验结果表明本文方法的性能均优于现有工作, 验证了本文所出方法的有效性. 本文方法能够提高性能的原因在于: 引入语义解析图嵌入实现了深层次语义的实体建模, 引入多粒度句嵌入实现了多粒度语义的实体建模. 另外, 通过基于解析图嵌入和加权图卷积网络的知识图谱补全模型框架实现了模块间的强独立性和低耦合性.

4.2　消融实验

为了评估本文知识图谱补全方法中各模块对性能的影响, 设计了消融实验对比方案, 如表 4 和 5 所示, 其中 "√" 和 "×" 分别表示多粒度句嵌入 "multi-granularity sentence embedding" 和语义依存解析图嵌入 "semantic dependency parsing graph embedding" 是否在当前模型中引入.

(1) Baseline. 选用 Conv-TransE 模型 [24].

(2) Bert-linear. 仅生成句嵌入, 即对 BERT 最后一层的分类器标记 [CLS] 嵌入经过全连接层的结果作为句嵌入. 未引入语义依存解析图嵌入.

(3) Bert-rel-att. 仅生成句嵌入, 通过依赖于关系的注意力机制对 BERT12 层的分类器标记 [CLS] 嵌入进行加权求和的结果作为句嵌入. 未引入语义依存解析图嵌入.

(4) Bert-cnn. 仅生成多粒度句嵌入, 即对 BERT12 层的分类器标记 [CLS] 嵌入经过卷积神经网络的结果作为多粒度句嵌入. 未引入语义依存解析图嵌入.

(5) SRL. 仅生成语义角色嵌入, 即句子的语义角色标注经过双向长短期记忆网络和注意力机制的结果作为语义角色嵌入. 未引入多粒度句嵌入、语义依存解析图嵌入.

(6) SyntGE (syntactic ggraph embedding). 仅生成句法依存解析图嵌入, 即句子的句法依存分析经过加权图卷积网络的结果作为句法依存解析图嵌入. 未引入多粒度句嵌入、语义依存解析图嵌入. 本文通过 StanfordNLP 的 Parse 工具 [62] 对实体描述文本中的句子进行句法依存分析.

(7) SemGE (semantic graph embedding). 生成语义依存解析图嵌入, 未引入多粒度句嵌入. 语义依存解析图嵌入是指句子语义依存分析经过加权图卷积网络的结果作为语义依存解析图嵌入.

表 4 数据集 UMLS 的消融实验结果

Table 4 Ablation experimental results on the UMLS dataset

Model	Multi-granularity sentence embedding	Semantic dependency parsing graph embedding	Hits@10 (%)	Hits@3 (%)	MR
(1) Baseline	×	×	99.29	95.62	1.4851
(2) Bert-linear	×	×	99.84	98.12	1.3796
(3) Bert-rel-att	×	×	99.53	97.26	1.4687
(4) Bert-cnn	√	×	99.68	96.48	1.3257
(5) SRL	×	×	99.68	97.73	1.3000
(6) SyntGE	×	×	99.68	97.26	1.4195
(7) SemGE	×	√	99.76	97.26	1.3375
(8) Bert-cnn+SRL	√	×	99.84	96.40	1.2812
(9) Bert-cnn+SyntGE	√	×	99.84	97.26	1.3125
(10) Bert-mean+SemGE	√	√	99.76	97.81	1.3171
(11) Bert-Multi-att2+SemGE	√	√	99.21	96.01	1.4828
(12) Bert-cnn+SemGE+avg-pooling	√	√	99.76	98.20	1.3234
(13) Bert-cnn+SemGE+max-pooling	√	√	99.84	98.59	1.2757
(14) Bert-cnn+SemGE	√	√	**99.92**	**98.43**	**1.2429**

表 5 数据集 FB15k-237 的消融实验结果

Table 5 Ablation experimental results on the FB15k-237 dataset

Model	Multi-granularity sentence embedding	Semantic dependency parsing graph embedding	Hits@10 (%)	Hits@3 (%)	MR
(1) Baseline	×	×	51.00	37.0	–
(2) Bert-linear	×	×	52.60	**38.09**	228.71
(3) Bert-rel-att	×	×	52.60	37.90	243.25
(4) Bert-cnn	√	×	52.72	37.90	233.43
(5) SRL	×	×	52.35	37.45	212.94
(6) SyntGE	×	×	52.34	37.61	224.22
(7) SemGE	×	√	52.77	37.64	**211.70**
(8) Bert-cnn+SRL	√	×	52.77	37.71	224.53
(9) Bert-cnn+SyntGE	√	×	52.71	37.71	222.90
(10) Bert-mean+SemGE	√	√	52.48	37.72	214.65
(11) Bert-Multi-att2+SemGE	√	√	52.30	37.52	225.64
(12) Bert-cnn+SemGE+avg-pooling	√	√	51.97	37.27	228.90
(13) Bert-cnn+SemGE+max-pooling	√	√	52.55	37.59	222.99
(14) Bert-cnn+SemGE	√	√	**52.80**	37.80	216.70

(8) Bert-cnn+SRL. 生成多粒度句嵌入、语义角色嵌入. 未引入语义依存解析图嵌入.

