

基于知信图卷积神经网络的开放域知识图谱自动构建模型

孙亚茹, 杨莹, 王永剑

(公安部第三研究所, 上海 201204)

摘要: 解决多源知识对齐和知识冗余问题是在开放数据域自动构建知识图谱的关键。建立一种融合知信学习与深度学习的知识图谱自动构建模型。分析图卷积神经网络(GCN)模型与知信学习之间的理论联系,以先验知识与深度学习相结合的方式构建实体语义联合空间,将先验知识对模型的干预形式化,并利用自动编码器实现一个细粒度的实体对齐和关系抽取模型。同时,采用GCN与多头注意力相结合的方式,缓解因结构数据中多跳推理造成实体依赖信息丢失的影响。在开源数据集 SemEval、FB15k 和收集整理的 MD 数据集上的实验结果表明,该模型针对关系抽取、实体对齐和三元组抽取任务的 F1 值分别达到 89.5%、86.6% 和 84.2%,较 BERT-Softmax 模型分别提升了 0.3、2.4 和 0.3 个百分点,具有更好的信息学习能力。

关键词: 开放数据域;知识图谱;知信学习;图卷积神经网络;注意力机制

开放科学(资源服务)标志码(OSID):



中文引用格式:孙亚茹,杨莹,王永剑.基于知信图卷积神经网络的开放域知识图谱自动构建模型[J].计算机工程,2022,48(10):116-122.

英文引用格式:SUN Y R, YANG Y, WANG Y J. Knowledge graph automatic construction model in open domain based on knowledge-informed graph convolutional neural network[J]. Computer Engineering, 2022, 48(10): 116-122.

Knowledge Graph Automatic Construction Model in Open Domain Based on Knowledge-Informed Graph Convolutional Neural Network

SUN Yaru, YANG Ying, WANG Yongjian

(The Third Research Institute of Ministry of Public Security, Shanghai 201204, China)

[Abstract] Solving the problem of multi-source knowledge alignment and knowledge redundancy is the key to automatically build a knowledge graph in the open data domain. To solve this problem, an automatic knowledge graph construction model that combines knowledge-informed learning with deep learning is proposed. The model is used to analyze the theoretical relationship between a Graph Convolutional Neural Network (GCN) model and knowledge-informed learning, construct an entity semantic joint space by combining prior knowledge with deep learning, formalize the intervention of prior knowledge on the model, and use an automatic encoder to achieve a fine-grained entity alignment and relationship extraction model. Furthermore, the GCN is combined with multi-head attention to mitigate the impact of entity dependency information loss caused by multi-hop reasoning in the structural data. The results of experimental conducted using the open-source datasets SemEval and FB15k as well as the collected and sorted MD datasets show that the F1 values of the model for relation extraction, entity alignment, and triplet extraction tasks reach 89.5%, 86.6%, and 84.2%, which are 0.3, 2.4, and 0.3 percentage points higher than those of the BERT-Softmax model, respectively. Thus, the proposed model has better information learning ability.

[Key words] open data domain; knowledge graph; knowledge-informed learning; Graph Convolutional Neural Network (GCN); attention mechanism

DOI: 10.19678/j.issn.1000-3428.0062902

0 概述

知识图谱^[1-3]可以理解为知识关联网络,主要通过实体、实体属性及实体间的关系来刻画。通过知识关联网络,可以完成知识智能查询、知识智能分

析、知识智能问答等重要的自然语言处理任务。以医疗领域数据为例,医疗知识图谱^[4-6]是实现医疗人工智能的基石,构建完善的医疗知识图谱,可以为人们提供更高效精准的医疗服务。在开放数据域中挖掘医疗知识,自动构建或补充现有的医疗知识图谱,

基金项目:公安部研究计划项目(C21361)。

作者简介:孙亚茹(1993—),女,硕士研究生,主研方向为自然语言处理、数据挖掘;杨莹(通信作者)、王永剑,副研究员、博士。

收稿日期:2021-10-11 修回日期:2021-11-28 E-mail: yangying@mcst.org.cn

更是现代化医疗人工智能的体现。
完善的医疗知识图谱应不受限于特定的医疗数据域以及特定的实体和关系,但现有的知识图谱在增加新的知识时会面临实体和关系无法对齐的问题,从而产生知识冗余^[7-8]。简单罗列实体和实体所属关系以及无逻辑支撑,是造成医疗知识图谱效率低、限制多、拓展性差的主要原因。

