
第 1 章

绪 论

“知识”二字，早在甲骨文中就已经出现：“知”造字本义是“谈论和传授狩猎作战的经验”；“识”造字本义是“辨识指认武器”。后来《说文解字》将“知”释义为“从口从矢，知理之速，如矢之疾也”，大意是“明白道理的人说话如同射箭一语中的”。古代“知识”尚指认识的人或事物，进入 20 世纪，“知识”有了现在的含义，《现代汉语词典》释义为“人们在社会实践中所获得的认识和经验的总和”。在中国文化中，人类的智慧从来与知识如影随形，“知”和“智”两字经常通用，孔子说“知之为知之，不知为不知，是知也”，最后一个“知”即通“智”字。

知识是人类智能的象征。知识对人工智能而言同样具有重要意义。自 1956 年达特茅斯研讨会首次提出人工智能以来，在多年的发展历史中，知识一直是人工智能的核心命题。实际上，作为人工智能学科的思想来源之一，英国著名哲学家伯特兰·罗素等倡导创立的分析哲学，就致力于采用各种形式化手段来探讨人类对世界的认识（即知识），现在计算机科学的重要理论基础数理逻辑就起源于此。早期人工智能研究多关注通用搜索机制来解决智能问题，但 MIT 著名学者约翰·麦卡锡早在 1958 年就发布了“有常识的程序”，首次在系统中考虑了关于世界的一般知识（即常识）。由于简单的搜索和规则方法无法解决大规模的困难和复杂问题，20 世纪 70 年代很多学者转而解决专门领域的智能任务，以斯坦福大学著名学者爱德华·费根鲍姆为首的学者通过收集领域专业知识研制了各类“专家系统”，在分析化学、医疗诊断等领域取得了喜人成绩。费根鲍姆在 1977 年发表文章，正式提出“知识工程”的思想，从此，以知识表示、获取和应用为主要内容的知识智能成为人工智能的重要研究方向。

进入 21 世纪，人工智能在数据和计算的双重加持下突飞猛进。在知识智能方面，2012 年搜索引擎巨头谷歌（Google）发布了知识搜索产品——谷歌知识图谱（Google Knowledge Graph），提出“Things, Not Strings”的理念。对于用户输入的查询，谷歌搜索引擎将不止返回匹配查询关键词的相关网页，还会根据查询中提及的人名、地名、机构名等实体

信息，展示这些实体的相关结构化信息。如图 1.1 所示，用户输入“姚明”时，谷歌搜索引擎不仅返回相关网页，而且会直接展示姚明的生日、身高等信息。在知识图谱的支持下，谷歌搜索引擎还能够回答用户提出的一些简单问题，如“姚明的生日是哪天？”等，显著提升搜索引擎的用户体验，因此微软必应、百度、搜狗等各大互联网公司纷纷推出知识搜索功能。与此同时，语音对话助手和智能音箱等新兴服务形态的出现，以及人工智能技术与医疗、教育、金融、法律等垂直领域的深度结合，点燃了人们对大规模知识图谱及在此之上的智能问答和推理等应用的旺盛需求，知识智能再次成为人工智能领域的热点方向。



图 1.1 谷歌知识图谱样例

1.1 知识图谱简介

如前所述，“知识图谱”本是谷歌知识搜索功能的产品名称，由于这个名字实在贴切上口，被学术界和工业界广泛使用，成为各类结构化知识库的统称。知识图谱，就是将人类知识结构化形成的知识系统，其中包含基本事实、通用规则和其他有关的结构化信息，可用于信息检索、推理决策等智能任务。知识图谱是人工智能研究和智能信息服务的基础核心技术，能够赋予智能体精准查询、深度理解与逻辑推理等能力，被广泛运用于搜索引擎、问答系统、智能对话系统及个性化推荐等知识驱动的任务。

为了高效地储存与利用结构化知识，人们结合专家手工标注与计算机自动标注等方式，面向开放领域和垂直领域构建了各种大规模知识图谱，如 WikiData^[188]、Freebase^[15]、DBpedia^[4]、YAGO^[81] 及 WordNet^[133] 等经典的知识库。以 WikiData 为例，截至 2019 年年初，其已经包含 5 700 多万个实体。与此同时，国内外各大互联网公司也均有知识图谱产品，如谷歌 (Google) 知识图谱、百度知心、搜狗知立方和微软 (Microsoft) Bing Satori 等。

如果按照类型划分，知识图谱可以包含语言知识、常识知识、世界知识、认知知识、专业知识等。例如，语言知识是使用人类语言应当具备的词法、句法、语义或语用等方面的知识，如 WordNet、HowNet 是典型的词法知识图谱；常识知识泛指普通人应当具备的基本知识信息，如 Cyc、ConceptNet 是典型的常识知识图谱；世界知识指的是现实世界中各实体间关系的事实知识，前述 WikiData、Freebase、DBpedia、YAGO 是典型的世界知识图谱；认知知识是人类理解世界所具备的知识，如隐喻知识等；专业知识则是各专业领域的特定知识，如化学、生物、医疗、金融、计算机等领域都有很多专业知识图谱。当然，这只是非常粗略的划分，并没有严格的界限，例如，HowNet 既包含词法级语言知识，也包含大量的常识知识。

知识图谱的特点是结构化，一般用三元组形式表示不同元素间的复杂关系，从而形成一个复杂的网络 (图谱)。在本书中，我们主要关注与自然语言处理密切相关的两种类型知识，即世界知识与语言知识。图 1.2 所示是世界知识与语言知识样例。

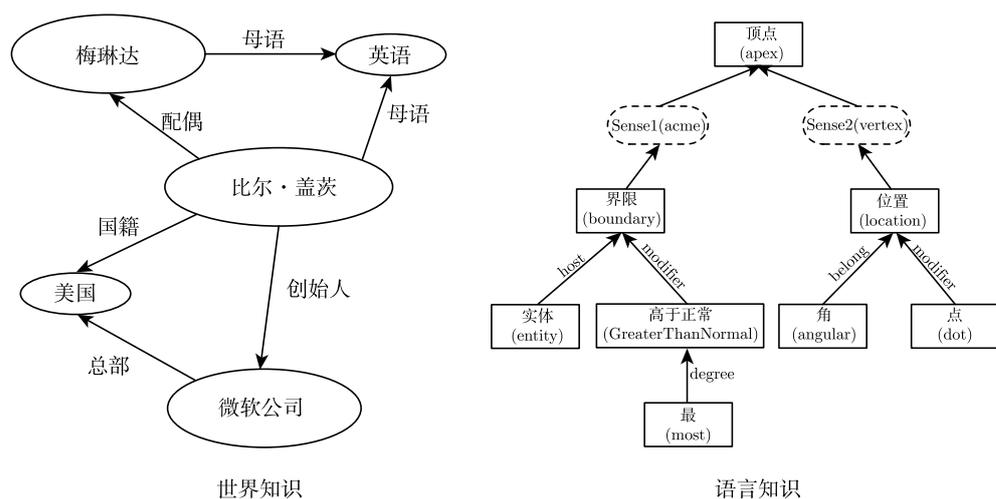


图 1.2 世界知识与语言知识样例

世界知识图谱将世界上的具象事物和抽象概念表示为实体 (entity)，将实体之间的联系表示为关系 (relation)，常以三元组事实 (triple fact) 的形式存储实体与实体之间的关系，类似于万维网联盟发布的资源描述框架 (Resource Description Framework, RDF)。以“比尔·盖茨是微软公司创始人”的世界知识为例，知识图谱将该知识储存为 (比尔·盖茨, 创始人, 微软公司)，其中比尔·盖茨称为首实体 (head entity)，微软公司称为尾实体 (tail entity)，创始人称为关系 (relation)。通过众多三元组构成的链接，世界知识图谱形成一张巨大的网络，其中网络节点是所有的实体，而节点之间的连边表示实体之间的关系。

语言知识图谱有很多不同的知识类型和标注方式。本书主要关注基于义原的词汇知识图谱 HowNet。在 HowNet 中，语言学家将语言的最小语义单位定义为义原 (语义原子的意思)，人工定义了一套包含 2 000 多个义原的集合，并用来标注每个词的词义。例如，“顶点”包含两个词义，分别是“最高点” (acme) 和“三角形两条线的交点” (vertex)，均可以用若干个义原及其依存关系来标注表示。这些结构化语言知识，也可以用三元组形式表示和存储，这些三元组构成的链接能够形成一个大的知识图谱。

以 Freebase、WikiData 为代表的世界知识图谱和以 HowNet、WordNet 为代表的语言知识图谱，过去已经在信息检索、自然语言处理等领域获得比较广泛的应用。进入深度学习时代，大数据及数据驱动深度学习技术成为推动人工智能发展的重要动力。这些大规模知识图谱是否依然有用，以及如何发挥作用，是我们接下来需要探讨的重要课题。

1.2 深度学习的优势和挑战

现在在人工智能领域，几乎无人不谈深度学习，这里以自然语言处理的发展趋势为代表来考察深度学习的优势和面临的挑战。

自然语言处理研究兴起于美国，在 20 世纪 50 年代电子计算机刚刚发明，人们就提出了利用计算机理解和处理人类语言的设想。当时，美国希望能够利用计算机将大量俄语材料自动翻译成英语，以监测苏联的科技发展情况。研究者最开始受到军事密码破译思想的启发，认为不同语言不过是对“相同语义”的不同编码而已，因此可以像译码技术破译密码那样来“破解”这些语言。1954 年，美国乔治敦大学和 IBM 公司成功将超过 60 句俄语自动翻译成英语，媒体纷纷报道认为这是一个巨大的进步，美国政府备受鼓

舞，加大了对自然语言处理研究的资助。当时的研究者也自信地认为只要制定好各种翻译规则，在 3~5 年内就能够完全解决自动翻译问题。然而，事与愿违，理解人类语言要远比破译密码复杂得多，研究进展异常缓慢，1966 年一份报告表明，自然语言处理经过十多年的研究结果远未达到预期，因此支持资金急剧下降，使自然语言处理特别是机器翻译陷入长达 20 年的低潮，这也标志着基于规则的理性主义思想在自然语言处理领域的败退。

直到 20 世纪 90 年代，随着电子计算机的计算能力的飞速提高和制造成本的大幅下降，研究者重新关注自然语言处理这个极富挑战的领域。此时研究者已经意识到简单语法规则的堆砌无法实现对人类语言的真正理解。同时人们发现，通过对大规模文本数据的自动学习和统计，能够更好地解决自动翻译等自然语言处理任务。这就是统计自然语言处理，标志着基于统计的经验主义思想在自然语言处理领域的兴起。

总结过去，人们主要探索了两种自然语言处理的思想，一种是基于规则的理性主义思想；另一种是基于统计的经验主义思想。理性主义思想认为，人类语言主要是由语法规则来产生和描述的，只要将人类语法规则整理出来，就能够理解人类语言并实现语言翻译等各种自然语言处理任务。经验主义思想则认为，可以从大规模语言数据中自动学习总结语言模型，只要有足够多的用于统计学习的语言数据，就能够理解人类语言。

近几十年来，互联网的普及在各领域积累下海量的唾手可得的大数据，计算机的存储和计算能力也在摩尔定律的支配下日新月异，为基于统计的经验主义思想提供了肥沃的发展土壤。深度学习则是统计学习方法的最新顶峰。深度学习是人工神经网络的复兴，在大数据和计算能力的支持下焕发了惊人的能量。与传统统计学习方法相比，深度学习有以下两个突出特点。

(1) 分布式表示。在深度学习中，自然语言处理的对象，如词、句等语言单元的语义信息，都以分布式表示 (distributed representation) 的方式进行处理，即表示为实值、稠密、低维向量。这相当于将语言单元映射到一个低维向量空间中，在这个空间中，各种元素的距离代表它们之间的语义关系，距离越近就表示语义越相近。深度学习技术可以通过大规模文本数据自动学习这些语言单元的向量表示，提供了非常广阔参数空间，可以习得自然语言中的复杂语义模式。传统自然语言统计模型则主要依赖以字词为起点的符号表示，如词袋模型 (bag-of-words model) 及其背后的独热表示 (one-hot representation) 思想，它们均假设所有词之间是语义独立的，极大地限制了传统模型的语义建模能力。

(2) 深层架构。深度学习的另外一个强大的原因是其深层架构，无论是卷积神经网络（CNN）中多层的卷积和池化，还是循环神经网络（RNN）中的随序列循环的状态层，这些深层架构为自动学习自然语言及各种任务的复杂语义模式提供了强大的建模能力。

自 2013 年 word2vec 问世以来，深度学习在短短几年时间里覆盖自然语言处理的各类重要任务，在信息检索、机器翻译、智能问答、自动对话等领域均取得了显著进展。然而，随着更加深入的探索，人们发现深度学习作为数据驱动的经验主义方法，仍然存在很多局限，面临诸多挑战。其中最致命的是，深度学习的鲁棒性和可解释性较差：有学者发现，针对一个深度学习模型，可以很容易地精心构造一些对抗样例（adversarial example）来欺骗该模型。最常见的例子是，对一幅主体为熊猫的图片，通过添加非常少量的噪声，虽然对人类而言这幅图与原图几乎一模一样，深度学习模型会将其错误地划分为其他类别。即使人们做出很多尝试，如提出生成式对抗网络（GAN）以及其他数据增广技术，但仍然无法彻底解决深度学习的鲁棒性和可解释性问题。归根到底，深度学习仍然是对已有数据的拟合，缺乏对数据的真正理解能力，因此只能做到机械地“举百反一”，而不能像人类那样“举一反三”。最近兴起的零次/少次学习（zero/few shot learning）就在努力赋予深度学习“举一反三”能力。

如何才能让计算机（或深度学习）具备理解与思考能力呢？我们应当参考人脑的工作模式。仍然以自然语言理解为例，当人们看到一句话“这个夏天就像蒸笼一样”，会产生哪些理解呢？如图 1.3 所示，假如此人母语是英语，且并不懂汉语，如果我们给他一本中英文词典和语法书（注意这代表某种类型的语言知识），则他可以借助这些语言知识，很容易地将这个结构简单的句子翻译成英语；此人还需要利用已有的常识知识和认知知识，才能理解将“夏天”比作“蒸笼”，是形容这个夏天很热；如果此人掌握一些关于现实的世界知识，则他还会想到用某些品牌的空调或风扇帮助降温；如果此人还有一些金融领域的行业知识，则他就能推测这个夏天空调大卖，空调厂商效益提升，可以提前做投资布局。

即使现有自然语言处理服务可以处理的简单样例，同样面临类似的问题。例如，用户检索“北京到上海的高铁”时，商业搜索引擎可以匹配“北京”“上海”“高铁”等实体，匹配预先定义好的模板，然后根据该模板对应指令查询后台数据库返回相关车次信息。这像一个事先编排好的操作流程，一切按照剧本来运行，一旦超出事先定义的范围就束手无策。而人类看到“北京到上海的高铁”，则会调动各种类型的知识来实现对这句话的

理解，会知道北京和上海是中国的两大直辖市，高铁是运行在陆地上的高速铁路，京沪高铁是连接北京和上海的最繁忙的铁路线。人类永远无法像搜索引擎那样记住繁多的京沪高铁车次，却能够利用有生以来不断积累的知识，轻而易举地理解这句话在现实世界中的意义，而不仅仅用来找到所有京沪高铁车次而已。只有真正理解这句话在现实世界中的全部意义，才是真正的智能；而专门为完成某种特定任务（搜索车次）研制的系统或算法，则无法做到随机应变。



图 1.3 对同一句话“这个夏天就像蒸笼一样”，在不同类型的知识支持下，实现不同程度的理解

可以看到，即使短短的一句话，只有在各种类型知识的支持下，人们才能进行不同层次的理解。这些知识是人类对包括自身在内的外部世界的认知，如同 Palm 发明人杰夫·霍金斯在他的 *On Intelligence*^① 中所描述的“世界模型”，是人脑对这个世界形成的理性认知模型。只有将自然语言置入这个“世界模型”中，才能实现真正的理解。

对于数据驱动的深度学习技术而言也是如此，如果没有相关知识所结成的“世界模型”的支持，那么深度学习技术只能从数据中机械地学习完成特定任务的语义模式，既不具备鲁棒性和可解释性，也无法明白言外之意，通晓弦外之音，实现对语言的深层理解与推理。因此，我们认为要想实现鲁棒可解释的人工智能和自然语言处理能力，需要各种类型知识图谱的支持。

^①作者几年前有幸参与翻译过该书，中文译名为《智能时代》。

1.3 深度学习 + 知识图谱 = ∞

过去的实践表明，数据驱动的经验主义思想是实现数据智能的可行路径，以深度学习为代表的经验主义思想，仍然探索如何更充分地大规模数据中学习和挖掘有用信息。例如，2018年提出的BERT预训练语言模型^[47]，就是利用大规模无标注文本数据学习一般的语言知识，2013年word2vec^[137]也是类似工作的典范，无疑它们对整个自然语言处理领域都产生了非常大的影响。近年来，机器翻译领域也在探索如何构建无监督机器翻译模型，也更充分地利用了无监督数据的做法。这些都属于数据驱动的方法。如何更充分地挖掘数据的价值，无论是有标注数据、弱标注数据，还是无标注数据，它们都是数据驱动方法的重要命题，仍然有很多开放问题等待解决，未来还会有更多的学习机制等待探索。

如前所述，单从数据学习无法实现有理解能力的人工智能。通往鲁棒可解释的人工智能之路，还需要人类知识的指导。为此，我们需要探索如何充分发挥实现经验主义思想和理性主义思想的优势，融合数据智能和知识智能。知识图谱是理性主义思想的最新成功实践。以自然语言处理为例，我们的目标就是，更好地将结构化知识图谱融入自然语言处理深度学习模型中。

实现深度学习与知识图谱的融合并非轻而易举，需要至少解决以下几个关键问题。

(1) **知识表示的问题**。深度学习采用分布式表示，而知识图谱采用符号表示，这两种表示方案难分轩轾，各有天地。如何在深度学习中充分利用大规模知识图谱，需要解决知识表示的问题。

(2) **知识获取的问题**。知识图谱形式丰富多样，根据历史实践经验，完全依靠手工标注费时费力，既极大地限制知识图谱的规模扩增，也无法有效保证知识图谱的内在一致性和可计算性。如何从互联网大规模的结构化数据、半结构数据和无结构数据中自动获取知识，辅以少量人工校验，是大规模知识图谱构建的必由之路，因此需要探索知识自动获取的技术。

(3) **知识计算的问题**。在面向大规模知识图谱建立了完善的知识表示后，需要系统探索如何面向不同自然语言处理任务和深度学习模型，将知识合理地嵌入与融合，实现知识指导的自然语言理解。

在过去几年中，我们系统探索了对自然语言理解至关重要的两类知识图谱即语言知识和世界知识的表示学习、自动获取和计算应用方法。接下来，我们以这两类知识为代

表，分别简单介绍其在不同方面的研究问题与解决思路。

1.3.1 知识的表示学习

大规模知识图谱是人类理性知识的总结，主要以符号形式进行表示和存储。例如，世界知识图谱采用三元组形式存储实体的关系事实，无论是实体还是关系，都采用独一无二的符号来进行标识。原始符号表示的大规模知识图谱在计算利用上面临以下挑战。

(1) 计算效率低下：基于图结构的知识表示虽然简洁直观，但是在利用知识图谱进行检索与多步推理时，常常需要设计专门的图算法以完成任务。这些图算法往往计算复杂度较高，在目前的大规模知识图谱上难以快速运行，且难以拓展至其他情况。

(2) 数据稀疏性强：大规模知识图谱中的实体与关系往往也存在着长尾分布，有很多实体只存在着极少数的关系与之相连。对这些稀疏的实体和关系，往往很难有效理解与推理。

为了解决计算效率低下与数据稀疏性强这两个问题，近年来人们提出知识表示学习 (Knowledge Representation Learning, KRL) 的技术方案，并被广泛研究与运用。知识表示学习基于分布式表示^[78]的思想，将实体（或关系）的语义信息映射到低维稠密实值的向量空间中，使得语义相似的两个对象之间的距离也相近。而传统对知识的符号表示其实相当于独热表示，即将知识表示成一个长向量，只有该知识对应的特定维度非零，而其他所有维度都为零。与独热符号表示相比，知识的分布式表示有以下优点。

(1) 分布式表示学习到的是低维向量。这使得实体与关系之间的语义联系能够在低维空间中得以高速计算，显著提高计算效率。

(2) 独热表示基于所有对象相互独立的假设，所有向量之间两两正交，丢失了大量对象之间的相似及关联信息。而分布式表示则能通过稠密低维向量之间的相似度计算表达对象之间的关系，较好地缓解了数据稀疏带来的问题。

