计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20160020
53(2):247-261,2016

知识表示学习研究进展

刘知远　孙茂松　林衍凯　谢若冰
(清华大学计算机科学与技术系　北京　100084)
(智能技术与系统国家重点实验室(清华大学)　北京　100084)
(清华信息科学与技术国家实验室(筹)　北京　100084)
(liuzy@tsinghua.edu.cn)

Knowledge Representation Learning: A Review

Liu Zhiyuan, Sun Maosong, Lin Yankai, and Xie Ruobing
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)
(*State Key Laboratory of Intelligent Technology and Systems (Tsinghua University), Beijing* 100084)
(*Tsinghua National Laboratory for Information Science and Technology, Beijing* 100084)

Abstract　Knowledge bases are usually represented as networks with entities as nodes and relations as edges. With network representation of knowledge bases, specific algorithms have to be designed to store and utilize knowledge bases, which are usually time consuming and suffer from data sparsity issue. Recently, representation learning, delegated by deep learning, has attracted many attentions in natural language processing, computer vision and speech analysis. Representation learning aims to project the interested objects into a dense, real-valued and low-dimensional semantic space, whereas knowledge representation learning focuses on representation learning of entities and relations in knowledge bases. Representation learning can efficiently measure semantic correlations of entities and relations, alleviate sparsity issues, and significantly improve the performance of knowledge acquisition, fusion and inference. In this paper, we will introduce the recent advances of representation learning, summarize the key challenges and possible solutions, and further give a future outlook on the research and application directions.

Key words　knowledge representation; representation learning; knowledge graph; deep learning; distributed representation

摘　要　人们构建的知识库通常被表示为网络形式,节点代表实体,连边代表实体间的关系.在网络表示形式下,人们需要设计专门的图算法存储和利用知识库,存在费时费力的缺点,并受到数据稀疏问题的困扰.最近,以深度学习为代表的表示学习技术受到广泛关注.表示学习旨在将研究对象的语义信息表示为稠密低维实值向量,知识表示学习则面向知识库中的实体和关系进行表示学习.该技术可以在低维空间中高效计算实体和关系的语义联系,有效解决数据稀疏问题,使知识获取、融合和推理的性能得到显著提升.介绍知识表示学习的最新进展,总结该技术面临的主要挑战和可能解决方案,并展望该技术的未来发展方向与前景.

关键词　知识表示;表示学习;知识图谱;深度学习;分布式表示

中图法分类号　TP391

收稿日期:2016-01-12;**修回日期**:2016-01-15

基金项目:国家"九七三"重点基础研究发展计划基金项目(2014CB340501);国家自然科学基金项目(61572273,61532010);清华大学自主科研计划基金项目(2015THZ)

This work was supported by the National Basic Research Program of China (973 Program) (2014CB340501), the National Natural Science Foundation of China (61572273,61532010), and Tsinghua University Initiative Scientific Research Program (2015THZ).

知识库将人类知识组织成结构化的知识系统.人们花费大量精力构建了各种结构化的知识库,如语言知识库 WordNet[1]、世界知识库 Freebase[2]等.知识库是推动人工智能学科发展和支撑智能信息服务应用(如智能搜索、智能问答、个性化推荐等)的重要基础技术.为了改进信息服务质量,国内外互联网公司(特别是搜索引擎公司)纷纷推出知识库产品,如谷歌知识图谱、微软 Bing Satori、百度知心以及搜狗知立方等.著名的 IBM Watson 问答系统和苹果 Siri 语音助理的背后,知识库也扮演着重要角色.如谷歌在介绍知识图谱时所说的"构成这个世界的是实体,而非字符串".可以说,知识库的兴起拉开了智能信息检索从字符串匹配跃迁至智能理解的序幕.

知识库描述现实世界中实体(entity)间的关系(relation).这些知识蕴藏在无(半)结构的互联网信息中,而知识库则是有结构的.因此,知识库的主要研究目标是:从无(半)结构的互联网信息中获取有结构知识,自动融合构建知识库、服务知识推理等相关应用.知识表示是知识获取与应用的基础,因此知识表示学习问题是贯穿知识库的构建与应用全过程的关键问题.

人们通常以网络的形式组织知识库中的知识,网络中每个节点代表实体(人名、地名、机构名、概念等),而每条连边则代表实体间的关系.因此,大部分知识往往可以用三元组(实体 1, 关系, 实体 2)来表示,对应着知识库网络中的一条连边及其连接的 2 个实体.这是知识库的通用表示方式,例如万维网联盟(W3C)发布的资源描述框架(resource description framework, RDF)技术标准[3],就是以三元组表示为基础的.特别是在谷歌提出知识图谱(knowledge graphs)的概念后,这种网络表示形式更是广受认可.然而,基于网络形式的知识表示面临诸多挑战性难题,主要包括如下 2 个方面:

1) 计算效率问题.基于网络的知识表示形式中,每个实体均用不同的节点表示.当利用知识库计算实体间的语义或推理关系时,往往需要人们设计专门的图算法来实现,存在可移植性差的问题.更重要的是,基于图的算法计算复杂度高、可扩展性差,当知识库达到一定规模时,就很难较好地满足实时计算的需求.

2) 数据稀疏问题.与其他类型的大规模数据类似,大规模知识库也遵守长尾分布,在长尾部分的实体和关系上,面临严重的数据稀疏问题.例如,对于长尾部分的罕见实体,由于只有极少的知识或路径涉及它们,对这些实体的语义或推理关系的计算往往准确率极低.

近年来,以深度学习[4]为代表的表示学习[5]技术异军突起,在语音识别、图像分析和自然语言处理领域获得广泛关注.表示学习旨在将研究对象的语义信息表示为稠密低维实值向量.在该低维向量空间中,2 个对象距离越近则说明其语义相似度越高.

顾名思义,知识表示学习是面向知识库中的实体和关系进行表示学习.该方向最近取得了重要进展,可以在低维空间中高效计算实体和关系的语义联系,有效解决数据稀疏问题,使知识获取、融合和推理的性能得到显著提升.

由于上述优点,知识表示学习引起了广泛关注和研究兴趣,但该方向仍然面临着诸多挑战.本文将介绍知识表示学习的最新进展,总结该技术面临的主要挑战和可能解决方案,并展望该技术的未来发展方向与前景.

1 知识表示学习简介

在正式介绍知识表示学习的主要模型和挑战之前,本节首先介绍表示学习的基本概念和理论基础,以及知识表示学习的重要意义.

1.1 表示学习的基本概念

如前所述,表示学习的目标是,通过机器学习将研究对象的语义信息表示为稠密低维实值向量.本文用黑斜体表示研究对象所对应的向量.以知识库中的实体 e 和关系 r 为例,我们将表示学习得到的向量表示为 $\boldsymbol{l}_e$ 和 $\boldsymbol{l}_r$.在该向量空间中,我们可以通过欧氏距离或余弦距离等方式,计算任意 2 个对象之间的语义相似度.

实际上,在表示学习之外,有更简单的数据表示方案,即独热表示(one-hot representation)[6].该方案也将研究对象表示为向量,只是该向量只有某一维非零,其他维度上的值均为 0.显而易见,为了将不同对象区分开,有多少个不同的对象,独热表示向量就有多长.独热表示是信息检索和搜索引擎中广泛使用的词袋模型(bag-of-words model)[7]的基础.以中文为例,假如网页中共有 W 个不同的词,词袋模型中的每个词都被表示为一个 W 维的独热表示向量.在此基础上,词袋模型将每个文档表示为一个 W 维向量,每一维表示对应的词在该文档中的重要性.

与表示学习相比,独热表示无需学习过程,简单高效,在信息检索和自然语言处理中得到广泛应用.

但是独热表示的缺点也非常明显.独热表示方案假设所有对象都是相互独立的.也就是说,在独热表示空间中,所有对象的向量都是相互正交的,通过余弦距离或欧氏距离计算的语义相似度均为0.这显然是不符合实际情况的,会丢失大量有用信息.例如,“苹果”和“香蕉”虽然是2个不同的词,但由于它们都属于水果,因此应当具有较高的语义相似度.显然,独热表示无法有效利用这些对象间的语义相似度信息.这也是词袋模型无法有效表示短文本、容易受到数据稀疏问题影响的根本原因.

与独热表示相比,表示学习的向量维度较低,有助于提高计算效率,同时能够充分利用对象间的语义信息,从而有效缓解数据稀疏问题.由于表示学习的这些优点,最近出现了大量关于单词[6]、短语[8-9]、实体[10]、句子[11-13]、文档[12]和社会网络[14-16]等对象的表示学习研究.特别是在词表示方面,针对一词多义[17-19]、语义组合[9,20-22]、语素或字母信息[23-25]、跨语言[26-28]、可解释性[29-32]等特点提出了相应表示方案,展现出分布式表示灵活的可扩展性.

1.2 表示学习的理论基础

表示学习得到的低维向量表示是一种分布式表示(distributed representation)[6].之所以如此命名,是因为孤立地看向量中的每一维,都没有明确对应的含义;而综合各维形成一个向量,则能够表示对象的语义信息.这种表示方案并非凭空而来,而是受到人脑的工作机制启发而来.

我们知道,现实世界中的实体是离散的,不同对象之间有明显的界限.人脑通过大量神经元上的激活和抑制存储这些对象,形成内隐世界.显而易见,每个单独神经元的激活或抑制并没有明确含义,但是多个神经元的状态则能表示世间万物.受到该工作机制的启发,分布式表示的向量可以看作模拟人脑的多个神经元,每维对应一个神经元,而向量中的值对应神经元的激活或抑制状态.基于神经网络这种对离散世界的连续表示机制,人脑具备了高度的学习能力与智能水平.表示学习正是对人脑这一工作机制的模仿.

