大规模结构化知识的表示学习、自动获取与计算应用

（申请清华大学工学博士学位论文）

培 养 单 位：计 算 机 科 学 与 技 术 系
学　　科：计 算 机 科 学 与 技 术
研　究　生：林　衍　凯
指 导 教 师：孙　茂　松　教　授

二〇一九年六月

Representation Learning, Automatical Acquisition and Computational Application for Large-scale Structured Knowledge

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Computer Science and Technology

by

Lin Yankai

Dissertation Supervisor : Professor Sun Maosong

June, 2019

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

知识图谱是人工智能研究和智能信息服务基础核心技术，能够赋予智能体精准查询、深度理解与逻辑推理等能力。目前，基于深度学习的自然语言处理技术只能从数据中机械地学习完成特定任务的语义模式，不具备鲁棒性和可解释性，做不到对语言的深层理解与推理。我们认为要想实现真正的自然语言理解，需要在现有深度学习技术的基础上融合知识图谱信息。实现自然语言处理与知识图谱的融合并非轻而易举，需要解决几个关键问题：

（1）**知识表示**。在深度学习模型中充分利用大规模知识图谱，需要首先解决知识图谱表示的问题。在这方面，我的工作包括：a. **考虑知识图谱复杂关系的知识表示**：我们提出了基于映射矩阵进行空间投影的知识图谱表示模型，用于处理知识图谱中的复杂关系。b. **考虑知识图谱复杂路径的知识表示**：我们认为实体之间多步的关系路径同样包含着丰富的关系推理信息，并提出了一种基于路径表示的知识图谱表示模型。c. **考虑知识图谱复杂属性的知识表示**：我们提出了一种同时学习知识图谱中实体、关系和特性表示的知识图谱表示模型，以提高知识图谱表示的质量。

（2）**知识获取**。如何从互联网大规模的结构化、半结构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。在这方面，我的工作包括：a. **基于选择注意力机制的关系抽取**：针对远程监督数据中存在大量的噪音的问题，我们提出了一个基于句子级别选择注意力机制的神经网络关系抽取模型，用于过滤错误标注的句子。b. **基于多语言注意力机制的关系抽取**：现有的关系抽取系统通常专注于如何更好地利用单语言数据，忽略了多语言数据对于关系抽取任务的帮助。针对这个问题，我们提出了一个基于多语言选择注意力机制的关系抽取模型。

（3）**知识应用**。面向不同自然语言处理任务，我们需要探索将知识合理地融合到该任务下的深度学习模型中，实现知识指导的自然语言理解。在这方面，我的工作包括：a. **基于知识的实体分类**：我们提出了基于知识注意力机制的实体分类模型，用于考虑命名实体和上下文之间的联系以及知识图谱中丰富的有关信息。b. **基于知识的开放域问答**：我们借鉴人类回答问题的模式提出了一个基于‘‘粗读-精读-总结’’模式的开放域问答系统。

我们的工作有效地解决了面向知识图谱的知识表示、知识获取、知识应用中的关键问题，为迈向真正的自然语言理解打下了坚实的基础。

关键词：知识计算；知识表示；知识获取；知识应用

Abstract

In the 21st century, with the deep integration of artificial intelligence technology and fields such as home, medicine, education, finance, and law, people have a strong demand for large-scale knowledge graphs, and their applications. Knowledge intelligence has become one of the most popular areas of artificial intelligence. Knowledge graphs can be viewed as knowledge systems which store structured human knowledge. Knowledge graph is the core of artificial intelligent service. It can give the intelligent agent the ability of precise query, deep understanding and logical reasoning. It is widely used in search engine, question answering, dialogue system, and recommendation system.

Until now, deep learning based natural language processing can only learn the semantic pattern for specific tasks from the data mechanically due to the lack of background knowledge. It is neither robust nor interpretable, and cannot be able to understand natural language. We believe that to achieve genius natural language understanding, it is necessary to integrate knowledge graph information into deep learning. Combining natural language processing and knowledge graphs is not trivial, and some key issues need to be addressed:

(1) **Knowledge Representation**. Deep learning based natural language processing models usually use distributed representations. To utilize large-scale knowledge graphs in deep learning models, we need first to represent knowledge graphs. In this part, my work consists of a. **Knowledge graph representation considering complex relations:** Due to the complexity of information in different scenarios, the unified entity representation greatly limits the modeling ability of TransE and its extensions. To solve this problem, we propose TransR model which models entities and relations in distinct spaces and performs the translation in the corresponding relation space. b. **Knowledge graph representation considering complex relational paths:** Most existing knowledge representation learning model only consider the direct relation between entities. We believe that the multi-step relational paths between entities also contains rich reasoning information, and propose a path-based TransE model. c. Most existing knowledge representation learning models cannot distinguish the differences of the relations between entities and the attributes of entities. Hence, it is impossible to accurately represent the interaction between entities, relationships, and attributes. To solve this problem, we propose a knowledge graph

representation learning model that simultaneously learns the representation of entities, relation, and attribute.

(2) **Knowledge Acquisition**. To automatically acquire relational facts from the large-scale structured, semi-structured and unstructured data on the Internet is the only way to build a large-scale knowledge graph. In this part, my work consists of: a. **Neural relation extraction with selective attention:** To address the noise problem of distantly supervised relation extraction data, we propose a neural relation extraction model based on the sentence level selective attention mechanism, which is used to filter the sentences with incorrect annotation. b. **Neural relation extraction with multi-lingual attention:** Most existing relation extraction systems concentrate on extracting relational facts on mono-lingual data, which cannot utilize diverse information hiding in the data with various languages. To address this problem, we propose a multi-lingual neural relation extraction system which employs multi-lingual attentions.

(3) **Knowledge Application**. For different natural language processing tasks, we explore how to integrate knowledge into task-specified deep learning models to achieve knowledge-driven natural language understanding. In this part, my work consists of a. **Knowledge-driven entity typing:** We propose a neural entity typing model based on the knowledge attention mechanism, which considers the relationship between name entities and contexts and the rich information in the knowledge graphs. b. **Knowledge-driven open-domain question answering:** We propose an open domain question answering system based on the pattern of "skimming-intensive reading-summarizing". Besides, we use knowledge representation learning to enhance the representation of the question and its relevant articles and perform multi-task learning with relation extraction to introduce the relations of entities in knowledge graphs to the model.

Our work addresses the key problems in knowledge representation, knowledge acquisition, and knowledge application, which would be the foundation of genius natural language understanding.

Key Words: Knowledge Computing; Knowledge Representation; Knowledge Acquisition; Knowledge Application

目 录

主要符号对照表

Knowledge Representation Learning	知识表示学习
Distributed Representation	分布式表示
One-hot Representation	独热表示
Relation Extraction	关系抽取
Recurrent Neural Networks	循环神经网络
Convolutional Neural Networks	卷积神经网络
Question Answering	自动问答
Entity Typing	实体分类
G	知识图谱
$\mathcal{E}$	实体集合
$\mathcal{R}$	关系集合
e	知识图谱中的一个实体
r	知识图谱中的一种关系
(h, r, t)	知识图谱中的一个关系事实三元组

第 1 章　引言

1.1　研究背景

知识是人类智能的象征。知识对人工智能而言也同样具有重要意义。自 1955 年达特茅茨研讨会首次提出人工智能以来，70 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脱胎于此。早期人工智能研究大多关注如何利用搜索机制来解决智能问题，麻省理工大学著名学者约翰 · 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常识信息。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，1970 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。费根鲍姆进而在 1977 年发表文章，正式提出 ‘‘知识工程’’ 的思想，至此，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之一。

进入 21 世纪，得益于互联网的海量数据以及计算资源的增长，以深度学习技术为代表的数据智能突飞猛进，在计算视觉、自然语言处理等各个领域取得了突破性的进展。而在知识智能方面，2012 年搜索引擎巨头谷歌发布了知识搜索产品 Google Knowledge Graph（谷歌知识图谱），提出 “Things，Not Strings’’ 的理念。谷歌公司认为：对于用户输入的查询，我们需要做到的不仅仅是基于关键词匹配返回在字符层面相似的网页，而是要做到真正理解用户的意图。因此，谷歌的搜索引擎会根据查询中提及的人名、地名、机构名等实体信息进行基于知识的检索，同时还向用户展示这些实体的相关信息。如图1.1所示，用户输入 ‘‘科比’’ 时，谷歌不仅返回相关网页，还会直接展示科比的生日、配偶、子女等信息。此外，通过使用知识图谱进行逻辑推理，谷歌还能够直接回答用户提出的一些简单问题，如 ‘‘科比的妻子是谁？’’ 等，显著提升搜索引擎的用户使用体验。与此同时，人工智能技术与家居、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以及在此之上的智能问答和推理等应用的旺盛需求，知识智能已经成为目前人工智能领域的最热门的方向之一。

图 1.1　谷歌知识图谱样例。

1.1.1　知识图谱

知识图谱，就是将人类知识结构化形成的知识系统。知识图谱是人工智能研究和智能信息服务基础核心技术，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、智能对话系统以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[1]，Freebase[2]，DBpedia[3]，YAGO[4] 等。以 WikiData 为例，截至 2019 年初已经包含超过 5700 多万实体。与此同时，国内外各大互联网公司也均有知识图谱产品，如谷歌知识图谱、百度知心、搜狗知立方等。

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）。知识图谱将世界上的事物和概念表示为实体（entity），将实体之间的联系表示为关系（relation），以关系事实三元组（triple）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）形式的关系事实三元组，其中 ‘‘比尔 · 盖茨’’ 被称为头实体（head entity），‘‘微软公司’’ 被称为尾实体（tail entity），‘‘创始人’’ 被称为关系（relation）。通过众多三元组构成的链接，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。

目前，以 Freebase、WikiData 为代表的知识图谱已经在信息检索、问答系统等

领域获得比较广泛的应用。进入深度学习时代，数据驱动的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，是我们接下来需要研究的内容。

1.1.2　数据驱动的自然语言处理

现在人工智能领域几乎无人不谈深度学习，我们以自然语言处理为例来探讨数据驱动的深度学习技术的优势和面临的挑战。在过去几十年，人们主要探索了两种自然语言处理的方法，一种是规则驱动的方法，另外一种是数据驱动的方法。规则驱动的方法认为，人类语言主要是由语言规则来产生和描述的，只要将人类语言规则整理出来，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。数据驱动的方法则认为，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，随着互联网的普及，各个不同领域积累下了海量的文本数据。与此同时，计算机的存储和计算能力遵循着摩尔定律逐步上升，为数据驱动的方法提供了肥沃的发展土壤。基于深度学习的自然语言处理技术则是数据驱动的方法的最新方向。与传统基于统计的自然语言处理技术相比，基于深度学习的自然语言处理技术有两个突出特点：

（1）分布式表示。相比于传统统计方法以字词的离散符号作为基础处理对象，在深度学习技术中，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。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，提供了非常广阔的参数空间，可以学习到自然语言中的复杂语义模式。

（2）深层架构。神经网络的深层架构赋予了深度学习模型强大的理解和推理能力，能够很好地从海量的训练数据中自动学习自然语言处理任务需要的相关能力。

从 2011 年 Hinton 提出深度学习的概念以来，深度学习在短短几年里横扫自然语言处理的各类重要任务，在信息检索、机器翻译、自动问答、对话系统等领域均取得了显著进展。然而，随着更加深入的探索，人们发现深度学习虽然已经取得了很好的效果，仍然存在着很大的局限性。其中最关键的一点是，基于深度学习的自然语言处理技术的鲁棒性和可解释性较差：有研究者发现，针对一个基于深度学习的自然语言处理模型，我们可以很容易地构造一些对抗样例（Adversarial Example）来欺骗该模型。例如，在自动问答系统中，研究者可以根据问题构造一个和问题结构十分相似但包含错误答案的句子放到文档中，深度学习模型往往会被对抗样例欺骗而提取出人工构造的错误答案。而这种对抗样例对于人类而言非