(3) 分布式表示能够将多源异质信息映射到同一语义空间中，建立多源跨模态的信息交互，且分布式表示也能更便捷地融入深度学习的模型框架中。近年来，知识表示学习在知识图谱补全、知识获取、融合与推理等任务上被广泛运用，显著地提高了知识驱动型应用的性能。

基于知识表示学习的思想，我们探索了一套融合知识的统一语义表示框架，如图 1.4 所示。我们选取了与文本理解最密切相关的语言知识和世界知识两种知识图谱，通过分布式表示学习的方式，将海量文本、语言知识和世界知识映射到统一的语义表示空间，

实现数据与知识的融合计算。

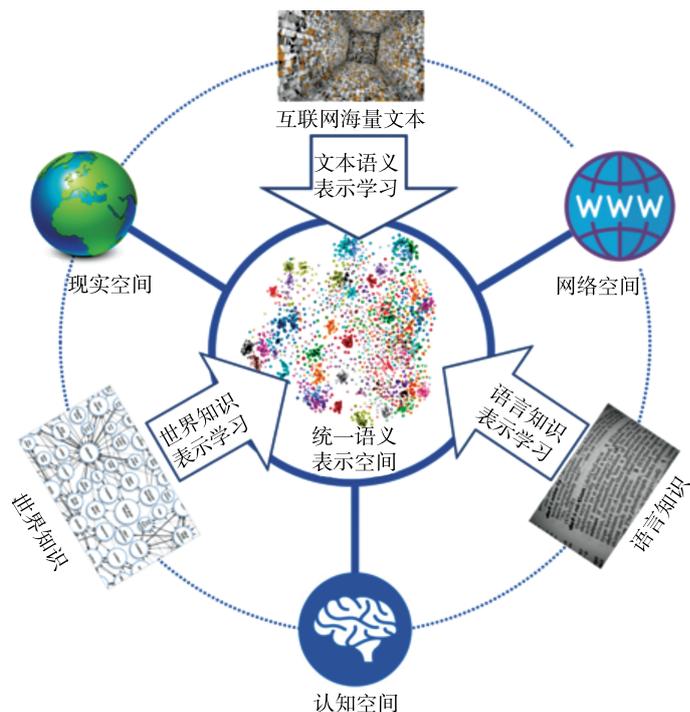


图 1.4 融合知识的统一语义表示框架

在融合语言知识的语言单元表示学习方面，我们探索了将 WordNet 词义知识、HowNet 义原知识等融入不同语言单元的表示学习，实现更加鲁棒有效的语义表示。在世界知识表示学习方面，我们系统探索了考虑各种外部信息来提升知识表示学习性能的可行性，如考虑复杂关系类型，考虑关系路径，考虑实体描述信息，综合考虑实体、属性与关系信息，考虑实体图像信息等。通过充分利用世界知识图谱和实体描述、类别和图像等外部信息，我们实现了高效知识表示学习，建立了低维语义表示空间与世界知识的关联。多粒度语言单元表示、融合语言知识的语言单元表示、世界知识表示形成了一套面向自然语言处理的统一语义表示体系，能够有力支持自然语言的深度学习与理解。

1.3.2 知识的自动获取

随着对这个世界探索的不断扩展与深入，人类知识日新月异。过去，人们主要依赖

专家手工标注知识库，如 Cyc、WordNet、HowNet 等，虽然品质精良，但是在知识图谱的规模、一致性和可计算性方面面临巨大挑战。近年来，人们探索出一条从互联网大规模数据中自动获取知识的技术路线，如现在搜索引擎产品中广泛应用的大规模世界知识图谱，就是自动获取与人工校验结合产生的杰作。

以世界知识为例，人们探索了从互联网各种类型数据获取实体关系事实的方法，如结构化的表格和列表数据、半结构的维基百科页面，以及无结构的文本数据等。其中，如何从无结构文本数据中自动获取这些结构化知识是最具挑战的任务，而且由于文本数据是人类传递信息和知识的主要载体，该任务对构建世界知识图谱也至关重要。

关系抽取是从文本中自动获取实体间关系事实的代表任务。该任务的目标是，给定一个包含两个实体的句子，从中抽出这两个实体之间的关系。例如，给定两个实体“比尔·盖茨”和“微软公司”，以及包含这两个实体的句子“比尔·盖茨建立并运营微软公司”，我们可以根据该句的语义信息，利用句子分类技术确定这两个实体之间是“创立者”关系。深度学习技术是目前解决关系抽取的最好方法，图 1.5 所示就是采用卷积神经网络

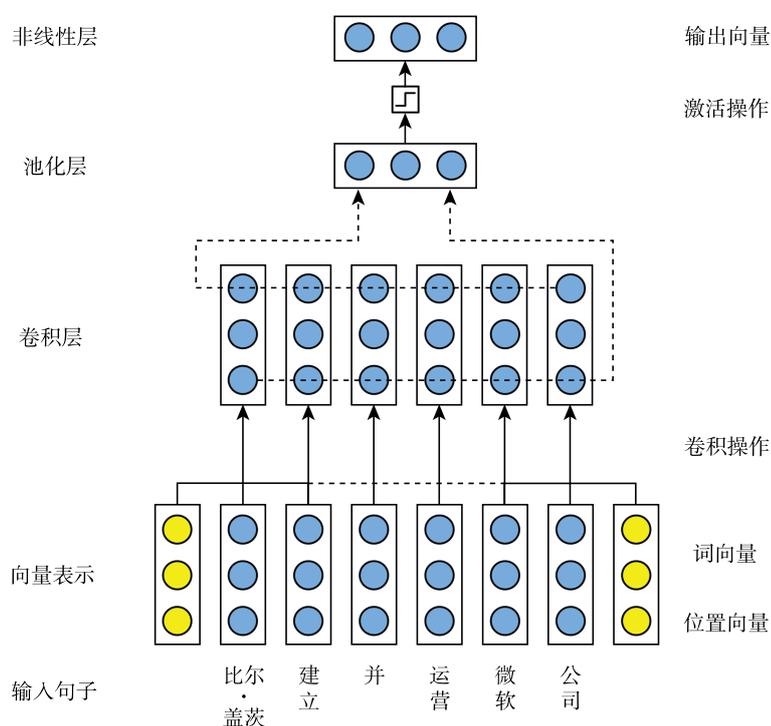


图 1.5 深度学习技术用于关系抽取

络 (CNN) 对句子语义进行编码表示用作分类特征的关系抽取示意图。当然这里还可以选用 LSTM、Transformer 等其他深度学习模型。

关系抽取作为典型的有监督分类任务，需要大规模标注数据来训练深度学习模型。由于为大规模知识手工标注训练数据费时费力，人们提出利用已有知识图谱自动标注大规模训练数据的思想，这种思想被称为远程监督，该思想假设包含某个实体对的所有句子都能够反映该实体对在知识图谱中的关系。如图 1.6 所示，远程监督认为同时包含“比尔·盖茨”和“微软公司”的 3 个句子都能够反映“创立者”的关系，会被自动标注为该关系的正例样本，作为关系分类训练数据。

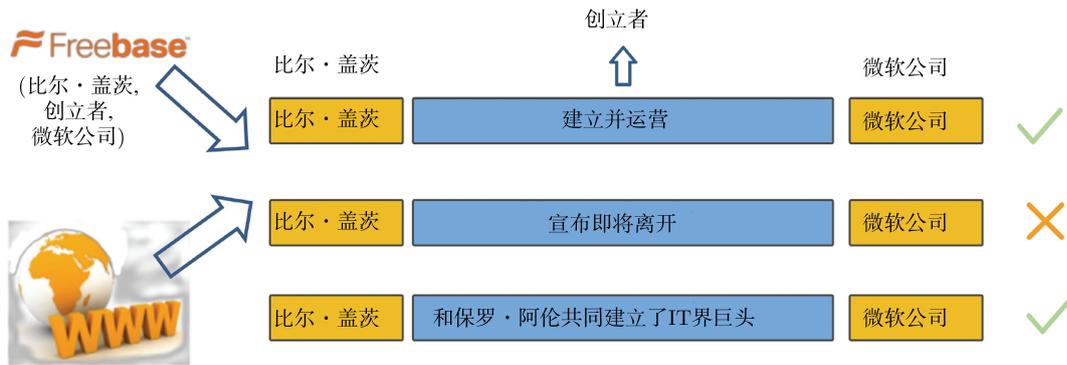


图 1.6 远程监督自动构建关系抽取训练数据集

但是远程监督会不可避免地引入噪声标注，例如图 1.6 中第二个句子的意思是“比尔·盖茨”将从“微软公司”退休，无法反映实体之间存在“创立者”的关系，这些噪声训练数据会显著影响关系抽取深度学习模型的性能。此外，如何充分利用已有知识图谱和文本数据进行关系抽取，如何充分利用多语言文本数据进行关系抽取，都是关系抽取深度学习模型面临的挑战性问题。

为了解决这些挑战性问题，我们提出利用注意力机制、对抗训练机制等技术来提升关系抽取深度学习模型的性能。例如，提出句级注意力机制缓解远程监督的噪声标注问题，提出跨语言注意力机制综合利用多语言数据进行关系抽取，提出知识注意力机制综合利用已有知识图谱和文本进行关系抽取，等等。综合这些改进方案，我们提出的关系抽取深度学习框架，能够面向典型数据信息自动获取高质量的结构化知识，通过进一步的知识融合和校验，可用于构建和扩展大规模知识图谱。

与上述世界知识的自动获取类似，我们也可以利用深度学习技术辅助进行语言知识的自动获取。我们面向 HowNet 的义原知识，也探索了一系列的技术，综合利用词的内部汉字、外部上下文及词典定义等信息，进行词汇的义原自动预测，可用于构建和扩展义原知识图谱。我们将在本书中分别介绍世界知识和语言知识的自动获取技术。

1.3.3 知识的计算应用

在构建完成大规模知识图谱后，我们关心如何在自然语言处理各种任务中充分利用这些知识。在进入深度学习时代之前，这些知识图谱往往采用以下使用方式。

(1) 作为信息资源。以世界知识图谱为例，商业搜索引擎会将知识图谱作为展示信息的重要来源，根据用户查询提及的实体名称，展示相关实体的结构化信息，提升用户的检索体验。

(2) 作为数据特征。以语言知识图谱为例，知识图谱中标注的词汇知识可以为词汇增加额外的特征信息，并作用于情感分析、文本分类等相关的自然语言处理任务上。

(3) 利用网络结构。以世界知识图谱为例，可以利用网络结构完成问答任务，例如，回答诸如“梁启超的儿子的妻子是谁？”这样需要在知识图谱中进行多步跳转的问题；也可以利用网络结构建立不同关系之间的推理规则，例如，发现“儿子的儿子”是“孙子”这样的推理规则，可以用来自动获取新的知识；也可以利用网络结构，如最短路径等计算不同节点之间的相似度；等等。可以看到，过去由于受到知识图谱表示与存储的限制，只能基于离散符号的形式进行使用。

大规模知识图谱是实现自然语言深层理解的重要基础，因此我们认为，在知识表示学习等新兴技术的支持下，我们可以在自然语言处理深度学习模型中找到大规模知识图谱更广阔的应用天地。我们可以充分利用知识图谱为大规模文本中提及的词汇和实体提供丰富的外部信息，在探索改进深度学习的鲁棒性、可解释性、弱标注学习、少次学习等能力的过程中发挥关键作用。近年来，我们在语言模型、信息检索、关系抽取、实体分类、阅读理解、开放问答、词典扩展等自然语言处理任务上探索了将知识图谱与深度学习技术融合的可行性技术方案，提出了一套知识指导的自然语言处理框架，如图 1.7 所示。在本书中，我们将面向世界知识和语言知识两种类型的知识图谱，分别介绍融入深度学习模型的相关尝试。

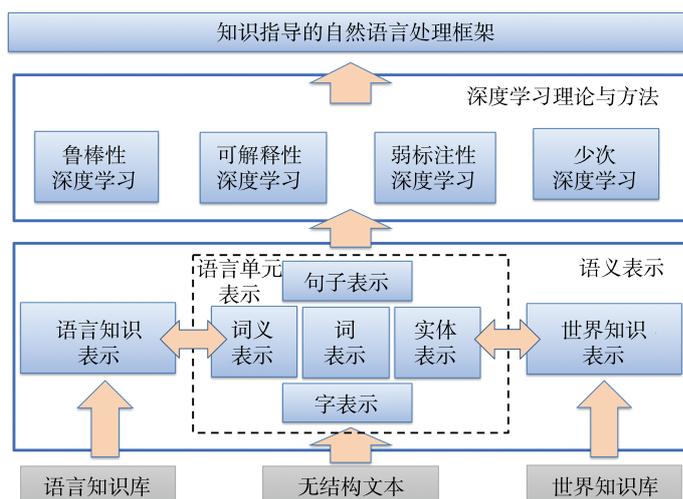


图 1.7 知识指导的自然语言处理框架

1.4 本书结构

本书第 1 章绪论整体阐述知识图谱和深度学习的研究背景和相关概念，介绍了新时期深度学习面临的关键挑战，以及将知识图谱与深度学习深度融合的重要意义和主要研究问题。

本书正文包括两大部分：本书第 1 篇包括第 2 章到第 4 章，主要面向以 Freebase 为代表的世界知识图谱，介绍世界知识的表示学习、自动获取与计算应用技术。本书第二篇包括第 5 章到第 7 章，主要面向以 HowNet 为代表语言知识图谱，介绍语言知识的表示学习、自动获取与计算应用技术。每个章节都会提供必要的研究背景和相关工作介绍，而主体内容是介绍我们在这些研究问题上的研究成果与思考。

在最后一章即第 8 章，我们对知识图谱与深度学习融合的未来发展方向进行了总结和展望，希望引起更多学者的兴趣，共同探索这个充满未知和希望的研究方向。

1.5 本章总结

我们认为，深度学习和知识图谱在未来将呈现一个不断互相增强、共同发展的态势。

如图 1.8 所示，一方面，深度学习技术的快速发展，使得从无结构文本自动抽取实体关系等结构化知识的性能显著提升，为大规模知识图谱的不断完善提供支持；另一方面，知识表示学习技术的兴趣与成熟，为将知识融入深度学习框架提供了可行性方案，能够进一步提升深度学习技术的鲁棒可解释能力，也有助于帮助深度学习技术克服远程监督的噪声标注问题及长尾分布数据的少次学习问题。

可以看到，深度学习和知识图谱的进步都会促进对方的进步，两者发展形成互利共生的技术体系，就如同巨人的双脚一样，互为支撑，协同进退，共同支持人工智能迈向鲁棒可解释的远大目标。

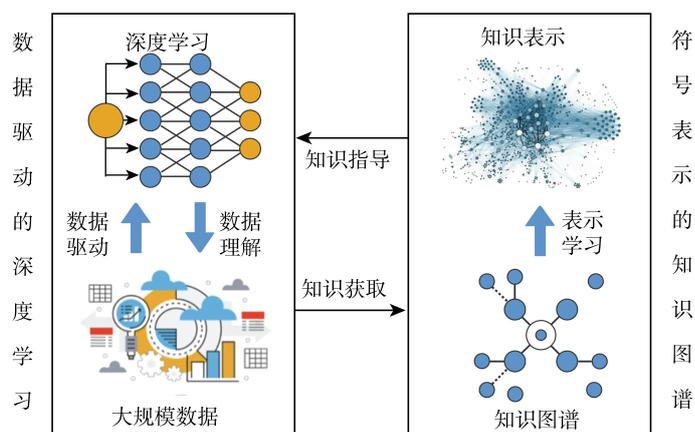


图 1.8 深度学习与知识图谱协同发展示意图

世界知识图谱

世界知识图谱将现实世界中的具象事物与抽象概念表示为实体，将实体之间的联系表示为关系，并最终用以（头实体, 关系, 尾实体）三元组为基本元素的结构来表述知识。伴随着 20 世纪末互联网的蓬勃发展及 21 世纪初信息技术的大规模普及，在大量世界知识被整理进入知识图谱的同时，新的世界知识也在不断高速产生。而伴随着知识爆炸式增长这一现象而来的便是如何对这些知识进行体系化的学习、获取及应用的问题。

在本篇中，我们将就此展开，从世界知识的表示学习、自动获取、计算应用 3 个角度对相关领域的研究进行梳理，并对其中的关键性问题给出我们的探索与尝试。具体来说：

- 世界知识的表示学习关注研究世界知识图谱抽象过程中所面临的计算复杂度与数据稀疏性问题，通过算法将知识图谱嵌入低维度连续空间之中，以便计算机对知识图谱进行理解与操作。
- 世界知识的自动获取关注构建高效的知识获取系统，以便从以自由文本为中心的多源数据中自动抓取世界知识，并将所抓取知识用于知识图谱的组织和扩展。
- 世界知识的计算应用关注如何使用知识图谱中的世界知识来辅助下游任务。我们将在实体对齐、实体类型分类、文档排序和信息检索 4 个任务场景中展现世界知识的重要作用。

上述三个部分内容将在后续章节中详细介绍。

第 2 章

世界知识的表示学习

2.1 章节引言

通过挖掘现实文本中的实体与关系信息，我们能够将世界知识组织成结构化的知识网络，诸如 Freebase^[15]、DBpedia^[4]、YAGO^[188] 等大规模的世界知识图谱蕴含着大量结构化世界知识。能够利用这些丰富的结构化信息将有助于我们在知识驱动下更好地完成各种场景下的具体任务。但是，正如我们在前文提到的那样，采用传统的特征提取方法来处理知识图谱将会面临计算效率低与结构稀疏等问题，这将在很大程度上影响知识图谱在具体任务场景上的部署与使用。因此，为了能够将知识图谱中丰富的结构化信息运用到下游应用中，对知识图谱进行表示学习，并进一步得到图谱中实体与关系的低维稠密向量表示是十分必要的。本章将主要介绍世界知识图谱的表示学习方法。

世界知识表示的核心在于对世界知识图谱中的实体与关系进行表示，并能够通过表示向量来捕捉实体与关系之间的关联。当下已有不少模型对世界知识图谱进行表示学习，其中平移模型 TransE^[19] 作为其中极具代表性的模型，将实体和关系映射至同一个低维向量空间，并将实体与实体之间的关系表示为实体向量之间的平移操作，在结构简单的同时能够取得显著的效果。但从长期以来的世界知识图谱表示学习研究上来看，世界知识图谱的一些特性对其表示学习的效果具有重要影响，其中较为典型的特性如下。

(1) 实体和关系类型多样。世界知识图谱中的实体覆盖了现实世界的方方面面，实体之间体现的关系也多种多样，规模庞大。

(2) 实体和关系在现实中有多种形式的体现。现实中的文本、声音、图像形式的信息均可能为实体和关系提供丰富的语义信息。

(3) 知识图谱具有高噪声、低完整度。由于现实中实体和关系的丰富多样，世界知

识图谱通常很难兼顾高质量与高覆盖率，所以世界知识图谱通常具有很多不准确或缺失的信息。

上述 3 点特性在很大程度上为知识表示学习带来了巨大挑战。鉴于上述特性带来的问题，我们有针对性地在世界知识图谱表示学习方面进行了研究，这些研究主要从以下两个方面进行深入与展开。

(1) 如何恰当地对世界知识图谱的结构进行建模，使实体和关系的表示能够充分融合知识图谱的结构信息，能够综合考虑世界知识图谱中关系与实体的复杂多样、图谱结构的稀疏与不完备，以及图谱构建过程带来的不准确信息。

(2) 如何恰当地在表示学习中融入世界知识图谱之外的信息，尤其是世界知识在各种不同形式的信息载体中的信息，从而构建多信息来源的知识表示模型。

在具体介绍针对上述问题提出的解决方法之前，我们先来回顾知识表示学习领域的相关现有方法，并指出我们的工作与现有方法之间的关系与联系，并对知识表示学习方法的整体脉络进行梳理。

2.2 相关工作

在介绍相关工作之前，我们首先简要地介绍知识表示学习中的常用符号。知识图谱通常以实体、关系及事实三元组进行组织，其中， \mathcal{E} 表示实体集合， \mathcal{R} 表示关系集合， \mathcal{T} 表示三元组集合。对于事实三元组中的任意事实，我们使用 (h, r, t) 进行表示，其中 h 代表头实体， t 代表尾实体， r 代表头尾实体之间的关系。根据定义，有 $h \in \mathcal{E}$ ， $t \in \mathcal{E}$ ， $r \in \mathcal{R}$ ，以及 $(h, r, t) \in \mathcal{T}$ 。知识表示学习旨在将实体与关系表示为低维连续空间之中的向量，在此使用粗体的符号 \mathbf{h} , \mathbf{t} , \mathbf{r} 表示头尾实体与关系对应的表示向量。