还值得一提的是,现实世界存在层次结构[33].一个对象往往由更小的对象组成.例如一个房屋作为一个对象,是由门、窗户、墙、天花板和地板等对象有机组合而成的,墙则由更小的砖块和水泥等对象组成,以此类推.这种层次或嵌套的结构反映在人脑中,形成了神经网络的层次结构.最近象征人工神经网络复兴的深度学习技术,其津津乐道的“深度”正是这种层次性的体现.

综上,我们在表1总结了现实世界与内隐世界的特点.可以说,分布式表示和层次结构是人类智能的基础,也是表示学习和深度学习的本质特点.

Table 1 Characteristics of Real World and Internal World

表1 现实世界与内隐世界的特点

Characteristics	Discreteness	Hierarchy
Real World	Discrete	Hierarchical
Internal World	Continuous	Hierarchical

1.3 知识表示学习的典型应用

知识表示学习是面向知识库中实体和关系的表示学习.通过将实体或关系投影到低维向量空间,我们能够实现对实体和关系的语义信息的表示,可以高效地计算实体、关系及其之间的复杂语义关联.这对知识库的构建、推理与应用均有重要意义.

知识表示学习得到的分布式表示有以下典型应用:

1) 相似度计算.利用实体的分布式表示,我们可以快速计算实体间的语义相似度,这对于自然语言处理和信息检索的很多任务具有重要意义.

2) 知识图谱补全.构建大规模知识图谱,需要不断补充实体间的关系.利用知识表示学习模型,可以预测2个实体的关系,这一般称为知识库的链接预测(link prediction),又称为知识图谱补全(knowledge graph completion).

3) 其他应用.知识表示学习已被广泛用于关系抽取、自动问答、实体链指等任务,展现出巨大的应用潜力.随着深度学习在自然语言处理各项重要任务中得到广泛应用,这将为知识表示学习带来更广阔的应用空间.

1.4 知识表示学习的主要优点

知识表示学习实现了对实体和关系的分布式表示,它具有以下主要优点:

1) 显著提升计算效率.知识库的三元组表示实际就是基于独热表示的.如前所分析的,在这种表示方式下,需要设计专门的图算法计算实体间的语义和推理关系,计算复杂度高、可扩展性差.而表示学习得到的分布式表示,则能够高效地实现语义相似度计算等操作,显著提升计算效率.

2) 有效缓解数据稀疏.由于表示学习将对象投影到统一的低维空间中,使每个对象均对应一个稠密向量,从而有效缓解数据稀疏问题,这主要体现在2个方面.一方面,每个对象的向量均为稠密有值的,因此可以度量任意对象之间的语义相似程度.而基于

独热表示的图算法,由于受到大规模知识图谱稀疏特性的影响,往往无法有效计算很多对象之间的语义相似度.另一方面,将大量对象投影到统一空间的过程,也能够将高频对象的语义信息用于帮助低频对象的语义表示,提高低频对象的语义表示的精确性.

3) 实现异质信息融合.不同来源的异质信息需要融合为整体,才能得到有效应用.例如,人们构造了大量知识库,这些知识库的构建规范和信息来源均有不同,例如著名的世界知识库有 DBPedia, YAGO, Freebase 等.大量实体和关系在不同知识库中的名称不同.如何实现多知识库的有机融合,对知识库应用具有重要意义.如果基于网络表示,该任务只能通过设计专门图算法来实现,效果较差,效率低下.而通过设计合理的表示学习模型,将不同来源的对象投影到同一个语义空间中,就能够建立统一的表示空间,实现多知识库的信息融合.此外,当我们在信息检索或自然语言处理中应用知识库时,往往需要计算查询词、句子、文档和知识库实体之间的复杂语义关联.由于这些对象的异质性,计算它们的语义关联往往是棘手问题.而表示学习亦能为异质对象提供统一表示空间,轻而易举实现异质对象之间的语义关联计算.

综上,由于知识表示学习能够显著提升计算效率,有效缓解数据稀疏,实现异质信息融合,因此对于知识库的构建、推理和应用具有重要意义,值得广受关注、深入研究.

2 知识表示学习的主要方法

知识表示学习是近年来的研究热点,研究者提出了多种模型,学习知识库中的实体和关系的表示.本节将主要介绍其中几种代表方法.为了介绍这些方法,我们首先定义几种符号,便于下文使用.首先,我们将知识库表示为 $G=(E,R,S)$,其中 $E=\{e_1,e_2,\cdots,e_{|E|}\}$ 是知识库中的实体集合,其中包含 $|E|$ 种不同实体;$R=\{r_1,r_2,\cdots,r_{|R|}\}$ 是知识库中的关系集合,其中包含 $|R|$ 种不同关系;而 $S\subseteq E\times R\times E$ 则代表知识库中的三元组集合,我们一般表示为 (h,r,t),其中 h 和 t 表示头实体和尾实体,而 r 表示 h 和 t 之间的关系.例如三元组(史蒂夫·乔布斯,创始人,苹果公司)就表示实体"史蒂夫·乔布斯"和"苹果公司"之间存在"创始人"的关系.

接下来,我们介绍知识表示学习的几个代表模型,包括距离模型、单层神经网络模型、能量模型、双线性模型、张量神经网络模型、矩阵分解模型和翻译模型等.

2.1 距离模型

结构表示(structured embedding, SE)[34]是较早的几个知识表示方法之一.在 SE 中,每个实体用 d 维的向量表示,所有实体被投影到同一个 d 维向量空间中.同时,SE 还为每个关系 r 定义了 2 个矩阵 $\boldsymbol{M}_{r,1},\boldsymbol{M}_{r,2}\in\mathbb{R}^{d\times d}$,用于三元组中头实体和尾实体的投影操作.最后,SE 为每个三元组 (h,r,t) 定义了如下损失函数:

$$f_r(h,t)=|\boldsymbol{M}_{r,1}\boldsymbol{l}_h-\boldsymbol{M}_{r,2}\boldsymbol{l}_t|_{L_1}.$$

我们可以理解为,SE 将头实体向量 $\boldsymbol{l}_h$ 和尾实体向量 $\boldsymbol{l}_t$ 通过关系 r 的 2 个矩阵投影到 r 的对应空间中,然后在该空间中计算两投影向量的距离.这个距离反映了 2 个实体在关系 r 下的语义相关度,它们的距离越小,说明这 2 个实体存在这种关系.

实体向量和关系矩阵是 SE 模型的参数.SE 将知识库三元组作为学习样例,优化模型参数使知识库三元组的损失函数值不断降低,从而使实体向量和关系矩阵能够较好地反映实体和关系的语义信息.

SE 能够利用学习得到的知识表示进行链接预测,即通过计算

$$\arg\min_r|\boldsymbol{M}_{r,1}\boldsymbol{l}_h-\boldsymbol{M}_{r,2}\boldsymbol{l}_t|_{L_1}$$

找到让两实体距离最近的关系矩阵,这就是它们之间的关系.

然而,SE 模型有一个重要缺陷:它对头、尾实体使用 2 个不同的矩阵进行投影,协同性较差,往往无法精确刻画两实体与关系之间的语义联系.

2.2 单层神经网络模型

单层神经网络模型(single layer model, SLM)[35]尝试采用单层神经网络的非线性操作,来减轻 SE 无法协同精确刻画实体与关系的语义联系的问题.SLM 为每个三元组 (h,r,t) 定义了如下评分函数:

$$f_r(h,t)=\boldsymbol{u}_r^{\mathrm{T}}g(\boldsymbol{M}_{r,1}\boldsymbol{l}_h+\boldsymbol{M}_{r,2}\boldsymbol{l}_t),$$

其中,$\boldsymbol{M}_{r,1},\boldsymbol{M}_{r,2}\in\mathbb{R}^{d\times k}$ 为投影矩阵,$\boldsymbol{u}_r^{\mathrm{T}}\in\mathbb{R}^k$ 为关系 r 的表示向量,$g()$ 是 tanh 函数.

虽然 SLM 是 SE 模型的改进版本,但是它的非线性操作仅提供了实体和关系之间比较微弱的联系.与此同时,却引入了更高的计算复杂度.

2.3 能量模型

语义匹配能量模型(semantic matching energy, SME)[36-37]提出更复杂的操作,寻找实体和关系之间的语义联系.在 SME 中,每个实体和关系都用低维向量表示.在此基础上,SME 定义若干投影矩阵,刻画

实体与关系的内在联系.更具体地,SME 为每个三元组(h,r,t)定义了 2 种评分函数,分别是线性形式:

$$f_r(h,t)=(\boldsymbol{M}_1\boldsymbol{l}_h+\boldsymbol{M}_2\boldsymbol{l}_r+\boldsymbol{b}_1)^{\mathrm{T}}(\boldsymbol{M}_3\boldsymbol{l}_t+\boldsymbol{M}_4\boldsymbol{l}_r+\boldsymbol{b}_2)$$

和双线性形式:

$$f_r(h,t)=(\boldsymbol{M}_1\boldsymbol{l}_h\otimes\boldsymbol{M}_2\boldsymbol{l}_r+\boldsymbol{b}_1)^{\mathrm{T}}(\boldsymbol{M}_3\boldsymbol{l}_t\otimes\boldsymbol{M}_4\boldsymbol{l}_r+\boldsymbol{b}_2),$$

其中 $\boldsymbol{M}_1,\boldsymbol{M}_2,\boldsymbol{M}_3,\boldsymbol{M}_4\in\mathbb{R}^{d\times k}$为投影矩阵;$\otimes$表示按位相乘(即 Hadamard 积);$\boldsymbol{b}_1,\boldsymbol{b}_2$ 为偏置向量.此外,也有研究工作用三阶张量代替 SME 的双线性形式[36].