常好识别，不会对人类回答问题造成很大干扰。归根到底，基于深度学习的自然语言处理仍然是对已有训练数据的拟合，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’’。

如何才能让现有的基于深度学习的自然语言处理模型具备真正的文本理解与推理能力呢？我们应当参考人脑的工作模式。例如，当人们看到一句话 ‘‘今晚的月色真美’’，会产生哪些理解呢？如果这个人缺乏对应的背景知识，他可能只能理解这句话的表面含义：今天月明星稀，月光特别美。但是，事实上要真正理解这句话，我们需要了解相关背景知识：这句话是来自于夏目漱石教英文课时的一个短文翻译，要求对文中男主角在月光下散步时情不自禁说出的 “I love you’’ 进行翻译。学生直接翻译为字面意思的 ‘‘我爱你’’，但是夏目漱石说，日本人对爱情的表达应当更婉转含蓄，只需要说 ‘‘月が绮丽ですね（今晚的月色真美）’’ 就足够了。因而，如果我们没有背景知识是无法真正理解一句话的深层含义的。

对于基于深度学习的自然语言处理模型而言也是如此，如果没有相关背景知识的支持，深度学习模型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，我们认为要想实现真正的自然语言理解，需要在现有基于深度学习的自然语言处理模型的基础上融合各种类型的背景知识。

1.1.3　知识驱动的自然语言处理

过去几十年的实践表明，数据驱动的方法是实现数据智能的可行路径。以深度学习为代表，数据驱动的方法旨在探索如何更充分地从大规模数据中学习和挖掘有用信息。在自然语言处理方面，从 2013 年 word2vec[5] 词向量学习算法，到 2018 年 BERT[6] 预训练语言模型，其主要思想都是利用大规模无标注文本数据学习通用的语言知识。这些数据驱动的方法对整个自然语言处理领域都产生了非常大的影响。如何更充分挖掘数据的价值，无论是有标注数据、弱标注数据还是无标注数据，都是数据驱动的方法的重要命题。这其中仍然有很多开放问题等待研究者去探索和解决。

然而，正如上一小节所讲，单纯数据驱动的方法无法实现有理解能力的自然语言处理。要实现真正的自然语言理解，还需要人类知识的指导。为此，我们需要探索如何充分融合数据智能和知识智能。我们的目标就是，更好地将结构化知识图谱融入目前基于深度学习自然语言处理模型中。

实现自然语言处理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**。基于深度学习的自然语言处理模型通常采用分布式表示，而知识图谱采用符号表示，这两种表示方案互有优劣。如何在自然语言处理模型中充分利用大规模知识图谱，需要首先解决知识表示的问题。

（2）**知识获取**。知识图谱形式非常丰富，根据以往经验，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、半结构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）**知识应用**。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，将知识合理地融合到该任务下的深度学习模型中，实现知识驱动的自然语言理解。

1.2　知识表示

大规模知识图谱是人类理性知识的总结，主要以离散符号形式进行表示和存储，例如采用三元组形式来存储实体之间的关系事实，无论是实体还是关系，都采用独一无二的符号来进行标识。离散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，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高效运行，且难以拓展至其它场景。

（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事实。对这些长尾的实体和关系，往往很难有效理解与推理。

为了解决计算效率低与数据稀疏两个问题，近年来人们提出知识表示学习（Knowledge Representation Learning）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[7]（Distributed Representation）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距离也相近。而传统对知识的离散符号表示其实相当于独热表示（One-hot Representation），即将知识表示成为一个长向量，只有该知识对应的特定维度非零，而其它所有维度都为零。与独热符号表示相比，对知识的分布式表示有以下优点：

（1）显著提升计算效率：分布式表示学习到的是实体与关系的低维向量表示。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效计算，相比于传统

的高维空间中的图算法可以显著提高计算效率。

（2）有效缓解数据稀疏：独热表示基于所有实体（或关系）相互独立的假设，所有表示向量之间两两正交，丢失了大量实体（或关系）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）实现异质信息融合：分布式表示能够将多源异质信息映射到同一语义空间中，建立多语言跨模态的信息交互，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链接预测、知识推理、自动问答、信息检索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。

1.3　知识获取

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图谱，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都面临巨大挑战。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知识图谱，就是自动获取与人工校验结合产生的杰作。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。该任务目标是，给定一个包含两个实体的句子，从中抽取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。基于深度学习技术的神经网络关系抽取模型是目前解决关系抽取最好方法。

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，人们提出远程监督思想，利用已有知识图谱自动标注大规模训练数据。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1.2所示，远程监督认为同时包含 ‘‘比尔 · 盖茨’’ 和 ‘‘微软公司’’ 的三个句子都能够反映 ‘‘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。但是远程监督会不可避免地引入噪音标注，例如上图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，这些噪音训练数据会显著影响关系抽取深度学习模型的性能。此外，如何充分利用多语言文本数据进行关系抽取也是关系抽取深度学习模型面临的挑战性问题。

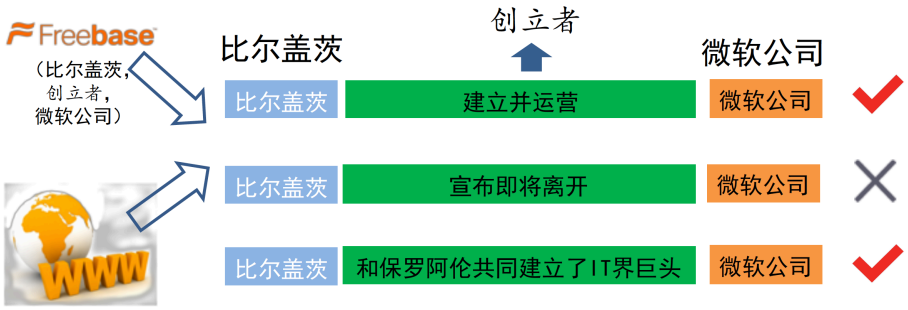

图 1.2　远程监督自动构建关系抽取数据集

1.4　知识应用

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。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。相关自然语言处理任务会在处理时特殊考虑知识图谱中标注的实体，为这些实体增加额外的特征信息，用于文本分类、信息检索等任务。

（3）利用网络结构。以自动问答系统为例，可以利用网络结构完成问答任务，例如回答诸如 ‘‘科比的妻子的父亲是谁？’’ 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 ‘‘父亲的父亲’’ 是 ‘‘爷爷’’ 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。

进入深度学习时代以后，如之前所探讨，为了实现真正的自然语言理解与推理，我们需要在基于深度学习的自然语言处理模型中融合大规模知识图谱的信息。因此我们认为，在知识表示学习等新兴技术的支持下，我们可以在基于深度学习的自然语言处理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实体以及其相互之间的关系提供丰富的外部信息，在探索改进深度学习的鲁棒性、可解释性等的过程中发挥关键作用。

1.5　本文主要内容

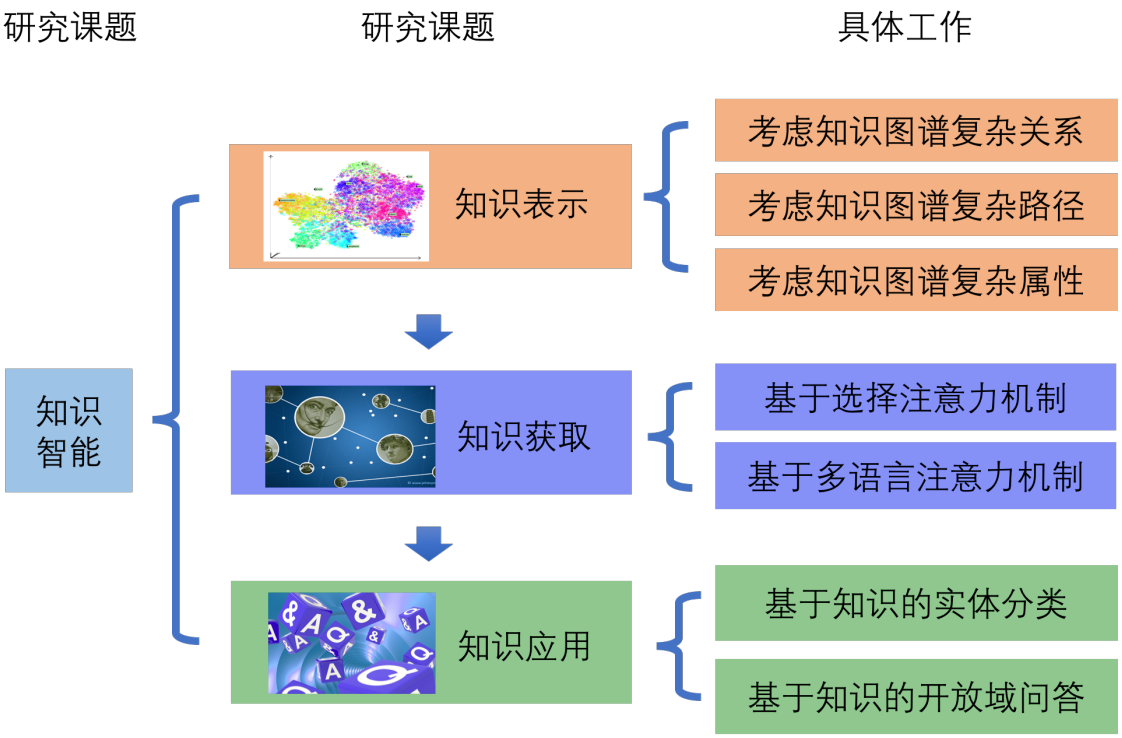

图 1.3　工作框架

如图1.3所示，本文针对知识智能中的三个关键问题：面向知识图谱的知识表示、知识获取和知识应用系统性的进行了以下七个工作：

- **考虑知识图谱复杂关系的知识表示学习**：在知识图谱中，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“奥巴马”既是美国总统，也是“米歇尔”的丈夫。由信息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现有知识表示学习模型对实体语义的建模能力。针对这一问题，我们提出了基于映射矩阵进行空间投影的 TransR 模型，用于处理知识图谱中的复杂关系，如一对多，多对多关系等。更进一步，我们采用聚类的方法对于每一种关系进行细分，分别学习向量表示。在实验上表明，我们的工作在多项知识驱动任务上均取得了显著的提升。
- **考虑知识图谱复杂路径的知识表示学习**：现有的大多数方法在知识表示学习中只考虑了实体之间的直接关系。我们认为实体之间多步的关系路径同样包含着丰富的关系推理信息，并提出了一种基于路径表示的知识表示模型。该模型在知识表示学习中将关系路径表示为实体之间的平移向量，主要贡献有：（1）由于并不是所有的关系路径都是有意义的，我们设计了一种基于路

径的资源分配算法用于计算不同关系路径的可信度；（2）我们用关系路径上的关系的不同语义组合向量来表示关系路径的向量。实验结果表明我们对知识图谱中的复杂路径进行考虑可以有效地提升模型在知识图谱链接预测上的效果。