(9) Bert-cnn+SyntGE. 生成多粒度句嵌入、句法依存解析图嵌入. 未引入语义依存解析图嵌入.

(10) Bert-mean+SemGE. 生成多粒度句嵌入, 对 BERT 不同层的 embedding 进行平均的结果作

为句嵌入, 引入语义依存解析图嵌入.

(11) Bert-Multi-att2+SemGE. 生成多粒度句嵌入, 通过多头注意力机制对 BERT12 层的分类器标记 [CLS] 嵌入进行加权求和的结果作为句嵌入. 引入语义依存解析图嵌入. 注意力头的个数为 2.

(12) Bert-cnn+SemGE+avg-pooling. 生成多粒度嵌入. 引入语义依存解析图嵌入, 即对实体描述的所有词嵌入进行均值池化, 生成实体描述的语义依存解析图嵌入.

(13) Bert-cnn+SemGE+max-pooling. 生成多粒度嵌入. 引入语义依存解析图嵌入, 即对实体描述的所有词嵌入进行最大池化, 生成实体描述的语义依存解析图嵌入.

(14) Bert-cnn+SemGE. 生成多粒度句嵌入、语义依存解析图嵌入.

根据表 4 和 5 消融实验结果可以看出, 模型 Bert-cnn 在数据集 UMLS 上的 Hits@10 为 99.68%, Hits@3 为 96.48% 和 MR 为 1.3257. 在数据集 FB15k-237 上的 Hits@10 为 52.72%, Hits@3 为 37.90% 和 MR 为 233.43. 在 UMLS 数据集上, 模型 Bert-cnn 的 MR 性能高于模型 Baseline, Bert-linear 和 Bert-rel-att. 在 FB15k-237 数据集上, 模型 Bert-cnn 的 Hits@10 实验性能高于这 3 种模型. 因此, 实验结果表明多粒度句嵌入所提取的 12 种粒度的语义特征有助于提高知识图谱补全的性能.

模型 SemGE 在数据集 UMLS 上的 Hits@10 为 99.76%, Hits@3 为 97.26%, MR 为 1.3375. 在数据集 FB15k-237 上的 Hits@10 为 52.77%, Hits@3 为 37.64%, MR 为 211.70. 在数据集 FB15k-237 上该模型的实验性能都高于 SRL 和 SyntGE. 在数据集 UMLS 上, 该模型的 Hits@10 性能高于 SRL 和 SyntGE. 由此, 实验结果表明语义依存解析图嵌入比语义角色嵌入、句法依存解析图嵌入更能够捕获实体描述文本的深层次语义特征.

本文模型 Bert-cnn+SemGE 在数据集 UMLS 上的 Hits@10 为 99.92%, Hits@3 为 98.43% 和 MR 为 1.2429. 在数据集 FB15k-237 上的 Hits@10 为 52.80%, Hits@3 为 37.80% 和 MR 为 216.70. 在数据集 UMLS 上, 本文模型的 Hits@10, Hits@3 和 MR 性能都优于表 4 的其他 13 种模型. 具体地, 与模型 Bert-linear 相比, 本文模型的 Hits@10, Hits@3 和 MR 分别提高了 0.08%, 0.31% 和 0.1367. 本文模型的 Hits@10, Hits@3 和 MR 比 Bet-rel-att 分别提高了 0.39%, 1.17% 和 0.2258. 在数据集 FB15k-237 上, 本文模型的 Hits@10 性能都优于表 5 的其他 13 种模型. 具体地, 与模型 Bert-linear 相比, 本文模型的 Hits@10 和 MR 分别提高了 0.2% 和 12.01. 本文模型的 Hits@10 和 MR 比 Bet-rel-att 分别提高了 0.2% 和 26.55. 由此, 实验结果表明本文提出的基于多粒度句嵌入和语义依存解析图嵌入的知识图谱补全方法的性能, 优于基于多粒度句嵌入和句法依存解析图嵌入的方法 (Bert-cnn+SyntGE)、基于多粒度句嵌入和语义角色嵌入的方法 (Bert-cnn+SRL), Bert-mean+SemGE, Bert-Multi-att2+SemGE, Bert-cnn+SemGE+avg-pooling, 以及 Bert-cnn+SemGE+max-pooling 模型. 同时, 实验结果表明引入语义依存解析图嵌入和多粒度句嵌入对知识图谱补全性能提升的有效性和贡献度.