2.4 双线性模型

隐变量模型(latent factor model, LFM)[38-39]提出利用基于关系的双线性变换,刻画实体和关系之间的二阶联系.LFM 为每个三元组(h,r,t)定义了如下双线性评分函数:

$$f_r(h,t)=\boldsymbol{l}_h{}^{\mathrm{T}}\boldsymbol{M}_r\boldsymbol{l}_t,$$

其中,$\boldsymbol{M}_r\in\mathbb{R}^{d\times d}$是关系 r 对应的双线性变换矩阵.与以往模型相比,LFM 取得巨大突破:通过简单有效的方法刻画了实体和关系的语义联系,协同性较好,计算复杂度低.

后来的 DISTMULT 模型[40]还探索了 LFM 的简化形式:将关系矩阵 $\boldsymbol{M}_r$ 设置为对角阵.实验表明,这种简化不仅极大降低了模型复杂度,模型效果反而得到显著提升.

2.5 张量神经网络模型

张量神经网络模型(neural tensor network, NTN)[35]的基本思想是,用双线性张量取代传统神经网络中的线性变换层,在不同的维度下将头、尾实体向量联系起来.其基本思想如图 1 所示:

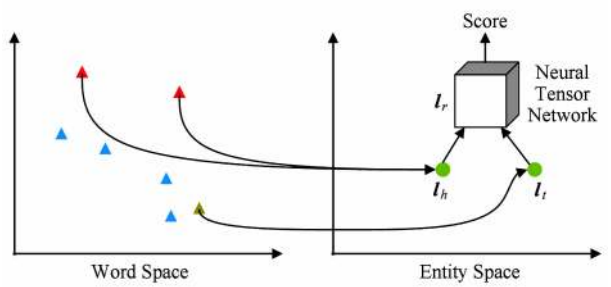


Fig. 1 Neural tensor network (NTN) model.

图 1 张量神经网络模型

NTN 为每个三元组(h,r,t)定义了如下评分函数,评价 2 个实体之间存在某个特定关系 r 的可能性:

$$f_r(h,t)=\boldsymbol{u}_r^{\mathrm{T}}g(\boldsymbol{l}_h\boldsymbol{M}_r\boldsymbol{l}_t+\boldsymbol{M}_{r,1}\boldsymbol{l}_h+\boldsymbol{M}_{r,2}\boldsymbol{l}_t+\boldsymbol{b}_r),$$

其中 $\boldsymbol{u}_r^{\mathrm{T}}$ 是一个与关系相关的线性层,$g()$是 tanh 函数,$\boldsymbol{M}_r\in\mathbb{R}^{d\times d\times k}$是一个三阶张量,$\boldsymbol{M}_{r,1},\boldsymbol{M}_{r,2}\in\mathbb{R}^{d\times k}$是与关系 r 有关的投影矩阵.可以看出,前述 SLM 是 NTN 的简化版本,是 NTN 将其中张量的层数设置为 0 时的特殊情况.

值得注意的是,与以往模型不同,NTN 中的实体向量是该实体中所有单词向量的平均值.这样做的好处是,实体中的单词数量远小于实体数量,可以充分重复利用单词向量构建实体表示,降低实体表示学习的稀疏性问题,增强不同实体的语义联系.

由于 NTN 引入了张量操作,虽然能够更精确地刻画实体和关系的复杂语义联系,但是计算复杂度非常高,需要大量三元组样例才能得到充分学习.实验表明,NTN 在大规模稀疏知识图谱上的效果较差[41].

2.6 矩阵分解模型

矩阵分解是得到低维向量表示的重要途径.因此,也有研究者提出采用矩阵分解进行知识表示学习.这方面的代表方法是 RESACL 模型[42-43].

在该模型中,知识库三元组构成一个大的张量 $\boldsymbol{X}$,如果三元组(h,r,t)存在,则 $X_{hrt}=1$,否则为 0.张量分解旨在将每个三元组(h,r,t)对应的张量值 X_{hrt}分解为实体和关系表示,使得 X_{hrt}尽量地接近于 $\boldsymbol{l}_h\boldsymbol{M}_r\boldsymbol{l}_t$.

可以看到 RESACL 的基本思想与前述 LFM 类似.不同之处在于,RESACL 会优化张量中的所有位置,包括值为 0 的位置;而 LFM 只会优化知识库中存在的三元组.

2.7 翻译模型

表示学习在自然语言处理领域受到广泛关注起源于 Mikolov 等人于 2013 年提出的 word2vec 词表示学习模型和工具包[8,44].利用该模型,Mikolov 等人发现词向量空间存在有趣的平移不变现象.例如他们发现:

$$\boldsymbol{C}(\text{king})-\boldsymbol{C}(\text{queen})\approx\boldsymbol{C}(\text{man})-\boldsymbol{C}(\text{woman}),$$

这里 $\boldsymbol{C}(w)$表示利用 word2vec 学习得到的单词 w 的词向量.也就是说,词向量能够捕捉到单词 king 和 queen 之间、man 和 woman 之间的某种相同的隐含语义关系.Mikolov 等人通过类比推理实验[8,44]发现,这种平移不变现象普遍存在于词汇的语义关系和句法关系中.有研究者还利用词表示的这种特性寻找词汇之间的上下位关系[45].

受到该现象的启发,Bordes 等人提出了 TransE 模型[41],将知识库中的关系看作实体间的某种平移向量.对于每个三元组(h,r,t),TransE 用关系 r 的向量 $\boldsymbol{l}_r$ 作为头实体向量 $\boldsymbol{l}_h$ 和尾实体向量 $\boldsymbol{l}_t$ 之间的平移.我们也可以将 $\boldsymbol{l}_r$ 看作从 $\boldsymbol{l}_h$ 到 $\boldsymbol{l}_t$ 的翻译,因此 TransE 也被称为翻译模型.

如图 2 所示,对于每个三元组(h,r,t),TransE 希望

$$\boldsymbol{l}_h+\boldsymbol{l}_r\approx\boldsymbol{l}_t.$$

TransE 模型定义了如下损失函数：

$$f_r(h,t)=|\boldsymbol{l}_h+\boldsymbol{l}_r-\boldsymbol{l}_t|_{L_1/L_2},$$

即向量 $\boldsymbol{l}_h+\boldsymbol{l}_r$ 和 $\boldsymbol{l}_t$ 的 L_1 或 L_2 距离.

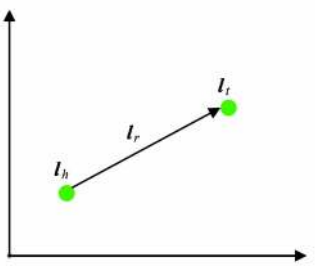


Fig. 2 TransE model.

图 2 TransE 模型

在实际学习过程中，为了增强知识表示的区分能力，TransE 采用最大间隔方法，定义了如下优化目标函数：

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

其中，S 是合法三元组的集合，S^- 为错误三元组的集合，$\max(x,y)$返回 x 和 y 中较大的值，γ 为合法三元组得分与错误三元组得分之间的间隔距离.

错误三元组并非随机产生的，为了选取有代表性的错误三元组，TransE 将 S 中每个三元组的头实体、关系和尾实体其中之一随机替换成其他实体或关系来得到 S^-，即：

$$S^-=\{(h',r,t)\}\cup\{(h,r',t)\}\cup\{(h,r,t')\}.$$

与以往模型相比，TransE 模型参数较少，计算复杂度低，却能直接建立实体和关系之间的复杂语义联系. Bordes 等人在 WordNet 和 Freebase 等数据集上进行链接预测等评测任务，实验表明 TransE 的性能较以往模型有显著提升. 特别是在大规模稀疏知识图谱上，TransE 的性能尤其惊人.

由于 TransE 简单有效，自提出以来，有大量研究工作对 TransE 进行扩展和应用. 可以说，TransE 已经成为知识表示学习的代表模型. 在第 3 节，我们将以 TransE 为例，介绍知识表示学习的主要挑战与解决方案.

2.8 其他模型

在 TransE 提出之后，大部分知识表示学习模型是以 TransE 为基础的扩展. 在 TransE 扩展模型以外，这里主要介绍全息表示模型(holographic embeddings, Hole)[46].

Hole 提出使用头、尾实体向量的“循环相关”操作来表示该实体对. 这里，循环相关 $*:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^d$ 操作如下：

$$[\boldsymbol{l}_h*\boldsymbol{l}_t]_k=\sum_{i=0}^{d-1}\boldsymbol{l}_{h_i}\boldsymbol{l}_{t_{(i+k)\bmod d}},$$

循环相关操作可以看作张量乘法特殊形式，具有较强的表达能力，具有以下 3 个优点：1)不可交换性. 循环相关是不可交换的，即 $\boldsymbol{l}_h*\boldsymbol{l}_t\neq\boldsymbol{l}_t*\boldsymbol{l}_h$. 而知识库中很多关系是不可交换的，因此该特点具有重要意义. 2)相关性. 循环相关操作得到的向量每一维都衡量了向量 $\boldsymbol{l}_h$ 和 $\boldsymbol{l}_t$ 的某种相似性. 例如，循环相关的第一位$[\boldsymbol{l}_h*\boldsymbol{l}_t]_0=\sum_{i=0}^{d-1}\boldsymbol{l}_{h_i}\boldsymbol{l}_{t_i}$ 相当于向量 $\boldsymbol{l}_h$ 和 $\boldsymbol{l}_t$ 的内积. 该性质处理头、尾实体比较相似的关系(例如“夫妻”关系)时具有重要意义. 3)计算效率高. 循环相关操作还可以使用如下公式进行优化：

$$\boldsymbol{l}_h*\boldsymbol{l}_t=F^{-1}(\overline{F}(\boldsymbol{l}_h)\odot F(\boldsymbol{l}_t)),$$

这里 $F(x)$，$F^{-1}(x)$为傅里叶变换与逆傅里叶变换，可以用快速傅里叶变换加速计算.