2.2.1 知识表示学习经典模型

1. 结构向量模型

结构向量 (Structured Embedding, SE) 模型^[20] 是知识表示学习较早的尝试。在结构向量模型中，所有实体被投影至同一个低维的向量空间中。针对每一个关系 r ，结构向量模型为头实体与尾实体分别设计了一个关系特化的映射矩阵 $\mathbf{M}_{r,1}$ 和 $\mathbf{M}_{r,2}$ ，这些映射

矩阵将会在训练过程中进行自动更新。模型对每个三元组 (h, r, t) 的评分函数定义为

$$E(h, r, t) = \|\mathbf{M}_{r,1}\mathbf{h} - \mathbf{M}_{r,2}\mathbf{t}\|_{L_1/L_2} \quad (2.1)$$

可以看出，结构向量模型将头实体与尾实体通过关系特化的映射矩阵投影至同一向量空间中，并希望映射后向量距离越小时三元组存在的可能性越高。

2. 语义匹配能量模型

语义匹配能量 (Semantic Matching Energy, SME) 模型^[17]与结构向量模型不同，使用低维向量表示实体及关系。在此之上，模型使用矩阵映射、点乘等操作，对实体与关系的联系进行评估。具体地，语义匹配能量模型设计了线性形式与双线性形式两种对三元组的评分函数。

$$E(h, r, t) = (\mathbf{M}_1\mathbf{h} + \mathbf{M}_2\mathbf{r} + \mathbf{b}_1)(\mathbf{M}_3\mathbf{t} + \mathbf{M}_4\mathbf{r} + \mathbf{b}_2) \quad (2.2)$$

$$E(h, r, t) = ((\mathbf{M}_1\mathbf{h} \otimes \mathbf{M}_2\mathbf{r}) + \mathbf{b}_1)((\mathbf{M}_3\mathbf{t} \otimes \mathbf{M}_4\mathbf{r}) + \mathbf{b}_2) \quad (2.3)$$

其中， \otimes 表示按位相乘； \mathbf{M}_1 、 \mathbf{M}_2 、 \mathbf{M}_3 、 \mathbf{M}_4 表示待学习的映射矩阵； \mathbf{b}_1 、 \mathbf{b}_2 表示偏置向量。此外，还有基于张量对语义匹配能量模型进行改进的模型^[18]。

3. 隐变量模型

隐变量模型 (Latent Factor Model, LFM)^[90,183]将实体表示成低维向量，将关系表示为双线性变换矩阵 \mathbf{M}_r ，在知识表示学习效果与计算复杂度方面都有显著改善。模型的评分函数为

$$E(h, r, t) = \mathbf{h}\mathbf{M}_r\mathbf{t} \quad (2.4)$$

此外，也有方法将关系矩阵限定为对角阵对隐变量模型进行改进的模型，在降低模型复杂度的情况下获得了更好的模型效果^[210]。

4. 矩阵分解模型

矩阵分解模型基于矩阵分解的方式进行知识表示学习，其中以 RESCAL 模型^[149,150]和 HoIE 模型^[148]为代表。RESCAL 模型使用一个三阶张量 $\mathbf{X} \in \mathbb{R}^{d \times d \times k}$ 表示三元组，

其中, d 是实体的数量; k 是关系的数量。如果 $\mathbf{X}_{htr} = 1$, 则表示三元组 (h, r, t) 存在。设 $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_k\}$, 则有矩阵分解

$$\mathbf{X}_i \approx \mathbf{A}\mathbf{R}_i\mathbf{A} \quad (2.5)$$

其中, \mathbf{A} 表示实体向量形成的矩阵; \mathbf{R}_i 表示第 i 个关系矩阵。HolE 模型则使用了循环相关运算改进 RESCAL 模型, 进一步提升了计算效率与知识表示效果。

5. 神经张量网络模型

神经张量网络 (Neural Tensor Network, NTN) 模型^[178] 基于单层神经网络模型做出改进, 使用双线性张量代替原模型中的线性变换层。神经张量网络模型对一个三元组 (h, r, t) 的评分函数为

$$E(h, r, t) = \mathbf{u}_r^\top \tanh(\mathbf{h}^\top \mathbf{M}_r \mathbf{t} + \mathbf{M}_{r,1} \mathbf{h} + \mathbf{M}_{r,2} \mathbf{t} + \mathbf{b}_r) \quad (2.6)$$

其中, \mathbf{M}_r 是三阶张量; $\mathbf{M}_{r,1}$ 和 $\mathbf{M}_{r,2}$ 是投影矩阵; \mathbf{u}_r 是一个针对关系 r 的特征表示; \mathbf{b}_r 是偏置向量。神经张量网络模型更精确地刻画了实体与之间的联系, 获得了更好的知识表示效果, 但是复杂的模型也使得计算复杂度提高, 同时对知识图谱的稠密程度有更高的要求。

由于计算复杂度较高, 上述经典模型通常难以在大规模世界知识图谱中兼顾效率和结果, 因此, 近年来世界知识图谱表示学习的研究工作, 包括本章详细讨论的工作, 主要基于下面介绍的平移模型展开的。

2.2.2 平移模型及其拓展模型

平移模型 (TransE)^[19] 是 Bordes 等研究者在 2013 年提出的知识表示学习算法。平移模型将实体和关系映射至同一个低维向量空间, 将实体与实体之间的关系表示为实体向量之间的平移操作。由于只考虑了向量之间的平移操作, 平移模型的计算复杂度大大降低, 并且学到的知识表示在知识图谱补全等任务方面的效果也得到了显著提升, 在知识表示任务中被广泛运用。

平移操作的思想在其他任务中早有运用。Mikolov 等研究者于 2013 年提出了词表示学习模型——word2vec^[127,130,131], 并发现 word2vec 学习到的词向量之间有着有趣的语

义平移现象，例如：

$$\mathbf{v}(\text{国王}) - \mathbf{v}(\text{男人}) \simeq \mathbf{v}(\text{王后}) - \mathbf{v}(\text{女人}) \quad (2.7)$$

其中， $\mathbf{v}(x)$ 表示 x 的词向量。这种语义平移现象表明了词和词之间的隐含语义关系被成功地编码进了词向量中。

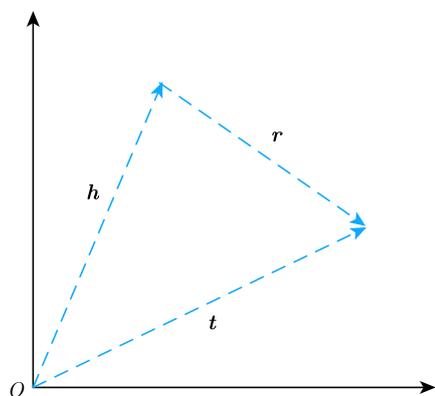


图 2.1 平移模型图示

受到词空间语义平移现象的启发，平移模型将这种隐含语义关系显式地用关系进行表示。具体地，对于给定的三元组 (h, r, t) ，平移模型将关系向量 \mathbf{r} 看作从头实体 \mathbf{h} 到尾实体 \mathbf{t} 的平移向量，如图 2.1 所示。基于以上平移假设，平移模型希望一个三元组内的实体与关系向量之间存在 $\mathbf{h} + \mathbf{r} \simeq \mathbf{t}$ 的关系。形式化地，模型对三元组 (h, r, t) 定义了如下评分函数：

$$E(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{L_1/L_2} \quad (2.8)$$

在实际训练中，平移模型使用最大间隔方法，定义了如下评分函数进行优化：

$$\mathcal{L} = \sum_{(h,r,t) \in \mathcal{T}} \sum_{(h',r,t') \in \mathcal{T}'} \max(\gamma + E(h, r, t) - E(h', r, t'), 0) \quad (2.9)$$

其中， \mathcal{T} 和 \mathcal{T}' 分别是正例三元组与负例三元组的集合； γ 是正负例三元组得分的间隔距离。平移模型通过最大化正负例三元组之间的得分差来优化知识表示。

尽管平移模型很好地兼顾了效率和结果，并因此被广泛运用于知识表示学习任务中，但在世界知识图谱上，平移模型仍存在着以下改进与优化的空间。

(1) 过于理想化的平移假设,在对知识图谱中的复杂关系进行建模时往往存在问题。根据 Bordes 的定义,知识图谱中存在一对多、多对一甚至多对多的复杂关系^[19]。以是 职业关系为例,有(莎士比亚,是 职业,作家)和(鲁迅,是 职业,作家)两条知识。但是莎士比亚和鲁迅仅应在谈论作家这一关系下有相似的表示,而在其他属性(如作品风格、年代、国籍等)关系上有较大区别。为了解决对复杂关系建模的问题,TransH 模型^[195]、TransD 模型^[91]等基于平移模型的改进方法相继被提出,使用向量空间投影与矩阵映射等方式,改进平移模型对复杂关系的建模效果。此外,TransG 模型^[202]提出使用高斯混合模型描述实体之间的关系,将每种语义以一个高斯分布进行刻画。KG2E 模型^[73]直接使用高斯分布表示实体和关系,以高斯分布的协方差表示实体与关系的不确定度。ManifoldE 模型^[201]则使用流形对知识表示进行建模,显著提高了知识表示相关任务的性能。TransR 模型^[113]通过关系特化的实体投影矩阵,在考虑不同关系的时候将实体映射到不同的语义空间中,使实体在不同关系中能够体现出不同的语义特征,从而增强了模型对复杂关系的处理能力。此外,KR-EAR 模型根据不同关系本身的语义和映射特点,从一般的关系中剥离出一类特殊的“属性”关系,对其单独使用分类模型建模,通过这种“分而治之”的方式改进了模型对复杂关系的处理。

(2) 平移模型仅仅关注知识图谱三元组的局部信息,而忽略了知识图谱网络的全局结构与关系之间的推理逻辑。知识图谱中的多步关系路径蕴含着丰富的信息,能够帮助我们进行知识推理。例如,如果我们知道(故宫,位于,北京)与(北京,是 首都,中国)两条知识,那么我们很容易能够推理出(故宫,位于,中国)这条知识。针对这个问题,KALE 模型^[71]在知识表示学习中引入了逻辑规则,提高了平移模型的效果,而 PTransE 模型^[111]则通过构建多步关系路径信息的低维嵌入表示来引入其中蕴含的知识推理信息。

(3) 平移模型只关注了知识图谱自身的网络结构信息,忽略了众多多源异质的丰富信息,如文本、实体类型与图像等信息。这些多源信息既能辅助知识表示学习模型更好地学习已有的结构化知识,也能提供在知识图谱结构信息之外的多模态信息。目前研究者对如何将多源信息融入知识表示学习的探索还仅仅处于初期阶段,我们在此方面提出了 DKRL、TKRL 和 IKRL 模型,创造性地给出了在知识表示学习中融入实体文本描述信息、实体层次类型信息和实体图像信息的方法。

在本章的剩余部分,将详细讨论这些改进工作。

2.3 基于复杂关系建模的知识表示学习

TransE 及其扩展模型通常将实体和关系映射到一个固定的共同空间中。在这样的设定下，两个不同的实体在该空间中的表示之间的距离在理想情况下应当能体现它们之间语义的相似性。然而，在世界知识中，同一个实体在不同的关系场景下具有的语义是有区别的，例如，“罗纳德 威尔逊 里根”既是美国总统，也是专栏作家及戏剧演员。由语义在不同场景中的复杂性可知，在固定空间中表示实体极大地限制了 TransE 及其扩展模型对实体语义的建模能力。

针对这一问题，我们提出了 TransR 模型。

2.3.1 算法模型

如图 2.2 所示，TransR 与传统模型的主要差异在于它为每种关系 r 定义了单独的语义空间，并使用不同的映射矩阵 M_r 定义从实体空间到各个关系空间的映射。具体地，对于一个给定的三元组 (h, r, t) ，TransR 首先使用关系特定的映射矩阵 M_r 将实体从实体空间映射到关系 r 所在的关系空间中，得到 h_r 和 t_r 。

$$h_r = M_r h, \quad t_r = M_r t \quad (2.10)$$

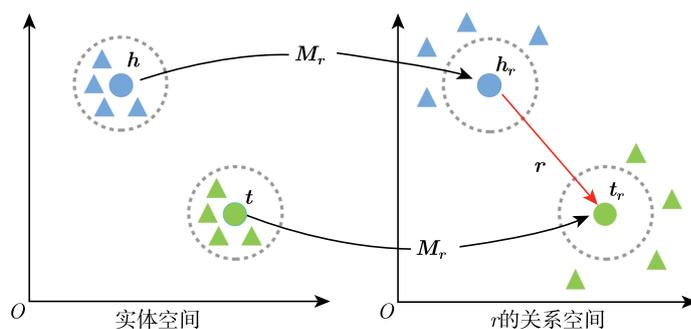


图 2.2 TransR 的简易示意图

在关系 r 所在的空间中， h_r 和 t_r 满足的关系转移约束与 TransE、TransH 相同，即使用类似的评分函数：

$$f_r(h, t) = \|h_r + r - t_r\|_{L_1/L_2} \quad (2.11)$$

此外，我们注意到，相同的关系在不同实体对中的语义通常具有一定的多样性。例如，关系“地区包含”可以出现在国家包含城市、城市包含大学和大洲包含国家等多种情形中。为了进一步增强模型对这种复杂关系的建模能力，我们又在 TransR 的基础上提出了 CTransR 模型。

CTransR 基于分段线性回归的思路，对 TransR 进行了进一步拓展。模型对头尾实体间的潜在关系进行聚类，并为每一个聚类的簇单独建立向量表示。具体地，对于特定的关系 r ，所有训练数据蕴含这个关系的实体对 (h, t) 将会根据 $(\mathbf{h} - \mathbf{t})$ 被聚类到若干组中，其中 \mathbf{h} 、 \mathbf{t} 为 TransE 得到的实体嵌入。我们假设同一组内的实体对所表现的关系 r 具有相近的特征，而不同组内表现的关系 r 可能有较大的差异。因而，对每一组实体对 c ，CTransR 学习了一个单独的关系嵌入 \mathbf{r}_c ，评分函数也相应地被修改为

$$f_r(h, t) = \|\mathbf{h}_{r,c} + \mathbf{r}_c - \mathbf{t}_{r,c}\|_{L_1/L_2} + \alpha \|\mathbf{r}_c - \mathbf{r}\|_2^2 \quad (2.12)$$

其中， $\|\mathbf{r}_c - \mathbf{r}\|_2^2$ 用来约束聚类成的关系向量 \mathbf{r}_c 与原始关系向量 \mathbf{r} 之间的距离，使不同簇表现的同一个关系仍具有一定程度上的相似性，而 α 用于调节这个约束对损失函数的影响。

2.3.2 实验分析

我们采用 WN18、WN11、FB15K 和 FB13 数据集，在链接预测、三元组分类及文本关系抽取任务上对 TransR 和 CTransR 进行了实验。下面就实验结果展开讨论。

1. 链接预测

链接预测是用来预测三元组 (h, r, t) 中缺失实体 h 或 t 的任务，并且在一系列工作中被使用过^[17, 19, 20]。在本任务中，对于每一个缺失的实体，模型将被要求用所有的知识图谱中的实体作为候选项进行计算，并进行排名，而不是单纯给出一个最优的预测结果。与之前的工作一样^[19, 20]，我们在 WN18 和 FB15K 上进行了实验。

在测试阶段，对于每个待测试三元组 (h, r, t) ，我们用知识图谱中的除去 h 与 t 之外的其他实体作为候选项来替换头实体或尾实体，并且按大小顺序给出这些实体的评分函数 f_r 。与 Bordes 等人的操作一样^[19]，我们使用了两种评测方式：① 正确的实体评分函数的平均排名 (mean rank)；② 正确的实体排名在前 10 的比例，即十命中率 (hits@10)。一个优秀的链接预测模型应当获得较低的平均排名和较高的十命中率。

实际上，一个被人为构建的负例三元组有可能是存在于知识图谱中的，这不应被视为负例。然而，上述的评测方法可能低估了这些三元组对评测结果的影响。因此，在对候选进行排名之前，我们先将这些三元组过滤掉，然后用上述的方法评测。我们将初始的评测方法称为“原始”，而将之后过滤的评测方法称为“过滤”。

因为采用了相同的数据集合，我们直接对比了模型与之前论文报告的结果。对于 TransR 和 CTransR 的实验，从 $\{0.1, 0.01, 0.001\}$ 中选择随机梯度下降算法 (SGD) 的学习率 λ ；从 $\{1, 2, 4\}$ 中选择边界值 γ ；从 $\{20, 50, 100\}$ 中选择实体和关系的维度 k 和 d ；从 $\{20, 120, 480, 1\ 440, 4\ 800\}$ 中选择同一批次训练的数据规模 B 。对于 CTransR，从 $\{0.1, 0.01, 0.001\}$ 中选择约束参数 α 。通过验证集上的平均排名评分来决定最好的参数。对于 WN18，我们采用了 L_1 距离，最优的参数为 $\lambda = 0.001$ ， $\gamma = 4$ ， $k = 50$ ， $d = 50$ ， $B = 1\ 440$ ， $\alpha = 0.001$ 。对于 FB15K，我们采用了 L_1 距离，最优的参数为 $\lambda = 0.001$ ， $\gamma = 1$ ， $k = 50$ ， $d = 50$ ， $B = 4\ 800$ ， $\alpha = 0.01$ 。对于这两个数据集合，我们均训练 500 轮。

WN18 和 FB15K 上的评测结果被罗列在表 2.1 中，包括两种不同的负例采样算法——unif（均衡采样）和 bern（基于头尾实体分布伯努利采样）。

表 2.1 WN18 和 FB15K 上的评测结果

数据集	WN18				FB15K			
	平均排名		十命中率 (%)		平均排名		十命中率 (%)	
	原始	过滤	原始	过滤	原始	过滤	原始	过滤
unstructured	315	304	35.3	38.2	1 074	979	4.5	6.3
RESCAL	1 180	1 163	37.2	52.8	828	683	28.4	44.1
SE	1 011	985	68.5	80.5	273	162	28.8	39.8
SME (linear)	545	533	65.1	74.1	274	154	30.7	40.8
SME (bilinear)	526	509	54.7	61.3	284	158	31.3	41.3
LFM	469	456	71.4	81.6	283	164	26.0	33.1
TransE	263	251	75.4	89.2	243	125	34.9	47.1
TransH (unif)	318	303	75.4	86.7	211	84	42.5	58.5
TransH (bern)	401	388	73.0	82.3	212	87	45.7	64.4
TransR (unif)	232	219	78.3	91.7	226	78	43.8	65.5
TransR (bern)	238	225	79.8	92.0	198	77	48.2	68.7
CTransR (unif)	243	230	78.9	92.3	233	82	44	66.3
CTransR (bern)	231	218	79.4	92.3	199	75	48.4	70.2

从表 2.1 中，我们可以看出：

(1) TransR 和 CTransR 比包括 TransE 和 TransH 在内的其他模型均要表现突出。这表明 TransR 在效率和复杂程度上找到了一个更好的权衡。

(2) CTransR 比 TransR 要表现优异，这表明我们应当构建更细粒度的模型来解决同一个关系下子关系复杂的多样性和相关性。CTransR 只是一个初步的尝试，之后我们会在工作中尝试使用更精细的模型来解决这个问题。

(3) bern 采样的效果在 TransH 和 TransR 上都比之前的采样有所提升，尤其是在拥有更多关系的 FB15K 上。

在表 2.2 中，我们将关系分类并且分别呈现了实验结果。^①在 FB15K 上，可以发现 TransR 在所有关系上都获得了最好的结果，尤其是：

(1) 预测单对单关系时，TransR 为实体与关系的复杂相关性提供了更精确的表示，正如图 2.2 所示的那样。

(2) 在预测单对多、多对单关系时，TransR 通过关系特定映射来区分相关实体的能力得到了充分体现，也取得了很大提升。

表 2.2 将关系分类后在 FB15K 上的评测结果

评测指标 关系类别	头实体预测十命中率 (%)				尾实体预测十命中率 (%)			
	1-to-1	1-to-N	N-to-1	N-to-N	1-to-1	1-to-N	N-to-1	N-to-N
unstructured	34.5	2.5	6.1	6.6	34.3	4.2	1.9	6.6
SE	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3
SME (linear)	35.1	53.7	19.0	40.3	32.7	14.9	61.6	43.3
SME (bilinear)	30.9	69.6	19.9	38.6	28.2	13.1	76.0	41.8
TransE	43.7	65.7	18.2	47.2	43.7	19.7	66.7	50.0
TransH (unif)	66.7	81.7	30.2	57.4	63.7	30.1	83.2	60.8
TransH (bern)	66.8	87.6	28.7	64.5	65.5	39.8	83.3	67.2
TransR (unif)	76.9	77.9	38.1	66.9	76.2	38.4	76.2	69.1
TransR (bern)	78.8	89.2	34.1	69.2	79.2	37.4	90.4	72.1
CTransR (unif)	78.6	77.8	36.4	68.0	77.4	37.8	78.0	70.3
CTransR (bern)	81.5	89.0	34.7	71.2	80.8	38.6	90.1	73.8