- **考虑知识图谱复杂属性的知识表示学习**：近年来在知识图谱的研究上取得了显著的进展，表示学习运用于知识图谱中，将所有实体与关系映射到一个低维连续向量空间中，解决了之前知识图谱学习时产生的稀疏性与效率问题。但是，目前已有的知识图谱表示学习方法在学习时将实体之间的关系和实体的特性混为一谈，使用同一种模型进行建模，无法精确的表示实体、关系和特性之间的相互联系。针对此问题，我们提出了一种同时学习知识图谱中实体、关系和特性表示的知识表示学习方法，解决现有技术中存在的无法精确表示实体、关系和特性之间联系的问题，以提高知识图谱表示的质量。实验结果表明，我们提出的知识表示学习框架在知识图谱链接预测的三个子任务上都取得了一致的提升。
- **基于选择注意力机制的文本关系抽取**：文本关系抽取任务目标为给定一个包含两个实体的句子，从中抽取出两个实体之间的关系。现有的深度学习神经网络需要大量的学习训练数据，而目前文本知识获取中远程监督标注的训练数据由于存在大量的噪音，无法很好的用于深度学习神经网络的训练中。远程监督数据标注技术假设包含一个实体对的所有句子都表示了该实体对之间的关系。对于事实关系三元组（比尔 · 盖茨，创立者，微软公司），远程监督假设包含 ‘‘比尔 · 盖茨’’ 和 ‘‘微软公司’’ 的所有句子都表达了两者之间的关系 ‘‘创立者’’。然而，句子 ‘‘比尔 · 盖茨宣布从今天开始卸任微软公司首席执行官’’ 虽然同时包含了两个实体，但是事实上讲述的是比尔盖茨卸任的事情，和关系 ‘‘创立者’’ 并没有关系。可以看到，远程监督数据由于其较强的假设，虽然标注了大量的训练数据，也引入了大量的噪音。针对这个问题，我们提出了一个基于句子级别选择注意力机制的神经网络关系抽取模型，动态地将错误标注的句子的权重降低。实验结果表明，我们提出的选择注意力机制可以极大地提升关系抽取的效果。
- **基于多语言注意力机制的文本关系抽取**：目前，基于深度学习的神经网络关系抽取已经在关系抽取任务上取得了不错的效果。然而，绝大部分现有的关系抽取系统专注于如何更好地利用单语言数据，忽略了多语言数据对于关系抽取任务的帮助。我们认为多语言数据对于关系抽取任务的帮助在于以下两点：（1）一致性，不同语言对于同一事实表达相似，可以帮助我们对远程监

督数据中的噪音进行过滤；（2）互补性，根据数据统计，超过 40% 的事实通常只由一种语言进行描述，而对于超过一半的关系，其在不同语言中的语料也有较大差异。因此，我们可以利用语料较多的语言帮助语料较少的语言训练关系分类器。针对这个问题，我们提出了一个基于多语言选择注意力机制的关系抽取模型框架。实验结果表明，我们提出的多语言注意力机制可以极大地提升关系抽取的效果。

- **基于知识的实体分类：**在命名实体分类中，现有工作都没有考虑到命名实体和上下文之间的联系，也没有考虑到知识图谱中丰富的有关信息。针对这个问题，我们提出了知识注意力机制。首先，句子中的实体提及会被消岐到知识图谱中的对应或者相近实体上，然后该实体的知识图谱表示向量将被用于构建注意力。这样的注意力可以更好地找到上下文中与实体提及相关的词语，从而更好的获取文本信息。实验证明，在两个不同的数据集上，该模型相比之前的基线模型，都可以显著地提高命名实体分类的表现。
- **基于知识的开放域问答：**开放域问答现有的方法一般通过结合信息检索和阅读理解技术来回答开放域的问题。比如说，我们有一个问题 ‘‘北京的人口是多少’’，首先我们用搜索引擎从文本库中检索出和这个问题相关的段落，然后对这些检索到的段落进行阅读理解，抽取出最终的答案即 ‘‘北京的人口’’。现有的开放域问答系统只是简单地利用了信息检索和阅读理解技术，没有综合利用检索到的段落，另外也没有考虑回答问题需要涉及的背景知识。针对这些问题，我们借鉴人类回答问题的模式提出了一个基于 ‘‘粗读-精读-总结’’ 模式的开放域问答系统。进一步地，我们采用知识表示学习来帮助表示问题及其相关文档，同时与关系抽取进行多任务学习来引入实体之间的关系信息。我们提出的模型在常用的开放域问答数据集 Quasar-T，SearchQA、TriviaQA 和 HotpotQA 上都取得了比较大的提升。

第 2 章　结构化知识的表示学习

2.1　引言

知识图谱描述现实世界中实体间的关系。这些关系事实蕴藏在无（半）结构的互联网文本信息中。知识图谱则是有结构的知识库。现有知识图谱主要以符号形式进行表示和存储，如采用 RDF（Resource Description Framework) 框架，也就是三元组的形式来存储实体、关系以及实体之间的关系事实，无论是实体还是关系，都采用独一无二的符号来进行标识。采用原始符号进行表示的大规模知识图谱，在计算利用上面临以下挑战：

- **计算效率问题**：基于图结构的知识表示简洁直观，每一个实体均采用不同的节点表示，实体间的关系通过图中的边表示。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，当知识图谱规模上升达到一定程度时，就无法很好地满足实时计算的需求。
- **数据稀疏问题**：在判定两实体间关系时，需要精确表示实体与关系的语义信息。离散符号表示方案无法捕捉到实体与关系的语义信息。因此现有方法只能专门抽取特征表示实体与关系的语义信息。由于这些特征往往与特定数据与任务有关，可扩展性较差。此外，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。对这些低频的实体和关系，简单地通过符号化表示往往很难有效理解与推理。

近年来，表示学习技术[7] 异军突起，在语音、图像和自然语言处理领域获得广泛关注。为了解决传统知识图谱表示计算效率低与数据稀疏两个问题，人们提出知识表示学习（Knowledge Representation Learning）的技术方案，并被提出广泛研究与运用。知识表示学习基于分布式表示的思想，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实体（或关系）之间距离相近。由于知识表示学习的以上优点，近些年来，知识表示学习已经成为学术界的一个研究热点。人们提出了很多不同种类的知识表示学习模型，这些模型已经在很多公开评测数据集上取得了不错的效果。但是，现有的知识表示学习模型均以三元组关系事实为单位进行学习，本章节将主要探索如何充分利用知识图谱信息，合理设计学习目标，实现知识表示的有效学习:

- **考虑知识图谱复杂关系**：按照知识图谱中关系两端连接实体的对应数目，我们可以将关系划分为一对一、一对多、多对一和多对多四种类型。例如 1-多

类型关系指的是，该类型关系中的一个左侧实体会平均对应多个右侧实体。现有知识表示学习算法在处理四种类型关系时的性能差异较大。针对这个问题，我们提出了基于空间转移的 TransR 模型对不同的知识/关系的结构类型进行精细建模。

- **考虑知识图谱复杂路径**：在知识图谱中，有些多步关系路径也能够反映实体之间的关系。为了突破现有知识表示学习模型孤立学习每个三元组的局限性，我们将借鉴循环神经网络（Recursive Neural Networks）的学术思想，提出考虑关系路径的表示学习方法。我们以平移模型 TransE 作为基础进行扩展，提出 Path-based TransE（PTransE）模型对知识图谱中的复杂关系路径进行建模。
- **考虑知识图谱复杂属性**：现有知识表示学习模型将所有关系都表示为向量，这在极大程度上限制了对关系的语义的表示能力。这种局限性在属性知识的表示上尤为突出。我们面向属性知识，研究利用分类模型表示属性关系，通过学习分类器建立实体与属性之间的关系，在既有知识图谱关系表示方案的基础上，探索具有更强表示能力的表示方案。

在本章节，我会先回顾知识表示学习领域的相关现有方法，并指出我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。在此基础上，我将在具体介绍我针对上述问题提出的解决方法。最后，我将对知识表示学习的未来方向进行展望。

2.2　相关工作

在本章节的相关工作介绍之前，我首先在此简要地介绍知识表示学习中的常用符号表示。知识图谱通常以实体、关系以及事实三元组进行组织，我们使用 E 表示实体集合，R 表示关系集合，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，而 r 代表头尾实体之间的关系。例如三元组（比尔 · 盖茨, 创始人, 微软公司）就表示实体 ‘‘比尔 · 盖茨’’ 和实体 ‘‘微软公司’’ 之间存在 ‘‘创始人’’ 的关系。根据定义，我们有 $h \in E$，$t \in E$，$r \in R$，以及 $(h,r,t) \in T$。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，在此我们使用粗体的符号 $\boldsymbol{h},\boldsymbol{t},\boldsymbol{r}$ 表示头尾实体与关系对应的表示向量。

接下来，我们开始介绍知识表示学习的几个代表模型，包括：结构向量模型、语义匹配能量模型、隐变量模型、神经张量网络模型、矩阵分解模型和平移模型，等等。

2.2.1　知识表示学习经典模型

结构向量模型（Structured embedding, SE）[8] 是知识表示学习较早的尝试。在结构向量模型中，每个实体用 d 维的向量表示，所有实体被投影至同一个低维的向量空间中。同时，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与关系相关的映射矩阵 $\boldsymbol{M}_{r,1}$ 和 $\boldsymbol{M}_{r,2}$，这些与关系相关的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) 的损失函数定义如下：

$$E(h,r,t) = ||\boldsymbol{M}_{r,1}\boldsymbol{h} - \boldsymbol{M}_{r,2}\boldsymbol{t}||. \tag{2-1}$$

可以看出，结构向量模型将头实体与尾实体通过与关系相关的映射矩阵投影至一个统一的向量空间中，后在该空间中计算两投影向量的距离。这个距离反映了两个实体在关系下的语义相关度，它们的距离越小，说明这两个实体越可能存在这种关系。然而，结构向量模型有一个重要缺陷：它对头、尾实体采用两个不同的矩阵进行映射，协同性较差，往往无法精确刻画头尾实体与关系之间的语义联系。

语义匹配能量模型（Semantic matching energy, SME）[9] 与结构向量模型不同，提出更复杂的操作，用于寻找实体和关系之间的语义联系。语义匹配能量模型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，对实体与关系的内在联系进行刻画。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价函数，分别是线性形式：

$$E(h,r,t) = (\boldsymbol{M}_1\boldsymbol{h} + \boldsymbol{M}_2\boldsymbol{r} + \boldsymbol{b}_1)^\top(\boldsymbol{M}_3\boldsymbol{t} + \boldsymbol{M}_4\boldsymbol{r} + \boldsymbol{b}_2), \tag{2-2}$$

和双线性形式：

$$E(h,r,t) = ((\boldsymbol{M}_1\boldsymbol{h} \otimes \boldsymbol{M}_2\boldsymbol{r}) + \boldsymbol{b}_1)^\top((\boldsymbol{M}_3\boldsymbol{t} \otimes \boldsymbol{M}_4\boldsymbol{r}) + \boldsymbol{b}_2), \tag{2-3}$$

其中 $\otimes$ 表示按位相乘，$\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3, \boldsymbol{M}_4$ 表示待学习的映射矩阵，$\boldsymbol{b}_1, \boldsymbol{b}_2$ 表示偏置向量。此外，人们还提出了基于三阶张量对语义匹配能量模型中的双线性形式进行改进的模型[10]。

隐变量模型（Latent factor model, LFM）[11-12] 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。隐变量模型将实体表示成为低维向量，将关系表示为双线性变换矩阵 $\boldsymbol{M}_r$，在知识表示学习效果与计算复杂度上都有显著改善。模型的评分函数如下：

$$E(h,r,t) = \boldsymbol{h}^\top\boldsymbol{M}_r\boldsymbol{t}. \tag{2-4}$$

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。此外，人们还提出了 DISTMUL 方法[13] 在隐变量模型的基础上进行改进，将关系矩阵 $\boldsymbol{M}_r$ 限定为对角阵。实验表明，这种简化不仅极大降低了模型复杂度，还获得了更好的模型效果。

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[14-15] 和 HOLE 模型[16] 为代表。RESCAL 模型使用一个三阶张量 $X \in \mathbb{R}^{d\times d\times k}$ 表示三元组，其中 d 是实体的数量，k 是关系的数量。如果 $X_{htr} = 1$，则表示三元组 (h, r, t) 存在。设 $X = \{X_1, \cdots, X_k\}$，则有矩阵分解：

$$\boldsymbol{X}_i \approx \boldsymbol{A}\boldsymbol{R}_i\boldsymbol{A}^\top. \tag{2-5}$$