对于每个三元组(h,r,t)，Hole 定义了如下评分函数：

$$f_r(h,t)=\sigma(\boldsymbol{l}_r^{\mathrm{T}}(\boldsymbol{l}_h*\boldsymbol{l}_t)),$$

这里 $\sigma(x)=\dfrac{1}{1+\mathrm{e}^{-x}}$为 sigmoid 函数.

由于该模型刚刚提出，尚未验证其效果，但是无疑为知识表示学习提供了全新的视角，值得关注.

3 知识表示学习的主要挑战与已有解决方案

以 TransE 为代表的知识表示学习模型，已经在知识图谱补全、关系抽取等任务中取得了瞩目成果. 但是，知识表示学习仍然面临很多挑战. 这里我们以 TransE 为代表模型，总结认为 TransE 面临的 3 个主要挑战，目前已有相关工作提出一些解决方案，具体介绍如下.

3.1 复杂关系建模

TransE 由于模型简单，在大规模知识图谱上效果明显. 但是也由于过于简单，导致 TransE 在处理知识库的复杂关系时捉襟见肘.

这里的复杂关系定义如下. 按照知识库中关系两端连接实体的数目，可以将关系划分为 1-1，1-N，N-1 和 N-N 四种类型[41]. 例如 N-1 类型关系指的是，该类型关系中的一个尾实体会平均对应多个头实体，即 $\forall i\in\{0,1,\cdots,m\}$，$(h_i,r,t)\in S$. 我们将 1-$N$，$N$-1 和 N-N 称为复杂关系.

研究发现,各种知识获取算法在处理 4 种类型关系时的性能差异较大[41]. 以 TransE 为例,在处理复杂关系时性能显著降低,这与 TransE 模型假设有密切关系. 根据 TransE 的优化目标,面向 1-N,N-1 和 N-N 三种类型关系,我们可以推出以下结论:如果关系 r 是 N-1 关系,我们将会得到 $\boldsymbol{l}_{h_0} \approx \boldsymbol{l}_{h_1} \approx \cdots \approx \boldsymbol{l}_{h_m}$. 同样,这样的问题在关系 r 是 N-1 关系时也会发生,得到 $\boldsymbol{l}_{t_0} \approx \boldsymbol{l}_{t_1} \approx \cdots \approx \boldsymbol{l}_{t_m}$.

例如,假如知识库中有 2 个三元组,分别是(美国,总统,奥巴马)和(美国,总统,布什). 这里的关系"总统"是典型的 1-N 的复杂关系. 如果用 TransE 从这 2 个三元组学习知识表示,如图 3 所示,将会使奥巴马和布什的向量变得相同.

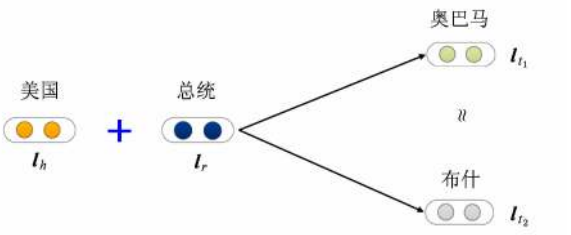


Fig. 3 The example of complex relations.

图 3 复杂关系示例

这显然不符合事实:奥巴马和布什除了作为美国总统这个身份上比较相似外,其他很多方面都不尽相同. 因此,由于这些复杂关系的存在,导致 TransE 学习得到的实体表示区分性较低.

那么应当如何实现表示学习对复杂关系的建模呢? 最近有大量关于 TransE 的扩展模型尝试解决这一挑战问题. 这里我们简要介绍其中 7 个代表模型.

3.1.1 TransH 模型[47]

为了解决 TransE 模型在处理 1-N,N-1,N-N 复杂关系时的局限性,TransH 模型提出让一个实体在不同的关系下拥有不同的表示.

如图 4 所示,对于关系 r,TransH 模型同时使用平移向量 $\boldsymbol{l}_r$ 和超平面的法向量 $\boldsymbol{w}_r$ 来表示它. 对于一个三元组(h,r,t),TransH 首先将头实体向量 $\boldsymbol{l}_h$ 和尾实体向量 $\boldsymbol{l}_t$ 沿法线 $\boldsymbol{w}_r$ 投影到关系 r 对应的超平面上,用 $\boldsymbol{l}_{h_r}$ 和 $\boldsymbol{l}_{t_r}$ 表示如下:

$$\boldsymbol{l}_{h_r} = \boldsymbol{l}_h - \boldsymbol{w}_r^{\mathrm{T}} \boldsymbol{l}_h \boldsymbol{w}_r,$$

$$\boldsymbol{l}_{t_r} = \boldsymbol{l}_t - \boldsymbol{w}_r^{\mathrm{T}} \boldsymbol{l}_t \boldsymbol{w}_r,$$

因此 TransH 定义了如下损失函数:

$$f_r(h,t) = \left\| \boldsymbol{l}_{h_r} + \boldsymbol{l}_r - \boldsymbol{l}_{t_r} \right\|_{L_1/L_2}.$$

需要注意的是,由于关系 r 可能存在无限个超平面,TransH 简单地令 $\boldsymbol{l}_r$ 与 $\boldsymbol{w}_r$ 近似正交来选取某一个超平面.

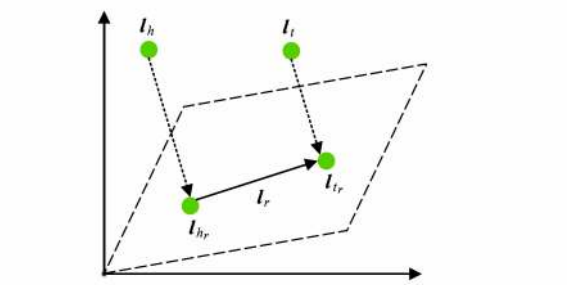


Fig. 4 TransH model.

图 4 TransH 模型

3.1.2 TransR / CTransR 模型[48]

虽然 TransH 模型使每个实体在不同关系下拥有了不同的表示,它仍然假设实体和关系处于相同的语义空间 $\mathbb{R}^d$ 中,这一定程度上限制了 TransH 的表示能力. TransR 模型则认为,一个实体是多种属性的综合体,不同关系关注实体的不同属性. TransR 认为不同的关系拥有不同的语义空间. 对每个三元组,首先应将实体投影到对应的关系空间中,然后再建立从头实体到尾实体的翻译关系.

如图 5 所示是 TransR 模型的简单示例. 对于每个三元组(h,r,t),我们首先将实体向量向关系 r 空间投影. 原来在实体空间中与头、尾实体(用圆圈表示)相似的实体(用三角形表示),在关系 r 空间中被区分开了.

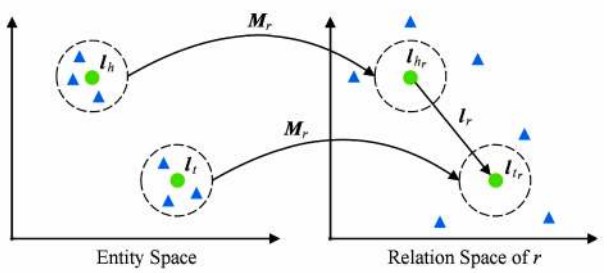


Fig. 5 TransR model.

图 5 TransR 模型

具体而言,对于每一个关系 r,TransR 定义投影矩阵 $\boldsymbol{M}_r \in \mathbb{R}^{d \times k}$,将实体向量从实体空间投影到关系 r 的子空间,用 $\boldsymbol{l}_{h_r}$ 和 $\boldsymbol{l}_{t_r}$ 表示如下:

$$\boldsymbol{l}_{h_r} = \boldsymbol{l}_h \boldsymbol{M}_r,$$

$$\boldsymbol{l}_{t_r} = \boldsymbol{l}_t \boldsymbol{M}_r,$$

然后使 $\boldsymbol{l}_{h_r} + \boldsymbol{l}_r \approx \boldsymbol{l}_{t_r}$. 因此,TransR 定义了如下损失函数:

$$f_r(h,t) = \left\| \boldsymbol{l}_{h_r} + \boldsymbol{l}_r - \boldsymbol{l}_{t_r} \right\|_{L_1/L_2}.$$

相关研究还发现,某些关系还可以进行更细致的划分. 例如 Freebase 中的"/location/location/contains"关系,可能是一个国家包含一个城市,可能

是一个国家包含一所大学,也可能是一个州包含一个城市等.如果将该关系做更细致的划分,就可以更精确地建立投影关系.

因此,Lin 等人进一步提出了 CTransR 模型,通过把关系 r 对应的实体对的向量差值 $\boldsymbol{l}_h-\boldsymbol{l}_t$ 进行聚类,将关系 r 细分为多个子关系 r_c.CTransR 模型为每一个子关系 r_c 分别学习向量表示,对于每个三元组 (h,r,t),定义了如下损失函数:

$$f_r(h,t)=\|\boldsymbol{l}_{h_r}+\boldsymbol{l}_{r_c}-\boldsymbol{l}_{t_r}\|_{L_1/L_2}.$$

3.1.3 TransD 模型[49]

虽然 TransR 模型较 TransE 和 TransH 有显著改进,它仍然有很多缺点:

1) 在同一个关系 r 下,头、尾实体共享相同的投影矩阵.然而,一个关系的头、尾实体的类型或属性可能差异巨大.例如,对于三元组(美国,总统,奥巴马),美国和奥巴马的类型完全不同,一个是国家,一个是人物.

2) 从实体空间到关系空间的投影是实体和关系之间的交互过程,因此 TransR 让投影矩阵仅与关系有关是不合理的.

3) 与 TransE 和 TransH 相比,TransR 由于引入了空间投影,使得 TransR 模型参数急剧增加,计算复杂度大大提高.