表 2.3 给出 FB15K 中“地区包含”关系的一些聚类示例。我们可以发现：聚类 1 是大洲包含国家，聚类 2 是国家包含城市，聚类 3 是区域包含城市，聚类 4 是国家包含大学。很明显，通过聚类，可以学习更精确和细粒度的关系嵌入，有助于进一步提高知识图谱的填充性能，这充分证实了我们在设计 TransR 与 CTransR 时的设想。

^①关系的映射方法遵循 Bordes 等人使用的规则 [19]。

表 2.3 〈头实体, 尾实体〉对于“地区包含”的聚类样例

	〈头实体, 尾实体〉
1	〈非洲, 刚果〉, 〈亚洲, 尼泊尔〉, 〈南美洲, 阿鲁巴岛〉, 〈大洋洲, 密克罗尼西亚联邦〉
2	〈美国, 坎卡基〉, 〈英国, 贝里圣埃德蒙兹〉, 〈英国, 达灵顿〉, 〈意大利, 佩鲁贾〉
3	〈佐治亚州, 查塔姆县〉, 〈爱达荷州, 博伊西〉, 〈阿肯色州, 波克县〉, 〈密苏里州, 杰克逊县〉, 〈内布拉斯加州, 卡斯县〉
4	〈瑞典, 隆德大学〉, 〈英国, 剑桥大学〉, 〈弗雷斯诺, 加州州立大学弗雷斯诺分校〉, 〈意大利, 米兰音乐学院〉

2. 三元组分类

三元组分类是一个判断给定三元组 (h, r, t) 正确与否的任务。这是一个二分类任务，已经在一系列工作中作为评测方式^[178, 194]。在这个任务上，我们采用 WN11、FB13 与 FB15K 来进行测试，并且与 Wang 等人的设置保持一致^[194]。

我们需要负例三元组来进行二分类测试。在 NTN^[178] 中，数据集合 WN11 和 FB13 已经有了负例三元组。但对于 FB15K 来说，却没有之前工作公开发布出的负例三元组，我们采用了 Socher 等人使用的负例生成算法^[178]进行负例构建。对于三元组分类，我们设置了一个特殊的阈值 δ_r 。对于三元组 (h, r, t) ，如果评分函数的结果低于 δ_r ，那么三元组将会被认为是正确的，反之则是错误的。 δ_r 则是通过最大化验证集上的分类精度来进行优化的。

对于 WN11 和 FB13，我们比较了 TransR 与 CTranR 及 Wang 等人汇报的结果^[194]。由于 FB15K 是根据 Socher 等人的策略自行生成的^[178]，因此评估结果无法直接与之前的结果进行比较。因此，我们自行实现 TransE 和 TransH，并使用 Socher 等人发布的 NTN 代码^[178]，在 FB15K 数据集上进行了评估与比较。

对于 TransR 的实验来说，我们从 $\{0.1, 0.01, 0.001, 0.0001\}$ 中为 SGD 选择学习率 λ ；从 $\{1, 2, 4\}$ 中选择边界值 γ ；从 $\{20, 50, 100\}$ 中选择实体和关系的维度 k 和 d ；从 $\{20, 120, 480, 1440, 4800\}$ 中选择同一批次训练的数据规模 B 。我们通过验证集的平均排名来决定最好的参数。对于 WN11，我们采用了 L_1 距离，最优的参数为 $\lambda = 0.001$ ， $\gamma = 4$ ， $k = 20$ ， $d = 20$ ， $B = 120$ ， $\alpha = 0.001$ 。对于 FB13，我们采用了 L_1 距离，最优的参数为 $\lambda = 0.0001$ ， $\gamma = 2$ ， $k = 100$ ， $d = 100$ ， $B = 480$ ， $\alpha = 0.01$ 。对于这两个数据集，我们均训练 1000 轮。

三元组分类的评估结果如表 2.4 所示。从表 2.4，我们观察到：

(1) 在 WN11 上, TransR 显著优于包括 TransE 和 TransH 在内的方法。

(2) TransE、TransH 和 TransR 都不能超过 FB13 上最具表现力的模型 NTN。相比之下, 在较大的数据集 FB15K 上, TransE、TransH 和 TransR 的性能要好于 NTN。结果可能与数据集的特征有关: FB15K 中有 1345 种关系类型, 而 FB13 中只有 13 种关系类型。同时, 两个数据集中的实体数量和三元组数量相近。正如 Wang 等人讨论到的^[194], FB13 中的知识图谱比 FB15K 甚至 WN11 更稠密。某种程度上, 最具表现力的模型 NTN 可以从 FB13 的稠密图中使用张量变换来学习复杂的相关性。相比之下, 更简单的模型能够更好地处理 FB15K 这样的稀疏图, 并具有良好的泛化能力。

(3) 此外, bern 采样技术提高了 TransE、TransH 和 TransR 在所有数据集上的性能。

表 2.4 三元组分类的评测结果 (%)

数据集	WN11	FB13	FB15K
SE	53.0	75.2	—
SME (bilinear)	70.0	63.7	—
SLM	69.9	85.3	—
LFM	73.8	84.3	—
NTN	70.4	87.1	68.5
TransE (unif)	75.9	70.9	79.6
TransE (bern)	75.9	81.5	79.2
TransH (unif)	77.7	76.5	79.0
TransH (bern)	78.8	83.3	80.2
TransR (unif)	85.5	74.7	81.7
TransR (bern)	85.9	82.5	83.9
CTransR (bern)	85.7	—	84.5

3. 文本关系抽取

关系抽取旨在从大规模文本中提取关系事实, 这是丰富知识图谱的重要信息来源。当前, 大量的方法^[82, 134, 159, 182]通过知识图谱作为远程监督信号, 对大量文本语料库中的句子进行自动标注, 然后提取文本特征来构建关系分类器。这些方法只使用纯文本来推断新的关系事实。与之不同的是, 知识图谱表示学习则基于现有的知识图谱结构进行链接预测来拓展新的关系事实。

所以, 同时利用纯文本和知识图谱来推断新的关系事实是很直接的想法。在 Weston 等人的工作中^[196], 模型将 TransE 和基于文本的提取模型相结合, 对候选事实进行评分

排序，取得了十分明显的提升效果。在 TransH^[194] 的工作中也发现了类似的效果改进。在本节中，我们将结合文本关系抽取模型来研究 TransR 的性能。

我们使用了 NYT+FB 数据集合，这个数据集合也被用在 Weston 等人^[196]的工作中。在这个数据集中，纽约时代周刊文本内容 (*New York Times Corpus*) 中的实体用 Stanford NER 来注释并链接到 Freebase 的实体之上。

在本实验中，我们实现了 Weston 等人^[196]提出的基于文本的提取模型（命名为 Sm2r）。对于知识图谱部分，Weston 等人^[196]使用了近 400 万个实体的 Freebase 子集，同时有 23 000 个关系类型。由于 TransH 尚未发布数据集，且 TransR 将需要花费很长时间才能从 400 万个实体的数据中学习得到嵌入表示。因而，我们自行生成了一个较小的数据集 FB40K，其中包含 NYT 中的所有实体和 1 336 个关系类型。为了测试公平，我们从 FB40K 中删除了出现在测试集中的所有三元组。与之前工作的结果^[194,196]相比，我们发现使用 FB40K 进行学习并不会显著降低 TransE 和 TransH 的有效性。因此，我们可以安全地使用 FB40K 来证明 TransR 的有效性。

采用与 Weston 等人^[196]相同的处理方法，我们将基于文本的关系抽取模型获得的预测评分与知识图谱表示获得的预测评分相加来进行排序，并获得 TransE、TransH 与 TransR 的精度-召回率曲线。我们将嵌入维度 $k, d = 50$ ，学习率 $\lambda = 0.001$ ，边界距离 $\gamma = 1.0$ ， $B = 960$ ，并且采用了 L_1 距离。精度-召回率曲线如图 2.3 所示。从图 2.3 中可

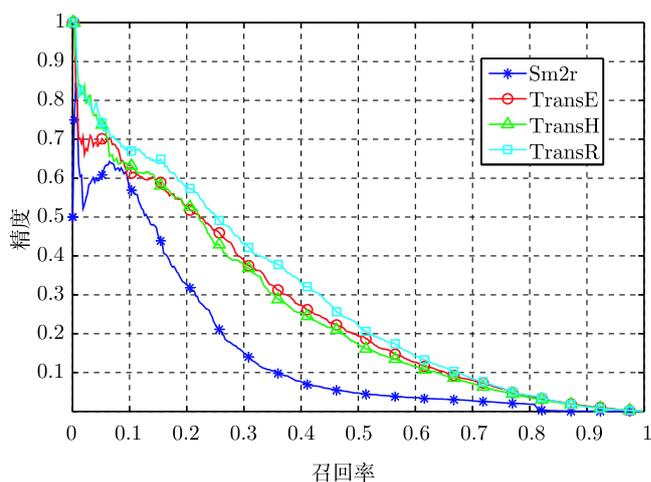


图 2.3 TransE、TransH 与 TransR 在关系抽取上的精度-召回率曲线

以观察到, 当召回范围为 $[0, 0.05]$ 时, TransR 优于 TransE, 与 TransH 相当; 当召回范围为 $[0.05, 1]$ 时, TransR 的表现超越了所有的模型, 包括 TransE 和 TransH。

2.3.3 小结

在本节中, 我们提出了一种基于复杂关系建模的知识表示模型——TransR。TransR 将实体和关系嵌入不同的空间中, 并通过每个关系特定的映射矩阵, 将实体投影到关系空间中来学习嵌入表示。此外, 我们还提出了 CTransR, 其引入了分段线性回归的思想来处理每个关系类型内部的复杂相关性, 可以进一步细化原有模型的关系表示。在实验中, 我们在 3 个任务上进行了模型评估, 包括链接预测、三元组分类与文本关系抽取。实验结果表明, 与 TransE 和 TransH 相比, TransR 取得了显著的提升效果。

2.4 基于关系路径建模的知识表示学习

TransE 及其扩展模型往往只考虑了实体之间的直接关系, 但事实上, 知识图谱中的实体之间的多步关系路径蕴含了丰富的语义信息, 考虑关系路径的特征对我们更好地进行知识表示具有重要意义。例如, 关系路径 $h \xrightarrow{\text{出生地}} e_1 \xrightarrow{\text{城市归属}} e_2 \xrightarrow{\text{省份归属}} t$ 隐含了 h 和 t 之间的国籍关系, 亦即 $(h, \text{国籍}, t)$ 。由此, 我们提出 PTransE 模型, 将知识图谱中的关系路径融入知识表示学习模型中。

2.4.1 算法模型

TransE 与 PTransE 的简易示意图如图 2.4 所示。可以看到, PTransE 仍然基于 TransE 的平移假设, 但以关系路径取代了 TransE 中的单个关系三元组, 即 PTransE 为关系三元组定义的评分函数考虑了实体间的多步关系路径信息:

$$G(h, r, t) = E(h, r, t) + E(h, P, t) \quad (2.13)$$

其中, $E(h, r, t)$ 用直接关系三元组刻画了关系和实体之间的相关性, 如 TransE 中定义的

$$E(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{L_1/L_2} \quad (2.14)$$

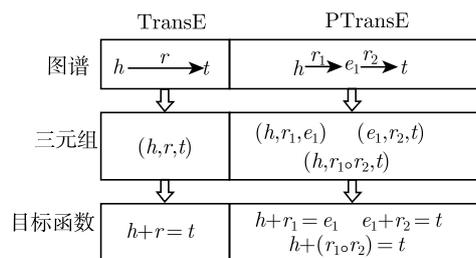


图 2.4 TransE 与 PTransE 的简易示意图

而 $E(h, P, t)$ 是 PTransE 模型的与众不同之处，它通过多步路径来刻画关系层面的推理信息。由于一对实体 h, t 在知识图谱中可能存在多个不同的关系路径（记实体间关系路径集合为 $P(h, t)$ ），不同的关系路径在体现实体间联系方面的可靠性可能各不相同，我们定义 $E(h, P, t)$ 为各关系路径下评分函数根据其可靠性加权平均的结果：

$$E(h, P, t) = \frac{1}{Z} \sum_{p \in P(h, t)} R(p | h, t) E(h, p, t) \quad (2.15)$$

其中， $Z = \sum_{p \in P(h, t)} R(p | h, t)$ 是归一化因子，而 $R(p | h, t)$ 、 $E(h, p, t)$ 分别衡量了关系路径可靠性和关系路径下实体对的能量。PTransE 模型设计的主要挑战便在于后两者的定义，即

- 如何评估关系路径 p 的可靠程度。
- 如何得到一条关系路径 p 的嵌入表示 \mathbf{p} 。

针对第一个挑战，我们提出了一种对路径进行约束的资源分配算法（Path-Constraint Resource Allocation, PCRA）来衡量关系路径的可靠性，其基本思想是：假设存在一定数量的资源，从头部实体 h 流出，且将沿着给定路径 p 流动，使用最终流向尾部实体 t 的资源总量来衡量路径 p 作为 h 和 t 之间连接路径的可靠性。从 h 开始，沿着关系路径 p ，可以将流动路径写为 $S_0 \xrightarrow{r_1} S_1 \xrightarrow{r_2} \dots \xrightarrow{r_l} S_l$ ，其中 $S_0 = \{h\}$ 且 $t \in S_l$ 。对于任意实体 $m \in S_i$ ，将它在关系 r_i 上的直接前驱记作 $S_{i-1}(\cdot, m)$ 。流向 m 的资源被定义为

$$R_p(m) = \sum_{n \in S_{i-1}(\cdot, m)} \frac{1}{|S_i(n, \cdot)|} R_p(n) \quad (2.16)$$

其中， $S_i(n, \cdot)$ 是 $n \in S_{i-1}$ 在关系 r_i 上的直接后继， $R_p(n)$ 是从实体 n 获取的资源。

由于两个实体间的关系路径数量可能非常庞大，我们在实验中仅考虑长度不超过 3、可靠度大于一个阈值的关系路径。

针对第二个挑战，我们尝试了 3 种不同的方法得到关系路径的表示。

- 和： $\mathbf{p} = \mathbf{r}_1 + \mathbf{r}_2 + \dots + \mathbf{r}_l$ 。
- 积： $\mathbf{p} = \mathbf{r}_1 \mathbf{r}_2 \dots \mathbf{r}_l$ 。
- 循环神经网络 (RNN)： $\mathbf{p} = \mathbf{c}_l$ ，其中 \mathbf{c}_l 是 RNN 在序列 $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_l$ 上产生的最后一个状态。

\mathbf{p} 可以被看作根据多步关系路径信息得到的对实体间关系的近似，因而，我们定义 PTransE 的评分函数 $E(h, p, t)$ 为 \mathbf{p} 与 \mathbf{r} 的接近程度，即

$$E(h, p, t) = \|\mathbf{p} - \mathbf{r}\|_{L_1/L_2} \quad (2.17)$$

2.4.2 实验分析

我们在实体预测、关系预测及文本中的关系抽取 3 个任务上对不同设置下的 PTransE 模型和基线模型进行了实验。

1. 实体预测

在实体预测这一子任务中，这里同样遵循 Bordes 等人 [19] 的实验设置。我们使用两种测量标准作为评价指标：① 正确的实体评分函数的平均排名 (mean rank)；② 正确的实体排名在前 10 的比例，即十命中率 (hits@10)。

理想情况下，PTransE 可以找到给定实体和所有候选实体之间的关系路径。然而，这是相当耗时且很难实际操作的。因为在这样的操作下，我们必须为每一个测试三元组迭代遍历 \mathcal{E} 中的每一个候选实体，并寻找相应的关系路径。这里，我们采用一种重排序方法：首先根据 TransE 计算的得分对所有候选实体进行排名，然后对排名在前 500 的实体使用 PTransE 进行评分，并重新排名。对于 PTransE 而言，根据验证集中的平均排名来找到最佳超参数。PTransE 的最佳参数配置是 $\lambda = 0.001$, $\gamma = 1$, $k = 100$ ，采用 L_1 距离。在训练时，限定训练轮数为 500 轮。

实体预测的评测结果如表 2.5 所示。基线模型有 RESCAL [149]、SE [20]、SME (线性) [17]、SME (双线性) [17]、LFM [90]、TransE [19]、TransH [194] 和 TransR [113]。对于

表 2.5 实体预测的评测结果

评测指标	平均排名		十命中率 (%)	
	原始	过滤	原始	过滤
RESCAL	828	683	28.4	44.1
SE	273	162	28.8	39.8
SME (linear)	274	154	30.7	40.8
SME (bilinear)	284	158	31.3	41.3
LFM	283	164	26.0	33.1
TransE	243	125	34.9	47.1
TransH	212	87	45.7	64.4
TransR	198	77	48.2	68.7
TransE (Our)	205	63	47.9	70.2
PTransE (ADD, 2-step)	200	54	51.8	83.4
PTransE (MUL, 2-step)	216	67	47.4	77.7
PTransE (RNN, 2-step)	242	92	50.6	82.2
PTransE (ADD, 3-step)	207	58	51.4	84.6

PTransE, 我们考虑关系路径表示的 3 种组合操作: 加法 (ADD)、乘法 (MUL) 和递归神经网络 (RNN)。在实验中, 我们考虑至多两步或三步的关系路径。我们也自发实现了 TransE, 在相同的参数配置下, 实现的 TransE 性能显著高出 Bordes 等人 [19] 在论文中报告的结果。

从表 2.5 中, 我们观察到:

(1) PTransE 显著优于包括 TransE 在内的其他基线模型。这表明关系路径为知识图谱的表示学习提供了一个很好的信息补充, 并且这些知识图谱中的关系路径已经被 PTransE 成功嵌入低维空间之中。例如, 乔治·沃克·布什和亚伯拉罕·林肯都是美国的总统, 其在 TransE 中展现出类似的嵌入。这可能会扰乱 TransE 对劳拉·威尔士·布什配偶的预测。相反, 由于 PTransE 刻画了关系路径, 它可以利用乔治·沃克·布什和劳拉·威尔士·布什之间的关系路径信息, 并据此做出更准确的预测。

(2) 对于 PTransE, 加法操作在均值排名和十命中率中均优于其他组合运算。原因是, 加法操作同时符合 TransE 和 PTransE 的学习目标。以 $h \xrightarrow{r_1} e_1 \xrightarrow{r_2} t$ 为例, 两个直接关系的优化目标 $h + r_1 = e_1$ 和 $e_1 + r_2 = t$ 可以很容易地导出路径的目标 $h + r_1 + r_2 = t$ 。

(3) 最多考虑两步和三步关系路径的 PTransE 能达到可比拟的结果, 这表明考虑过长的关系路径对最终的结果影响不大。

我们对关系进行分类, 分为单对单 (1-to-1)、单对多 (1-to- N)、多对单 (N -to-1)、多对多 (N -to- N) 4 类, 并对 PTransE 和一些基线模型在四类关系上的结果进行了更细致的评测, 结果如表 2.6 所示。观察结果同样表明, 在所有的关系类型中, PTransE 显著且一致地取得了提升效果。

2. 关系预测

关系预测旨在预测给定的两个实体之间的关系。我们同样在 FB15K 上进行该项评测。由于我们实现的 TransE 已经在实体预测的所有基线模型比较中取得了最佳性能, 所以我们直接将 PTransE 与我们自发实现的 TransE 进行比较。评估结果如表 2.7 所示。因为与实体相比, 关系的数量要小得多, 且 TransE 和 PTransE 的十命中率指标均超过了 95%, 因而列举一命中率而不是十命中率来进行比较。在表 2.7 中, 我们给出了诸多模型的结果, 包括不含逆向关系的 TransE (TransE)、含有逆向关系的 TransE (+Rev)、简单考虑关系路径的 TransE (+Rev+Path)。针对 PTransE, 我们也对去除图谱结构信息 (-TransE) 与去除路径信息 (-Path) 两种情况进行了相应测试。用于关系预测的