其中 $\boldsymbol{A}$ 表示实体向量形成的矩阵，而 $\boldsymbol{R}_i$ 表示第 i 个关系矩阵。可以看到 RESACL 模型的基本思想与前述隐变量模型类似。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；而隐变量模型只会优化知识图谱中存在的三元组。在此基础上，HOLE 模型进一步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。

神经张量模型（Neural tensor network, NTN）[17] 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，在不同的维度下将头、尾实体向量联系起来。神经张量模型对一个三元组 (h, r, t) 的评分函数如下：

$$E(h, r, t) = \boldsymbol{u}_r^\top \tanh(\boldsymbol{h}^\top\boldsymbol{M}_r\boldsymbol{t} + \boldsymbol{M}_{r,1}\boldsymbol{h} + \boldsymbol{M}_{r,2}\boldsymbol{t} + \boldsymbol{b}_r). \tag{2-6}$$

其中 $\boldsymbol{M}_r$ 是三阶张量，$\boldsymbol{M}_{r,1}$ 和 $\boldsymbol{M}_{r,2}$ 是投影矩阵。神经张量模型更精确地刻画了实体与关系之间的联系，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。值得注意的是，与以往模型不同，神经张量模型还提出了采用了实体中的单词向量对实体进行表示，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。由于神经张量模型引入了复杂的张量操作，虽然其提高了精确地刻画实体和关系的复杂语义联系的能力，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，导致了该模型在大规模稀疏知识图谱上的效果较差。

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，近年来知识表示学习的研究工作，包括本章中详细讨论的工作，都是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）[18] 是 Bordes 等研究人员在 2013 年提出的知识表示学习算法。TransE 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，TransE 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上的效果也得到了显著的提升，在知识表示任务中被广为运用。

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[5,19-20]，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：

$$\boldsymbol{v}(king) - \boldsymbol{v}(man) \simeq \boldsymbol{v}(queen) - \boldsymbol{v}(woman). \tag{2-7}$$

其中，$\boldsymbol{v}(x)$ 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。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。

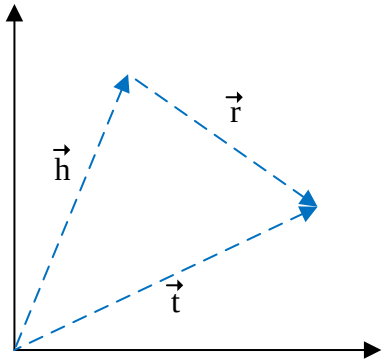

图 2.1　TransE 模型图示

受到词空间语义平移现象的启发，TransE 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)，TransE 模型将关系向量 $\boldsymbol{r}$ 看做是从头实体 $\boldsymbol{h}$ 到尾实体 $\boldsymbol{t}$ 的平移向量，如图2.1所示。基于以上平移假设，TransE 模型希望一个三元组内的实体与关系向量之间存在 $\boldsymbol{h} + \boldsymbol{r} \simeq \boldsymbol{t}$ 的关系。形式化地，模型对三元组 (h, r, t) 定义了如下的能量函数：

$$E(h, r, t) = ||\boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t}||_{L1/L2}. \tag{2-8}$$

即 $\boldsymbol{h} + \boldsymbol{r}$ 向量和 $\boldsymbol{t}$ 的 $L1$ 或 $L2$ 距离。

在实际训练中，为了增强模型对于关系事实三元组的区分能力，TransE 模型使用最大间隔方法，定义了如下评分函数进行优化：

$$L = \sum_{(h,r,t)\in T} \sum_{(h',r,t')\in T'} \max(\gamma + E(h,r,t) - E(h',r,t'), 0), \tag{2-9}$$

其中，T 和 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边界距离值。TransE 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。其中，负例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 模型将每个关系事实三元组的头实体、关系和尾实体其中之一随机替换成其他实体或关系：

$$T^- = \{(h',r,t)\} \cup \{(h,r',t)\} \cup \{(h,r,t')|(h,r,t) \in T\}. \tag{2-10}$$

与以往模型相比，TransE 模型模型参数较少，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 WN18 和 FB15K 等数据集上进行链接预测等评测任务，实验表明 TransE 模型的效果相比于以往模型有显著提升。特别是在大规模稀疏知识图谱上，TransE 模型的效果尤其惊人，远远超过了已有的经典模型。

尽管 TransE 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，在世界知识图谱上，TransE 模型仍存在着以下改进与优化的空间：

第一，过于理想化的平移假设，在对知识图谱中的复杂特性（关系和属性）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[18]。以职业关系为例，我们有（莎士比亚，职业，作家）和（鲁迅，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，而在其它属性，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[21]、TransD 模型[22] 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[23] 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[24] 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[25] 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。我们提出的 TransR 模型[26] 通过与关系相关的实体投影矩阵，在考虑不同关系的时候将实体映射到不同关系专属的语义空间中，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，我们的 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属性的处理。

第二，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，首都，中国）两条知识，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[27] 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，而我们提出的 PTransE 模型[28] 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。

在本章的剩余部分，我们将详细讨论我们的改进工作。

2.3　知识图谱复杂关系建模

TransE 模型由于其模型通过简单的平移操作对实体和关系之间的相互联系进行了刻画，使得其在大规模知识图谱上有很好的效果。但是也由于过于简单，导致 TransE 模型在处理知识图谱的复杂关系时捉襟见肘。这里的复杂关系定义如下。按照知识图谱中关系两端连接实体的数目，可以将关系划分为一对一、一对多、多对一和多对多四种类型。例如多对一类型关系指的是，该类型关系中的一个尾实体平均会对应多个头实体。我们将一对多、多对一和多对多的关系称为复杂关系。

实验结果表示，平移算法在处理四种类型关系时的性能差异较大，尤其在处理复杂关系时模型效果显著降低，这与 TransE 模型的假设有着密切的联系。TransE 模型认为关系 r 可以表示为头实体和尾实体之间的一个平移向量。对于每一个事实关系三元组 (h,r,t)，该模型希望 $\mathbf{h}+\mathbf{r}\approx\mathbf{t}$。根据 TransE 模型的优化目标，在面向一对多、多对一和多对多三种类型的复杂关系时，我们可以推出以下结论：如果关系是一对多关系，我们将会得到 $\mathbf{t}_0\approx\mathbf{t}_1\approx\cdots\approx\mathbf{t}_m$。同样的，这样的问题在关系是多对一、多对多关系时也同样会出现。

例如，对于知识图谱中的事实三元组（美国，总统，奥巴马）和事实三元组（美国，总统，乔治 · 布什）。可以看到，这里 ‘‘总统’’ 关系就是一个经典的一对多关系。如果我们采用 TransE 模型对这两个事实三元组进行建模，如图2.2所示，我们会得到 ‘‘奥巴马’’ 和 ‘‘乔治 · 布什’’ 拥有同样的向量表示。

然而，在知识图谱中，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 ‘‘奥巴马’’ 既是美国总统，也是 ‘‘米歇尔’’ 的丈夫。由信息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，我们提出了基于映射矩阵进行空间投影的 TransR

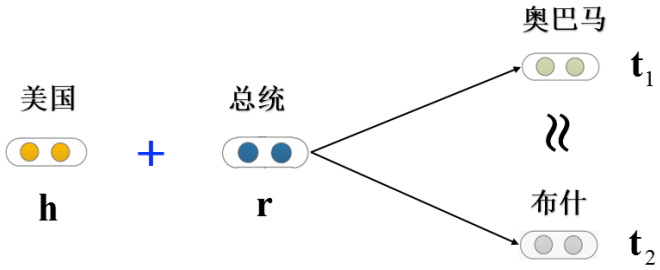

图 2.2　复杂关系 ‘‘总统’’ 的一个样例

模型。

2.3.1　算法模型

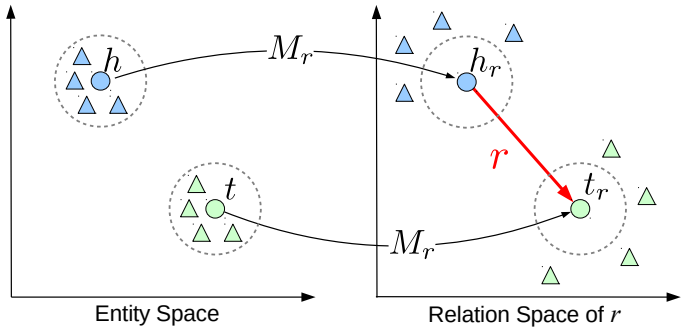

图 2.3　TransR 的简易示意图

TransR 模型观察到：一个实体可以在不同的关系中将会展示其在不同身份下的特性。也就是说，一个实体是多种身份的综合体，不同关系关注实体的不同身份。因此，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平移关系来建模头尾实体在该关系空间中的关联性。因此，虽然 ‘‘奥巴马’’ 和 ‘‘乔治 · 布什’’ 这样的实体在实体空间中可能彼此相距很远，但它们在某些特定关系（如总统）的空间中是相似且彼此接近的；同样地，对于 ‘‘北京’’ 和 ‘‘上海’’ 这样的实体在实体空间中可能彼此十分接近，但它们在某些特定关系（如首都）的空间中是极其不相似且相距很远的。

如图2.3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 $\boldsymbol{M}_r$ 定义从实体空间到每个关系空间的映射。因此，对于每个关系事实三元组 (h, r, t)，我们首先通过关系 r 的映射矩阵将实体向量向关系 r 对应的关系空间进行投影。在投影之后，原来在实体空间中与头、尾

实体（用彩色圆圈表示）相似的实体（用彩色三角形表示），在关系空间中被区分开了。

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 $\boldsymbol{M}_r$ 将实体从实体空间映射到关系 r 所在的关系空间中，得到 $\boldsymbol{h}_r$ 和 $\boldsymbol{t}_r$：

$$\boldsymbol{h}_r = \boldsymbol{h}\boldsymbol{M}_r, \quad \boldsymbol{t}_r = \boldsymbol{t}\boldsymbol{M}_r \tag{2-11}$$

在关系 r 所在的空间中，$\boldsymbol{h}_r$ 和 $\boldsymbol{t}_r$ 满足的关系转移约束与 TransE、TransH 相同，即我们使用类似的评分函数：

$$E(h, r, t) = \|\boldsymbol{h}_r + \boldsymbol{r} - \boldsymbol{t}_r\| \tag{2-12}$$

此外，我们注意到，相同的关系在不同实体对中的语义通常也具有一定的多样性。有些通用关系往往出现在多个场景下，例如 ‘‘包含’’ 关系既出现在知识 (国家，包含，城市) 中，也可能出现在知识 (城市，包含，大学) 中，由于这两类知识中对应的实体类型不同（前者是国家-城市，后者是城市-大学），如果始终用同一个向量表示 ‘‘包含’’ 关系，会降低其关系表示的区分度。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。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，所有训练数据中蕴含这个关系的实体对 (h, t) 将会根据 $(\boldsymbol{h} - \boldsymbol{t})$ 被聚类到若干组中，其中 $\boldsymbol{h}, \boldsymbol{t}$ 为 TransE 得到的实体嵌入。我们假设同一组内的实体对表现的关系 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 $\boldsymbol{r}_c$，评分函数也相应地被修改为：

$$E(h, r, t) = \|\boldsymbol{h}_{r,c} + \boldsymbol{r}_c - \boldsymbol{t}_{r,c}\| + \alpha\|\boldsymbol{r}_c - \boldsymbol{r}\|, \tag{2-13}$$

其中 $\|\boldsymbol{r}_c - \boldsymbol{r}\|_2^2$ 用来约束聚类成的关系向量 $\boldsymbol{r}_c$ 与原始关系向量 $\boldsymbol{r}$ 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。