为了解决这些问题,Ji 等人提出了 TransD 模型.如图 6 所示:

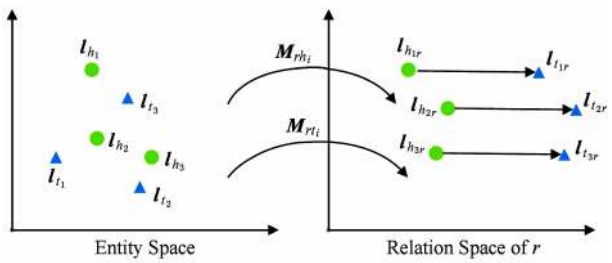


Fig. 6 TransD model.

图 6 TransD 模型

给定三元组 (h,r,t),TransD 模型设置了 2 个分别将头实体和尾实体投影到关系空间的投影矩阵 $\boldsymbol{M}_{rh}$ 和 $\boldsymbol{M}_{rt}$,具体定义如下:

$$\boldsymbol{M}_{rh}=\boldsymbol{l}_{r_p}\boldsymbol{l}_{h_p}+\boldsymbol{I}^{d\times k},$$

$$\boldsymbol{M}_{rt}=\boldsymbol{l}_{r_p}\boldsymbol{l}_{t_p}+\boldsymbol{I}^{d\times k},$$

这里 $\boldsymbol{l}_{h_p},\boldsymbol{l}_{t_p}\in\mathbb{R}^d,\boldsymbol{l}_{r_p}\in\mathbb{R}^k$,下标 p 代表该向量为投影向量.显然,$\boldsymbol{M}_{rh}$ 和 $\boldsymbol{M}_{rt}$ 与实体和关系均相关.而且,利用 2 个投影向量构建投影矩阵,解决了原来 TransR 模型参数过多的问题.最后,TransD 模型定义了如下损失函数:

$$f_r(h,t)=\|\boldsymbol{l}_h\boldsymbol{M}_{rh}+\boldsymbol{l}_r-\boldsymbol{l}_t\boldsymbol{M}_{rt}\|_{L_1/L_2}.$$

3.1.4 TranSparse 模型[50]

知识库中实体和关系的异质性和不平衡性是制约知识表示学习的难题:

1) 异质性.知识库中某些关系可能会与大量的实体有连接,而某些关系则可能仅仅与少量实体有连接.

2) 不均衡性.在某些关系中,头实体和尾实体的种类和数量可能差别巨大.例如,"国籍"这个关系的头实体是成千上万不同的人物,而尾实体只有几百个国家.

为了解决实体和关系的异质性,TranSparse 提出使用稀疏矩阵代替 TransR 模型中的稠密矩阵,其中矩阵 $\boldsymbol{M}_r$ 的稀疏度由关系 r 连接的实体对数量决定.这里头、尾实体共享同一个投影矩阵 $\boldsymbol{M}_r$.投影矩阵 $\boldsymbol{M}_r(\theta_r)$ 的稀疏度 θ_r 定义如下:

$$\theta_r=1-(1-\theta_{\min})N_r/N_{r^*},$$

其中,$0\leqslant\theta_{\min}\leqslant1$ 为计算稀疏度的超参数,N_r 表示关系 r 连接的实体对数量,r^* 表示连接实体对数量最多的关系.这样,投影向量可定义为

$$\boldsymbol{l}_{h_r}=\boldsymbol{l}_h\boldsymbol{M}_r(\theta_r),$$

$$\boldsymbol{l}_{t_r}=\boldsymbol{l}_t\boldsymbol{M}_r(\theta_r).$$

为了解决关系的不平衡性问题,TranSparse 对于头实体和尾实体分别使用 2 个不同的投影矩阵 $\boldsymbol{M}_r^h(\theta_r^h)$ 和 $\boldsymbol{M}_r^t(\theta_r^t)$.两者的稀疏度定义如下:

$$\theta_r^l=1-(1-\theta_{\min})N_r^l/N_{r^*}^l,$$

其中,N_r^l 表示关系 r 在位置 l 处连接不同实体的数量(l 可能是头实体或尾实体),$N_{r^*}^l$ 表示 N_r^l 中最大的数.这样,投影向量可定义为

$$\boldsymbol{l}_{h_r}=\boldsymbol{l}_h\boldsymbol{M}_r^h(\theta_r^h),$$

$$\boldsymbol{l}_{t_r}=\boldsymbol{l}_t\boldsymbol{M}_r^t(\theta_r^t).$$

TranSparse 对于以上 2 种形式,均定义如下损失函数:

$$f_r(h,t)=\|\boldsymbol{l}_{h_r}+\boldsymbol{l}_r-\boldsymbol{l}_{t_r}\|_{L_1/L_2}.$$

3.1.5 TransA 模型[51]

Xiao 等人认为 TransE 及其之后的扩展模型均存在 2 个重要问题:1)损失函数只采用 L_1 或 L_2 距离,灵活性不够;2)损失函数过于简单,实体和关系向量的每一维等同考虑.

为了解决这 2 个问题,Xiao 等人提出 TransA 模型,将损失函数中的距离度量改用马氏距离,并为每一维学习不同的权重.对于每个三元组 (h,r,t),TransA 模型定义了如下评分函数:

$$f_r(h,t)=(\boldsymbol{l}_h+\boldsymbol{l}_r-\boldsymbol{l}_t)^{\mathrm{T}}\boldsymbol{W}_r(\boldsymbol{l}_h+\boldsymbol{l}_r-\boldsymbol{l}_t),$$

其中 $\boldsymbol{W}_r$ 为与关系 r 相关的非负权值矩阵.

如图 7 所示,(h_1,r_1,t_1)和(h_2,r_2,t_2)两个合法三元组,t_3 是错误的尾实体.如果使用欧氏距离,如图 7(a)所示,错误的实体 t_3 会被预测出来.而如图 7(b)所示,TransA 模型通过对向量不同维度进行加权,正确的实体由于在 x 轴或者 y 轴上距离较近,从而能够被正确预测.

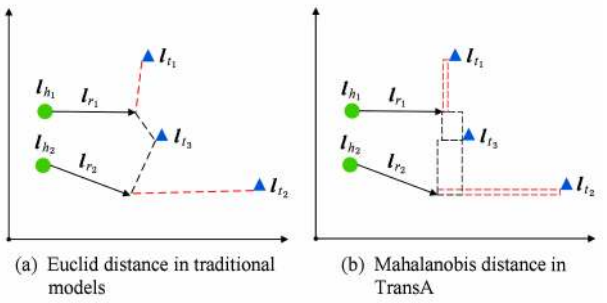


Fig. 7 Comparison between traditional models and TransA.

图 7 传统模型和 TransA 模型比较

3.1.6 TransG 模型[52]

TransG 模型提出使用高斯混合模型描述头、尾实体之间的关系.该模型认为,一个关系会对应多种语义,每种语义用一个高斯分布来刻画,即:

$$\boldsymbol{l}_t-\boldsymbol{l}_h \mid \boldsymbol{l}_r \sim \sum_{m=1}^{M}\pi_{r,m}N(\boldsymbol{\mu}_{r,m},\boldsymbol{I}),$$

其中 $\boldsymbol{I}$ 表示单位矩阵.

TransG 模型与传统模型的对比如图 8 所示.其中三角形表示正确的尾实体,圆形表示错误的尾实体.图 8(a)中为传统模型示例,由于将关系 r 的所有语义混为一谈,导致错误的实体无法被区分开.而如图 8(b)所示,TransG 模型通过考虑关系 r 的不同语义,形成多个高斯分布,就能够区分出正确和错误实体.

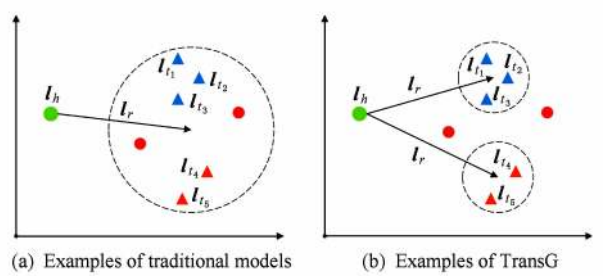


Fig. 8 Comparison between traditional models and TransG.

图 8 传统模型和 TransG 模型比较

3.1.7 KG2E 模型[53]

He 等人认为,知识库中的关系和实体的语义本身具有不确定性,这在过去模型中被忽略了.因此,He 等人提出 KG2E,使用高斯分布来表示实体和关系.其中高斯分布的均值表示的是实体或关系在语义空间中的中心位置,而高斯分布的协方差则表示该实体或关系的不确定度.

图 9 为 KG2E 模型示例,每个圆圈代表不同实体与关系的表示,它们分别与"比尔·克林顿"构成三元组,其中圆圈大小表示的是不同实体或关系的不确定度,可以看到"国籍"的不确定度远远大于其他关系.

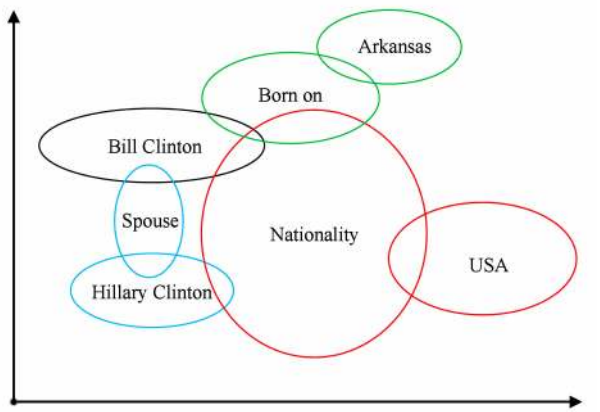


Fig. 9 KG2E model.