PTransE 的最佳参数配置与用于实体预测中最佳参数是一致的： $\lambda = 0.001$, $\gamma = 1$, $k = 100$, 且采用了 L_1 距离。

表 2.6 对关系分类后在 FB15K 上的评测结果

评测指标 关系类别	头实体预测十命中率 (%)				尾实体预测十命中率 (%)			
	1-to-1	1-to- N	N -to-1	N -to- N	1-to-1	1-to- N	N -to-1	N -to- N
SE	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3
SME (linear)	35.1	53.7	19.0	40.3	32.7	14.9	61.6	43.3
SME (bilinear)	30.9	69.6	19.9	38.6	28.2	13.1	76.0	41.8
TransE	43.7	65.7	18.2	47.2	43.7	19.7	66.7	50.0
TransH	66.8	87.6	28.7	64.5	65.5	39.8	83.3	67.2
TransR	78.8	89.2	34.1	69.2	79.2	37.4	90.4	72.1
TransE (Our)	74.6	86.6	43.7	70.6	71.5	49.0	85.0	72.9
PTransE (ADD, 2-step)	91.0	92.8	60.9	83.8	91.2	74.0	88.9	86.4
PTransE (MUL, 2-step)	89.0	86.8	57.6	79.8	87.8	71.4	72.2	80.4
PTransE (RNN, 2-step)	88.9	84.0	56.3	84.5	88.8	68.4	81.5	86.7
PTrasnE (ADD, 3-step)	90.1	92.0	58.7	86.1	90.7	70.7	87.5	88.7

从表 2.7 中, 我们观察到:

表 2.7 关系预测的评测结果

评测指标	平均排名		一命中率 (%)	
	原始	过滤	原始	过滤
TransE (Our)	2.8	2.5	65.1	84.3
+Rev	2.6	2.3	67.1	86.7
+Rev+Path	2.4	1.9	65.2	89.0
PTransE (ADD, 2-step)	1.7	1.2	69.5	93.6
-TransE	135.8	135.3	51.4	78.0
-Path	2.0	1.6	69.7	89.0
PTransE (MUL, 2-step)	2.5	2.0	66.3	89.0
PTransE (RNN, 2-step)	1.9	1.4	68.3	93.2
PTransE (ADD, 3-step)	1.8	1.4	68.5	94.0

(1) PTransE 在关系预测上要显著优于 TransE+Rev+Path, 预测错误率下降了 41.8%。

(2) 即使 TransE 本身, 考虑关系路径的 TransE+Rev+Path 与不考虑路径的 TransE+Rev 相比, 在测试中也可以减少 17.3% 的预测错误率。这表明对关系路径进行建模有利于关系预测。

(3) 仅考虑关系路径而没有图谱特征的模型 (PTransE-TransE) 的答案平均排名非常高。考虑其中的原因, 主要是测试三元组中并非所有实体对之间都有关系路径, 这将导致模型对关系进行随机猜测, 正确答案的排名期望值为 $|\mathcal{R}|/2$ 。与之相对的是 PTransE-TransE 的一命中率结果比较合理, 这说明建模关系路径对知识图谱表示学习是具有重要意义的。与 TransE 相比, 不考虑图谱特征的 PTransE-TransE 还是具有劣势的, 这表明虽然建模关系路径有利于获取实体之间的关系, 但图谱本身的实体表示为关系预测提供了关键信息, 是必不可少的特征。

3. 文本关系抽取

文本关系抽取旨在从纯文本中提取关系事实以丰富现有知识图谱。已有的工作主要立足于使用大规模知识图谱作为远程监督信号, 自动地给句子打标签以获得训练实例, 进而用自动标注的数据训练关系分类器。所有这些方法仅基于纯文本来推理新的事实。TransE 曾被用来与基于文本的关系抽取模型进行结合, 并取得了显著的效果 [196], TransH [194] 和 TransR [113] 也进行了同样的工作。在此任务中, 我们探索利用 PTransE 与文本关系抽取模型进行结合, 从文本中提取实体之间的关系。

我们使用 Riedel 等人 [159] 发布的纽约时报语料库 (NYT) 作为训练实测和测试数据。NYT 将 Freebase 与《纽约时报》的文章对齐, 其中有 53 类关系 (包括无关系, 记作 NA 和 121 034 个用于训练的实体对。我们在 NYT 与 Freebase 对齐的基础上, 对图谱进行了拓展, 构建出 FB40K, 其中包括 NYT 中提及的所有实体和 1 336 种关系。

在实验中, 我们实现了 Weston 等人 [196] 提出的关系抽取模型 Sm2r, 并将其作为基线方法。将 TransE 和 PTransE 用于学习时, 我们设置实体与关系的维度为 $k = 50$, 学习率为 $\lambda = 0.001$, 训练边界值为 $\gamma = 1.0$, 并采用了 L_1 距离。我们还与 Surdeanu 等人 [182] 提出的 MIMLRE 模型进行比较, 这是使用远程监控的经典算法。各模型在关系抽取任务上的精度-召回率曲线如图 2.5 所示。

从图 2.5 中, 我们可以看到:

通过与基于文本的模型 Sm2r 相结合, PTransE 的精度明显优于 TransE, 且高于朴素的 Sm2r 算法的结果。这表明编码关系路径对从文本中提取关系也具有效果。需要注意的是, 此处使用的 TransE 没有考虑逆向关系和关系路径, 因此性能提升不明显。我们认为, 导致这种现象的主要原因如下: 在知识图谱补全任务中, 每个测试实体对之间至少含有一个有效关系。相反, 在这个任务中, 许多的测试实体对之间没有关系 (即关

系为 NA)，但这些无关系实体对之间存在若干关系路径。TransE 在训练阶段并不会如 PTransE 一般对关系路径进行编码，这会在测试阶段考虑关系路径时导致预测无关系的性能更差，抵消了对于确实存在关系的三元组的改进。这表明对关系路径进行编码并非十分容易，同时证实了 PTransE 的有效性。

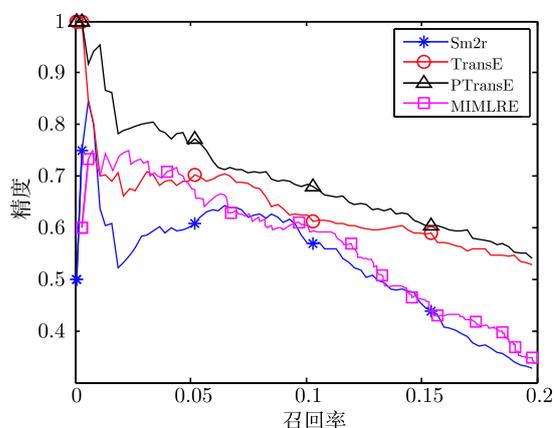


图 2.5 各模型在关系抽取任务上的精度-召回率曲线

4. 关系推理的个案研究

实验结果表明，PTransE 可以实现高性能的知识图谱补全和文本关系抽取。在本节中，我们给出一些在关系路径上进行关系推断的例子。如图 2.6 所示，两个实体阿甘正传

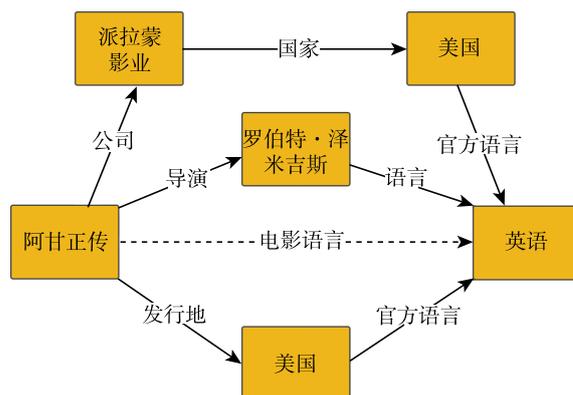


图 2.6 Freebase 中的一个推理例子

和英语通过 3 条关系路径连接起来，这使我们更有把握预测两个实体之间的关系——电影语言。

2.4.3 小结

在本节中，我们提出了 PTransE，这是一种知识图谱的新型表示学习方法。PTransE 通过编码关系路径将实体和关系嵌入一个低维空间之中。为了利用关系路径，我们提出了路径约束的资源分配算法来衡量关系路径可靠性，并且使用关系的语义复合算法来表示路径以便对模型进行优化。我们评估了 PTransE 在知识图谱补全任务和文本关系抽取任务上的表现。实验结果表明，与 TransE 和其他基线模型相比，PTransE 表现出了一致且显著的改进效果，验证了基于关系路径建模的知识表示学习模型的有效性。

2.5 基于属性关系建模的知识表示学习

TransE 模型的最大局限是无法准确地建模一对多、多对一和多对多关系，这促使我们认真审视知识图谱关系的多样性特征，以寻求针对此问题的改进方案。通过观察发现，在许多大规模知识图谱中，关系可以划分为两大类，一类关系指示实体的特征（尾实体通常是抽象概念，如性别和职业），而另一类关系则指示实体之间的联系（头尾实体都是真实世界的物体），我们将二者分别命名为属性和关系。一些典型的属性、关系与它们相应的映射属性之间的关系如表 2.8 所示。表 2.8 同时还列出了每个头实体对应的尾实体的期望数量及每个尾实体对应的头实体的期望数量。

表 2.8 一些典型的属性、关系与它们相应的映射属性之间的关系

关系类型	关系	尾实体期望数量	头实体期望数量
属性	国籍	1.05	1 551.90
	性别	1.00	637 333.33
	种族	1.12	41.52
	宗教	1.09	107.40
关系	父母	1.58	1.67
	首都	1.29	1.42
	作者	1.02	2.17
	成立者	1.37	1.31

如表 2.8 所示，属性是一对多关系和多对一关系的主要来源。例如，在属性性别中，属性值男性是一个与数百万表示人的实体相关的存在。对于这些关系，TransE 及其扩展（如 TransH^[195] 和 TransR^[113]）是无法在实体及其属性值之间充分构建位移的，这也是之前模型对于一对多关系和多对一关系效果始终欠缺的因素之一。因此，我们认为将两种类型的实体关系用不同的方法分别建模将有助于知识表示模型对复杂关系的建模能力。在此思想的引导下，我们提出了 KR-EAR 模型。

2.5.1 算法模型

如图 2.7 所示， A_1 和 A_2 是两个属性，分别有值域 V_1 和 V_2 。在传统知识表示模型中 [图 2.7(a)]，属性 A_1 和 A_2 被视作普通的关系 r_a 和 r_b ，而与此不同的是，KR-EAR [图 2.7(b)] 将属性预测视作分类问题。

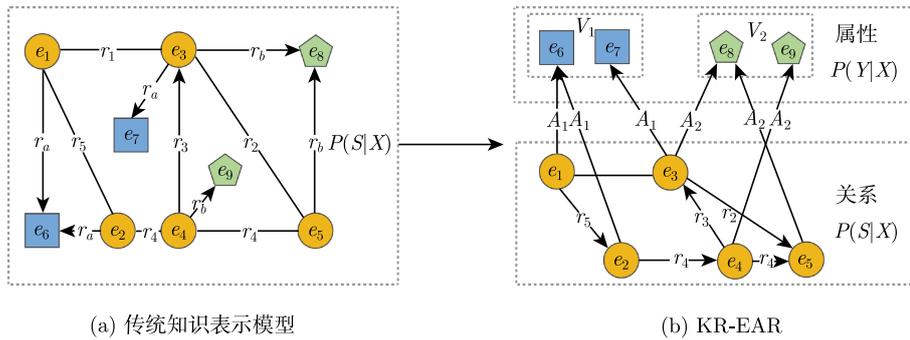


图 2.7 传统知识表示模型和 KR-EAR 的例子

具体地，一般的实体关系集合被划分为关系集合 \mathcal{R} 和属性集合 $\mathcal{A} = \{A_1, A_2, \dots, A_{|\mathcal{A}|}\}$ ，其中每个属性集合 A_i 对应一个值域集合 V_i ，训练集也被划分为关系三元组集 $S \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ 和 $\mathcal{Y} \subseteq \mathcal{E} \times \mathcal{A} \times \mathcal{V}$ 。假设给定嵌入 \mathbf{X} 时关系三元组和属性三元组条件独立，则训练集上的似然值为

$$\begin{aligned}
 P(S, Y | \mathbf{X}) &= P(S | \mathbf{X})P(Y | \mathbf{X}) \\
 &= \prod_{(h,r,t) \in S} P((h, r, t) | \mathbf{X}) \prod_{(e,a,v) \in Y} P((e, a, v) | \mathbf{X})
 \end{aligned}
 \tag{2.18}$$

大量现有表示学习模型可以直接被用于建模关系三元组部分的条件概率 $P((h, r, t) | \mathbf{X})$ ，包括 TransE、TransR、TransD、TranSparse、KG2E、PTransE 等。我们在实验中尝试了

TransE 和 TransR 两种模型, 在此我们不做过多赘述。

而 $P((e, a, v) | \mathbf{X})$ 使用分类模型进行建模, 即

$$P((e, a, v) | \mathbf{X}) \propto P(v | e, a, \mathbf{X}) = \frac{\exp(h(\mathbf{e}, \mathbf{a}, \mathbf{v}))}{\sum_{\hat{v} \in V_a} \exp(h(\mathbf{e}, \mathbf{a}, \hat{v}))} \quad (2.19)$$

其中, 评分函数为

$$h(\mathbf{e}, \mathbf{a}, \mathbf{v}) = -\|f(\mathbf{W}_a \mathbf{e} + \mathbf{b}_a) - \mathbf{V}_{av}\|_{L_1/L_2} + b \quad (2.20)$$

其中, \mathbf{e} 、 \mathbf{V}_{av} 分别为实体和属性值的嵌入, $f()$ 为非线性激活函数。

属性和关系的划分方便了对属性之间关联的建模。我们在实验中尝试了将 $P((e, a, v) | \mathbf{X})$ 替换为

$$P((e, a, v) | \mathbf{X}) \propto P(v | e, a, \mathbf{X})P((e, a, v) | Y(e)) \quad (2.21)$$

其中, $P((e, a, v) | Y(e))$ 为在给定实体 e 的其他属性时属性三元组 (e, a, v) 的概率, 它被定义为 Softmax 函数。

$$P((e, a, v) | Y(e)) = \frac{\exp(z(\mathbf{e}, \mathbf{a}, \mathbf{v}, Y(e)))}{\sum_{\hat{v} \in V_a} \exp(z(\mathbf{e}, \mathbf{a}, \hat{v}, Y(e)))} \quad (2.22)$$

其中, $z()$ 是一个衡量属性间预测相关性的得分函数。它综合了 (e, a, v) 与每一个包含在 $Y(e)$ 中的属性三元组之间的相关性。

$$z(\mathbf{e}, \mathbf{a}, \mathbf{v}, Y(e)) \propto \sum_{(e, \hat{a}, \hat{v}) \in Y(e)} P((a, v) | (\hat{a}, \hat{v}))(\mathbf{A}_a \cdot \mathbf{A}_{\hat{a}}) \quad (2.23)$$

其中, $(\mathbf{A}_a \cdot \mathbf{A}_{\hat{a}})$ 是 \mathbf{A}_a 与 $\mathbf{A}_{\hat{a}}$ 的点积, 表示属性 A_a 与 $A_{\hat{a}}$ 之间的相关程度。 $P((a, v) | (\hat{a}, \hat{v}))$ 是由训练数据的每个实体得到的在给定 (\hat{a}, \hat{v}) 时属性值 (a, v) 的条件概率, 表示属性值 (a, v) 和 (\hat{a}, \hat{v}) 间的相关性。

2.5.2 实验分析

我们使用数据集 FB24K 评估我们的模型和基线模型在知识图谱补全任务中的表现。我们将知识图谱补全分为实体预测、关系预测和属性预测 3 个子任务, 并分别对其展示和讨论实验结果。

1. 实体预测

我们沿用了文献 [19] 中的两种评测指标：所有正确实体的平均排名 (mean rank) 和排名前十的候选实体中正确实体的比例 (hits@10)。值得一提的是，对于一个特定的三元组 (h, r, t) ，其负采样的三元组也可能存在于知识图谱中，其也应被视为有效。对于将其其他有效三元组排得比 (h, r, t) 更高的可能性，上述两种评测指标可能是不公平的。因此，我们在进行排序前过滤掉了所有其他有效三元组。同文献 [19] 一样，我们将经过了过滤的版本命名为“过滤”，将未经过滤的版本命名为“原始”。实体预测的评测结果如表 2.9 所示。从表 2.9 中，我们观察到：

(1) KR-EAR 在平均排名指标下稳定地显著优于基线方法，包括 TransE、TransH 和 TransR。这表明 KR-EAR 为实体和关系学习更加好的嵌入表达。

(2) 在十命中率指标下，KR-EAR (TransE) 优于 TransE，KR-EAR (TransR) 优于 TransR。这表明 KR-EAR 可以很好地利用传统的知识表示模型。

表 2.9 实体预测的评测结果

属性 评测指标	头实体预测				尾实体预测				合计			
	平均排名		十命中率 (%)		平均排名		十命中率 (%)		平均排名		十命中率 (%)	
	原始	过滤	原始	过滤	原始	过滤	原始	过滤	原始	过滤	原始	过滤
TransE	385	277	20.2	39.2	134	124	51.4	66.7	259	200	35.8	53.0
TransH	416	309	17.7	35.4	147	138	50.0	65.0	282	224	33.9	50.2
TransR	394	285	20.5	41.2	125	116	53.4	71.0	260	200	37.0	56.1
KR-EAR(TransE)	295	198	22.7	39.6	77	69	54.2	69.5	186	133	38.5	54.5
KR-EAR(TransR)	268	170	23.4	43.0	75	66	55.7	71.5	172	118	39.5	57.3

2. 关系预测

关系预测旨在推断两个给定实体之间的可能关系。对每一个测试三元组 (h, r, t) ，我们用知识图谱中每个可能的关系 \hat{r} 替换它的关系 r ，并计算得分。之后，我们按照得分升序计算知识图谱中所有候选关系的名次。类似地，我们使用了平均排名和一命中率两种评测指标。Krompass 等人 [103] 的研究显示，通过潜在变量模型，类型约束通常可以支持多关系数据建模。我们认为实体的类型信息是实体属性的特例。在 KR-EAR 中，我们可以很容易地采用头尾实体的属性之间的约束进行关系预测，这被称为 CRA。关系预测的评估结果如表 2.10 所示。

表 2.10 关系预测的评测结果

评测指标	平均排名		一命中率 (%)	
	原始	过滤	原始	过滤
TransE	3.1	2.8	65.9	83.8
TransH	3.4	3.1	64.9	84.1
TransR	3.4	3.1	65.2	84.5
KR-EAR(TransE)	2.4	2.1	67.9	86.2
+ CRA	1.8	1.6	70.9	88.7
KR-EAR(TransR)	2.6	2.2	66.8	89.0
+ CRA	1.9	1.6	71.5	90.4

从表 2.10 中，我们观察到：

(1) KR-EAR 在平均排名和一命中率下再次优于基线方法，而 TransE、TransH 和 TransR 在这个子任务中取得了接近的结果。

(2) 对于 KR-EAR (TransE) 和 KR-EAR (TransR)，CRA 可以进一步将一命中率提高 2.5% 和 1.4%，同时可以降低平均排名。这证明了在关系预测中考虑实体属性的有效性。

3. 属性预测

属性预测的目标是预测实体的缺失属性。这个任务在先前的一些研究中被当作实体预测的一部分 [19, 113, 195]。对于每个测试三元组 (e, a, v) ，我们用每个可能的属性值 \hat{v} 替代 v ，并计算相应得分 $\sigma(h(e, a, \hat{v}))$ 。之后，我们按照得分升序计算知识图谱中所有候选值的名次。

注意到，通过将属性值按照 $\sigma(h(e, a, \hat{v}))$ 、 $\sigma(z(e, a, \hat{v}, Y(e)))$ 排序，KR-EAR 还可以加入对属性相关性 (AC) 的考虑。我们使用两种属性预测的评估指标：平均排名和一命中率。属性预测的评测结果如表 2.11 所示。

从表 2.11 中，我们可以看出：

(1) KR-EAR 仍然明显优于基线方法。这验证了在传统知识表示模型中将属性预测建模为分类而不是位移的必要性。

(2) 对于 KR-EAR (TransE) 和 KR-EAR (TransR)，考虑属性相关性可以分别将一命中率提升 1.4% 和 1.0%。这表明属性相关性在属性预测中是有效的。