2.3.1.1　训练方法和实现细节

和 TransE 模型相同，我们同样定义了一个基于边界距离的损失函数作为我们的训练目标，具体函数如下：

$$L = \sum_{(h,r,t)\in T} \sum_{(h',r,t')\in T^-} \max\left(0, E(h,r,t) + \gamma - E(h',r,t')\right) \tag{2-14}$$

这里 $\max(x, y)$ 用来在 x 和 y 中选取一个最大值，γ 是边界距离值，T 是正例的三元组集合而 T^- 是负例的三元组集合。知识图谱中只存在正例三元组，所以我们在正例三元组集合 $(h, r, t) \in T$ 中替换实体来构建负例三元组集合 $(h', r, t') \in T'$。当替换三元组的实体时，我们可以简单采用平均采样（uniform）的方法，也就是以均匀的概率替换头尾实体，我们把这个方法命名为 “unif’’。此外，Wang 等人[29] 提出我们应当采用不同概率来替换头尾实体。原因在于，对于单对多、多对单、多对多这样的关系，采样 ‘‘多’’ 测的实体作为负例会比采样 ‘‘单’’ 测的实体更有可能采样到本身正确的事实三元组。因此，我们会对其中 ‘‘单’’ 测的实体给予更大的采样概率来生成负例，我们将这个 Wang 等人提出的新的方法命名为 “bern’’。

在训练过程中，TransR 和 CTransR 的模型均采用随机梯度下降（stochastic gradient descent，SGD）来进行优化。为了使得模型更容易收敛，我们用 TransE 训练得到的实体与关系向量来初始化 TransR 中的实体与关系向量，并且用单位矩阵来初始化各个关系映射矩阵。

2.3.2　实验分析

2.3.2.1　数据集与实验设置

我们在两个经典知识图谱数据集上 WordNet[30] 和 Freebase[2] 对我们提出的 TransR 和 CTransR 方法进行了评测。

WordNet 是一个词语级别的语义知识图谱。在 WordNet 中，每个实体都是一个包含若干个词语的**同义词集合**，并对应一个单独的**词义**。而 WordNet 中的关系表示各个同义词集合之间的词汇级别语义关联，比如 “上位词’’、“下位词’’、“局部词’’、“整体词’’，等等。在本文实验中，我们采用了两个来源于 WordNet 的数据集 WN18 和 WN11。其中 WN18 包含了 18 种关系类型，曾被用在 Bordes 等人的论文中的链接预测实验[10]；而 WN11 包含了 11 中关系类型，被用在 Socher 等人的论文中三元组分类实验[17]。

与 WordNet 不同，Freebase 是一个提供这个世界上的通用知识的世界知识图谱。例如，三元组（乔布斯, 创始人, 苹果公司）表述了在人物实体 “乔布斯’’ 与组

织实体 “苹果公司’’ 之间存在 “建立’’ 关系，即乔布斯创建了苹果公司。在本论文中，我们同样采用了两个来源于 Freebase 的数据集合 FB15K 和 FB13。其中 FB15K 也曾被用在 Bordes 等人的论文中的链接预测实验中[10]，而 FB13 也曾在 Socher 等人的论文中的三元组分类实验中被使用[17]。表格2.1中列出了上述数据集的详细统计数据。

表 2.1　数据集 WN18、FB15K、WN11、FB40K 的统计数据

数据集	# 关系	# 实体	# 训练集	# 验证集	# 测试集
WN18	18	40,943	141,442	5,000	5,000
FB15K	1,345	14,951	483,142	50,000	59,071
WN11	11	38,696	112,581	2,609	10,544
FB13	13	75,043	316,232	5,908	23,733
FB40K	1336	39,528	370,648	6,7946	96,678

2.3.2.2　链接预测

链接预测是知识图谱补全中的一个经典任务，可以通过预测三元组 (h,r,t) 中缺失的头实体 h 或尾实体 t 来对知识表示学习模型的效果进行评测。链接预测任务已经用于很多现有知识表示学习模型的效果评测[8-9,18]。在本任务中，对于每一个缺失的实体，知识表示学习模型将所有的知识图谱中的实体作为候选实体，计算每一个候选实体作为答案的评分并进行排名，而不是仅仅给出一个最优的预测结果。与之前的工作[8,18] 一样，我们在 WN18 和 FB15K 上进行了实验。

在评测阶段，对于每个待测试关系事实三元组 (h,r,t)，我们用知识图谱中的所有实体作为候选实体来替换头实体或尾实体，然后计算这些替换后的三元组的评分并进行排序。参考 Bordes 等人论文中的实验方法[18]，我们采用了两种评测指标：(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，即前十命中率（Hits@10）。显而易见，一个效果好的知识表示学习模型应当在链接预测任务中获得较低的平均排名和较高的前十命中率。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关系事实三元组，这种三元组本质上是正例而不应当被视作负例。然而，上述的评测方法会受到这些三元组的影响，导致低估了评测的知识表示学习模型的效果。因此，我们可以在对候选实体进行排名之前先将这些三元组过滤掉，然后再用上述的方法评测。我们将原本的评测方法称为 “Raw’’ 方法，而将之后过滤的评测方法称为 “Filter’’ 方法。

在本实验中，由于使用了相同的评测数据，我们将我们模型的结果和已有知

表 2.2　链接预测的评测结果

数据集	WN18				FB15K			
评测指标	Mean Rank		Hits@10 (%)		Mean Rank		Hits@10 (%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
Unstructured[9]	315	304	35.3	38.2	1,074	979	4.5	6.3
RESCAL[14]	1,180	1,163	37.2	52.8	828	683	28.4	44.1
SE[8]	1,011	985	68.5	80.5	273	162	28.8	39.8
SME (linear)[9]	545	533	65.1	74.1	274	154	30.7	40.8
SME (bilinear)[9]	526	509	54.7	61.3	284	158	31.3	41.3
LFM[12]	469	456	71.4	81.6	283	164	26.0	33.1
TransE[18]	263	251	75.4	89.2	243	125	34.9	47.1
TransH (unif)[29]	318	303	75.4	86.7	211	84	42.5	58.5
TransH (bern)[29]	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	44.0	66.3
CTransR (bern)	**231**	**218**	79.4	**92.3**	199	**75**	**48.4**	**70.2**

识表示学习模型论文中报告的结果进行了比较。对于 TransR 和 CTransR 的实验超参数，我们通过验证集上的平均排名评分来选择最优的超参数。对于 SGD 的学习率 λ，我们搜索了 $\{0.1, 0.01, 0.001\}$；对于边界距离值 γ，我们搜索了 $\{1, 2, 4\}$；对于实体和关系的维度 k 和 d，我们搜索了 $\{20, 50, 100\}$；对于训练时的 batch 大小 B，我们搜索了 $\{20, 120, 480, 1440, 4800\}$。对于 CTransR 的约束参数 α，我们搜索了 $\{0.1, 0.01, 0.001\}$。对于 WN18，我们采用了 L_1 距离，最优的参数为 $\lambda = 0.001$，$\gamma = 4$，$k = 50$，$d = 50$，$B = 1440$，$\alpha = 0.001$。对于 FB15K，我们同样采用了 L_1 距离，最优的参数为 $\lambda = 0.001$，$\gamma = 1$，$k = 50$，$d = 50$，$B = 4800$，$\alpha = 0.01$。对于这两个数据集，我们均在所有训练数据上训练了 500 轮。

表格2.2列出了 TransR 和 CTransR 模型和现有知识表示学习模型在 WN18 和 FB15K 数据集上的对比结果的评测结果。从表中我们可以观察到：

(1) 在 WN18 和 FB15K 两个数据集上，TransR 和 CTransR 模型均比包括 TransE 和 TransH 在内现有知识表示学习模型展现出了更好的效果。这表明 TransR 通过将实体表示投影到特定关系的子空间中，实现了对知识图谱中复杂关系的更好的建模，在效率和复杂程度上找到了一个更好的权衡。

(2) CTransR 模型比 TransR 模型效果更好，这表明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可以帮助建立更好的知识表示学习模

表 2.3　将关系分类后在 FB15K 上的评测结果 (%)。其中 1-to-1 表示一对一关系，1-to-N 表示一对多关系，N-to-1 表示多对一关系，N-to-N 表示多对多关系

任务	头实体预测 (Hits@10)				尾实体预测 (Hits@10)			
关系分类	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[9]	34.5	2.5	6.1	6.6	34.3	4.2	1.9	6.6
SE[8]	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3
SME (linear)[9]	35.1	53.7	19.0	40.3	32.7	14.9	61.6	43.3
SME (bilinear)[9]	30.9	69.6	19.9	38.6	28.2	13.1	76.0	41.8
TransE[18]	43.7	65.7	18.2	47.2	43.7	19.7	66.7	50.0
TransH (unif)[29]	66.7	81.7	30.2	57.4	63.7	30.1	83.2	60.8
TransH (bern)[29]	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**

型。本文中我们实现的 CTransR 模型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) 通过采用 “bern’’ 采样，TransH 和 TransR 模型的效果在比 “unif’’ 采样都有提升, 尤其是在拥有更多复杂关系的 FB15K 上。这说明 “bern’’ 采样方式对于复杂关系的学习有帮助。

如上文所讨论，我们将复杂关系分类四类：一对一关系，一对多关系，多对一关系，多对多关系。关系的分类方法我们遵循 Bordes 等人使用的规则[18]：对于一个关系，如果在包含这个关系的事实三元组中，一个头实体平均对应超过 1.5 个尾实体，那么这个关系就是一个一对多或者多对多关系，以此类推。经过数据统计，我们得到在 FB15K 数据集上：一对一关系占 26.2%，一对多关系占 22.7%，多对一关系占 28.3%，多对多关系占 22.8%。

在表格2.3中分别报告了知识表示学习模型在不同关系上的实验结果。在 FB15K 上，我们可以发现 TransR 和 CTransR 模型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 和 CTransR 模型为实体与关系的复杂相关性提供了更精确的表示，正如图2.3所示的那样;

(2) 在预测单对多、多对单关系时，TransR 和 CTransR 模型通过关系特定映射来在不同关系中区分相关实体的能力得到了充分体现，其模型效果也取得了很大提升。

表 2.4　Freebase 中〈头实体, 尾实体〉对于 ‘‘包含’’ 关系的聚类样例

	〈头实体, 尾实体〉
1	⟨Africa, Congo⟩, ⟨Asia, Nepal⟩, ⟨Americas, Aruba⟩, ⟨Oceania, Federated States of Micronesia⟩
2	⟨United States of America, Kankakee⟩, ⟨England, Bury St Edmunds⟩, ⟨England, Darlington⟩, ⟨Italy, Perugia⟩
3	⟨Georgia, Chatham County⟩, ⟨Idaho, Boise⟩, ⟨Iowa, Polk County⟩, ⟨Missouri, Jackson County⟩, ⟨Nebraska, Cass County⟩
4	⟨Sweden, Lund University⟩, ⟨England, King’s College at Cambridge⟩, ⟨Fresno, California State University at Fresno⟩, ⟨Italy, Milan Conservatory⟩

我们在表格2.4给出 FB15K 数据集中 ‘‘包含’’ 关系的一些聚类示例。我们可以发现：聚类 #1 是关于大陆包含国家, 聚类 #2 是国家包含城市，聚类 #3 是区域包含乡村, 聚类 #4 是国家包含大学。很明显，通过聚类，我们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表示学习模型的性能，这充分证实了我们在设计 TransR 与 CTransR 时的设想。

2.3.2.3　三元组分类

三元组分类任务旨在判断一个给定三元组 (h, r, t) 正确与否。三元组分类任务是一个二分类任务，已经被用于很多现有知识表示学习模型的效果评测[17,29]。在这个任务上，我们采用了数据集 WN11、FB13 与 FB15K 来进行测试，并且与 Wang 等人保持一致的实验设置[29]。