图卷积神经网络(Graph Convolutional Neural Network,GCN)具有强大的图结构建模表达能力,是结构化输入的通用逼近器。因开放域中的数据无类别界限,若要达到精准识别的效果,需要结合先验知识。这种以知识驱动参与神经网络模型的训练方式,被称为知信学习^[9]。将先验知识与深度学习相结合,对于模型在开放域中达到精准识别起到至关重要的作用。文献[10]对先验知识与神经网络结合的初步尝试,表明该方式可以达到与现代神经网络相似的数据拟合能力。

本文采用知信学习的设计思想,在先验知识与GCN之间建立关联,提出自适应医疗知识图谱构建方法 Ad-MKG,实现在开放域中自适应地进行实体对齐和关系消歧。在此基础上,从关系抽取、实体对齐、三元组抽取 3 个方面评估 Ad-MKG 的有效性。

1 相关工作

现有的医疗知识图谱多采用半自动构建方法,主要依赖人工定义实体与关系的规范、人工知识降重和知识消歧^[11-12]。中医药学语言系统的语义网络框架在中医药术语系统的质量保证和国际推广工作中发挥重要的作用,但其中包含复杂、庞大的语义概念,这为知识的构建和扩增增加了难度。虽然之后 128 种语义类型被去掉了设置不合理的类型,精简成 58 种,但是其抽象形式仍不利于知识的动态扩增。对此,韩普等^[12]将医疗知识重新划分为 6 类 14 种实体关系,通过多源医疗实体链接融合促进了医疗知识的融合。

针对知识间表象关联、无实质性逻辑支撑等问题,各种不同的实体对齐和关系抽取方法相继被提出,如

文献[13-14]采用分析文本依赖树的方法来挖掘文本中的非局部语法关系,文献[15-16]在完整树中实体间的最短依赖路径上应用神经网络,文献[17]将完整依赖树裁剪为实体间的最低公共祖先作为模型的输入,以降低完整树中的无关信息对模型噪声的影响,文献[18]则在修剪过的依赖树上应用GCN。然而,基于规则的裁剪策略会忽略文本中的一些重要信息,如在跨句子关系抽取任务中,裁剪后的依赖树可能会丢失关键的依赖路径和中间载体。获取的实体信息与关系信息的丰富程度严重影响知识融合的准确率,是知识图谱自动构建性能优化的关键点。因此,如何使模型在完整树中学习有用的信息而剔除无关的信息,是完成实体和关系抽取任务的关键问题。

2 知信图卷积神经网络

本文提出的 Ad-MKG 模型主要由 3 个模块组成,即知识库构建、数据预处理和知信图卷积语义分析,其中,知信图卷积语义分析模块包含图编码层、注意力引导层、知信牵引层和三元组判别层。Ad-MKG 模型框架如图 1 所示,具体步骤如下:

- 1)通过爬虫技术从多源获取所需数据:获取实体与别名实体对作为先验知识存放在知识库中,以及获取模型所需训练的医疗数据。训练数据详情将在下文实验部分详述。
- 2)对整理好的训练数据样本,通过数据预处理模块将文本转换为模型可处理的数据:先通过依存句法分析把文本转换成关联矩阵,再将文本关联矩阵和实体作为知信图卷积语义分析的输入。
- 3)在知信图卷积语义分析模块的处理过程中:首先通过GCN编码对结构化数据进行初步信息挖掘;然后结合多头注意力机制捕获实体间的关联依赖信息,通过将先验知识与锚点相结合的方式,使得模型计算输出的实体特征在空间域中不偏移;最后对GCN输出的实体和关系的特征信息进行分析判别:若判别是相同实体,则输出三元组存储知识图谱;若判别不是相同实体,则输出不同三元组存储知识图谱。