4.3 参数实验

本文设计了参数敏感性分析实验, 分析本文知识图谱补全方法对实验参数的敏感性.

(1) 参数实验: 卷积核个数 (number of channels). 图 3(a) 和 (b) 给出了 Conv-TransE 中卷积神经网络的不同卷积核个数的实验结果曲线图. 卷积核个数取值为 {50, 100, 200, 300}. 其中, “Bert-cnn+SyntGE” 表示基于多粒度句嵌入、句法依存解析图嵌入的方法, 简称为基于句法依存解析图嵌入的方法. “Bert-cnn+SemGE” 表示本文研制的基于多粒度句嵌入、语义依存解析图嵌入的方法, 简称为基于语义依存解析图嵌入的方法.

在 UMLS 数据集上, 基于句法依存解析图嵌入和语义依存解析图嵌入的 Hits@10 结果随着卷积核个数的增大总体呈下降的趋势. 这两种方法都在卷积核个数为 50 时取得最高值, 即 Hits@10 分别

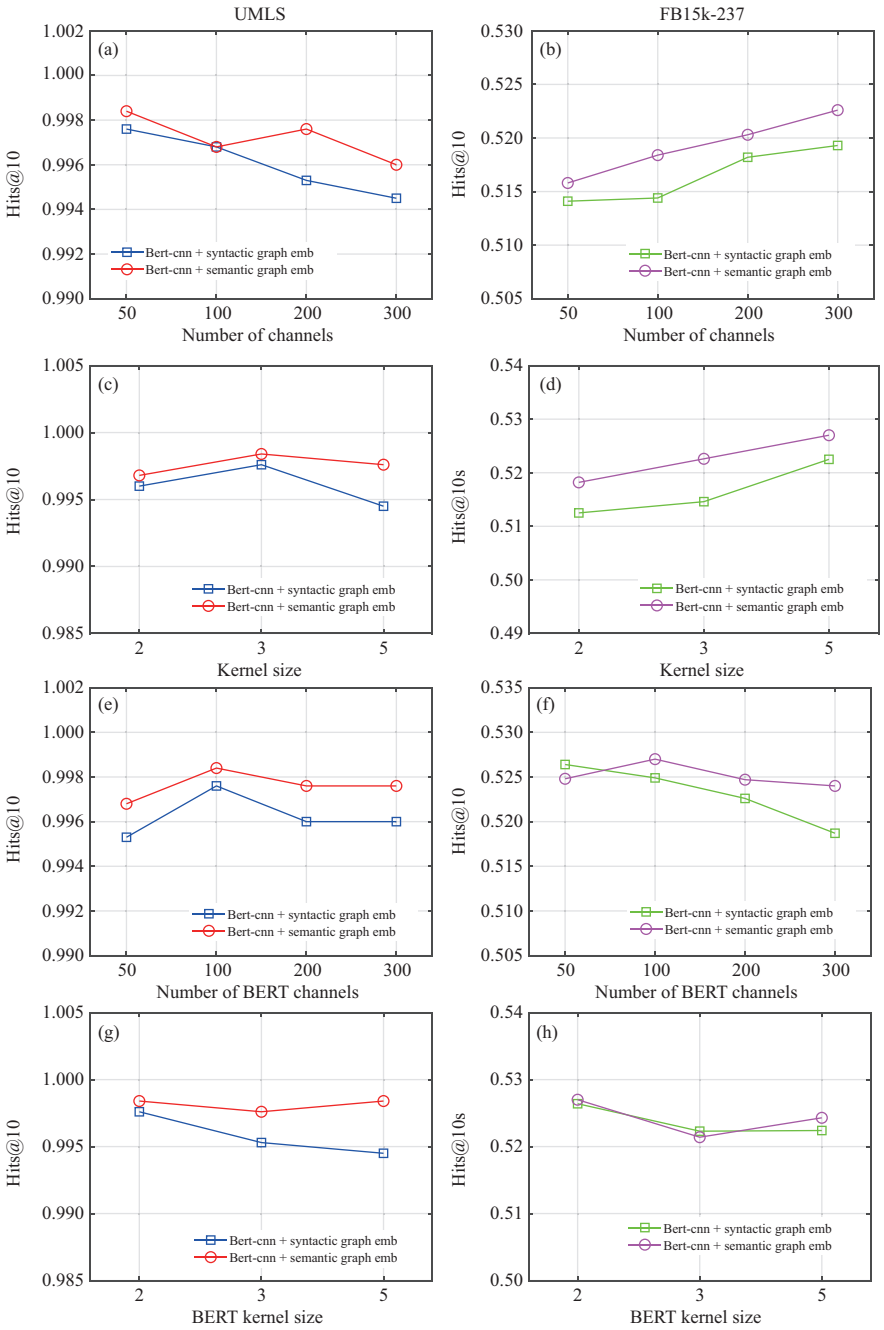

图 3 (网络版彩图) 数据集 **UMLS** 和 **FB15k-237** 上的参数敏感性分析. **(a) (b) Conv-TransE** 卷积核个数效果分析; **(c) (d) Conv-TransE** 卷积核大小效果分析; **(e) (f) BERT** 卷积核个数效果分析; **(g) (h) BERT** 卷积核大小效果分析

Figure 3 (Color online) Parameter sensitivity analyses on datasets of UMLS and FB15k-237. (a) (b) Effect analysis of the number of channels of Conv-TransE; (c) (d) effect analysis of the kernel size of Conv-TransE; (e) (f) effect analysis of the number of BERT channels; (g) (h) effect analysis of the BERT kernel size

为 99.76%, 99.84%, 它们的最高值与最低值的差距分别为 0.31% 和 0.24%. 在 FB15k-237 数据集上, 这两种方法的 Hits@10 结果均随着卷积核个数的增大总体呈上升的趋势, 在卷积核个数为 300 时达到最

高值, 即 Hits@10 分别为 51.93% 和 52.26%, 比最低值增大了 0.52% 和 0.68%.