我们需要负例三元组来进行二分类测试。在 NTN[17] 中, 数据集 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[17] 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。

对于 WN11 和 FB13，我们比较了我们的模型以及 Wang 等人汇报的结果[29]。由于 FB15K 是根据 Socher 等人的策略自行生成的[17]，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[17]，在 FB15K 数据集上进行了评估与比较。

下面介绍 TransR 的模型参数选择。对于 SGD 的学习率 λ，我们搜索了

表 2.5　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**

$\{0.1, 0.01, 0.001, 0.0001\}$；对于边界距离值 γ，我们搜索了 $\{1, 2, 4\}$；对于实体和关系的维度 k 和 d，我们搜索了 $\{20, 50, 100\}$；对于训练时的 batch 大小 B，我们搜索了 $\{20, 120, 480, 1440, 4800\}$。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，我们均在所有训练数据上训练 500 轮。

三元组分类的结果如表格2.5所示。从表格2.5，我们观察到：

(1) 在 WN11 上，TransR 模型显著优于包括 TransE 和 TransH 模型在内的所有知识表示学习模型。

(2) TransE、TransH 和 TransR 在 FB13 数据集上的效果都差于模型结构最复杂的 NTN 模型。相比之下，在较大的数据集 FB15K 上，TransE、TransH 和 TransR 的性能要好于 NTN 模型。其原因可能与数据集的特征有关：FB15K 中有 1,345 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却相近。正如 Wang 等人讨论到的现象[29]，FB13 中的知识图谱远比 FB15K 以及 WN11 更稠密，也就是对于同一个关系拥有更有的实体和三元组。某种程度上，模型结构最复杂的 NTN 模型可以从 FB13 的稠密图中使用张量变换来学习复杂关系内部的特性。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 模型在所有数据

集上的性能。

2.3.2.4　文本中的关系抽取

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，人们提出了很多远程监督关系抽取模型[31-34]。这些模型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是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，我们可以很直接想到同时利用纯文本和知识图谱来推断新的关系事实。在 Weston 等人的工作中[35]，他们将 TransE 模型和基于文本的关系抽取 Sm2r 模型相结合，对候选关系事实进行评分排序，在文本关系抽取任务中取得了十分明显的效果提升。在 TransH[29] 的工作中也发现了在文本关系抽取中考虑 TransH 模型的信息可以有类似的效果改进。在本节中，我们将研究 TransR 模型对于文本关系抽取模型的帮助效果。

在这个实验中，我们使用了 NYT+FB 数据集。这个数据集合也被用在 Weston 等人的工作中[35]。在这个数据集中，我们通过 Stanford NER 工具来标注纽约时报语料库（New York Times Corpus）中的实体，并将发现的命名实体链接到 Freebase 中的实体。

在本实验中，我们实现了 Weston 等人[35] 提出的基于文本的关系抽取 m2r 模型。对于知识图谱部分，Weston 等人[35] 使用了近 4 百万个实体的 Freebase 子集，同时有 23000 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四百万个实体的数据中学习到实体和关系的表示。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 1336 个关系类型。为了测试公平性，我们从 FB40K 中删除了所有出现在测试集中的所有三元组。与之前工作的结果[29,35] 相比，我们发现使用 FB40K 进行学习并不会显着降低 TransE 和 TransH 的有效性。因此，我们可以使用 FB40K 来检验 TransR 的有效性。

采用与 Weston 等人[35] 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表示学习模型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。在实验中，我们令关系和实体的向量维度 $k, d = 50$，学习率 $\lambda = 0.001$，边界距离值 $\gamma = 1.0$，$B = 960$，并且在评分函数中采用了 L_1 距离。精度——召回率曲线如图2.4所示。从图中我们观察到，当召回范围为 $[0, 0.05]$ 时，TransR 模型显著优于 TransE，与 TransH 相当；当召回范围为

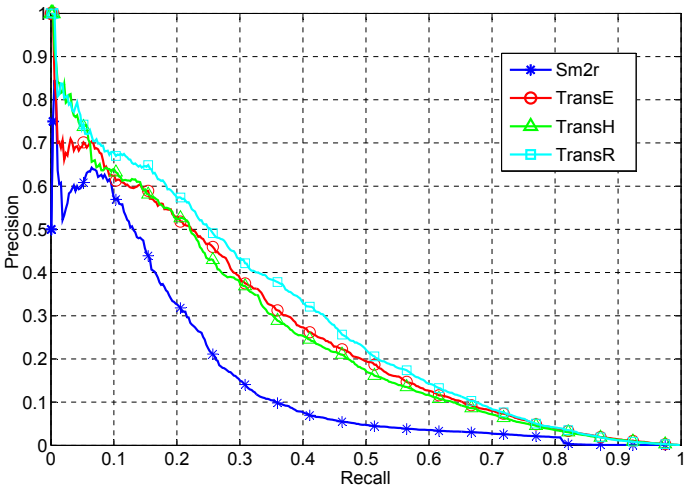

图 2.4　TransE、TransH 与 TransR 在关系抽取任务上的精度——召回率曲线

[0.05, 1] 时，TransR 模型的表现超越了所有的模型，包括 TransE 和 TransH 模型。

最近表示学习的想法也被广泛地用于表示单词和文本的工作中[5,19-20,36]，这些都可以在未来用于基于文本的关系抽取任务中。

2.4　知识图谱复杂路径建模

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也蕴含了丰富的语义信息。例如，关系路径 $h \xrightarrow{BornInCity} e_1 \xrightarrow{CityInState} e_2 \xrightarrow{StateInCountry} t$ 隐含了 h 和 t 之间的 *Nationality* 关系，亦即 (h, *Nationality* ,t)。

目前，关系路径信息已经在社会计算和推荐系统领域中被广泛使用。这些工作大多将关系和关系路径视为一个个独立的离散符号，并使用基于图的算法来处理它们，如随机游走算法[37] 等。随机游走算法已被用于专家发现[38] 和信息检索[39] 等任务中并取得了很好的效果。事实上，人们已经尝试将关系路径用于大规模知识图谱中的推理运算中并取得了一定的效果，比如路径排序算法[38]。此外，路径排序算法也被用于利用知识图谱结构信息来进行关系抽取[40-41]。之后，Neelakantan 等人[42] 进一步通过循环神经网络来学习关系路径的表示向量，使得模型能够具有更强的表达能力和鲁棒性。然而，这些方法专注于建模关系路径涉及的关系，而不考虑路径上的实体信息。因此，有效整合实体和关系路径两种信息才能更好地为知识图谱补全和关系抽取任务学习更加合理的实体与关系的表示。

为了突破现有 TransE 等模型孤立学习每个三元组的局限性，我们提出考虑关系路径的知识表示学习方法，以 TransE 作为基础进行扩展，提出 Path-based TransE（PTransE）模型，将知识图谱中的关系路径融入到知识表示学习模型中。PTransE

模型面临的挑战在于：

（1）**关系路径置信度**：并不是所有的实体间的关系路径都是可靠且对于知识表示学习有意义的。例如，对于关系路径 $h \xrightarrow{\texttt{Friend}} e_1 \xrightarrow{\texttt{Profession}} t$，事实上这条路径并没有反映头实体 h 和尾实体 t 之间的语义关系。一个人的职业可能跟他的朋友的职业完全无关。因此，我们的模型可能不能直接考虑所有的路径。在实验中，我们发现对于那些链接特别多尾实体的路径可能对于一个特定实体对的关系预测没有很大帮助。为此，我们在 PTransE 模型中提出 Path-Constraint Resource Allocation 图算法度量关系路径的可靠性，然后选择可靠程度高的路径用于 PTransE 模型的表示学习中。

（2）**关系路径表示**：为了在知识表示学习模型中考虑关系路径，我们需要将关系路径同样表示为低维稠密向量。也就是说 PTransE 模型需要建立关系路径的向量表示，参与从头实体到尾实体的关系推理过程。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。具体地，给定关系路径 $p = (r_1, \ldots, r_l)$，其语义由 r_1 到 r_l 所有关系确定，我们定义二元组合操作 $(\circ)$，然后通过递归地合并关系路径上的多种关系来获得关系路径向量，也就是 $\mathbf{p} = \mathbf{r}_1 \circ \ldots \circ \mathbf{r}_l$。PTransE 尝试了三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。

通过关系路径选择算法和关系路径表示，PTransE 模型将路径表示为头实体和尾实体之间的平移向量，可以有效地学习实体、关系以及关系路径的表示。

2.4.1　算法模型

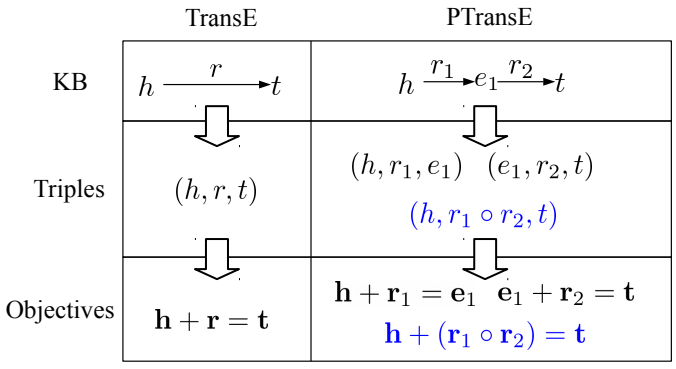

图 2.5　TransE 与 PTransE 的简易示意图

PTransE 与 TransE 的主要思想对比如图2.5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能量函数考虑了实体间的多步关系路径信息。具体地，假定头实体 h 和尾实体 t 之间的多条关系路径集合为 $P(h,t)=\{p_1,\ldots,p_N\}$，其中 $p=(r_1,\ldots,r_l)$ 表示关系路径 $h \xrightarrow{r_1} \ldots \xrightarrow{r_l} t$。对于每个关系三元组 (h,r,t)，其能量函数定义为：

$$G(h,r,t)=E(h,r,t)+E(h,P,t), \tag{2-15}$$

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：

$$E(h,r,t)=|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}|_{L_1/L_2}, \tag{2-16}$$

而 $E(h,P,t)$ 则是 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置信，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 $E(h,P,t)$ 为各关系路径下能量函数根据其可靠性加权平均的结果：

$$E(h,P,t)=\frac{1}{Z}\sum_{p\in P(h,t)} R(p|h,t)E(h,p,t), \tag{2-17}$$

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

- 如何评估一条关系路径 p 的可靠程度。
- 如何得到一条关系路径 p 的向量表示 $\boldsymbol{p}$。

2.4.1.1　关系路径置信度

针对第一个挑战，我们提出了一种基于路径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。网络资源分配算法最初是针对个性化推荐系统[43], 提出的，并且已成功用于信息检索任务中用以测量两个对象之间的相关性[44]。在这里，我们将其扩展为基于路径约束的资源分配算法以测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，我们可以将流动路径写为 $S_0 \xrightarrow{r_1} S_1 \xrightarrow{r_2} \ldots \xrightarrow{r_l} S_l$，其中 $S_0=\{h\}$ 且 $t\in S_l$。对于任意实体

$m \in S_i$，我们将它在关系 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：

$$R_p(m) = \sum_{n \in S_{i-1}(\cdot,m)} \frac{1}{|S_i(n,\cdot)|} R_p(n) \tag{2-18}$$

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。

对于每一条关系路径 p，我们令其头实体 h 最开始的资源为 $R_p(h) = 1$。我们从头实体 h 通过路径 p 递归地执行资源分配算法，尾实体 t 最终获得了 $R_p(t)$ 的资源。尾实体获得的资源大小代表了其可以从头实体获取到的信息大小。因此，我们采用尾实体资源 $R_p(t)$ 来衡量路径 p 对于实体对 (h,t) 的置信度，也就是，$R(p|h,t) = R_p(t)$。