表 2.11 属性预测的评测结果

评测指标	平均排名		一命中率 (%)	
	原始	过滤	原始	过滤
TransE	10.7	5.6	36.5	55.9
TransH	10.7	5.6	38.5	57.9
TransR	9.0	3.9	42.7	65.6
KR-EAR(TransE)	8.3	3.2	47.2	69.0
+AC	7.5	3.0	49.4	70.4
KR-EAR(TransR)	8.3	3.2	47.6	69.8
+AC	7.5	3.0	49.8	70.8

表 2.12 给出了一些 KR-EAR 在 FB24K 训练集上得到的属性相关性的例子。可以发现，在给定一个属性时，预测得到的相关属性常常反映符合常理的相关性。这表明 KR-EAR 可以有效捕获属性之间的相关性。

表 2.12 属性相关性的例子

属性	相关属性
职业	婚姻状况, 国籍, 性别, 语言, 种族
电影发行地	电影的国家, 电影的语言, 电影的发行日期, 电影的类型
地区的时区	地区所属国家, 地区通行货币

2.5.3 小结

在本节中，我们将现有的知识图谱关系分为属性和关系，并提出了一个具有实体、属性和关系的新的知识表示模型 (KR-EAR)。另外，我们还在 KR-EAR 中对实体、属性之间的相关性进行了编码。在实验中，我们对模型在 3 个子任务（预测实体、关系和属性）中的表现进行了评估。通过显式地建模，KR-EAR 在 3 个子任务中都稳定地显著优于最新的知识表示模型。

2.6 融合实体描述信息的知识表示学习

实体描述是对实体自身信息的浓缩精练的文字性介绍，通常由几句话或者一个段落组成。实体描述的来源十分丰富，既可以在知识图谱构建过程中由人工填写，也可以提取自百科类数据库的既有信息，还能从丰富的网页文本中抽取。实体描述蕴含着丰富的

实体信息，这些信息能够作为知识图谱中置信度较高的结构化信息的辅助信息，帮助模型构建更准确的知识表示。图 2.8 给出了在维基百科中关于威廉·莎士比亚的一段实体描述节选。从图中我们可以看到，威廉·莎士比亚的实体描述中蕴含了该实体各个方面的丰富信息，如国籍、职业、作品、出生地等，这种实体描述可以很好地作为知识图谱中已有结构化信息的辅助与补充，提供更加深入的细节描述，甚至能挖掘知识图谱中可能遗漏的新知识。可以想到，如果我们能够将此部分信息智能地抽取并融入知识表示学习中，知识表示的性能将会得到明显的提升。

威廉·莎士比亚

威廉·莎士比亚……是英国文学史上最杰出的戏剧家，是西方艺术史上最杰出的作家之一，也是全世界最卓越的文学家之一。他流传下来的作品包括38部戏剧、2首长叙事诗、154首十四行诗和其他诗歌……莎士比亚于雅芳河畔斯特拉特福出生并长大，在18岁时与安妮·哈瑟维结婚，共生育了三个孩子……莎士比亚崇尚高尚情操，常常描写牺牲与复仇，包括《李尔王》《奥赛罗》《哈姆雷特》和《麦克白》等，其被认为是英语最佳范例。

图 2.8 实体描述节选示例

我们从上述思路出发，围绕以下两个主要挑战，提出了融合实体描述的知识表示学习 (Description-embodied Knowledge Representation Learning, DKRL) 模型。

- 如何从实体描述中自动抽取高质量的实体相关知识。实体描述往往在长度、有效信息、噪声形式、描述质量等方面存在较多差异，这要求模型考虑差异化的文本输入，自动从实体描述中获取高质量的知识信息，同时减少文本中无用信息的干扰。
- 当模型抽取到高质量的实体相关知识后，如何将这基于实体描述的文本信息与基于知识图谱三元组的结构化信息融合，即如何实现文本空间与知识空间的对齐与联合训练。

2.6.1 算法模型

图 2.9 给出了 DKRL 模型整体架构的简单图示。如图 2.9 所示，DKRL 模型以实体描述为输入，通过实体描述编码器得到实体基于描述表示。在顶层，基于描述的表示与基于结构的表示在统一的平移模型框架下进行学习。

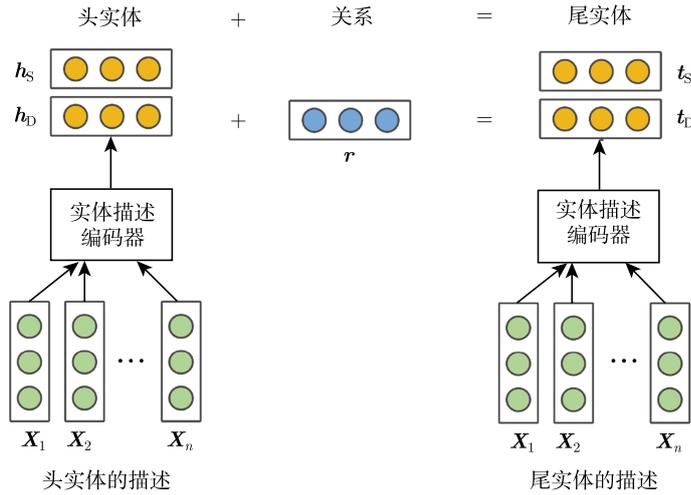


图 2.9 DKRL 模型整体架构的简单图示

DKRL 沿用了 TransE 中的平移假设思想。为了对文本空间与知识空间的表示进行联合训练，DKRL 的能量函数对二者进行了融合：

$$E(h, r, t) = \alpha_1 \|h_S + r - t_S\| + \alpha_2 \|h_S + r - t_D\| + \alpha_3 \|h_D + r - t_S\| + \alpha_4 \|h_D + r - t_D\| \quad (2.24)$$

其中， $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ 是控制各项权值的超参数，下标 S、D 用于区分基于知识图谱和基于实体描述文本的实体表示。在能量函数中， $\|h_S + r - t_S\|$ 部分与平移模型的能量函数类似，而 $\|h_S + r - t_D\|$ 、 $\|h_D + r - t_S\|$ 和 $\|h_D + r - t_D\|$ 3 项则加入了基于描述的实体向量影响。这样，通过混合项的软限制，DKRL 模型将实体的两种向量映射到了同一语义空间中，并共享相同的关系向量。

在训练过程中，DKRL 使用随机生成负例，使用最大间隔方法，定义如下优化目标函数：

$$\mathcal{L} = \sum_{(h,r,t) \in \mathcal{T}} \sum_{(h',r',t') \in \mathcal{T}'} \max(\gamma + E(h, r, t) - E(h', r', t'), 0) \quad (2.25)$$

其中， \mathcal{T} 、 \mathcal{T}' 分别为知识图谱中的关系三元组正例集和正例三元组随机替换任一元素得到的三元组负例集。

在模型中尝试使用了如下两种不同的实体描述信息表示模型。

- 连续词袋模型 (CBOW): 根据经典的文本特征, 如词频、TF-IDF 值等, 筛选出实体描述中最具信息量的 k 个关键词, 将它们词向量的平均值作为实体的文本空间表示。
- 卷积神经网络模型 (CNN): 如图 2.10 所示, 实验中使用两层卷积层与对应的池化层及非线性层对实体描述进行建模, 最后一层池化层将信息编码为一个固定长度的向量。我们在第一个池化层使用了最大池化策略, 而在第二层使用了平均池化策略, 这是由于我们认为第一层对应句子内的局部特征, 需要更强的去除冗余和噪声的能力, 而第二层对应文本描述全局的特征, 需要每个句子都对最终结果产生贡献。

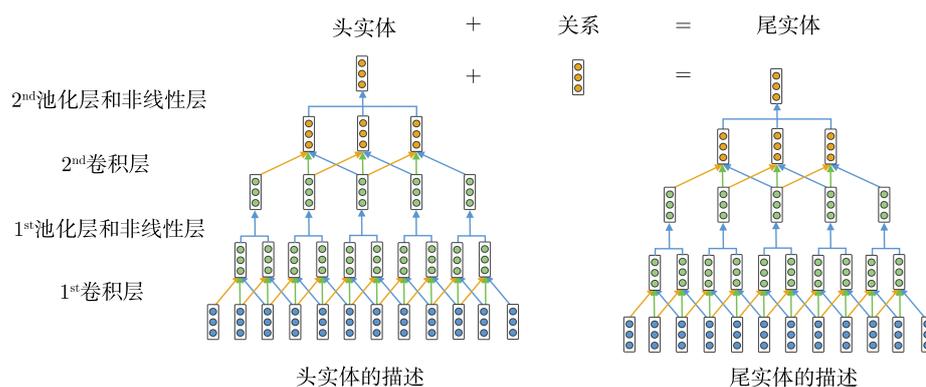


图 2.10 卷积神经网络模型示意图

2.6.2 实验分析

我们使用 FB15K^[19] 作为实验所用的知识图谱数据。为了测试 DKRL 模型在零样本场景下的知识表示构建效果, 我们在 FB15K 的基础上使用 Freebase 对测试集进行扩展, 构建了 FB20K 数据集。FB20K 的关系集、训练集、验证集与 FB15K 相同, 但测试集包括了大量训练集中未出现的实体。我们将 FB20K 的测试样例分为 4 组: $e-e$ 组表示头实体和尾实体都在训练集中; $d-e$ 组表示仅尾实体出现在训练集中; $e-d$ 组表示仅头实体出现在训练集中; $d-d$ 组表示头实体和尾实体都未出现在训练集中。

我们使用两种实体描述编码器对实体描述进行建模, 使用连续词袋的模型记录为 DKRL(CBOW), 使用卷积神经网络的模型记录为 DKRL(CNN)。对于连续词袋模型, 使

用 TF-IDF 值作为各关键词的评分进行排序，并选择前 20 个关键词构建实体表示（如果描述长度少于 20，则选择描述中全部的词作为关键词）。

为了充分展示 DKRL 模型学到的基于描述的知识表示的效果，我们还探索了模型在零样本环境下的知识表示学习任务表现。零样本学习 (zero-shot learning) 是人工智能，特别是深度学习领域的研究热点之一，着重关注于对新样例的建模。在大数据时代，文字、图像或知识等领域上的模型通常是数据驱动的，而这往往需要在训练时对某类样例进行成百上千的学习。例如，在知识表示学习任务上，以往的模型大多要求测试集中的实体曾出现在训练集中，以学得实体的知识表示。然而，在信息爆炸时代，每天都有新实体与新知识的出现，传统知识表示模型无法自动获得新实体的表示，可拓展性差。我们提出的 DKRL 模型能够根据实体的描述信息（实体描述信息广泛存在于互联网文本中），构建知识图谱中没有的新实体基于描述的知识表示，是零样本知识表示学习的新思路。

我们使用 FB20K 数据集模拟零样本知识表示学习的环境，在知识图谱补全与实体类型分类两个任务上进行零样本环境的评测。FB20K 与 FB15K 有着相同的训练集与验证集，而在测试集中，FB20K 有着 5 019 个训练集中未出现的新实体。对于这些新实体，我们使用其基于描述的表示作为知识表示。在知识图谱补全任务中，我们主要关注对存在新实体的三元组的补全情况。在实体类型分类任务中，我们主要关注新实体的分类情况。

1. 知识图谱补全结果与分析

知识图谱补全是知识表示学习的经典评测任务^[17,19,113]，可以看作知识图谱上的链接预测任务。具体地，知识图谱补全任务要求模型在给定三元组 (h, r, t) 的某一实体或关系丢失的情况下，利用另外两个对象正确预测出丢失的实体或关系。根据具体丢失的对象，知识图谱补全可以细分为实体预测与关系预测两个任务。知识图谱补全任务旨在验证所学知识表示的准确性，同时与现实中间答系统的实际任务相呼应。例如，补全三元组 $(?, \text{编写}, \text{哈姆雷特})$ ，即相当于回答“谁编写了《哈姆雷特》这本书？”这个问题。

我们使用 DKRL 模型和基线模型在训练集上学习到的知识向量，对测试集样例进行知识图谱补全。对于测试集中的每个三元组，我们分别隐藏其头实体、关系和尾实体来进行实体预测和关系预测。由于 DKRL 模型和基线模型都基于平移假设，所以我们直接使用评分函数 $S(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$ 进行补全。例如，给定三元组的 h 和 r ，我们首先

使用 $\mathbf{h} + \mathbf{r}$ 得到一个向量，然后在所有实体集合 \mathcal{E} 中按照和 $\mathbf{h} + \mathbf{r}$ 向量的距离进行排序，得到预测结果。为了更好地展示两种实体表示的效果，在评测中，DKRL(CBOW) 模型和 DKRL(CNN) 模型均只使用实体基于描述的进行预测，而 DKRL(CNN)+TransE 模型使用实体基于描述和基于结构的两种知识表示联合进行预测。

在评测中，我们参考了 Bordes 论文中提出的评测方法，使用了平均排名和 N 命中率两个指标进行评测^[19]。其中，平均排名表示所有测例中正确答案的平均排序，关注知识表示的整体效果；而 N 命中率表示排名前 N 的答案中有正确答案的测例比例，关注在实际任务中模型补全的实用性。特别地，我们在关系预测任务中使用一命中率，在实体预测任务中使用十命中率对模型进行评测。评测结果分别如表 2.13 和表 2.14 所示。

表 2.13 FB15K 上实体预测的评测结果

评测指标	平均排名		十命中率 (%)	
	原始	过滤	原始	过滤
TransE	210	119	48.5	66.1
DKRL(CBOW)	236	151	38.3	51.8
DKRL(CNN)	200	113	44.3	57.6
DKRL(CNN)+TransE	181	91	49.6	67.4

表 2.14 FB15K 上关系预测的评测结果

评测指标	平均排名		一命中率 (%)	
	原始	过滤	原始	过滤
TransE	2.91	2.53	69.5	90.2
DKRL(CBOW)	2.85	2.51	65.3	82.7
DKRL(CNN)	2.91	2.55	69.8	89.0
DKRL(CNN)+TransE	2.41	2.03	69.8	90.8

从结果中，我们可以得到以下一些结论。

(1) 在实体预测与关系预测两个任务上，DKRL(CNN)+TransE 模型在所有评测指标上都取得了最好的结果。这个结果说明了实体描述确实蕴含着丰富的文本信息，而这些文本信息能有助于构建更好的知识表示。同时，这也说明了卷积神经网络编码器能够智能地选择实体描述中有信息量的特征，自动过滤可能的噪声。具体地，平均排名评测指标上的提升，说明 DKRL 模型学到的知识表示整体效果较好；而 N 命中率评测指标上的提升，则说明 DKRL 模型对于知识图谱的补全任务的推荐质量更高。

(2) 对比两种实体描述编码器的模型结果，发现 DKRL(CNN) 模型的结果在实体预测的平均排名、十命中率指标及关系预测的一命中率指标上都显著地优于 DKRL(CBOW) 模型。在关系预测的平均排名指标上，二者的结果也是相近的。这是因为与连续词袋模型相比，卷积神经网络不仅抓住了词级别的语义信息，而且通过卷积操作获得了局部的词序信息。另外，最大池化操作使得编码器能够更关注局部特征的强信号，避免长篇描述中的噪声干扰。综上所述，虽然 DKRL(CBOW) 模型更加简单快速，但 DKRL(CNN) 模型能够在知识图谱补全任务上获得更好的效果。

(3) 进一步地，我们还通过样例分析证实了有时仅通过结构信息是难以抓住实体相关的知识细节的。例如，如果知识图谱中没有直接提及，我们很难仅通过有限的三元组信息回答一个球员到底是前锋还是守门员这种细节问题；而如果通过实体描述，我们就可以在文字信息中挖掘可能相关的知识细节（如球员在比赛中的表现），从而给出正确的回答。

(4) 融合实体描述的知识表示在某些评测指标上，相比基线模型的提升似乎不太明显。这是因为对于一些三元组的补全任务，仅依靠结构信息就能达到不错的效果。另外，由于实体描述中存在较长的文本，仅使用卷积神经网络还是很难智能地抓住实体相关的每个细节信息。使用更好的实体描述编码器，或者使用更好的实体基于描述的表达方法，可能会是模型未来的提升点。除此之外，DKRL 模型在新实体的零样本知识表示构建任务上展示出了它的能力，而这是传统基于结构信息的知识表示学习模型难以做到的。

2. 实体类型分类结果与分析

实体类型分类任务旨在预测一个实体的所有类型，可以看作一个多标签分类的问题，在以往的工作中也常被研究^[142]。目前，大多数大规模百科或知识图谱都维护有自己的实体类型 (entity type) 信息。例如，在 Freebase 中，莎士比亚这个实体即有作家、艺术家和获奖提名名人等实体类型。实体类型分类任务能够从实体类型的粒度上证明知识表示的有效性。

我们首先从 Freebase 中抽取 FB15K 训练集中所有实体的全部实体类型，共得到 4 054 个实体类型。由于这些类型分布具有长尾性，有相当的实体类型仅仅出现了很少的次数，所以我们按照这些实体类型在对应数据集实体中出现的频次从高到低排序，选择频次前 50 的实体类型作为待分类的实体类型集（去除了话题这个类型，因为它几乎出现在所有实体的类型列表中）。前 50 的实体类型覆盖了 13 445 个实体，随机将这些实体分

为训练集与测试集，其中训练集含有 12 113 个实体，测试集含有 1 332 个实体。FB15K 中部分实体类型及其对应频次如表 2.15 所示。

表 2.15 FB15K 中实体类型及其对应频次

排名	实体类型	频次
1	人	4507
2	奖项提名者	4042
3	获奖者	3571
4	电影演员	3381
5	地点	3184

在训练时，我们使用基线模型和 DKRL 模型训练得到的知识表示作为实体的特征向量，使用逻辑斯蒂回归算法 (logistic regression) 作为分类器，并使用一对其他 (one-versus-rest) 的训练策略训练多标签分类的逻辑斯蒂回归分类器。在评测中，DKRL(CBOW) 模型和 DKRL(CNN) 模型同样只使用实体基于描述的代表作为特征向量。为了充分显示 DKRL 模型的优势，我们还实现了一个基于文本的经典词袋模型 (BOW) 作为补充基线模型。这个模型基于词袋模型，使用实体描述中词的 TF-IDF 值构建特征向量，然后使用回归算法进行分类。

在评测指标上，我们参考相关工作的设定，使用平均准确率均值 (Mean Average Precision, MAP) 作为评测指标^[142]。MAP 是多标签分类任务中常用的评测指标之一，在信息检索等领域被广泛使用，其公式为

$$\text{MAP} = \frac{\sum_{i=1}^{|Q|} \text{AveP}(i)}{|Q|} \quad (2.26)$$

其中， $|Q|$ 是测试集的样例数， $\text{AveP}(i)$ 是第 i 个测试样例的平均准确率，有

$$\text{AveP}(i) = \frac{\sum_{k=1}^n (P_i(k) \times \text{correct}_i(k))}{m_i} \quad (2.27)$$

其中， n 是候选类型的个数， m_i 是第 i 个测试样例真实拥有的实体类型数。在计算 $\text{AveP}(i)$ 时，首先将各候选类型的得分从高到低排序；然后依序判断是否是正确答案。 $P_i(k)$ 是第 i 个测试样例截至第 k 个候选类型时的准确率，而 $\text{correct}_i(k)$ 函数是一个指示函数，当且仅当第 k 个候选类型是正确答案时 $\text{correct}_i(k)$ 为 1，其余情况下为 0。

FB15K 上实体类型分类的评测结果如表 2.16 所示，我们分析 DKRL 模型在 FB15K 数据集上的结果，得到了以下结论。

(1) DKRL(CBOW) 模型和 DKRL(CNN) 模型在 MAP 评测指标上均能显著超过两个基线模型。这个结果从另一个侧面证明了实体描述中的丰富文本信息能够帮助模型建立更好的知识表示，从而为实体类型分类任务提供更加准确的指导信息。我们认为 DKRL 模型在实体类型任务上的优势主要来源于联合知识图谱的结构信息与实体描述的文本信息。这是由于与 DKRL 模型相比，TransE 模型使用的知识表示中仅编码了知识图谱三元组的结构信息，而 BOW 模型也仅编码了实体描述中的文本信息，所以 DKRL 模型学习到的知识表示在类型分类任务上更加有效。

(2) DKRL(CNN) 模型在实体类型分类任务上的表现优于 DKRL(CBOW) 模型。这个结果进一步证明了卷积神经网络模型能够更准确地对实体描述信息进行编码与表示。在实验过程中，我们发现，对于一些描述较长的测试样例，使用 DKRL(CNN) 模型的准确率通常要高于 DKRL(CBOW)。