(2) 参数实验: 卷积核大小 (kernel size). 图 3(c) 和 (d) 给出了 Conv-TransE 中卷积神经网络的不同卷积核大小的实验结果曲线图. 卷积核大小取值为 {2, 3, 5}. 在 UMLS 数据集上, 基于句法依存解析图嵌入、基于语义依存解析图嵌入的 Hits@10 均在卷积核大小为 3 时取得最高值, 即 Hits@10 分别为 99.76%, 99.84%, 与最低值相比分别增大了 0.31%, 0.16%. 在 FB15k-237 数据集上, 这两种方法的 Hits@10 在 kernel size 为 5 时达到最高值, Hits@10 分别为 52.25%, 52.7%, 它们的最高值与最低值的差距分别为 1%, 0.88%.

(3) 参数实验: BERT 卷积核个数 (number of BERT channels). 图 3(e) 和 (f) 给出了 BERT 中不同卷积核个数的实验结果曲线图. 卷积核个数取值为 {50, 100, 200, 300}. 在 UMLS 数据集上, 基于句法依存解析图嵌入、基于语义依存解析图嵌入 Hits@10 均在卷积核个数为 100 时获得最高值, Hits@10 分别为 99.76%, 99.84%, 与最低值相比分别增大了 0.23%, 0.16%. 在 FB15k-237 数据集上, 基于句法依存解析图嵌入的 Hits@10 在卷积核个数为 50 时取得最佳值, 即 Hits@10 为 52.64%, 与最低值的差距为 0.77%. 基于语义依存解析图嵌入的 Hits@10在卷积核个数为 100 时达到最佳值, 即 Hits@10 为 52.7%, 与最低值的差距为 0.3%.

(4) 参数实验: BERT 卷积核大小 (BERT kernel size). 图 3(g) 和 (h) 给出了不同 BERT 卷积核大小的实验结果曲线图. 卷积核大小取值为 {2, 3, 5}. 在 UMLS 数据集上, 基于句法依存解析图嵌入的 Hits@10 在 BERT 卷积核大小为 2 时取得最高值, 即 Hits@10 为 99.76%, 其最高值与最低值的差距为 0.31%. 基于语义依存解析图嵌入的 Hits@10 其最高值与最低值的差距为 0.08%. 在 FB15k-27 数据集上, 基于句法依存解析图嵌入的 Hits@10 在 BERT 卷积核大小为 2 时获得最高值 52.64%, 比最低值增大了 0.41%. 基于语义依存解析图嵌入的 Hits@10 在 BERT 卷积核大小为 2 时达到最佳值, 即 Hits@10 为 52.7%, 其最高值与最低值的差距为 0.56%.