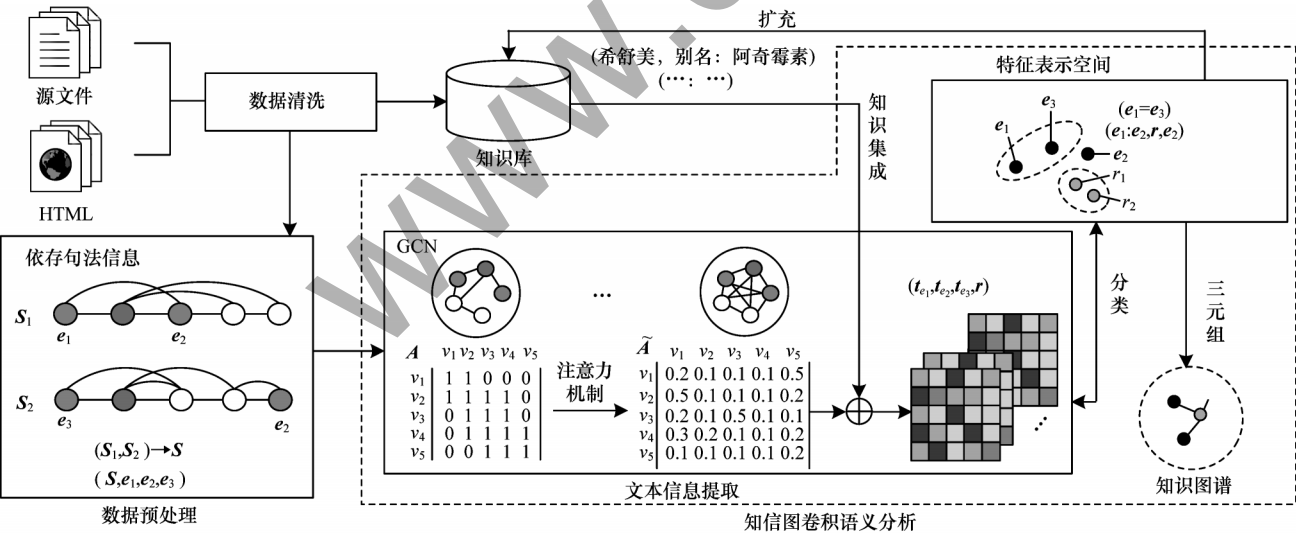


图 1 Ad-MKG 模型框架
Fig.1 Framework of Ad-MKG model

Ad-MKG模型的核心部分为知信图卷积语义分析模块,下文将结合具体的输入和输出,具体描述该模块各层的计算过程。

2.1 图编码层

图编码层用于将数据预处理模块输出的文本依存信息编码成图结构数据,以支持模型从图结构数据中有效捕获实体关联信息。

数据预处理模块将输入的文本 $\{s_1, s_2, \dots, s_n\}$ 处理成输出 (S, e_1, e_2, e_3) ,其中 S 表示文本的依存信息, e_1 、 e_2 和 e_3 表示不同实体。 S 包含文本的token以及关联矩阵。关联矩阵刻画了文本中词与词之间的关联,若词间有关联则元素值为1,若无关联则为0。在图编码层,词变成了图中的节点,词间的关系表示图中的边。若文本中有 n 个词,则图中有 n 个节点,可将该图表示成一个 $n \times n$ 的邻接矩阵 A ,其中 A_{ij} 和 A_{ji} 表示节点 i 和 j 之间存在一条边,初始值为0或1。GCN通过邻接节点来表征本节点,在 L 层的GCN中,给定输入集合 $\{h_1^0, h_2^0, \dots, h_n^0\}$,输出 $\{h_1^L, h_2^L, \dots, h_n^L\}$ 。第 l 层节点 i 的输出向量 h_i^l 由第 $l-1$ 层节点 i 及其相邻节点表示如下:

$$h_i^l = \sigma \left(\sum_{j=1}^n A_{ij} W^{(l)} h_j^{(l-1)} + b^{(l)} \right) \quad (1)$$

其中: $W^{(l)}$ 是做线性变换的权重矩阵; $b^{(l)}$ 是偏差向量; h_i^0 是初始输入的单词向量 $x_i (x_i \in \mathbb{R}^d, d$ 是输入特征的维度)。

经过 L 层GCN对每个节点向量的处理,得到节点的隐藏表示,利用这些词表征可以得到一个句子的特征表示:

$$h_{\text{sent}} = f(h^{(L)}) = f(\text{GCN}(h^{(0)})) \quad (2)$$

其中: $h^{(L)} \in \mathbb{R}^{n \times d}$,表示 L 层所有的隐藏表示(L 是超参数);函数 $f: \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^d$ 将 n 个向量转变成一个句子向量。同理,第 i 个实体 h_{ei} 的计算公式如下:

$$h_{ei} = f(h_{ei}^{(L)}) \quad (3)$$

2.2 注意力引导层

注意力机制是一种有效计算数据中哪些重要部分影响结果的方法,其中多头注意力机制可以从数据的多个层次挖掘影响结果的重要信息。本文采用多头注意力机制对文本全局信息特征进行把控,将图编码层中的邻接矩阵通过注意力机制转换成边-权连接图,使得边-权连接图能够深度刻画节点和节点之间的信息交互。