表 2.16 FB15K 上实体类型分类的评测结果

评测指标	MAP
TransE	87.9
BOW	86.3
DKRL(CBOW)	89.3
DKRL(CNN)	90.1

3. 零样本知识图谱补全

在零样本知识图谱补全任务中，我们仅关注测试集中存在新实体的三元组的结果，即 FB20K 数据集中的 $d-e$ 、 $e-d$ 和 $d-d$ 3 类三元组。由于传统的平移模型无法对新实体进行表示，所以我们仅测试了 DKRL(CBOW) 模型和 DKRL(CNN) 模型的表示效果。在测试时，对于所有新实体，我们均使用其基于描述的表示作为知识表示，而对于测试集中出现在训练集中的实体，我们使用其两种表示分别进行测试。综上所述，我们在测试中一共使用了 4 种模型设定，其中 CBOW 表示测试中所有实体均使用基于描述的表示，而 Partial-CBOW 表示在训练集中有的实体使用基于结构的表示。同理，CNN 和 Partial-CNN 也分别表示在卷积神经网络模型下使用实体基于描述和基于结构的表示的模型。我们在实体预测与关系预测任务上分别进行了测试，并且展示的不同三元组类的十命中率（对实体）和一命中率（对关系）的实验结果。

FB20K 上零样本实体预测的评测结果如表 2.17 所示, FB20K 上零样本关系预测的评测结果如表 2.18 所示。

表 2.17 FB20K 上零样本实体预测的评测结果 (%)

三元组类型	$d - e$	$e - d$	$d - d$	总计
Partial-CBOW	26.5	20.9	67.2	24.6
CBOW	27.1	21.7	66.6	25.3
Partial-CNN	26.8	20.8	69.5	24.8
CNN	31.2	26.1	72.5	29.5

表 2.18 FB20K 上零样本关系预测的评测结果 (%)

三元组类型	$d - e$	$e - d$	$d - d$	总计
Partial-CBOW	49.0	42.2	0.0	46.2
CBOW	52.2	47.9	0.0	50.3
Partial-CNN	56.6	52.4	4.0	54.8
CNN	60.4	55.5	7.3	58.2

从这些结果中, 我们可以得到以下结论。

(1) 在实体预测与关系预测两个任务上, 基于 CNN 的模型在整体结果与各个子分类上均取得了最佳结果。对于训练集中未出现的新实体, 仅依靠其基于描述的表达, 即能在关系预测任务上达到将近 60% 的准确率。这说明 DKRL 模型学得的基于描述的表达即使在零样本情境下也能获得不错的效果, 体现了实体描述文本信息的重要性。需要指出的是, $d - d$ 类三元组中的头实体和尾实体均为新实体, 是理论上最为困难的测试样例类。然而 $d - d$ 类在零样本实体预测任务中取得远超其他三元组分类的准确率。这是因为 $d - d$ 类中有很大部分是自环型的关系, 即实体自身相连的关系, 所以实体预测反而更加容易。

(2) CNN 模型在两个任务上的表现均优于 CBOW 模型。具体地, CNN 模型在实体预测任务上比 CBOW 模型提高约 4.2%, 在关系预测任务上提高约 7.9%。这说明卷积神经网络模型与连续词袋模型相比考虑了词序信息, 能够更好地抓住实体描述中关键的文本信息。

(3) Partial-CBOW 模型和 Partial-CNN 模型在两个任务上也得到了不错的效果。这是因为 DKRL 模型的能量函数是由实体两种表示的四种组合联合形成的。通过 $\|\mathbf{h}_S + \mathbf{r} - \mathbf{t}_D\|$ 、 $\|\mathbf{h}_D + \mathbf{r} - \mathbf{t}_S\|$ 两个混合项的软限制, 实体基于描述和基于结构的表示倾向被学习到同一个语义空间中, 理想状态下能够相互替换。我们在实验过程中尝试过

使用更为严格的限制（如强制实体的两种知识表示相等）强化这种语义空间的可替代性，但是这会使得模型在其他任务上的表现效果下降。最后我们在模型中仅保留了混合项的软限制，在确保知识表示质量的前提下，一定程度维护了实体两种表示的可替代性。

4. 零样本实体类型分类

在零样本实体类型分类实验中，仿照上一节实体类型分类的实验设计，将前 50 类型覆盖的全部 13 445 个实体（这些实体在 FB20K 的训练集中出现过）作为训练集，使用实体基于描述的表达作为实体特征向量，采用逻辑斯蒂回归训练分类器。将 FB20K 中的 4 050 个被前 50 类型覆盖的新实体作为测试集，依旧使用 MAP 作为评测指标。我们测试了 DKRL(CBOW) 模型和 DKRL(CNN) 模型，由于平移模型无法对新实体进行表示，我们依然沿用了上一节的 BOW 模型作为基线模型。

FB20K 上实体类型分类的评测结果如表 2.19 所示。从结果中可以发现，DKRL(CNN) 模型取得了最佳结果，比 DKRL(CBOW) 模型和 BOW 模型分别提高了 9.9% 和 4.4%。这个结果再次验证了卷积神经网络构建的基于描述的实体表示较好地对结构信息与文本信息进行编码。

表 2.19 FB20K 上实体类型分类的评测结果 (%)

评测指标	MAP
BOW	57.5
DKRL(CBOW)	52.0
DKRL(CNN)	61.9

2.6.3 小结

在本节中，我们提出的 DKRL 模型实现了融合实体描述的知识表示学习任务。我们基于平移模型的思想建立知识表示学习框架，并创新性地为每个实体设置了两种知识表示——基于结构的表示与基于描述的表示。我们使用连续词袋模型与卷积神经网络模型两种实体描述编码器对实体描述进行建模，从实体描述中学习实体基于描述的知识表示。模型在知识图谱补全和实体类型分类等任务上进行了评测，均取得了超过基线模型的效果，同时在零样本情境下验证了模型对新实体的表示能力。实验结果充分证明了实体描述蕴含的丰富文本信息有助于建立更好的知识表示，同时说明融合实体描述的知识表示学习模型能够准确地对实体描述和三元组结构信息进行联合建模。

2.7 融合层次类型信息的知识表示学习

传统的知识表示方法 (如 TransE 及其扩展模型) 仅使用知识图谱之间的关系信息作为监督信号对实体嵌入施加约束, 但实际上实体层次类型也能够为表示学习提供大量重要的信息, 加深模型对实体与三元组的理解。由于实体层次类型信息往往具有人工定义的结构化体系, 它可以被看作一种较为准确的先验知识。一些广泛使用的大规模知识图谱 (如 Freebase、DBpedia 等) 及常见的百科类数据库 (如维基百科等) 都维护着一套自己的实体层次类型信息。

因此, 我们尝试将实体层次类型信息应用在世界知识表示学习中。我们认为, 特定关系下的实体应该有更需要突出的实体类型, 而实体在突出不同类型时应该有不同的知识表示。在图 2.11 所示的 Freebase 实体层次类型信息样例中, 我们发现在写作这个关系下, 威廉·莎士比亚更应突出书籍作者这个类型, 而《罗密欧与朱丽叶》更应突出书籍作品这个类型, 这些相对重要的实体类型在图 2.11 中以实线进行连接。所以, 对于图 2.11 中的三元组, 应该由威廉·莎士比亚在书籍作者上的知识表示与《罗密欧与朱丽叶》在书籍作品上的知识表示进行交互。为了利用实体的类型信息指导知识表示学习, 使实体在不同的情境下具有不同的嵌入表示, 我们提出了融合实体层次类型信息的知识表示学习模型 (TKRL)。

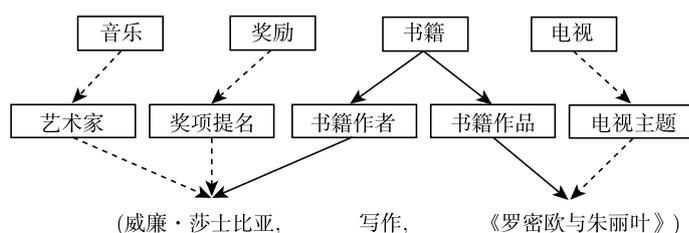


图 2.11 Freebase 实体层次类型信息示例

2.7.1 算法模型

如图 2.12 所示, 类似 TransR, TKRL 基于平移假设思想, 将实体以不同的方式投影到各自的关系空间中, 但不同的是, TKRL 对各个实体的投影矩阵可能不同, 且投影矩阵的构建受到实体层次类型指导, 这可以从 TKRL 的能量函数中看出。

$$E(h, r, t) = \|\mathbf{M}_{r,h}\mathbf{h} + \mathbf{r} - \mathbf{M}_{r,t}\mathbf{t}\| \quad (2.28)$$

其中, $\mathbf{M}_{r,h}$ 、 $\mathbf{M}_{r,t}$ 为关于关系 r 为头尾实体 h 、 t 构建的投影矩阵, 它们由层次类型编码器根据层次类型信息给出。具体地, 对于三元组 (h, r, t) , 根据知识图谱中关系特化的实体类型信息, h 在关系 r 下的头实体位置所应该突出的类型集合为 $C_{r,h} = \{c_{r,h}^1, c_{r,h}^2, \dots, c_{r,h}^n\}$, 则

$$\mathbf{M}_{r,h} = \frac{1}{n} \left(\mathbf{M}_{c_{r,h}^1} + \mathbf{M}_{c_{r,h}^2} + \dots + \mathbf{M}_{c_{r,h}^n} \right) \quad (2.29)$$

其中, \mathbf{M}_c 为层次类型 c 的投影矩阵。层次类型编码器为尾实体 t 构建的投影矩阵可以同理得到。

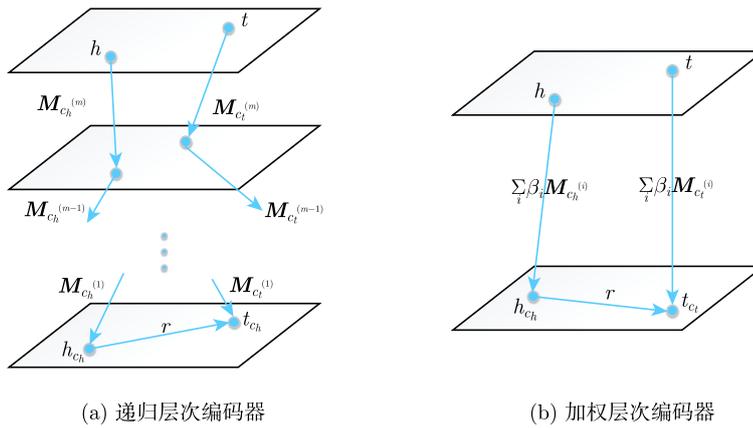


图 2.12 融合层次类型信息的知识表示学习模型图

假设有层次类型 $c = \{c^{(k-1)} \rightarrow \dots \rightarrow c^{(1)}\}$, 其在层次类型结构中对应该一条由子类型构成的链, 其中从粒度最细到粒度最粗的子类型依次为 $c^{(1)}$ 、 $c^{(2)}$ 、 \dots 、 $c^{(k-1)}$, 其对应的投影矩阵分别为 $\mathbf{M}_{c^{(i)}}$, 我们尝试使用以下两种策略在 TKRL 中构建层次类型 c 的投影矩阵 \mathbf{M}_c 。

- 递归层次编码: $\mathbf{M}_c = \prod_{i=1}^{k-1} \mathbf{M}_{c^{(i)}} = \mathbf{M}_{c^{(1)}} \mathbf{M}_{c^{(2)}} \dots \mathbf{M}_{c^{(k-1)}}$;
- 加权层次编码: $\mathbf{M}_c = \sum_{i=1}^{k-1} \beta_i \mathbf{M}_{c^{(i)}} = \beta_1 \mathbf{M}_{c^{(1)}} + \dots + \beta_{k-1} \mathbf{M}_{c^{(k-1)}}$, 其中 β_i 为子类型 $c^{(i)}$ 的权重, 随 i 增大等比例减小, 即 $\beta_i = \beta_1 \lambda^{i-1}$, $\sum_i \beta_i = 1, 0 < \lambda < 1$, 粒度越粗的子类型, 其权重越低。

子类型的投影矩阵被作为模型的参数直接学习得到。

2.7.2 实验分析

1. 知识图谱补全结果与分析

在实验中，我们使用了两种层次类型编码器对实体类型的层次结构进行建模，使用递归层次编码器的模型记录为 TKRL(RHE)，使用加权层次编码器的模型记录为 TKRL(WHE)。我们将训练中使用了软类型限制的模型加上 (+STC) 的标志，将测试中使用了类型限制的模型加上 (+TCE) 的标志。

我们使用融合实体层次类型的知识表示学习模型和基线模型在训练集上学到的实体与关系向量，对测试集中的三元组样例进行知识图谱补全。对于测试集中的每个样例，我们分别隐藏其头实体、关系和尾实体，用于进行实体预测和关系预测。在测试时，将全部实体（或关系）作为候选集，根据各模型的能量函数为每一个候选三元组进行评分。具体地，对于平移模型，使用 $E(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$ 的评分函数；对于 TransR 模型，我们使用 $E(h, r, t) = \|\mathbf{M}_r \mathbf{h} + \mathbf{r} - \mathbf{M}_r \mathbf{t}\|$ 的评分函数；而对于 TKRL 模型，我们使用 $E(h, r, t) = \|\mathbf{M}_{r,h,1} \mathbf{h} + \mathbf{r} - \mathbf{M}_{r,t,2} \mathbf{t}\|$ 的评分函数。该实验仍然使用平均排名和 N 命中率作为评测指标。我们在实体预测与关系预测两个任务上进行了测试，FB15K 上实体预测的评测结果如表 2.20 所示。FB15K 上关系预测的评测结果如表 2.21 所示。

表 2.20 FB15K 上实体预测的评测结果

评测指标	平均排名		十命中率 (%)	
	原始	过滤	原始	过滤
RESCAL	828	683	28.4	44.1
SE	273	162	28.8	39.8
SME(linear)	274	154	30.7	40.8
SME(bilinear)	284	158	31.3	41.3
LFM	283	164	26.0	33.1
TransE	238	143	46.4	62.1
TransR	199	77	47.2	67.2
TKRL(RHE)	184	68	49.2	69.4
TKRL(WHE)	186	68	49.2	69.6
TKRL(RHE+STC)	202	89	50.4	73.1
TKRL(WHE+STC)	202	87	50.3	73.4

分析实体预测与关系预测的评测结果，我们可以得到以下结论。

(1) 在实体预测与关系预测中，TKRL(RHE) 模型与 TKRL(WHE) 模型在所有评测指标上超过了所有的基线模型。这说明实体层次类型蕴含着丰富信息，能够帮助模型学到更加准确的知识表示，同时说明我们设计的递归层次编码器与加权层次编码器能够很好地对实体类型的层次结构进行建模，将实体层次类型信息顺利地融入知识表示中。

表 2.21 FB15K 上关系预测的评测结果

评测指标	平均排名		一命中率 (%)	
	原始	过滤	原始	过滤
TransE	2.79	2.43	68.4	87.2
TransR	2.49	2.09	70.2	91.6
TKRL(RHE)	2.12	1.73	71.1	92.8
TKRL(WHE)	2.22	1.83	70.8	92.5
TKRL(RHE+STC)	2.38	1.97	68.7	90.7
TKRL(WHE+STC)	2.47	2.07	68.3	90.6

(2) 我们将 TKRL 模型与同样基于平移假设的基线模型进行比较。与 TransE 模型相比，TKRL 模型在所有评测指标上的提升效果都十分显著。这从侧面证实了实体在不同类型下应该有不同知识表示的假设，一定程度上解决了平移模型框架对于复杂关系建模难的问题。另外，TKRL 模型与 TransR 模型相比也获得了不少提升。TransR 模型使用了关系特化的映射矩阵，使实体在不同关系下拥有不同的表示，而使用模型充分利用了实体层次类型信息及关系特化的类型信息，使得各个层次类型之间能通过层次结构和共享的子类型产生关联。实体预测的实验结果也证明，这种基于类型先验知识的映射矩阵融合了更多的信息，从而在实体表示上取得了更好的结果。

(3) TKRL(WHE+STC) 模型在实体预测十命中率(过滤)的评价指标上取得了最好的实验结果，在 TKRL(WHE) 模型的基础上又提升了约 3.8%。而训练中的软类型限制方法在递归层次编码器模型上也能带来相似的提升效果。这是因为在训练时，软类型限制能够增大拥有同类型的实体被选中组成负例三元组的概率。这样的训练方式与之前的方法相比，能够使得拥有相同类型的实体向量之间的差异度更高，从而在测试时得以对相似实体进行更精确的区分。然而，使用软类型限制的方法在平均排名指标上会稍稍下降。这是因为类型限制方法不可避免地减弱了相似实体在向量空间中的聚类现象，增加了极端错误的可能性，而平均排名指标较容易受到极端错误结果的影响，最终导致了排名指标结果的降低。

(4) 比较基于递归层次编码器和基于加权层次编码器的模型，我们发现，前者在关系预测上表现较好，而后者在实体预测上表现较好，但是两者相差较小。这一部分是因为我们在实验中使用的是基于 Freebase 的实体类型层次结构，而这些层次结构大多较浅，无法充分体现两种层次编码器的差别。但是在运行速度上，基于加权层次编码器的模型明显快于基于递归层次编码器的模型。

2. 测试中的实体类型限制分析

测试中的实体类型限制也能提高模型在知识图谱补全任务上的性能。由于测试中的实体类型限制方法受限于实体类型的完整度与准确度，为了显示模型的鲁棒性，我们仅将此部分结果单独作为辅助实验，并对测试中实体类型限制的效果进行详尽分析。

我们在 TKRL(RHE+STC) 模型和 TKRL(WHE+STC) 模型上加入测试中的实体类型限制，并相应地得到了 TKRL(RHE+STC+TCE) 模型和 TKRL(WHE+STC+TCE) 模型。需要注意的是，与 krompaß 等研究者的测试设定不同^[103]，我们在测试中仍使用实体集合中的全部实体作为候选集（即与传统知识表示学习的测试设定相同）。为了更充分地展示我们模型的优势，我们为基于平移假设的基线模型也增加了训练和测试中的实体类型限制。测试中的实体类型限制在实体预测上的效果如表 2.22 所示。

表 2.22 测试中的实体类型限制在实体预测上的效果

评测方法	平均排名		十命中率 (%)	
	原始	过滤	原始	过滤
TransE+TCE	212	116	46.9	63.4
TransR+TCE	182	60	47.9	68.6
TransE+STC+TCE	203	104	49.8	69.9
TransR+STC+TCE	185	63	48.5	71.7
TKRL (RHE+STC+TCE)	169	56	51.4	75.4
TKRL (WHE+STC+TCE)	170	55	51.3	75.6

将表 2.22 的结果与表 2.20 中对应模型的结果进行比较，我们可以得到以下结论。

(1) 包括 TKRL 模型、平移模型和 TransR 模型在内的所有模型，在加入测试中的实体类型限制时，实体预测的效果都有提升。在十命中率上，TKRL(WHE+STC+TCE) 模型的过滤评测指标上达到了 75.6%，比不使用测试时类型限制的模型进一步提升了 2.2%。这是因为测试中的实体类型限制去除了违反类型限制的候选实体，在实体类型较完整与准确时，能够缩小候选范围并较少地引入错误，从而提升实体预测的效果。

(2) 对于基于平移假设的基线模型，在训练中使用软类型限制能够进一步提升测试中类型限制的效果。以 TransE 模型为例，使用测试中的类型限制能够使得 TransE 模型在十命中率 (过滤) 评测指标上的结果提高 1.3%。而再加上训练时的软类型限制，TransE+STC+TCE 模型能够进一步提高 6.5%。这是因为训练中的软类型限制能够使得相似实体间的差异明显化，在联合使用测试中的实体类型限制时能够获得更好的效果。

(3) 对比所有使用了训练/测试中实体类型限制方法的模型，TKRL 模型仍然能够取得最佳的结果。这充分说明了根据实体层次类型建立类型特化映射矩阵的重要性与使用模型的鲁棒性。

3. 长尾分布数据上的结果与分析

由于真实世界中的知识图谱往往具有长尾分布，所以我们构造了 FB15K+ 数据集，用以测试模型在长尾分布下实体预测与关系预测的效果。与 FB15K 相比，FB15K+ 中增加了 510 个关系，而这些关系往往都具有较低频次。我们根据各关系在 FB15K+ 的三元组中出现的频次对测试集划分成组，展示了模型在实体预测与关系预测中各组测试集的 N 命中率 (过滤) 评测指标的结果。为了便于展示，我们仅对 TransE 模型、TransR 模型与 TKRL(WHE) 模型的结果进行比较。其结果如表 2.23 所示。