2.4.1.2　关系路径表示

除了关系路径置信度之外，我们还需要为公式 (2-15) 中的路径三元组 (h,p,t) 定义能量函数 $E(h,p,t)$。和公式 (2-16) 中 TransE 的能量函数类似，我们首先需要在语义空间中将关系路径 p 表示为稠密向量。

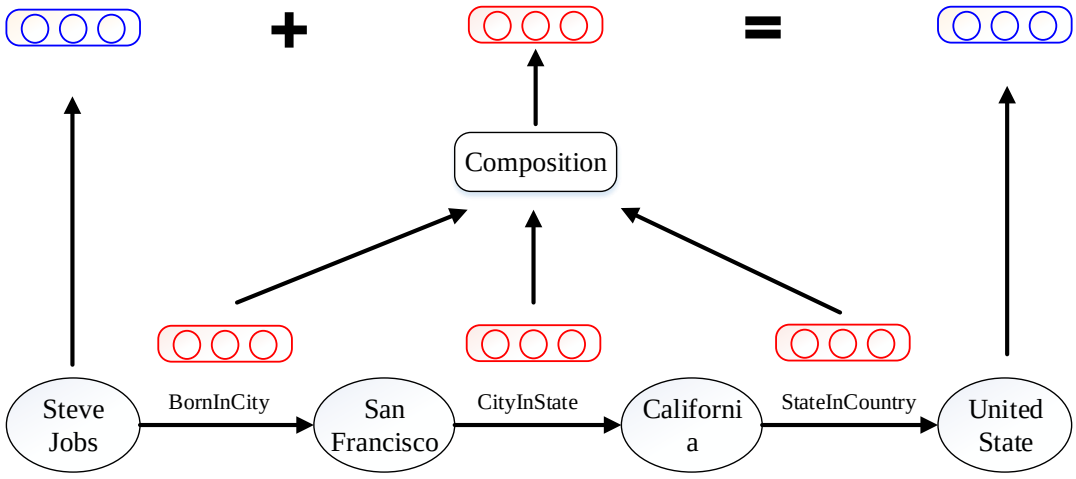

图 2.6　路径向量通过路径上所有关系的向量进行语义组合得到

关系路径的语义很大程度上依赖于它所涉及的关系。因此，通过路径上涉及的所有关系的表示的语义组合来构建路径表示是合理的。如图2.6所示，路径表示 $\mathbf{p}$ 由 `BorninCity`，`CityInState` 和 `StateInCountry` 的表示组合而成。

具体来说，对于一个关系路径 $p = (r_1, \ldots, r_l)$，我们定义一个语义组合操作 $\circ$ 并获得关系路径表示 $\mathbf{p} = \mathbf{r}_1 \circ \ldots \circ \mathbf{r}_l$。我们尝试了三种不同的方法得到关系路径的表示：

- **加和**：加和语义组合操作通过对所有关系的向量进行求和来获得关系路径的表示向量，将其形式化为：

$$\boldsymbol{p} = \boldsymbol{r}_1 + \cdots + \boldsymbol{r}_l \tag{2-19}$$

- **乘积**：乘积语义组合操作通过对所有关系的向量进行累积乘积来获得路径的表示向量，其被形式化为：

$$\boldsymbol{p} = \boldsymbol{r}_1 \circ \cdots \circ \boldsymbol{r}_l \tag{2-20}$$

  加和和乘积语义组合操作都很简单，并且在短语和句子的语义组合中进行了广泛的研究[45]。
- **循环神经网络**：循环神经网络是最近提出的基于神经网络的语义组合模型[46]，使用矩阵 $\mathbf{W}$ 实现语义组合操作：

$$c_i = f(W[c_{i-1}; r_i]), \tag{2-21}$$

  其中，f 是一个非线性函数（一般可以是 tanh 或者 sigmoid 函数等），$[a; b]$ 表示两个向量的拼接。通过令 $c_1 = r_1$ 然后在关系路径上递归地执行循环神经网络操作，我们最后得到 $\boldsymbol{p} = \boldsymbol{c}_l$，其中 $\boldsymbol{c}_l$ 是 RNN 在序列 $\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_l$ 上产生的最后一个状态。事实上，RNN 已经被尝试用于知识图谱中关系路径的表示[42]。

对于一个多步关系路径三元组 (h, p, t)，我们可以简单参考 TransE 的来定义其能量函数：$E(h, p, t) = ||\mathbf{h} + \mathbf{p} - \mathbf{t}||$。但是，由于我们对于事实三元组 (h, r, t) 同样优化 $|\mathbf{h} + \mathbf{r} - \mathbf{t}|_{L_1/L_2}$，因此我们有 $\mathbf{r} \approx \mathbf{t} - \mathbf{h}$。因此，我们可以将关系路径 $\boldsymbol{p}$ 看做根据多步关系路径信息得到的对实体间关系 r 的近似，故我们定义 PTransE 的能量函数 $E(h, p, t)$ 为 $\boldsymbol{p}$ 与 $\boldsymbol{r}$ 的接近程度：

$$E(h, p, t) = |\boldsymbol{p} - \boldsymbol{r}|_{L_1/L_2}, \tag{2-22}$$

该能量函数在关系路径 p 和关系 r 一致时拥有较低的分数，不需要考虑实体的向量信息。

2.4.1.3　训练方法和实现细节

我们定义 PTransE 模型的训练目标函数如下：

$$L = \sum_{(h,r,t)\in\mathbf{T}} \Big[L(h, r, t) + \frac{1}{Z} \sum_{p \in P(h,t)} R(p|h, t) L(p, r) \Big]. \tag{2-23}$$

其中，和 TransE 模型类似，我们采用基于边界距离的函数来定义 $L(h,r,t)$ 和 $L(p,r)$：

$$L(h,r,t) = \sum_{(h',r',t')\in \mathbf{T}^-} [\gamma + E(h,r,t) - E(h',r',t')]_+, \tag{2-24}$$

和

$$L(p,r) = \sum_{(h,r',t)\in \mathbf{T}^-} [\gamma + E(p,r) - E(p,r')]_+, \tag{2-25}$$

其中 $[x]_+ = \max(0,x)$，用来在 x 和 0 中选取一个最大值，γ 是边界距离值，T 是正例的三元组集合而 T' 是负例的三元组集合。

在训练过程中，PTransE 模型采用随机梯度下降来进行优化。在实际优化过程中，我们对实体和关系表示向量添加如下约束：

$$\|\mathbf{h}\|_2 \le 1, \quad \|\mathbf{r}\|_2 \le 1, \quad \|\mathbf{t}\|_2 \le 1. \quad \forall h,r,t. \tag{2-26}$$

此外，在实验过程中还有一些具体问题需要考虑，这些具体问题会显著地影响模型的效果：

- **反向关系添加**：在很多情况下，我们会需要考虑现有知识图谱中关系的反向关系，而这样的反向关系在现有知识图谱中很可能不存在。例如，对于关系路径 $e_1 \xrightarrow{\texttt{BornInCity}} e_2 \xleftarrow{\texttt{CityOfCountry}} e_3$，我们希望预测关系事实 $(e_1, \texttt{Nationality}, e3)$。但是，由于基于路径约束的资源分配算法只能考虑从头实体到尾实体的正向关系，该关系事实无法被预测出来。因此，我们对于知识图谱中的每一个关系添加一个反向关系，也就是，对于每一个关系事实三元组 (h,r,t)，我们向知识图谱中添加其对应的反向三元组 (t,r^{-1},h)。这样，我们可以考虑关系路径 $e_1 \xrightarrow{\texttt{BornInCity}} e_2 \xrightarrow{\texttt{CityOfCountry}^{-1}} e_3$ 来进行知识表示学习。
- **关系路径选择的约束**：在知识图谱中，很多实体对应的关系和事实数目庞大，直接枚举头部和尾部实体之间的所有可能的关系路径会非常耗时和占用空间，因而我们无法直接在 PTransE 的学习中考虑所有可能的关系路径。例如，如果平均每个实体指向超过 100 个关系（这在 Freebase、Wikidata 等大规模知识图谱中很常见），那么理论上一个实体对之间可能数十亿个 4 步路径。即使对于 2 步或 3 步路径，直接考虑所有这些路径也是非常耗时的。为了提高计算效率，本章节中我们将路径长度限制在最多 3 步，并仅考虑那些置信度分值大于 0.01 的关系路径。

2.4.2　实验分析

我们在一个现有的大规模通用知识图谱 Freebase[2] 上对 PTrasnE 模型进行了评测，具体地，我们采用了从 Freebase 中抽取的两个子数据集，即 FB15K 和 FB40K 来进行测试。这两个数据集的统计信息在上一节已经进行介绍，在这里不再赘述。

在实验中，我们采用了两种评测任务来对 PTransE 以及基线方法进行对比：（1）链接预测，仅基于已有知识图谱的基础上，来预测给定三元组中缺失的实体或关系；（2）文本关系抽取，结合关系抽取模型和知识表示学习模型的信息来共同抽取实体之间的关系。

2.4.2.1　链接预测

链接预测任务旨在 (h,t,r) 中的某一元素缺失时，能够补全整个三元组。这一任务曾在许多现有知识表示学习工作中被用于评测[8-9,18]。

在评测阶段，对于每一个候选三元组 (h,r,t)，我们定义如下评分函数用于评测：

$$S(h,r,t) = G(h,r,t) + G(t,r^{-1},h), \tag{2-27}$$

而评分函数 $G(h,r,t)$ 可被进一步定义为

$$G(h,r,t) = ||\boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t}|| + \frac{1}{Z}\sum_{p \in P(h,t)} \Pr(r|p)R(p|h,t)||\boldsymbol{p} - \boldsymbol{r}||,$$

这一评分函数和算法模型一节中定义的评分函数很相近。区别在于，我们认为某一关系路径 p 的可靠性也和 r 的推测强度有关。在这里，我们根据训练数据对推测强度进行量化：$\Pr(r|p) = \Pr(r,p)/\Pr(p)$。

在评测中，我们将链接预测任务分为两个子任务，即实体预测与关系预测。

实体预测　在实体预测这一子任务中，我们遵循了 Bordes 等人[18] 的实验设置，并使用了两种评价指标：(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，即前十命中率（Hits@10）。作为比较，我们选取了 Bordes 与 Wang 等人[18,29] 相关工作中的所有方法作为我们的基准模型，且由于评测数据集相同，我们直接使用了他们论文中所报告的结果。

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。因为如果要考虑所有可能的关系路径，我们必须为每一个测试三元组迭代遍历实体集合 E 中的每一个候选实体，并寻找相应的关系路径。这里，我们提出了一种重排序方法：我们首先根据 TransE 计算的

表 2.6　实体预测的测试结果

Metric	Mean Rank		Hits@10 (%)	
	Raw	Filter	Raw	Filter
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**

评分对所有候选实体进行排序，然后对排名在前 500 的实体使用 PTransE 进行评分，并重新排序。对于 PTransE 而言，我们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 $\lambda = 0.001, \gamma = 1, k = 100$，采用了 L_1 距离。在训练时，我们限定训练轮数为 500 轮。

实体预测的评估结果显示在表格2.6中。基准模型包括 RESCAL[14]、SE[8]、SME（线性）[9]、SME（双线性）[9]、LFM[12]、TransE[18]、TransH[29] 和 TransR[26]。对于 PTransE，我们考虑关系路径表示的三种组合操作：加和（ADD）、乘积（MUL）和递归神经网络（RNN）。在实验中，我们考虑了 2 步和 3 步的关系路径。此外，我们重新实现了 TransE，在相同的参数配置下，我们实现的 TransE 性能显著高出 Bordes 等人[18] 在论文中报告的结果。