图 9 KG2E 模型

KG2E 使用 $\boldsymbol{l}_h-\boldsymbol{l}_t$ 来表示头、尾实体之间的关系.这里 $\boldsymbol{l}_h-\boldsymbol{l}_t$ 可以用一个概率分布来表示:

$$P_e\sim N(\boldsymbol{\mu}_h-\boldsymbol{\mu}_t,\Sigma_h+\Sigma_t),$$

而关系 r 同样是一个高斯分布 $P_r\sim N(\boldsymbol{\mu}_r,\Sigma_r)$.因此,可以根据 2 个概率分布 P_e 和 P_r 的相似度来估计三元组的评分.KG2E 考虑 2 种计算概率相似度的办法:KL 距离和期望概率.

KL 距离是一种不对称相似度,其得分函数定义如下:

$$f_r(h,t)=\int_{x\in\mathbf{R}^{k_e}}N(x;\boldsymbol{\mu}_r,\Sigma_r)\log\frac{N(x;\boldsymbol{\mu}_e,\Sigma_e)}{N(x;\boldsymbol{\mu}_r,\Sigma_r)}\mathrm{d}x=$$

$$\frac{1}{2}\{tr(\Sigma_r^{-1}\Sigma_e)+(\boldsymbol{\mu}_e-\boldsymbol{\mu}_r)^{\mathrm{T}}\Sigma_r^{-1}(\boldsymbol{\mu}_e-\boldsymbol{\mu}_r)-$$

$$\log\frac{\det(\Sigma_e)}{\det(\Sigma_r)}+k_e\}.$$

期望概率是一种对称相似度,其得分函数定义如下:

$$f_r(h,t)=\int_{x\in\mathbf{R}^{k_e}}N(x;\boldsymbol{\mu}_r,\Sigma_r)N(x;\boldsymbol{\mu}_e,\Sigma_e)\mathrm{d}x=$$

$$\frac{1}{2}\{(\boldsymbol{\mu}_e-\boldsymbol{\mu}_r)^{\mathrm{T}}(\Sigma_e+\Sigma_r)^{-1}(\boldsymbol{\mu}_e-\boldsymbol{\mu}_r)+$$

$$\log(\det(\Sigma_e+\Sigma_r))+k_e\log 2\pi\}.$$

需要注意的是,为了防止过拟合,KG2E 使用了对参数进行了强制限制:

$$\forall\, l\in E\cup R, c_{\min}\boldsymbol{I}\leqslant\Sigma_l\leqslant c_{\max}\boldsymbol{I}, c_{\min}>0.$$

3.1.8 小结

可以看到,在 TransE 之后,在如何处理复杂关系建模的挑战问题上,提出了 TransH, TransR, TransD, TranSparse, TransA, TransG 和 KG2E 等多种模型,从不同角度尝试解决复杂关系建模问题,可谓百花齐放.在相关数据集合上的实验表明,这些方法均较 TransE 有显著的性能提升,验证了这些方法的有效性.

3.2 多源信息融合

知识表示学习面临的另外一个重要挑战,是如何实现多源信息融合.现有的知识表示学习模型如 TransE 等,仅利用知识图谱的三元组结构信息进行表示学习,尚有大量与知识有关的其他信息没有得到有效利用,例如:

1) 知识库中的其他信息,如实体和关系的描述信息、类别信息等.

2) 知识库外的海量信息,如互联网文本蕴含了大量与知识库实体和关系有关的信息.

这些海量的多源异质信息可以帮助改善数据稀疏问题,提高知识表示的区分能力.如何充分融合这些多源异质信息,实现知识表示学习,具有重要意义.

在融合上述信息进行知识表示学习方面,已经有一些研究工作,但总体来讲还处于起步状态,这里简单介绍其中 2 个代表性工作.

3.2.1 DKRL 模型

考虑实体描述的知识表示学习模型(description-embodied knowledge representation learning, DKRL)[54]提出在知识表示学习中考虑 Freebase 等知识库中提供的实体描述文本信息.在文本表示方面,DKRL 考虑了 2 种模型:一种是 CBOW[8,44],将文本中的词向量简单相加作为文本表示;一种是卷积神经网络(convolutional neural network, CNN)[55-56],能够考虑文本中的词序信息.

如图 10 和图 11 所示,DKRL 可以利用 CBOW 和 CNN 根据实体描述文本得到实体表示,然后将该实体表示用于 TransE 的目标函数学习.

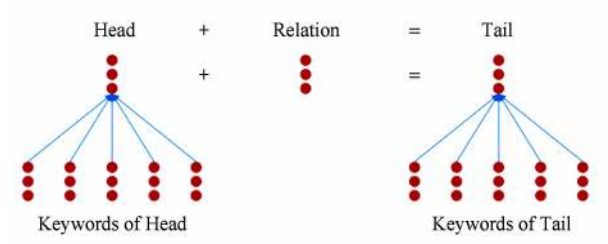


Fig. 10 DKRL (CBOW) model.

图 10 DKRL(CBOW)模型

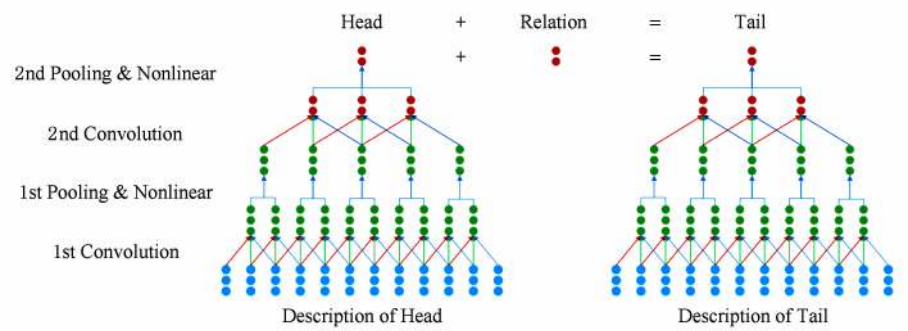


Fig. 11 DKRL (CNN) model.

图 11 DKRL(CNN)模型

DKRL 的优势在于,除了能够提升实体表示的区分能力外,还能实现对新实体的表示.当新出现一个未曾在知识库中的实体时,DKRL 可以根据它的简短描述产生它的实体表示,用于知识图谱补全等任务.这对于不断扩充知识图谱具有重要意义.

3.2.2 文本与知识库融合的知识表示学习[47]

Wang 等人提出在表示学习中考虑文本数据,利用 word2vec 学习维基百科正文中的词表示,利用 TransE 学习知识库中的知识表示.同时,利用维基百科正文中的链接信息(锚文本与实体的对应关系),让文本中实体对应的词表示与知识库中的实体表示尽可能接近,从而实现文本与知识库融合的表示学习.Wang 等人还将类似的想法用于融合实体描述信息[57].

3.2.3 小结

已有工作表明,多源信息融合能够有效提升知识表示的性能,特别是可以有效处理新实体的表示问题.但是,也可以看出,多源信息融合的知识表示学习仍处于非常起步的阶段,相关工作较少,考虑的信息源非常有限,有大量的信息(如实体类别等)未被考虑,具有广阔的研究前景.

3.3 关系路径建模

在知识图谱中,多步的关系路径也能够反映实

体之间的语义关系. Lao 等人曾提出 Path-Constraint Random Walk[58],Path Ranking Algorithm[59]等算法,利用两实体间的关系路径信息预测它们的关系,取得显著效果,说明关系路径蕴含着丰富的信息.

为了突破 TransE 等模型孤立学习每个三元组的局限性,Lin 等人提出考虑关系路径的表示学习方法,以 TransE 作为扩展基础,提出 Path-based TransE(PTransE)模型.

图 12 展示的是 PTransE 考虑 2 步关系路径的示例. PTransE 模型面临的挑战在于:

1) 并不是所有的实体间的关系路径都是可靠的. 为此,PTransE 提出 Path-Constraint Resource Allocation 图算法度量关系路径的可靠性.

2) PTransE 需要建立关系路径的向量表示,参与从头实体到尾实体的翻译过程. 这是典型的组合语义问题,需要对路径上所有关系的向量进行语义组合产生路径向量. PTransE 尝试了 3 种代表性的语义组合操作,分别是相加、按位相乘和循环神经网络. 相关数据实验表明,相加的组合操作效果最好.

几乎同时,也有其他研究团队在知识表示学习中成功考虑了关系路径的建模[60]. 关系路径的表示学习也被用来进行基于知识库的自动问答[61].

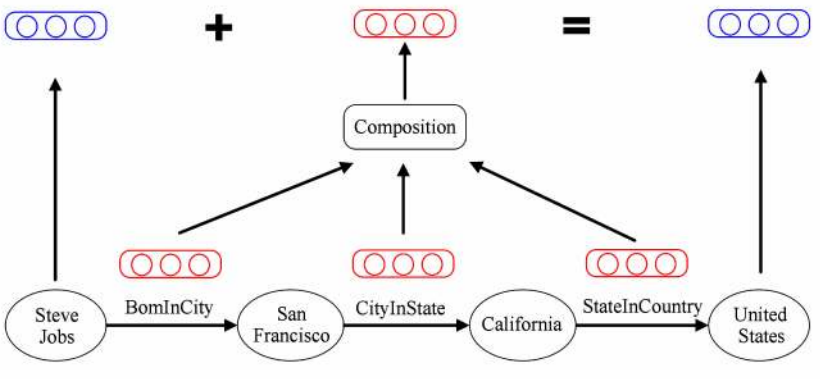


Fig. 12 PTransE model.

图 12 PTransE 模型

PTransE 等研究的实验表明,考虑关系路径能够极大提升知识表示学习的区分性,提高在知识图谱补全等任务上的性能. 关系路径建模工作还比较初步,在关系路径的可靠性计算、关系路径的语义组合操作等方面,还有很多细致的考察工作需要完成.