邻接矩阵通过注意力矩阵 \tilde{A} 转化成边-权连接图。在图1中, \tilde{A}_{ij} 表示节点 i 到节点 j 的边的权重,描述了单个序列2个任意位置之间的相互作用。注意力矩阵 \tilde{A} 的计算公式如下:

$$\tilde{A} = \text{Softmax} \left(\frac{QW_i^Q \times (KW_i^K)^T}{\sqrt{d}} \right) \quad (4)$$

其中: Q 和 K 是 l 层的图特征表示; \tilde{A} 继续参与下一GCN层的计算,其大小与原始邻接矩阵的大小相同,不会涉及额外的计算开销; $W_i^Q \in \mathbb{R}^{d \times d}$, $W_i^K \in \mathbb{R}^{d \times d}$,均为参数矩阵。

相应地,初始输入的原始图转换成了全连接的边-权图,第 l 层节点 i 的输出向量 h_i^l 计算公式如下:

$$h_i^l = \sigma \left(\sum_{j=1}^n \tilde{A}_{ij} W^{(l)} h_j^{(l-1)} + b^{(l)} \right) \quad (5)$$

2.3 知信牵引层

知信牵引层以先验知识引导模型学习拟合数据的规律,实现文本中的实体对齐。本文以医药实体别名对作为先验知识,将成对的医药实体作为医药特征的锚点。在实体特征输出时,设置一个惩罚项 γ 来奖励或惩罚模型对实体对齐和关系抽取学习的行为。在注意力引导层输出实体特征和句子特征时,首先会判别句子中是否含有锚点,若有则增加锚点在实体特征和句子特征中的权重,否则惩罚该实体和句子整体的表示。以第 l 层第 i 个词的信息特征 h_i^l 为例,计算公式如下:

$$h_i^l = \gamma \sigma \left(\sum_{j=1}^n \tilde{A}_{ij} W^{(l)} h_j^{(l-1)} + b^{(l)} \right) \quad (6)$$

实体对齐是指不同名字的实体在语义表征空间中具有相同的信息。本文通过直接计算实体特征向量在特征空间的表示,找出文本中与实体 h_e 表征最相似的实体表征 h_e' ,计算公式如下:

$$h_e' = \underset{h_i \in H_e}{\text{argmax}} \| h_e \|_2 \cdot \| h_i \|_2 \quad (7)$$

其中: H_e 表示实体特征的集合。通过式(7)可得到潜在的实体对齐对 $\langle h_e', h_e \rangle$ 。

2.4 三元组判别层

三元组的提取依赖实体对和实体间关系的确定。知信牵引层确定了实体对齐对,对去除掉别名实体的实体对判别关系,构成三元组。本层通过一个维度变换函数计算GCN编码后的句子的整体表达,如式(8)所示:

$$h_{\text{sent}} = f(h^{(L)}) = f(\text{GCN}(h^{(0)})) \quad (8)$$

关系抽取任务可看作是对描述文本中实体对的关系进行分类。实体 h_{ei} 与实体 h_{ej} 的关系 r_{ij} 通过一层前馈神经网络FFNN(\cdot)计算得到:

$$r_{ij} = \text{FFNN}(h_{ei}, h_{ej}, h_{\text{sent}}) \quad (9)$$

本文通过Softmax函数对 r_{ij} 进行关系类别的预测,计算公式如下:

$$\hat{r} = P(r_{ij} | h_{ei}, h_{ej}, h_{\text{sent}}) = \text{Softmax}(\text{MLP}(h_{ei}, h_{ej}, h_{\text{sent}})) \quad (10)$$

对三元组的提取,笔者期望模型对于给定的实体集合 $\{h_{e1}, h_{e2}, \dots, h_{en}\}$ 可以判别出哪些是相同实体,并输出不同实体 h_{ei} 和 h_{ej} 之间的关系 r_{ij} ,然后组成三元组的形式 $\langle h_{ei}, r_{ij}, h_{ej} \rangle$ 。因此,目标函数定义为:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m \lg(\hat{r}_i) + E_{\langle h_e', h_e \rangle} \left(\underset{h_i \in H_e}{\text{argmax}} \| h_e \|_2 \cdot \| h_i \|_2 \right) \quad (11)$$

其中: m 为关系标签的个数。可采用梯度下降法对目标函数求参,并采用超参数学习率 α 更新参数,计算公式如下:

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta) \quad (12)$$

3 实验与分析

3.1 数据集

实验将从关系抽取、实体对齐和三元组抽取 3 个方面测试和评估本文提出的 Ad-MKG 模型。为便于对比现有的任务模型,采用 SemEval 2010 Task 8 作为关系抽取评估数据集,采用 FB15k 作为三元组抽取评估数据集。同时,为验证知识图谱自动构建的流程,采用收集整理得到的数据集 Medical Dataset(MD)进行实体对齐和三元组抽取。数据集介绍具体如下:SemEval 2010 Task 8^[19]训练集包含 8 000 个样本,测试集包含 717 个样本;FB15k 数据集是从 Facebook 中抽取出的数据集,训练集包含 480 000 个样本,测试集包含 59 000 个样本;MD 数据集训练文本依据爬取到的实体别名对清洗得到。根据医疗领域中行文的语义网络结构,定义 5 种实体、5 种关系和 7 种属性,整理训练得到的数据集包含 10 000 个样本,其中训练集有 80 000 个样本,测试集有 20 000 个样本。

3.2 对比模型

在关系抽取实验中,选取以下 4 个具有代表性的模型进行对比:支持向量机(Support Vector Machine, SVM)^[20],最短路径 LSTM(Shortest Path LSTM, SDP-LSTM)^[21],注意力引导的图卷积神经网络(Attention Guided GCN, AGGCN)^[22],基于 BERT 的关系抽取模型 R-BERT^[23]。

在实体对齐和三元组抽取实验中,选择 BERT-Softmax、TransE^[24]和 TransR^[24]作为对比模型。

3.3 实验设置

实验中主要涉及的超参数有 GCN 层数 L 、多头注意力头数 N 和单词向量维度 d 。从 $L = \{2, 3, 4\}$ 中选择 GCN 层数,从 $N = \{1, 2, 3, 4, 5\}$ 中选择多头注意力头数,从 $d = \{100, 200, 300\}$ 中选择单词维度,进行 10 次独立运行的实验,选择具有中间验证结果的 F1 值模型,并报告其测试的 F1 值。通过测试集上的实验结果可以发现, $(L = 2, N = 2, d = 300)$ 的组合设置在 MK 数据集上取得了最好的效果。该模型是在 NVIDIA GeForce GTX 1050 下采用 CUDA10.2 训练的,包含 100 个训练周期,每个周期时长约为 209 s,初始词典大小为 5 000,dropout 率设置为 0.1。同时,采用 Adam 优化器训练参数,初始化的学习率为 0.001,选择带指数衰减的学习率设置,衰减率为 0.9。

3.4 实验结果

3.4.1 关系抽取任务

在 SemEval 数据集上进行关系抽取任务的实验,实验结果如表 1 所示,其中加粗数据表示最优值。由表 1 可以看出:深度学习模型性能优于基于特征的模型,基于图结构的深度学习模型性能优于基于序列学习的模型,这说明基于图结构的深度学习方法可以更深层次地捕获实体之间的信息;基于实体信息挖掘完成关系抽取的 R-BERT 模型较 AGGCN 的 Macro-F1 值

提升了 3.5 个百分点,这是因为 C-BERT 依赖性能较稳定的预训练模型 BERT;本文所提出的 Ad-MKG 模型取得了最好的性能,表明先验知识的参与可以有效提升模型挖掘实体间关联信息的能力。

表 1 SemEval 数据集中关系抽取任务的实验结果

Table 1 Experimental results for relational extraction task in SemEval dataset %

模型	Macro-F1 值
SVM ^[20]	82.2
SDP-LSTM ^[21]	83.7
AGGCN ^[22]	85.7
R-BERT ^[23]	89.2
Ad-MKG	89.5

3.4.2 实体对齐任务

在 MK 数据集上进行实体对齐任务的实验,实验结果如表 2 所示,其中加粗数据表示最优值。由表 2 可以看出,Ad-MKG 模型依然取得了较好的结果,其 F1 值较 BERT-Softmax 模型提升了 2.4 个百分点。

表 2 MK 数据集中实体对齐任务的实验结果

Table 2 Experimental results for entity alignment task in MK dataset %

模型	准确率	召回率	F1 值
BERT-Softmax	85.1	83.4	84.2
Ad-MKG	87.8	85.4	86.6

进一步地,从数据规模对模型性能的影响出发进行评估和分析。将 MK 数据集中的训练集比例划分为 20%、40%、60%、80%、100%,实验结果如图 2 所示。由图 2 可以看出:数据规模对 Ad-MKG 有大影响,随着数据规模不断增加,Ad-MKG 的学习效果越来越好,但是对基于强大预训练模型的 BERT-Softmax 影响较小,这说明 BERT-Softmax 对数据规模有较好的扰动性,而 Ad-MKG 对样本数据有较强的依赖性。