表 2.23 长尾分布数据集 FB15K+ 上的实体预测与关系预测结果

关系频次	实体预测十命中率 (%)			关系预测一命中率 (%)		
	TransE	TransR	TKRL(WHE)	TransE	TransR	TKRL(WHE)
≤ 10	28.0	32.4	38.1 (+5.7)	13.2	17.0	21.5 (+4.5)
≤ 100	49.9	54.5	57.9 (+3.4)	45.7	50.5	54.3 (+3.8)
$\leq 1\ 000$	66.1	69.1	71.6 (+2.5)	70.9	75.4	77.8 (+2.4)
全部	61.9	67.2	69.2 (+2.0)	80.4	88.8	89.7 (+0.9)

从表 2.23 我们可以得到以下结论。

(1) TKRL(WHE) 模型在所有频次测试组中的实体预测和关系预测结果均好于 TransE 模型和 TransR 模型，这充分证明了融合实体类型信息知识表示学习模型在真实的长尾分布数据上相对于基线模型也具有显著优势，也进一步说明了模型的有效性与鲁棒性。

(2) 通过观察发现，TKRL(WHE) 模型在低频关系下实体预测和关系预测的结果相比高频提升更大。在表 2.23 中，TKRL(WHE) 模型的结果后面标注了其相对 TransR 模型的提升结果。在关系频次小于等于 10 的分组下，TKRL(WHE) 模型在实体预测和关系预

测上分别比 TransR 模型提高了 5.7% 和 4.5%；而在全部测试集上的结果，TKRL(WHE) 模型在实体预测和关系预测上仅比 TransR 模型提高了 2.0% 和 0.9%。这是因为尽管 TKRL 模型与 TransR 模型都使用了映射矩阵的方法构建实体在不同情境下的表示，但是 TKRL 的构建基于实体层次类型，而 TransR 的构建基于关系。在对拥有低频关系的实体进行建模时，TransR 模型较容易受到数据稀疏及训练不充分的影响；而 TKRL 模型能够充分利用实体层次类型之间的联系，对低频关系下的实体也能较好地建立表示。

4. 三元组分类结果与分析

使用 FB15K 的测试集进行三元组分类的评测，然而知识图谱中并没有显式的负例三元组。我们参考 Socher 等研究者在三元组分类任务上的设定^[178]，基于测试集自动构建负例。具体地，对于测试集中每一个正例三元组 (h, r, t) ，随机替换掉其中的头实体或者尾实体，组成负例三元组 (h', r, t) 或者 (h, r, t') 。为了使三元组分类任务更具挑战性，更大程度上展示模型的性能差异，我们使用了实体类型限制的方法，强制要求替换的负例实体和被替换的原实体拥有相同的类型。例如，(莎士比亚, 写作, 哈姆雷特) 的负例中头实体应该也是一位作家，因为拥有错误类型的负例三元组很容易会被知识表示学习模型检测出来，从而失去评测意义。在实验中，对于验证集和测试集中的每一个三元组，我们以均等概率随机替换头实体或尾实体，产生对应的负例，这保证了三元组分类任务中正负例的数量一致。

我们使用 TransE 模型和 TransR 模型作为基线模型，与 TKRL 模型进行比较。在评测时，我们仍然使用在知识图谱补全任务中各模型对应的三元组评分函数 $E(h, r, t)$ 进行判定。具体地，我们会为每一种关系设定一个阈值 δ_r 。当 $E(h, r, t) > \delta_r$ 时，三元组被判断为负例；当 $E(h, r, t) < \delta_r$ 时，三元组被判断为正例。我们在验证集上进行优化，得到各关系所对应的阈值 δ_r ，用于对测试集进行评测。FB15K 上三元组分类的评测结果如表 2.24 所示。

从结果中，我们可以发现以下结论。

(1) TKRL 模型的准确率超过了所有的基线模型，其中 TKRL(WHE+STC) 模型得到了最好的结果。这从另一个角度说明了融合实体类型的知识表示学习模型能够学到更准确的知识表示，从而帮助三元组分类等任务。

(2) 加上训练中的软类型限制方法后，递归层次编码器模型和加权层次编码器模型都能获得进一步的提升。这说明了软类型限制能够学习到相似实体之间的差异，从而在

三元组分类任务上获得更加精确的结果。

表 2.24 FB15K 上三元组分类的评测结果

模型	准确率 (%)
TransE	85.7
TransR	86.4
TKRL(RHE)	86.9
TKRL(WHE)	87.1
TKRL(RHE+STC)	88.4
TKRL(WHE+STC)	88.5

2.7.3 小结

在本节中，我们提出了 TKRL 模型，融合实体层次类型信息辅助知识表示学习。我们在平移模型的思想做出改进，提出实体在不同类型下应该具有不同知识表示的假设，构建类型特化的映射矩阵，并使用递归层次编码器和加权层次编码器对类型的层次结构进行建模。我们还提出了软类型限制的策略，并在训练与测试中分别使用了类型限制的方法，进一步提高知识表示的精确度。模型在知识图谱补全和三元组分类等任务上进行了评测，均取得了超过基线模型的效果，同时在具有长尾分布的数据集上验证了模型对低频关系的表示效果。实验结果充分证明了实体层次类型蕴含了丰富的信息，能够帮助建立更精确的知识表示，同时说明提出的融合实体层次类型信息知识表示学习模型能够有效地联合编码并学习到实体类型的知识。

2.8 融合实体图像信息知识表示学习

图像信息是人类能够自然接收并认知的信息之一，以图像形式储存的信息往往更加灵活和丰富。研究表明，人类每日从外界进行的信息获取与交互行为中，很大一部分是通过视觉来完成的。通过图像信息的帮助，我们往往得以从多角度全方位地理解实体在文字或结构化信息之外的知识细节。主体部分描述指定实体的实体图像能够提供对应实体自身外形、行为和其他相关实体的视觉信息。实体图像信息来源十分丰富，一些世界知识图谱（如维基百科等）往往会有对应实体的图像信息，而专门的图像数据集更是储存着海量的实体相关图像信息。

受此启发，我们提出了融合实体图像信息的知识表示学习模型（IKRL）。如图 2.13 所示，对于一个三元组，IKRL 模型以其头实体和尾实体的所有实体图像作为输入，经过实体图像编码器得到各个图像在知识空间的表示。由于一个实体往往有多张质量良莠不齐的图像，IKRL 使用注意力机制自动评估图像的质量，并根据各图像的质量对它们的嵌入表示进行整合得到实体基于图像的表达。

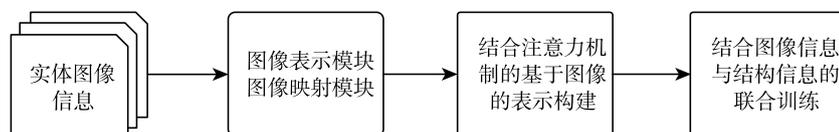


图 2.13 融合实体图像信息的知识表示学习模型流程图

2.8.1 算法模型

类似 DKRL，IKRL 也需要融合两种不同的信息以实现图像空间和知识空间的对齐和联合训练，故 IKRL 也采用了类似式 2.24 的能量函数。

$$E(h, r, t) = \alpha_1 \|h_S + r - t_S\| + \alpha_2 \|h_S + r - t_I\| + \alpha_3 \|h_I + r - t_S\| + \alpha_4 \|h_I + r - t_I\| \quad (2.30)$$

其中，下标 S、I 用于区分基于知识图谱和基于实体图像的实体表示。同时，IKRL 采用了与 DKRL 相同的带间隔的目标函数（见式 2.25）。

对于实体 e_k ，为了从其图像集合 $I_k = \{\text{img}_1^{(k)}, \text{img}_2^{(k)}, \dots, \text{img}_n^{(k)}\}$ 得到基于图像的实体表示，IKRL 首先使用图 2.14 所示的图像编码器对每个图像 $\text{img}_i^{(k)}$ 分别进行编码得到 $p_i^{(k)}$ ，然后应用基于注意力机制的多实例学习（MIL）方法，将多个图像得到的编码

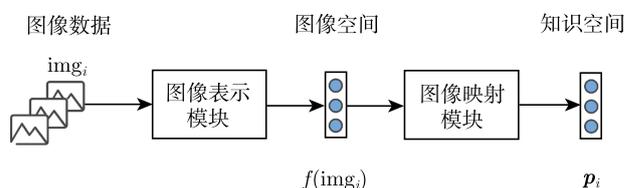


图 2.14 实体图像编码器示例图

$\mathbf{p}_1^{(k)}, \mathbf{p}_2^{(k)}, \dots, \mathbf{p}_n^{(k)}$ 整合为最终的嵌入表示 $\mathbf{e}_I^{(k)}$, 即

$$\mathbf{e}_I^{(k)} = \sum_{i=1}^n \frac{\text{att}(\mathbf{p}_i^{(k)}, \mathbf{e}_S^{(k)}) \cdot \mathbf{p}_i^{(k)}}{\sum_{j=1}^n \text{att}(\mathbf{p}_j^{(k)}, \mathbf{e}_S^{(k)})} \quad (2.31)$$

其中

$$\text{att}(\mathbf{p}_i^{(k)}, \mathbf{e}_S^{(k)}) = \frac{\exp(\mathbf{p}_i^{(k)} \cdot \mathbf{e}_S^{(k)})}{\sum_{j=1}^n \exp(\mathbf{p}_j^{(k)} \cdot \mathbf{e}_S^{(k)})} \quad (2.32)$$

其中, $\mathbf{e}_S^{(k)}$ 表示 $e^{(k)}$ 基于知识图谱的表示。 $\mathbf{e}_S^{(k)}$ 与 $\mathbf{e}_I^{(k)}$ 均将被用于式 (2.30) 中进行实体表示的训练。

2.8.2 实验分析

1. 知识图谱补全结果与分析

为了对 IKRL 模型的效果进行实验, 我们自己构建了一个包含丰富实体图像信息知识图谱数据集。为了保证构建数据集的三元组和图像质量, 我们选择从 ImageNet 中获取实体图像, 并从 WN18 数据集中选择头实体和尾实体均在 ImageNet 中有实体图像的三元组, 构建 WN9-IMG 跨模态知识图谱数据集。

由于 IKRL 模型为每个实体构建了基于结构和基于图像的两种实体表示向量, 在测试时, 我们根据使用实体表示的不同报告了 3 种不同设定下的结果: IKRL(SBR) 模型代表测试时仅使用基于结构的表示进行预测; IKRL(IBR) 模型代表测试时仅使用基于图像的表示进行预测; 与前两者不同, IKRL(UNION) 模型代表测试时使用简单的加权策略, 联合使用两种实体表示进行预测。WN9-IMG 上实体预测的评测结果如表 2.25 所示。

表 2.25 WN9-IMG 上实体预测的评测结果

评测指标	平均排名		十命中率 (%)	
	原始	过滤	原始	过滤
TransE	143	137	79.9	91.2
TransR	147	140	80.1	91.7
IKRL(SBR)	41	34	81.1	92.9
IKRL(IBR)	29	22	80.2	93.3
IKRL(UNION)	28	21	80.9	93.8

从结果中，我们可以得到以下结论。

(1) 所有的 IKRL 模型在全部评测指标上的实体预测结果都超过了基线模型，其中 IKRL(UNION) 模型得到了最好的结果。这说明实体图像中的丰富视觉信息能够帮助我们更深入地理解实体，同时说明我们的模型能够成功地将这些信息编码进实体表示中，提升实体预测的性能。

(2) IKRL(SBR) 模型和 IKRL(IBR) 模型在所有评测指标上都超过了基线模型。IKRL(IBR) 模型基于实体图像信息直接构建实体的表示向量，融合了实体视觉信息与实体在知识图谱中的结构信息，因此表示效果与基线模型相比得到了提升。而 IKRL(SBR) 模型虽然使用的是基于结构的实体表示，但是这些表示向量通过能量函数中两种表示的混合项训练，也间接学习到了一部分图像信息，从而使其表示能力得到提高。

(3) 所有的 IKRL 模型在平均排名评测指标上都得到了显著的提升。这是因为平均排名评测指标关注实体表示在向量空间中的整体效果，对错误样例比十命中率指标更加敏感。传统基于平移假设的模型仅考虑了知识图谱的结构信息，当所预测的信息丢失时很容易错得比较离谱。而我们使用的模型通过融合实体图像信息，能够间接地发现知识图谱中没有直接提出的实体潜在关系，从而使得知识表示的整体效果得到提高。

(4) 实验中 IKRL 模型是基于原始的平移模型框架进行训练的，但其在实体预测上的效果仍优于 TransR 模型，这反映了 IKRL 模型的有效性和鲁棒性。融合实体图像也能较容易地运用于基于平移模型的改进模型上。

2. 注意力机制的影响与分析

为了更深入地展示注意力机制在模型中起到的作用，我们还具体分析了在构建基于图像的表示时 3 种不同策略的影响。我们使用 IKRL(ATT) 表示使用注意力机制加权图像表示的策略（即表 2.25 中给出的模型），使用 IKRL(MAX) 表示仅考虑注意力最大的图像表示的策略，使用 IKRL(AVG) 表示使用所有图像表示均值的策略。为了更全面地进行比较，我们展示了在以上 3 种策略下基于结构的表示和基于图像的表示在实体预测任务上的评测结果。WN9-IMG 上各组合策略的实体预测结果如表 2.26 所示。

从结果中，我们可以得到以下结论。

(1) IKRL(ATT) 模型的两表示在所有评测指标上都得到了最好的结果。这是由于注意力机制能够智能地从多个实体图像中选择高质量的图像样例，构建实体基于图像表示，在充分利用实体图像多样性的同时，保证实体表示尽可能少地受到低质量图像的

噪声影响。

表 2.26 WN9-IMG 上各组合策略的实体预测结果

表示类型 评测方法	基于图像的表达				基于结构的表示			
	平均排名		十命中率 (%)		平均排名		十命中率 (%)	
	原始	过滤	原始	过滤	原始	过滤	原始	过滤
IKRL(MAX)	59	52	79.8	92.1	62	55	81.0	92.3
IKRL(AVG)	29	22	79.3	92.9	43	36	80.7	92.8
IKRL(ATT)	29	22	80.2	93.3	41	34	81.1	92.9

(2) 尽管使用了简单的 IKRL(MAX) 策略和 IKRL(AVG) 策略，融合实体图像信息的知识表示模型仍然能超过表 2.25 中基线模型的结果。这进一步说明实体图像信息对知识表示构建的重要性及 IKRL 模型编码图像信息的鲁棒性。

(3) IKRL(AVG) 策略比 IKRL(MAX) 策略的整体结果更好，在平均排名指标上尤其如此。这是因为 IKRL(AVG) 策略综合考虑了所有的实体图像信息，虽然不可避免地会引入一些噪声，但是仍比仅考虑质量最高图像的 IKRL(MAX) 策略具有优势。

(4) 通过比较可以发现，和 IKRL(AVG) 策略相比，IKRL(ATT) 策略的优势似乎并不特别明显。这是由于我们构建评测数据集时，为了保证实体图像的质量，选择从 ImageNet 中直接抽取实体图像的缘故。WN9-IMG 数据集中的图像质量整体较高，一定程度上导致了基于注意力机制的策略优势不明显，因为几乎所有图像都能提供实体相关的正确信息，彼此应得的注意力相当。为了充分展示注意力机制对模型的正面影响，我们在结果样例分析中进一步探索，给出了注意力机制从多个实体图像中成功分辨出高信息量图像的实例。

3. 三元组分类结果与分析

我们在 WN9-IMG 数据集上进行三元组分类的评测。在测试时，我们依靠验证集上的分类效果为每一种关系设置一个关系特化的阈值 η_r 。以 IKRL 模型为例，当其三元组的评分函数 $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\| > \eta_r$ 时，三元组被判断为负例；而当 $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\| < \eta_r$ 时，三元组被判断为正例。其他模型也按照自己的评分函数进行分类判断。为了使比较充分，我们测试了 IKRL(MAX)、IKRL(AVG)、IKRL(ATT)3 种策略下的实验结果。WN9-IMG 上三元组分类的评测结果如表 2.27 所示。

表 2.27 WN9-IMG 上三元组分类的评测结果

模型	准确率 (%)
TransE	95.0
TransR	95.3
IKRL (MAX)	96.3
IKRL (AVG)	96.6
IKRL (ATT)	96.9

从结果中，我们可以发现：

(1) 所有 IKRL 模型都获得了比基线模型更好的效果，这从另一个角度重新证实了实体图像信息的重要性，也说明 IKRL 模型能够联合编码实体结构信息与视觉信息，具有较高的鲁棒性。

(2) 在 3 种组合策略中，IKRL(ATT) 策略获得了最好的结果。这说明注意力模型能够自动选取信息量更高的实体图像组成基于图像的实体表示，在保证图像质量的前提下充分利用实体图像的多样性，获得实体相关更完整的视觉信息。而 IKRL(AVG) 策略也获得了比 IKRL(MAX) 策略更好的效果，说明综合考虑多个实体图像带来了模型效果的提升。

4. 图像-知识空间的语义平移规则

word2vec 模型得到的是有趣的语义平移规则，受此工作启发，我们尝试探索 IKRL 模型学到的跨模态知识表示是否也具有这种语义平移现象。具体地，我们使用实体基于图像的进行语义平移规则的探索。实验结果表明，在跨模态的图像-知识空间中，语义平移规则是普遍存在的。图像-知识空间的语义平移规则示例如图 2.15 所示。



图 2.15 图像-知识空间的语义平移规则示例

以图 2.15 所示的示例为例，设 $IBR(x)$ 表示实体 x 的基于图像表示，我们首先计算 $IBR(\text{柜子}) - IBR(\text{抽屉}) + IBR(\text{琴键})$ ，得到一个向量，然后以整个实体集为候选，根据实体与此向量的距离进行排序，发现最相近的实体向量是 $IBR(\text{钢琴})$ 。与之前的工作不同，由于我们是在跨模态图像-知识空间进行的测试，实体之间的向量差 $IBR(\text{柜子}) - IBR(\text{抽屉})$ 、 $IBR(\text{钢琴}) - IBR(\text{琴键})$ 与真实的关系属于成功对应，将语义平移规则中隐藏的联系显式地表现出来，展示了 IKRL 模型的有效性。

2.8.3 小结

在本节中，我们提出了融合实体图像信息的知识表示学习模型——IKRL 模型，以构建知识的跨模态表示。特别地，模型为每个实体设置了基于结构的表示和基于图像的表示，在平移模型的学习框架下进行联合训练。我们使用基于深度神经网络的图像表示模块抽取实体图像的特征，然后将图像特征映射到知识空间中，并针对实体图像的多样例学习提出基于注意力的模型，最终构建实体基于图像的表示。模型在知识图谱补全和三元组分类等任务上进行了测试，实验结果验证了实体图像蕴含的丰富视觉信息的重要性，也说明我们的模型能够较好地多实例实体图像中智能准确地抽取实体的相关知识以构建实体表示。我们还结合实例分析，探索了图像-知识空间的语义平移现象和注意力机制对模型的影响。

2.9 本章总结

本章介绍了针对世界知识图谱的表示学习方法。由于世界知识图谱具有信息复杂、规模庞大、完整度低的特点，我们在平移模型的基础上，从世界知识图谱结构建模和多源信息融合两个不同的角度为世界知识图谱的表示学习提供了解决的思路。

在世界知识图谱结构建模方面：

(1) TransR 模型和 CTransR 模型通过为每种关系定义单独的语义空间、允许同一个实体在不同关系场景下具有不同的语义，增强了对复杂关系的建模能力。

(2) PTransE 模型通过引入知识图谱中实体间的多步关系路径，更加充分地利用了知识图谱的结构信息。

(3) KR-EAR 模型根据关系自身的特点将一般意义下的关系分为了属性和关系两类, 并通过为属性这类关系设计特殊的模型, 改善了知识表示学习一对多和多对一关系处理效果欠佳的问题。

在多源信息融合方面:

(1) DKRL 和 IKRL 分别对实体的描述文本和图像进行了编码, 通过在能量函数中进行知识空间和文本/图像空间的对齐将描述和图像信息引入知识表示学习。

(2) TKRL 使用知识图谱中实体的层次类型信息来构建实体的投影矩阵并根据特定关系突出不同的实体类型, 将实体类型信息引入知识表示学习。

通过一系列实验, 我们验证了对关系进行深入建模的有效性, 也展示了不同类型的外部信息在知识表示学习中发挥的突出作用, 这些都将有助于更好地表示知识图谱中的世界知识。