(5) 参数实验: 加权图卷积神经网络随机失活率 (WGCN dropout rate). 对于图卷积神经网络随机失活率, 图 4(a) 和 (b) 给出了不同随机失活率的实验结果曲线图. 该参数取值为 $\{0.0, 0.1, \ldots, 0.7\}$. 在 UMLS 数据集上, 基于句法依存解析图嵌入和基于语义依存解析图嵌入的 Hits@10 的最高值与最低值的差距都为 0.31%. 在 FB15k-237 数据集上, 前一种方法的 Hits@10 最高值与最低值的差距为 0.47%. 后一种方法的最高值比最低值增大了 0.61%.

(6) 参数实验: 学习率 (learning rate). 针对学习率进行实验, 图 4(c) 和 (d) 给出了不同学习率的实验结果曲线图. 学习率取值为 {0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05}. 在 UMLS 数据集上, 基于句法依存解析图嵌入、基于语义依存解析图嵌入的 Hits@10 最高值与最低值的差距分别为 84.38%, 30.7%. 在 FB15k-237 数据集上, 这两种方法的 Hits@10 最高值与最低值的差距分别为 30.21%, 31.64%.

(7) 参数实验: 嵌入维度 (embedding size). 图 4(e) 和 (f) 给出了实体和关系的不同嵌入维度的实验结果曲线图. 嵌入维度取值为 {50, 100, 200, 300}. 在 UMLS 数据集上, 基于句法依存解析图嵌入、基于语义依存解析图嵌入的 Hits@10 最高值与最低值的差距分别为 1.48% 和 0.55%. 在 FB15k-237 数据集上, 这两种方法的 Hits@10 最高值与最低值的差距分别为 1.85% 和 3.77%.

(8) 参数实验: 隐含层随机失活率 (hidden dropout rate). 针对隐含层随机失活率, 图 4(g) 和 (h) 给出了不同随机失活率的实验结果曲线图. 该参数取值为 {0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7}. 在 UMLS 数据集上, 基于句法依存解析图嵌入、基于语义依存解析图嵌入的 Hits@10 最高值与最低值的差距分别为 3.99% 和 2.11%. 在 FB15k-237 数据集上, 这两种方法的 Hits@10 最高值与最低值的差距分别为 6.33% 和 6.11%. 综上分析, 本文的知识图谱补全方法对学习率比较敏感, 对其他参数敏感性微小.