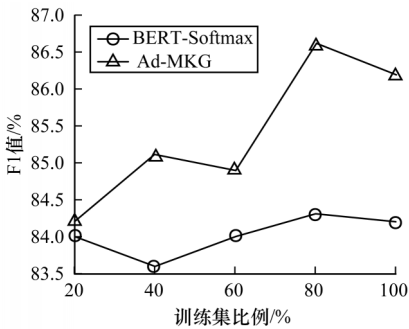


图 2 不同数据集规模下实体对齐任务的实验结果

Fig.2 Experimental results for entity alignment task under different dataset sizes

3.4.3 三元组抽取任务

三元组抽取任务是指从文本中识别出成对的实体和实体间的关系。对三元组抽取结果进行对比,实验结果如表 3 所示,其中加粗数据表示最优值。

由表3可以看出,Ad-MKG依然取得了较好的结果,这说明Ad-MKG更关注于文本中关键的节点关联信息。但同时从表中发现,Ad-MKG精确率高但召回率低,而BERT-Softmax相对稳定,这可能与文本中存在多个实体有关,并且文本的倾向性也会误导模型对实体关系的表征计算。

表3 MK数据集中三元组抽取任务的实验结果

Table 3 Experimental results for triple tuple extraction task in MK dataset %

模型	准确率	召回率	F1值
BERT-Softmax	84.1	83.7	83.9
Ad-MKG	86.7	81.8	84.2

为直观分析影响模型性能的因素,可视化Ad-MKG和BERT-Softmax在5种关系类别上的准确率,同时给出每种关系在数据集中的个数,实验结果如图3所示。由图3可以看出:在样本数量最多的drugs_of关系下模型性能最差,说明实体的模糊含义对模型准确率的影响最大;而在样本数量最少的belongs_to关系下模型性能最好,说明明确的特征表征使得模型在判别时对处理的数据有清晰的判别界限。

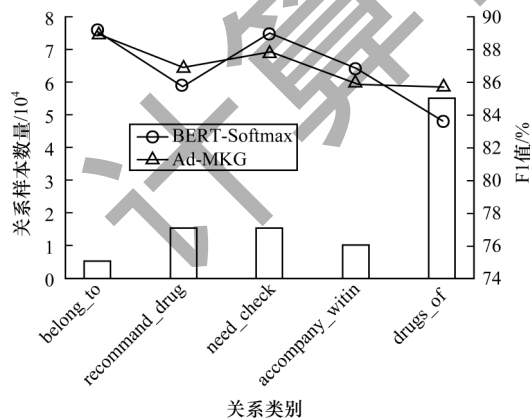


图3 不同关系样本数量下三元组抽取任务的实验结果

Fig.3 Experimental results for triple pull task under different relationship sample sizes

在FB15k数据集中进行三元组抽取任务的实验,将本文提出的Ad-MKG模型与TransE和TransR模型进行对比,实验结果如表4所示,其中加粗数据表示最优值。由表4可以看出,Ad-MKG模型依然取得了最优的结果。

表4 FB15k数据集中三元组抽取任务的实验结果

Table 4 Experimental results for triple tuple extraction task in FB15k dataset %

模型	准确率
TransE	83.1
TransR	83.7
Ad-MKG	84.5

3.4.4 消融实验

对模型各模块进行消融实验,以分析Ad-MKG知信图卷积语义分析模块中各层对模型整体性能的影响。评测依据三元组提取性能,实验结果如表5所示。由表5可以看出,注意力引导层与知信牵引层对模型性能的影响较大。在去除图编码层后,本文利用多层感知机对数据信息进行处理,F1值下降了3.1个百分点,说明图编码层对模型提取的特征信息质量起到了重要的保障作用。目前,注意力机制在很多模型^[18-19]中已被证明是有效的,在Ad-MKG模型中,注意力机制也同样发挥了重要作用。在去除注意力引导层后,F1值下降了1.5个百分点。在同时去除注意力引导层和知信牵引层后,模型性能明显下降,说明注意力引导层发挥了提升模型捕获信息特征的作用。同时由消融实验结果也可以看出,将先验知识与深度学习模型相结合,有利于减小模型在特征域中的位置偏移。

表5 消融实验结果

Table 5 Ablation experiment results %

模型	准确率	召回率	F1值
Ad-MKG	86.7	81.8	84.2
去除图编码层	83.3	79.1	81.1
去除注意力引导层	84.6	80.9	82.7
去除知信牵引层	82.7	77.6	80.1
去除注意力引导层和知信牵引层	78.3	76.5	77.4