4 知识表示学习未来研究方向展望

近年来知识表示学习已经崭露头角,在很多任务中展现了巨大的应用潜力. 对于 TransE 等模型面临的挑战,也已经提出了很多改进方案. 然而,知识表示学习距离真正实用还很远,本节将对知识表示学习的未来方向进行展望.

4.1 面向不同知识类型的知识表示学习

如 3.1 节所述,已有工作将知识库的关系划分为 1-1,1-*N*,*N*-1 和 *N*-*N* 四类,并面向复杂关系建模开展了大量研究工作. 研究表明,面向不同类型的关系,需要设计专门的知识表示模型.

然而,1-1,1-*N*,*N*-1 和 *N*-*N* 的关系类型划分略显粗糙,无法直观地解释知识的本质类型特点. 我们需要面向知识表示任务,有针对性地设计知识类型划分标准.

近期发表在 Science 等权威期刊的认知科学研究成果[62-63]总结认为,人类知识包括以下 4 种结构:1)树状关系,表示实体间的层次分类关系(如生物界的分类系统等);2)二维网格关系,表示现实世界的空间信息(如地理位置信息等);3)单维顺序关系,表示实体间的偏序关系(如政治家的左右倾谱系分布等);4)有向网络关系,表示实体间的关联或因果关系(如疾病之间的传染关系等).

认知科学关于人类知识类型的总结,与许多知识库的组织形式有一定契合之处,但不完全相同. 例如 Freebase 等大部分知识库采用有向网络结构(即三元组形式)组织人类知识[2];WordNet 则首先将同义词聚集成同义词集合(Synset),然后再以同义词集合为单位用有向网络结构表示集合之间的关系(如上下位关系、整体-部分关系等)[1]. 在大部分知识库中,树状关系等类型的知识均用有向网络表示,这并不利于在知识表示中充分利用不同类型知识的结构特点.

认知科学对人类知识类型的总结,有助于对知识图谱中知识类型的划分和处理. 未来有必要结合

人工智能和认知科学的最新研究成果,有针对性地设计知识类型划分标准,开展面向不同复杂关系类型的知识表示学习研究.

4.2 多源信息融合的知识表示学习

在多元融合的知识表示学习方面,相关工作还比较有限,主要是考虑实体描述的知识表示学习模型,以及文本与知识库融合的知识表示学习,这些模型无论是信息来源,还是融合手段都非常有限.

我们认为在多源信息融合的知识表示学习方面,有以下 3 个方面的工作需要开展:

1) 融合知识库中实体和关系的其他信息. 知识库中拥有关于实体和关系的丰富信息,如描述文本、层次类型等. 有机融合这些信息,将显著提升知识表示学习的表示能力.

2) 融合互联网文本信息. 互联网海量文本数据是知识库的重要知识来源. 人们提出远程监督(distant supervision)[64-68]、开放信息抽取(open information extraction)[69-73]等技术,从开放文本中抽取知识. 这个过程也自然而然地建立起了知识库和文本之间的联系,如何充分利用这些联系融合互联网文本信息,意义重大. 值得一提的是,目前大部分工作主要关注面向实体表示的融合. 实际上,若干研究工作已经利用卷积神经网络(CNN)建立起了关系表示[74-76],这为面向关系表示的信息融合提供了技术基础,最终实现融合文本信息和知识库的知识表示.

3) 融合多知识库信息. 人们利用不同的信息源构建了不同的知识库. 如何对多知识库信息进行融合表示,对于建立统一的大规模知识库意义重大. 融合多源知识库信息,主要涉及实体融合、关系融合与事实融合.

首先,由于存在大量别名现象,需对多信息源进行实体对齐和关系对齐. 这在分布式表示中,是典型的多表示空间投影问题,可以采用学习匹配(learning to match)[77]等思想,利用 PSI(polynomial semantic indexing)[78], SSI (supervised semantic indexing)[79]等技术,建立多源表示空间投影关系,实现实体对齐与关系对齐. 此外,还可以充分利用多表示空间之间的一致性,实现多空间协同映射(collective fusion).

然后,在实体对齐和关系对齐的基础上,可对多信息源获取的知识进行融合. 由于大量知识来自海量互联网文本,无法确保获取知识的真实性,存在大量互相矛盾的知识. 可以综合考虑信息源可信性、多信息源一致性等要素,利用 TrustRank[80]等可信性度量技术,检测实体间知识的矛盾并分别度量其可信性,实现知识融合,建成统一的大规模知识库及其融合表示.

4.3 考虑复杂推理模式的知识表示学习

考虑关系路径的知识表示学习,实际上是充分利用了两实体间的关系和关系路径之间的推理模式,来为表示学习模型提供更精确的约束信息. 例如,根据三元组(康熙,父亲,雍正)和(雍正,父亲,乾隆)构成的"康熙"和"乾隆"之间"父亲+父亲"的关系路径,再结合三元组(康熙,祖父,乾隆),PTransE 实际上额外提供了"父亲+父亲=祖父"的推理模式,从而提升知识表示的精确性.

实际上,关系路径只是复杂推理模式中的一种特殊形式,它要求头实体和尾实体必须保持不变. 但实际上,知识库中还有其他形式的推理模式,例如三元组(美国,总统,奥巴马)和(奥巴马,是,美国人)之间就存在着推理关系,但是两者的头、尾实体并不完全一致. 如果能将这些复杂推理模式考虑到知识表示学习中,将能更进一步提升知识表示的性能.

在该问题中,如何总结和表示这些复杂推理模式是关键难题. 目前来看,一阶逻辑(first-order logic, FOL)是对复杂推理模式的较佳表示方案,未来我们需要探索一阶逻辑的分布式表示及其融合到知识表示学习中的技术方案.

4.4 其他研究方向

除了以上 3 个主要研究方向,还有很多关于知识表示学习的研究工作亟待开展,简单总结如下:

1) 面向大规模知识库的在线学习和快速学习. 大规模知识库稀疏性很强. 初步实验表明,已有表示学习模型在大规模知识库上效果堪忧,特别是对低频实体和关系的表示效果较差,而且知识库规模不断扩大,我们需要设计高效的在线学习方案. 除了充分融合多源信息降低稀疏性之外,我们还可以探索如何优化表示学习的样例顺序,借鉴 Curriculum Learning[81]等算法思想,优先学习核心知识,然后学习外围知识,也许能够一定程度改善表示效果.

2) 基于知识分布式表示的应用. 知识表示学习还处于起步阶段,在知识获取、融合和推理等方向均有广阔的应用空间. 我们需要在若干重要任务上探索和验证知识表示学习的有效性. 例如,关系抽取任务如果能够基于知识表示学习有效利用知识库信息,将能够极大提升抽取性能和覆盖面. 再如,我们可以充分利用表示学习在信息融合上的优势,实现跨领域和跨语言的知识融合. 此外,人脑强大的学习与推理能力[82],说明在低维语义空间中进行知识的学习与推理极具潜力,相关机理值得深入探索.

5 结束语

通过以上对知识表示学习相关代表方法的梳理,我们认为知识表示学习具有重要意义:现有知识库的构建与应用主要依赖于离散符号表示,而分布式表示方案则为实体与关系的语义信息的统一精确表示提供了可行方案.在分布式表示学习的支持下,将极大推动知识的自动获取、融合与推理能力,实现知识库更加广泛而深入的应用.

本文还对知识表示学习面临的主要挑战、已有解决方案以及未来研究方向进行了总结.我们认为知识表示学习虽然展现出巨大的潜力,但距离广泛应用还有很长的路要走.可以说,大规模知识表示学习是人工智能学科发展的学术前沿问题,是智能信息处理和服务发展的基础技术保障.知识表示学习技术将推动人工智能学科、智能信息服务产业以及创新社会管理与社会服务的发展.

期待更多研究者加入到知识表示学习的研究行列中来,希望本文对于知识表示学习在国内的研究发展提供一些帮助.

参 考 文 献

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

[2] Bollacker K, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge [C] //Proc of KDD. New York: ACM, 2008: 1247-1250

[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] Bengio Y. Learning deep architectures for AI [J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1-127

[5] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828

[6] Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning [C] //Proc of ACL. Stroudsburg, PA: ACL, 2010: 384-394

[7] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval [M]. Cambridge, UK: Cambridge University Press, 2008

[8] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C] //Proc of NIPS. Cambridge, MA: MIT Press, 2013: 3111-3119

[9] Zhao Y, Liu Z, Sun M. Phrase type sensitive tensor indexing model for semantic composition [C] //Proc of AAAI. Menlo Park, CA: AAAI, 2015: 2195-2202

[10] Zhao Y, Liu Z, Sun M. Representation learning for measuring entity relatedness with rich information [C] //Proc of IJCAI. San Francisco, CA: Morgan Kaufmann, 2015: 1412-1418

[11] Hu B, Lu Z, Li H, et al. Convolutional neural network architectures for matching natural language sentences [C] //Proc of NIPS. San Francisco, CA: Morgan Kaufmann, 2014: 2042-2050

[12] Le Q V, Mikolov T. Distributed representations of sentences and documents [C] //Proc of ICML. New York: ACM, 2014: 873-882

[13] Blunsom P, Grefenstette E, Kalchbrenner N, et al. A convolutional neural network for modelling sentences [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2014. arXiv: 1402.2188

[14] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations [C] //Proc of KDD. New York: ACM, 2014: 701-710

[15] Tang J, Qu M, Wang M, et al. LINE: Large-scale information network embedding [C] //Proc of Int World Wide Web Conferences Steering Committee. New York: ACM, 2015: 1067-1077

[16] Yang C, Liu Z, Zhao D, et al. Network representation learning with rich text information [C] //Proc of IJCAI. San Francisco, CA: Morgan Kaufmann, 2015

[17] Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes [C] //Proc of ACL. Stroudsburg, PA: ACL, 2012: 873-882