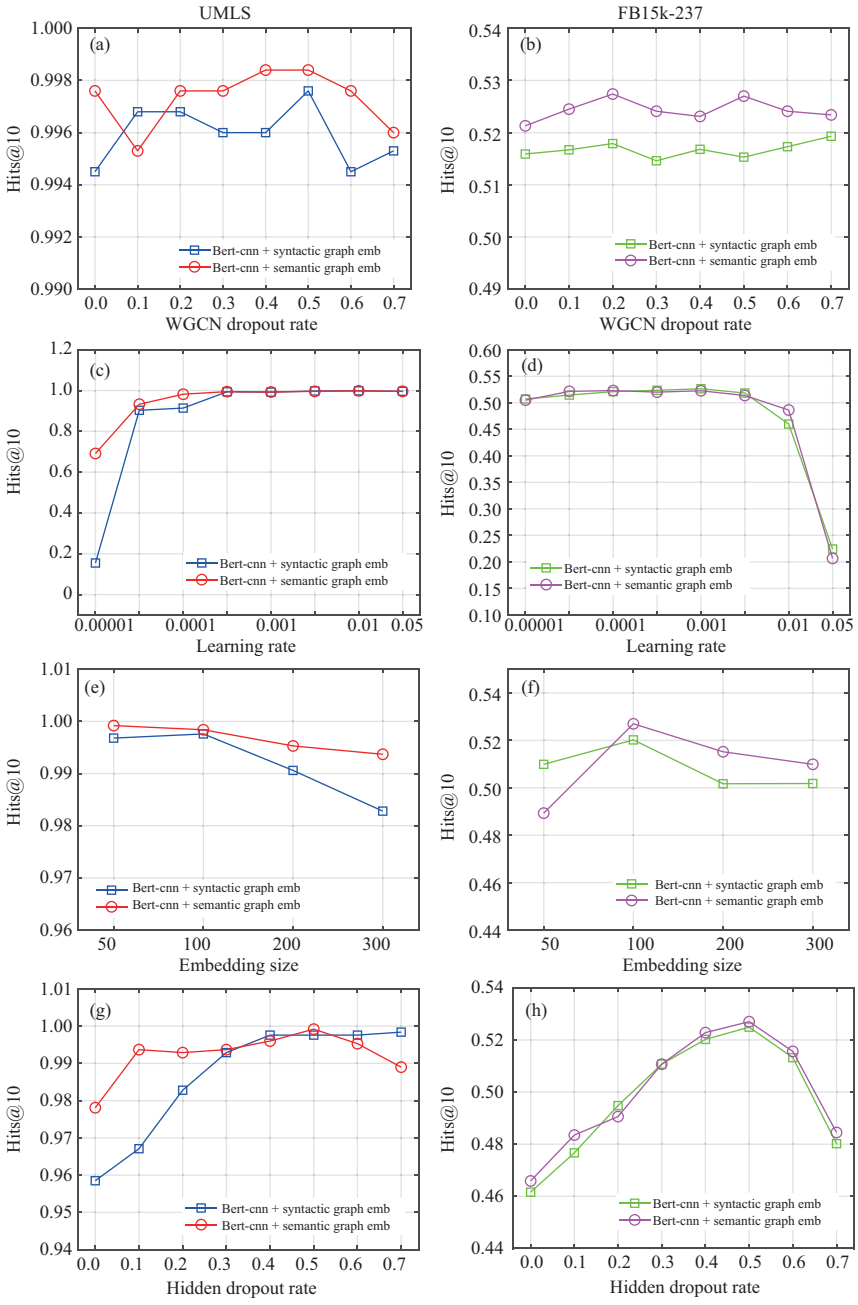

图 4 (网络版彩图) 数据集 UMLS 和 FB15k-237 上的参数敏感性分析. (a) (b) 加权图卷积网络随机失活率效果分析; (c) (d) 学习率效果分析; (e) (f) 嵌入维度效果分析; (g) (h) 隐藏层随机失活率效果分析

Figure 4 (Color online) Parameter sensitivity analyses on datasets of UMLS and FB15k-237. (a) (b) Effect analysis of WGCN dropout rate; (c) (d) effect analysis of learing rate; (e) (f) effect analysis of embedding size; (g) (h) effect analysis of hidden dropout rate

5 结论

知识图谱补全是通用领域知识图谱和专业领域知识图谱能够广泛应用的关键技术之一. 知识图谱

补全不仅能够不断补全现有知识图谱缺失的知识, 而且能够不断扩充新知识. 本文提出了一种基于解析图嵌入和加权图卷积网络的知识图谱补全方法. 该方法的实体表示学习特点是: 第一, 引入语义依存解析图嵌入, 通过加权图卷积网络模型对实体描述文本的语义依存分析进行建模. 第二, 引入基于BERT 的实体描述文本的多粒度句嵌入, 提取 BERT 中 12 种粒度的语义特征. 本文的实体表示学习方法能够捕获实体描述文本的多粒度语义、深层次语义特征. 在两个公开数据集上进行了知识图谱补全实验, 实验结果表明了本文知识图谱补全方法优于现有方法, 验证了本文方法的有效性. 下一步将研究如何利用文本、图片和视频等多模态信息进行知识图谱补全.

参考文献

1 Wang H W, Zhang F Z, Wang J L, et al. Ripplenet: propagating user preferences on the knowledge graph for recommender systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, 2018. 417–426

2 Qin C, Zhu H S, Zhuang F Z, et al. A survey on knowledge graph-based recommender systems. Sci Sin Inform, 2020, 50: 937–956 [秦川, 祝恒书, 庄福振, 等. 基于知识图谱的推荐系统研究综述. 中国科学: 信息科学, 2020, 50: 937–956]

3 Lukovnikov D, Fischer A, Lehmann J, et al. Neural network based question answering over knowledge graphs on word and character level. In: Proceedings of the 26th International Conference on World Wide Web, Perth, 2017. 1211–1220

4 Wang Z Y, Yu Q, Wang N, et al. Survey of intelligent question answering research based on knowledge graph. Comput Eng Appl, 2020, 56: 1–11 [王智悦, 于清, 王楠, 等. 基于知识图谱的智能问答研究综述. 计算机工程与应用, 2020, 56: 1–11]

5 Chen X J, Jia S B, Xiang Y. A review: knowledge reasoning over knowledge graph. Expert Syst Appl, 2020, 141: 112948

6 Zhao J, Liu K, He S Z, et al. Knowledge Graph. Beijing: Higher Education Press, 2018 [赵军, 刘康, 何世柱, 等. 知识图谱. 北京: 高等教育出版社, 2018]

7 Qi G L, Gao H, Wu T X, The research advances of knowledge graph. Technol Intell Eng, 2017, 3: 4–25 [漆桂林, 高桓, 吴天星. 知识图谱研究进展. 情报工程, 2017, 3: 4–25]

8 Wu Y B, Yang F, Lai G H, et al. Research progress of knowledge graph learning and reasoning. J Chinese Comput Syst, 2016, 37: 2007–2013 [吴运兵, 杨帆, 赖国华, 等. 知识图谱学习和推理研究进展. 小型微型计算机系统, 2016, 37: 2007–2013]