3.4.5 实体特征质量评估

为直观地评估模型所得到的实体特征质量,对各模型进行样例测试,二维可视化模型输出的特征。选取可以对齐的实体和不同的实体进行实验,测试样例如表6所示。

表6 实体特征测试样例

Table 6 Examples of entities features test

类型	是否相似	实体名
疾病	不相似	黑色素斑
	不相似	钩虫病
	不相似	低血糖症
药品	不相似	谷氨酸钠注射液
	相似	肝复康丸,健民肝复康丸
	相似	阿奇霉素,希舒美

首先测试Ad-MKG和BERT-Softmax这2个模型是否能够正确判别出属于疾病和药品的实体(测试1),然后测试同类别中相似实体和不似实体之间的判别结果(测试2),实验结果如图4所示,其中形状相同的表示是同一种类别。由图4(a)和图4(b)可以看出,Ad-MKG模型能够正确区分疾病和药品实体,且界限明晰,达到了预期的效果,而BERT-Softmax模型存在少许的错误,如将“黑色素斑”判别为药品。由图4(c)和图4(d)可以看出,在相似实体测试中,

BERT-Softmax 模型不能准确区分相似实体“阿奇霉素”和“希舒美”,而 Ad-MKG 模型仍达到了预期的效

果,但是对别名实体对不能较好区分,如“阿奇霉素”和“希舒美”。

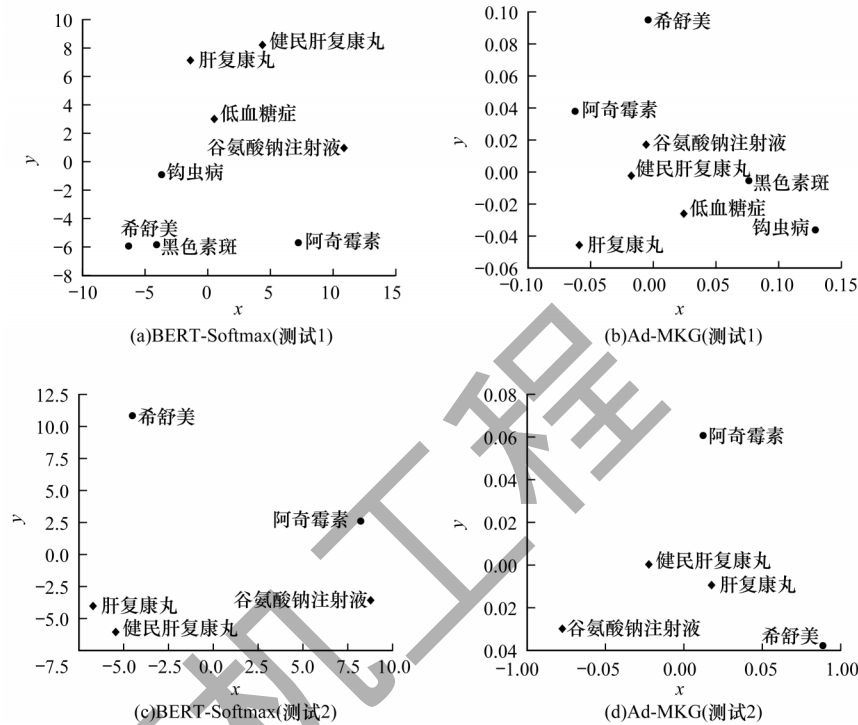


图4 BERT-Softmax 和 Ad-MKG 的样例实体特征空间

Fig.4 Example entities feature space for BERT-Softmax and Ad-MKG

4 结束语

针对知识间表象关联、无实质性逻辑支撑等问题,本文提出一种自适应的开放域知识图谱自动构建模型。该模型将先验知识与深度学习模型的训练过程相结合,联合语义空间实现知识对齐。以医疗领域数据为例,在关系抽取、实体对齐和三元组抽取3个任务上的实验结果验证了该模型的可行性和有效性。多源异构信息融合是完善知识图谱自动构建的关键点,后续将研究多源异构知识信息对知识图谱构建的影响,进一步提升本文模型的泛化能力。

参考文献

- [1] SHI W, ZHENG W G, YU J X, et al. Keyphrase extraction using knowledge graphs[J]. Data Science and Engineering, 2017, 2(4): 275-288.
- [2] XIAO G H, CORMAN J. Ontology-mediated SPARQL query answering over knowledge graphs[J]. Big Data Research, 2021, 23: 1-10.
- [3] 金婧, 万怀宇, 林友芳. 融合实体类别信息的知识图谱表示学习[J]. 计算机工程, 2021, 47(4): 77-83.
JIN J, WAN H Y, LIN Y F. Knowledge graph representation learning fused with entity category information[J]. Computer Engineering, 2021, 47(4): 77-83. (in Chinese)
- [4] 丁辰晖, 夏鸿斌, 刘渊. 融合知识图谱与注意力机制的短文本分类模型[J]. 计算机工程, 2021, 47(1): 94-100.
DING C H, XIA H B, LIU Y. Short text classification model combining knowledge graph and attention mechanism[J]. Computer Engineering, 2021, 47(1): 94-100. (in Chinese)