[18] Reisinger J, Mooney R J. Multi-prototype vector-space models of word meaning [C] //Proc of HLT-NAACL. Stroudsburg, PA: ACL, 2010: 109-117

[19] Tian F, Dai H, Bian J, et al. A probabilistic model for learning multi-prototype word embeddings [C] //Proc of COLING. New York: ACM, 2014: 151-160

[20] Socher R, Bauer J, Manning C D, et al. Parsing with compositional vector grammars [C] //Proc of ACL. Stroudsburg, PA: ACL, 2013: 455-465

[21] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2013: 1642-1653

[22] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces [C] //Proc of EMNLP-CoNLL. Stroudsburg, PA: ACL, 2012: 1201-1211

[23] Luong M, Socher R, Manning C. Better word representations with recursive neural networks for morphology [C] //Proc of CoNLL. Stroudsburg, PA: ACL, 2013: 104-113

[24] Botha J A, Blunsom P. Compositional morphology for word representations and language modelling [C] //Proc of ICML. New York: ACM, 2014

[25] Chen X, Xu L, Liu Z, et al. Joint learning of character and word embeddings [C] //Proc of IJCAI. San Francisco, CA: Morgan Kaufmann, 2015: 1236-1242

[26] Klementiev A, Titov I, Bhattarai B. Inducing crosslingual distributed representations of words [C] //Proc of COLING. New York: ACM, 2012: 1459-1474

[27] Lauly S, Larochelle H, Khapra M, et al. An autoencoder approach to learning bilingual word representations [C] //Proc of NIPS. San Francisco, CA: Morgan Kaufmann, 2014: 1853-1861

[28] Shi T, Liu Z, Liu Y, et al. Learning cross-lingual word embeddings via matrix co-factorization [C] //Proc of ACL. Stroudsburg, PA: ACL, 2015: 567-574

[29] Luo H, Liu Z, Luan H, et al. Online learning of interpretable word embeddings [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2015: 1687-1692

[30] Murphy B, Talukdar P P, Mitchell T M. Learning effective and interpretable semantic models using non-negative sparse embedding [C] //Proc of COLING. New York: ACM, 2012: 1933-1950

[31] Fyshe A, Talukdar P P, Murphy B, et al. Interpretable semantic vectors from a joint model of brain-and text-based meaning [C] //Proc of ACL. Stroudsburg, PA: ACL, 2014: 489-499

[32] Faruqui M, Tsvetkov Y, Yogatama D, et al. Sparse overcomplete word vector representations [C] //Proc of ACL. Stroudsburg, PA: ACL, 2015: 1491-1500

[33] Hawkins J, Blakeslee S. On Intelligence [M]. London: Macmillan, 2007

[34] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases [C] //Proc of AAAI. Menlo Park, CA: AAAI, 2011: 301-306

[35] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [C] //Proc of NIPS. Cambridge, MA: MIT Press, 2013: 926-934

[36] Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data [J]. Machine Learning, 2014, 94(2): 233-259

[37] Bordes A, Glorot X, Weston J, et al. Joint learning of words and meaning representations for open-text semantic parsing [C] //Proc of AISTATS. Cadiz, Spain: JMLR, 2012: 127-135

[38] Jenatton R, Roux N L, Bordes A, et al. A latent factor model for highly multi-relational data [C] //Proc of NIPS. Cambridge, MA: MIT Press, 2012: 3167-3175

[39] Sutskever I, Tenenbaum J B, Salakhutdinov R. Modelling relational data using Bayesian clustered tensor factorization [C] //Proc of NIPS. Cambridge, MA: MIT Press, 2009: 1821-1828

[40] Yang B, Yih W, He X, et al. Embedding entities and relations for learning and inference in knowledge bases [C] //Proc of Int Conf on Learning Representations (ICLR). 2015

[41] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of NIPS. Cambridge, MA: MIT Press, 2013: 2787-2795

[42] Nickel M, Tresp V, Kriegel H. A three-way model for collective learning on multi-relational data [C] //Proc of ICML. New York: ACM, 2011: 809-816

[43] Nickel M, Tresp V, Kriegel H. Factorizing YAGO: Scalable machine learning for linked data [C] //Proc of WWW. New York: ACM, 2012: 271-280

[44] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. Proc of ICLR. arXiv: 1301.3781. 2013

[45] Fu R, Guo J, Qin B, et al. Learning semantic hierarchies via word embeddings [C] //Proc of ACL. Stroudsburg, PA: ACL, 2014: 1199-1209

[46] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs [J]. arXiv preprint arXiv: 1510.04935. 2015

[47] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes [C] //Proc of AAAI. Menlo Park, CA: AAAI, 2014: 1112-1119

[48] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion [C] //Proc of AAAI. Menlo Park, CA: AAAI, 2015

[49] Ji G, He S, Xu L, et al. Knowledge graph embedding via dynamic mapping matrix [C] //Proc of ACL. Stroudsburg, PA: ACL, 2015: 687-696

[50] Ji G, Liu K, He S, et al. Knowledge graph completion with adaptive sparse transfer matrix [J]. AAAI. 2016

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

[52] Xiao H, Huang M, Hao Y, et al. TransG: A generative mixture model for knowledge graph embedding [J]. arXiv preprint arXiv: 1509.05488. 2015

[53] He S, Liu K, Ji G, et al. Learning to represent knowledge graphs with Gaussian embedding [C] //Proc of CIKM. New York: ACM, 2015: 623-632

[54] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions [C] //Proc of AAAI. Menlo Park, CA: AAAI, 2016

[55] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning [C] //Proc of ICML. New York: ACM, 2008: 160-167

[56] Collobert R, Weston J, Bottou L E O, et al. Natural language processing (almost) from scratch [J]. JMLR, 2011, 12: 2493-2537

[57] Zhong H, Zhang J, Wang Z, et al. Aligning knowledge and text embeddings by entity descriptions [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2015: 267-272

[58] Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks [J]. Machine Learning, 2010, 81(1): 53-67

[59] Lao N, Mitchell T, Cohen W W. Random walk inference and learning in a large scale knowledge base [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2011: 529-539

[60] Garc I A-Dur A N A, Bordes A, Usunier N. Composing relationships with translations [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2015: 286-290

[61] Gu K, Miller J, Liang P. Traversing knowledge graphs in vector space [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2015

[62] Tenenbaum J B, Kemp C, Griffiths T L, et al. How to grow a mind: Statistics, structure, and abstraction [J]. Science, 2011, 331(6022): 1279-1285

[63] Kemp C, Tenenbaum J B. Structured statistical models of inductive reasoning [J]. Psychological Review, 2009, 116 (1): 20

[64] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data [C] //Proc of ACL-IJCNLP. Stroudsburg, PA: ACL, 2009: 1003-1011

[65] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2012: 455-465

[66] Sebastian R, Yao L, Mccallum A. Modeling relations and their mentions without labeled text [C] //Proc of ECML-PKDD. Berlin: Springer, 2010: 148-163

[67] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations [C] //Proc of ACL-HLT. Stroudsburg, PA: ACL, 2011: 541-550

[68] Takamatsu S, Sato I, Nakagawa H. Reducing wrong labels in distant supervision for relation extraction [C] //Proc of ACL-HLT. Stroudsburg, PA: ACL, 2012: 721-729

[69] Etzioni O, Cafarella M, Downey D, et al. Web-scale information extraction in knowitall:(preliminary results)[C] //Proc of WWW. New York: ACM, 2004: 100-110

[70] Yates A, Cafarella M, Banko M, et al. Textrunner: Open information extraction on the Web [C] //Proc of HLT-NAACL. Stroudsburg, PA: ACL, 2007: 25-26

[71] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning. [C] //Proc of AAAI. Stroudsburg, PA: ACL, 2010: 3-10

[72] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding [C] //Proc of the 2012 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012: 481-492

[73] Wu F, Weld D S. Open information extraction using Wikipedia [C] //Proc of ACL. Stroudsburg, PA: ACL, 2010: 118-127

[74] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2015: 1753-1762

[75] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network [C] //Proc of COLING. New York: ACM, 2014: 2335-2344

[76] Dos Santos C I C N, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks [C] //Proc of ACL-IJCNLP. Stroudsburg, PA: ACL, 2015: 626-634

[77] Li H, Xu J. Semantic matching in search [J]. Foundations and Trends® in Information Retrieval, 2013, 7(5): 343-469

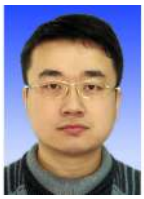
[78] Bai B, Weston J, Grangier D, et al. Polynomial semantic indexing [C] //Proc of NIPS. San Francisco, CA: Morgan Kaufmann, 2009: 64-72

[79] Bai B, Weston J, Grangier D, et al. Supervised semantic indexing [C] //Proc of CIKM. New York: ACM, 2009: 187-196

[80] Gy O Ngyi Z A N, Garcia-Molina H, Pedersen J. Combating Web spam with trustrank [C] //Proc of VLDB. San Francisco, CA: Morgan Kaufmann, 2004: 576-587

[81] Bengio Y, Louradour J E R O, Collobert R, et al. Curriculum learning [C] //Proc of ICML. New York: ACM, 2009: 41-48

[82] Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction [J]. Science, 2015, 350(6266): 1332-1338



Liu Zhiyuan, born in 1984. PhD, assistant researcher. Senior member of China Computer Federation. His research interests include natural language processing, representation learning, and computational social sciences.

Sun Maosong, born in 1962. PhD, professor and PhD supervisor. Senior member of China Computer Federation. His research interests include natural language processing, Chinese computing, Web intelligence, and computational social sciences.

Lin Yankai, born in 1991. PhD candidate. Student member of China Computer Federation. His research interests include knowledge graphs and representation learning.

Xie Ruobing, born in 1992. Master candidate. Student member of China Computer Federation. His research interests include knowledge graphs and representation learning.