9 Ding J H, Jia W J. The research advances of knowledge graph completion algorithm. Inform Commun Technol, 2018, 12: 56–62 [丁建辉, 贾维嘉. 知识图谱补全算法综述. 信息通信技术, 2018, 12: 56–62]

10 Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2014. 601–610

11 Bollacker K D, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Vancouver, 2008. 1247–1250

12 Ji S X, Pan S R, Cambria E, et al. A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst, 2021. doi: 10.1109/TNNLS.2021.3070843

13 Bordes A, Usunier N, García-Durán A, et al. Translating embeddings for modeling multi-relational data. In: Proceedings of the 27th Conference on Neural Information Processing Systems, Lake Tahoe, 2013. 2787–2795

14 Wang Z, Zhang J W, Feng J L, et al. Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, Québec City, 2014. 1112–1119

15 Ji G L, He S Z, Xu L H, et al. Knowledge graph embedding via dynamic mapping matrix. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Beijing, 2015. 687–696

16 Xiao H, Huang M L, Hao Y, et al. TransG: a generative mixture model for knowledge graph embedding. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, 2016. 2316–2325

17 Ji G L, Liu K, He S Z, et al. Knowledge graph completion with adaptive sparse transfer matrix. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016. 985–991

18 Jia Y T, Wang Y Z, Lin H L, et al. Locally adaptive translation for knowledge graph embedding. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016. 992–998

19 Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D knowledge graph embeddings. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, 2018. 1811–1818

20 Nguyen D Q, Nguyen T D, Nguyen D Q, et al. A novel embedding model for knowledge base completion based on convolutional neural network. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, 2018. 327–333

21 Yuan J, Gao N, Xiang J. TransGate: knowledge graph embedding with shared gate structure. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, 2019. 3100–3107

22 Guan S P, Jin X L, Wang Y Z, et al. Shared embedding based neural networks for knowledge graph completion. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, 2018. 247–256

23 Schlichtkrull M S, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks. In: Proceedings of the 15th International Conference on Extended Semantic Web Conference, Heraklion, 2018. 593–607

24 Shang C, Tang Y, Huang J, et al. End-to-end structure-aware convolutional networks for knowledge base completion. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, 2019. 3060–3067

25 Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks. Mach Learn, 2010, 81: 53–67

26 Lin Y K, Liu Z Y, Luan H B, et al. Modeling relation paths for representation learning of knowledge bases. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Lisbon, 2015. 705–714

27 Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, 2013. 926–934

28 Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016. 2659–2665

29 Shi B, Weninger T. Open-world knowledge graph completion. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, 2018. 1957–1964

30 Yao L, Mao C S, Luo Y. KG-BERT: BERT for knowledge graph completion, 2019. ArXiv:1909.03193

31 Kim B, Hong T, Ko Y, et al. Multi-task learning for knowledge graph completion with pre-trained language models. In: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, 2020. 1737–1743

32 Wang S, Du Z J, Meng X F. Research progress of large-scale knowledge graph completion technology. Sci Sin Inform, 2020, 50: 551–575 [王硕, 杜志娟, 孟小峰. 大规模知识图谱补全技术的研究进展. 中国科学: 信息科学, 2020, 50: 551–575]

33 Nickel M, Rosasco L, Poggio T A. Holographic embeddings of knowledge graphs. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016. 1955–1961

34 Yang B D, Yih W T, He X D, et al. Embedding entities and relations for learning and inference in knowledge bases. In: Proceedings of the 3rd International Conference on Learning Representations, San Diego, 2015

35 Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, 2017. 2168–2178

36 Li S J, Chen S D, Hao Y X, et al. A novel knowledge representation model based on convolutional neural network. High Technol Lett, 2009, 30: 901–907 [李少杰, 陈曙东, 郝悦星, 等. 基于卷积神经网络的高效知识表示模型. 高技术通讯, 2020, 30: 901–907]

37 Guo L B, Zhang Q H, Ge W Y, et al. DSKG: a deep sequential model for knowledge graph completion. In: Proceedings of the 3rd China Conference on Knowledge Graph and Semantic Computing, Tianjin, 2018. 65–77

38 Nathani D, Chauhan J, Sharma C, et al. Learning attention-based embeddings for relation prediction in knowledge graphs. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, 2019. 4710–4723

39 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780

40 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations, Toulon, 2017

41 Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks. In: Proceedings of the 6th International Conference on Learning Representations, Vancouver, 2018