- [5] LIU W H, YIN L, WANG C, et al. Medical knowledge graph in Chinese using deep semantic mobile computation based on IoT and WoT[J]. Wireless Communications and Mobile Computing, 2021, 2021: 1-13.
- [6] 刘勘, 张雅琴. 基于医疗知识图谱的并发症辅助诊断[J]. 中文信息学报, 2020, 34(10): 85-93, 104.
LIU K, ZHANG Y Q. Medical knowledge graph based auxiliary diagnosis of complications[J]. Journal of Chinese Information Processing, 2020, 34(10): 85-93, 104. (in Chinese)
- [7] VASHISHTH S, JAIN P, TALUKDAR P. CESI: canonicalizing open knowledge bases using embeddings and side information[C]//Proceedings of the 2018 World Wide Web Conference. Washington D. C., USA: IEEE Press, 2018: 1317-1327.
- [8] GALÁRRAGA L, HEITZ G, MURPHY K, et al. Canonicalizing open knowledge bases[C]//Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2014: 1679-1688.
- [9] VON RUEDEN L, MAYER S, BECKH K, et al. Informed machine learning—a taxonomy and survey of integrating prior knowledge into learning systems[J/OL]. IEEE Transactions on Knowledge and Data Engineering: 1-20 [2021-10-25]. <https://ieeexplore.ieee.org/document/9429985/citations#citations>.
- [10] WEIN S, MALLONI W M, TOMÉ A M, et al. A graph neural network framework for causal inference in brain networks[J]. Scientific Reports, 2021, 11(1): 8061.
- [11] 俞思伟, 范昊, 王菲, 等. 基于知识图谱的智能医疗研究[J]. 医疗卫生装备, 2017, 38(3): 109-111, 126.

- YU S W, FAN H, WANG F, et al. Research on intelligent medicine based on knowledge graph[J]. Chinese Medical Equipment Journal, 2017, 38(3): 109-111, 126. (in Chinese)
- [12] 韩普, 马健, 张嘉明, 等. 基于多数据源融合的医疗知识图谱框架构建研究[J]. 现代情报, 2019, 39(6): 81-90.
- HAN P, MA J, ZHANG J M, et al. The framework construction of medical knowledge graph based on multi-data source fusion[J]. Journal of Modern Information, 2019, 39(6): 81-90. (in Chinese)
- [13] BUNESCU R C, MOONEY R J, MANNING C. A shortest path dependency kernel for relation extraction [C]// Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2017: 724-731.
- [14] PENG N Y, POON H, QUIRK C, et al. Cross-sentence N-ary relation extraction with graph LSTMs[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 101-115.
- [15] XU K, FENG Y S, HUANG S F, et al. Semantic relation classification via convolutional neural networks with simple negative sampling[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2015: 536-540.
- [16] XU Y, MOU L L, LI G, et al. Classifying relations via long short term memory networks along shortest dependency paths[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2015: 536-540.
- [17] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, USA: Association for Computational Linguistics, 2016: 90-94.
- [18] ZHANG Y H, QI P, MANNING C D. Graph convolution over pruned dependency trees improves relation extraction [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2018: 2205-2215.
- [19] HENDRICKX I, KIM S N, KOZAREVA Z, et al. SemEval-2010 task 8: multi-way classification of semantic relations between pairs of nominals [C]//Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Stroudsburg, USA: Association for Computational Linguistics, 2009: 33-38.
- [20] ZHANG Y H, ZHONG V, CHEN D Q, et al. Position-aware attention and supervised data improve slot filling [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2017: 35-45.
- [21] XU Y, MOU L L, LI G, et al. Classifying relations via long short term memory networks along shortest dependency paths [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2015: 1785-1794.
- [22] GUO Z J, ZHANG Y, LU W. Attention guided graph convolutional networks for relation extraction [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 241-251.
- [23] WU S C, HE Y F. Enriching pre-trained language model with entity information for relation classification [C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2019: 2361-2364.
- [24] LIN Y K, LIU Z Y, SUN M S, et al. Learning entity and relation embeddings for knowledge graph completion [C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2015: 2181-2187.

编辑 金胡考