42 Gardner M, Talukdar P P, Krishnamurthy J, et al. Incorporating vector space similarity in random walk inference over knowledge bases. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Doha, 2014. 397–406

43 Bao K F. Research on algorithms of knowledge graph completion based on multi-source information representation learning. Dissertation for Master Degree. Shanghai: East China Normal University, 2018 [鲍开放. 基于多源信息表示学习的知识图谱补全算法研究. 硕士学位论文. 上海: 华东师范大学, 2018]

44 Wen Y. Research on knowledge representation learning based on entity description and entity similarity. Dissertation for Master Degree. Wuhan: Central China Normal University, 2020 [文洋. 基于实体描述和实体相似性的知识表示学习研究. 硕士学位论文. 武汉: 华中师范大学, 2020]

45 Ding J H. Research on knowledge representation learning algorithms for knowledge graph completion. Dissertation for Master Degree. Shanghai: Shanghai Jiao Tong University, 2019 [丁建辉. 面向知识图谱补全的知识表示学习算法研究. 硕士学位论文. 上海: 上海交通大学, 2019]

46 Xiao H, Huang M L, Meng L, et al. SSP: semantic space projection for knowledge graph embedding with text descriptions. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, 2017. 3104–3110

47 McCormick C. BERT word embeddings tutorial. 2019. https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/

48 Sun C, Qiu X P, Xu Y G, et al. How to fine-tune BERT for text classification? In: Proceedings of the 18th China National Conference on Chinese Computational Linguistics, Kunming, 2019. 194–206

49 Sahoo D, Pham Q, Lu J, et al. Online deep learning: learning deep neural networks on the fly. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, 2018. 2660–2666

50 Yang Y, Zhou D W, Zhan D C, et al. Adaptive deep models for incremental learning: considering capacity scalability and sustainability. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, 2019. 74–82

51 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324

52 Fukushima K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern, 1980, 36: 193–202

53 Language Cloud (Language Technology Platform Cloud). 2020 [语言云 (语言技术平台云). 2020] http://www.ltp-cloud.com

54 He H. HanLP: han language processing. 2021. https://github.com/hankcs/HanLP

55 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Sardinia, 2010. 249–256

56 Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res, 2004, 32: 267–270

57 Toutanova K, Chen D. Observed versus latent features for knowledge base and text inference. In: Proceedings of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality, Beijing, 2015. 57–66

58 Zhong X. Understanding knowledge graph. 2019 [钟翔. 一文了解知识图谱. 2019] https://zhuanlan.zhihu.com/p/67827902

59 Lin Y K, Liu Z Y, Sun M S, et al. Learning entity and relation embeddings for knowledge graph completion. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, 2015. 2181–2187

60 Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction. In: Proceedings of the 33nd International Conference on Machine Learning, New York City, 2016. 2071–2080

61 Che F H, Zhang D W, Tao J H, et al. ParamE: regarding neural network parameters as relation embeddings for knowledge graph completion. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, 2020

62 Chen D Q, Manning C D. A fast and accurate dependency parser using neural networks. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Doha, 2014. 740–750

Knowledge graph completion based on parsing graph embedding and a weighted graph convolutional network

Meiqiu LUO, Chunxia ZHANG*, Cheng PENG, Xin ZHANG, Guisuo GUO & Zhendong NIU

School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081*, China*
* Corresponding author. E-mail: cxzhang@bit.edu.cn

Abstract Knowledge graph completion is an important research issue in knowledge graph construction, knowledge engineering, and natural language processing. A knowledge graph is a knowledge support for realizing accurate knowledge services in general and professional fields. It is also an important breakthrough foundation in information retrieval, question-and-answer interactions, and information recommendation. The low quality and small scale of the knowledge graph are the main bottlenecks that hinder its wide applications. The purpose of knowledge graph completion is to build a large-scale and high-quality knowledge graph for continuously updating and expanding the knowledge graph. Aiming at the difficulty of knowledge graph completion methods to extract deep semantic features from auxiliary information, such as unstructured texts, this study proposes a knowledge graph completion method based on parsing graph embedding and weighted graph convolutional network. This method uses the weighted graph convolutional network to model the semantic dependency parsing of the entity description and construct the semantic dependency parsing graph embedding. Furthermore, it introduces a multi-grained sentence-embedding generation method of the entity description, which is intended to build entity representation learning that can capture multi-grained semantics and deep-level semantic features. The experimental results on two public datasets show that the proposed knowledge graph completion approach outperforms the existing methods, thereby demonstrating its effectiveness and superiority.

Keywords knowledge graph completion, parsing graph embedding, weighted graph convolutional network, semantic dependency parsing, entity representation learning