从表格2.6中我们观察到：

（1）PTransE 显著优于包括 TransE 在内的其他基准模型。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 成功嵌入到低维空间之中。例如，*George W. Bush* 和 *Abraham Lincoln* 都是 *the United States* 的总统，他们在 TransE 中展现出类似的向量表示。这可能会扰乱 TransE 对于 *Laura Bush* 配偶的预测。相反，由于 PTransE 刻画了关系路径，它可以利用 *George W. Bush* 和 *Laura Bush* 之间的关系路径信息，并据此作出更准确的预测。

表 2.7　将关系分类后在 FB15K 上的评测结果 (%)

任务	头实体预测 (Hits@10)				尾实体预测 (Hits@10)			
关系分类	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**

（2）对于 PTransE，加和操作在 Mean Rank 和 Hits@10 中均优于其他组合运算。我们认为原因在于加和操作同时符合 TransE 和 PTransE 的学习目标。以 $h \xrightarrow{r_1} e_1 \xrightarrow{r_2} t$ 为例，两个直接关系在加和操作下的优化目标 $\boldsymbol{h}+\boldsymbol{r}_1=\boldsymbol{e}_1$ 和 $\boldsymbol{e}_1+\boldsymbol{r}_2=\boldsymbol{t}$，可以很容易地从两个式子中得到关系路径的优化目标 $\boldsymbol{h}+\boldsymbol{r}_1+\boldsymbol{r}_2=\boldsymbol{t}$；

（3）PTransE 模型在考虑两步和三步关系路径的情况下能达到比较接近的结果，这表明考虑更长的关系路径对最终的结果可能不会有进一步的帮助，这也佐证了我们对关系路径长度进行限制的想法。

我们将关系进行分类，分类为一对一（1-to-1）、一对多（1-to-N）、多对一（N-to-1）、多对多（N-to-N）四类，并对 PTransE 和一些基准模型在四类关系上的结果进行了更细致的评测，结果呈现在表格2.7中。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效果提升。

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准模型比较中取得了最佳性能，所以我们直接将 PTransE 与我们自己实现的 TransE 进行比较。评估结果显示在表格2.8中。因为与实体相比，关系的数量要小的多，且 TransE 和 PTransE 的 Hits@10 指标均超过了 95%，因而我们列举 Hits@1 而不是 Hits@10 来进行比较。在表中，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），考虑关系路径的 PTransE（+Rev+Path）。针对 PTransE，我们也单独对仅考虑关系路径（-TransE）与仅考虑公式 (2-16) 中的

表 2.8　关系预测的评测结果

评测指标	Mean Rank		Hits@1 (%)	
	Raw	Filter	Raw	Filter
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**

情况（-Path）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。

从表格2.8中我们观察到：

（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%。即使 TransE 本身，考虑了关系路径的 TransE+Rev+Path 与 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；

（2）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之高。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 PTransE-TransE 模型在预测这一类实体对时对关系进行随机猜测，正确答案的排名期望值为 $|R|/2$。与之相对的，是 PTransE-TransE 的 Hits@1 结果比较合理，这说明了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，是知识表示学习中必不可少的特征。

2.4.2.2　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。现有的工作使用大规模的知识图谱作为远程监督信号，用以自动标注句子作为训练实例，并根据从句子中提取的特征建立关系分类器。所有这些方法仅基于纯文本来推理新的关系事实。TransE 曾被用来与基于文本的关系抽取模型进行结合，并使之取得了显著的效果提升[35]，TransH[29] 和 TransR[26] 也进行了同样的工作。在此任务中，

我们探索利用 PTransE 模型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[32] 发布的纽约时报语料库（NYT）作为训练和测试数据。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，对知识图谱进行了拓展，构建出 FB40K，其中包括了 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[35] 提出的关系抽取 Sm2r 模型作为基线方法。在训练 TransE 和 PTransE 模型时，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 $\gamma = 1.0$，以及采用了 L_1 距离。我们还与 Surdeanu 等人[34] 提出的 MIMLRE 模型进行比较。评估曲线如图2.7所示。

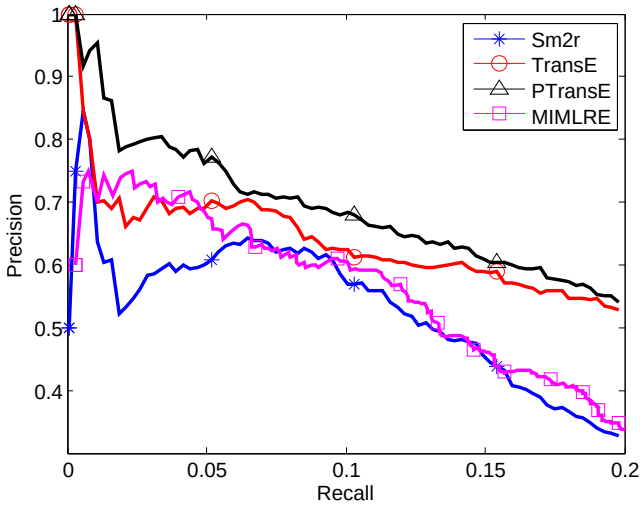

图 2.7　各模型在关系抽取任务上的精度——召回率曲线

从图2.7中，我们可以看到：在与基于文本的模型 Sm2r 相结合进行关系抽取时，PTransE 模型的效果明显优于 TransE，且远高于最基础的 Sm2r 模型的结果。这表明对关系路径进行建模对于从文本中提取关系也具有效果。需要注意的是，此处使用的 PTransE 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为：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却存在若干关系路径。TransE 在训练阶段并不会如 PTransE 一般对关系路径进行建模，这会导致在测试阶段考虑关系路径时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一件简单的工作，同时也证明了 PTransE 的有效性。

2.4.2.3　关系推理的个案研究

实验结果表明，PTransE 可以在链接预测和文本关系抽取任务中取得很好的效果。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，用于解释为什么考虑关系路径可以在这些任务上有如此巨大帮助。如图2.8所示，两个实体 *Forrest Gump* 和 *English* 通过三条关系路径连接起来，这三条路径的信息可以使我们更有把握预测两个实体之间的关系 `LanguageOfFilm`。

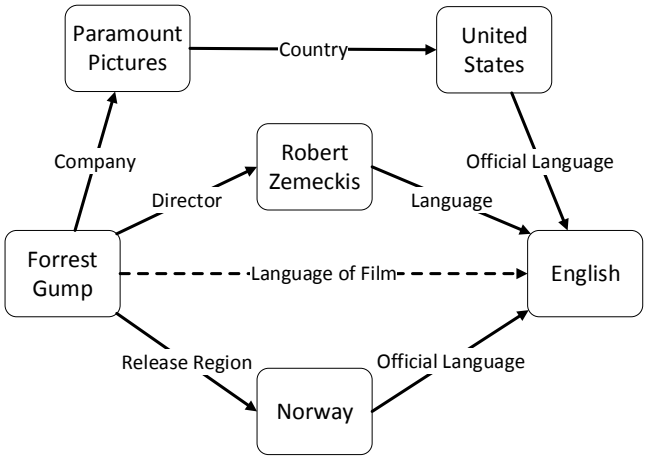

图 2.8　Freebase 中的一个推理的例子

2.5　知识图谱复杂属性建模

TransE 及其扩展模型的最大问题是无法准确地建模一对多、多对一和多对多关系，这促使我们重新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。

通过观察我们发现，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，例如 ‘‘性别’’ 和 ‘‘职业’’），而另一类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**。我们在表格2.9中列出了一些典型的关系和属性的例子。此外，我们同时还列出了每个头实体对应的尾实体的期望数量以及每个尾实体对应的头实体的期望数量，分别表示为 E_t 和 E_h。

如表中所示，属性是一对多和多对一关系的主要来源。例如，在属性 ‘‘性别’’ 中，属性值 ‘‘男’’ 是一个与数百万个表示人的实体相关的存在。对于这些关系，TransE 及其扩展模型（如 TransH[21] 和 TransR[26]），是无法在实体及其属性值之间

表 2.9　一些典型的属性和关系与它们相应的映射属性之间的关系

类型	名称	E_t	E_h
属性	国籍	1.05	1,551.90
	性别	1.00	637,333.33
	种族	1.12	41.52
	宗教信仰	1.09	107.40
关系	父母	1.58	1.67
	首都	1.29	1.42
	作者	1.02	2.17
	创始人	1.37	1.31

通过构建平移关系来有效建模的，这也是现有知识表示学习模型对于一对多和多对一关系效果始终欠缺的因素之一。

显然，属性和关系表现出相当明显的特征：（1）对于关系，一般其关系事实中的头实体和尾实体可以是知识图谱中的所有实体。在这些关系事实中，实体一般仅和有限数量的实体建立特定关系；（2）对于属性，属性值通常来自知识图谱中某一种特定实体（一般是抽象实体）。例如，关系 ‘‘性别’’ 通常只有两个属性 ‘‘男’’ 和 ‘‘女’’。每个属性值会由许多实体共享。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。在 KR-EAR 模型中，每个实体都具有各种属性，实体之间通过关系相连接。我们将所有实体和关系都都表示成低维稠密向量，通过以下方式进行学习的：（1）对于实体之间的关系，我们将其表示为实体之间的平移向量，如 TransE 模型（也可以使用 TransH 和 TransR 模型）；（2）根据实体的向量表示来推断实体的属性。

2.5.1　算法模型

我们首先介绍本章节中使用的定义和符号。令 $G = (E, T, Y)$ 表示一个知识图谱，其中 $E = \{e_1, e_2, \cdots, e_{|E|}\}$ 为实体集合，T 是关系三元组集合，Y 是属性三元组集合。

定义 1：关系三元组： $T \subseteq E \times R \times E$ 是表示实体相互之间关系的关系三元组集合，其中 R 是关系集合。

定义 2：属性三元组： $Y \subseteq E \times A \times V$ 是表示实体属性的属性三元组集合，其中 $A = \{A_1, A_2, ...A_{|A|}\}$ 是属性集合其对于每一种属性 $A_i \in A$ 都有其对应的属性值集合 $V_i \in V$。例如，一个人的性别可以能是 ‘‘男’’ 或者 ‘‘女’’。

给定一个知识图谱 G，我们旨在学习到一个可以用于预测实体间的关系以及

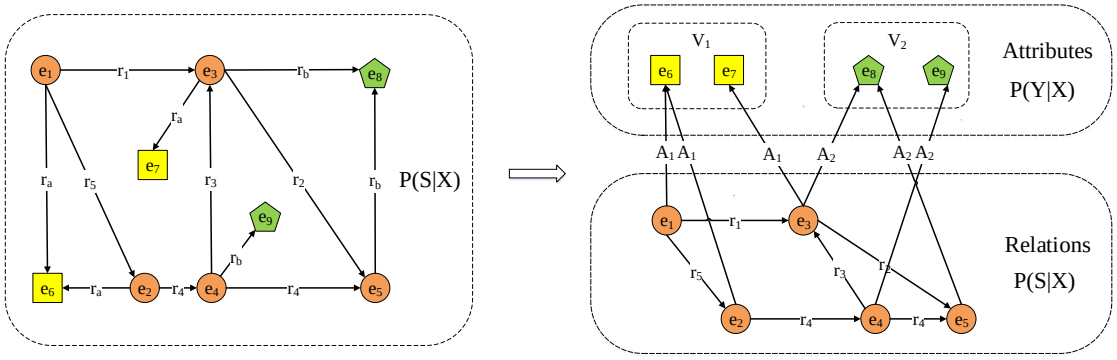

图 2.9　KR-EAR 和传统知识表示模型的例子

实体的属性的统一模型。假设给定实体、关系和属性的表示 $\mathbf{X}$，我们将目标函数定义为所有关系三元组和属性三元组的联合概率。假设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：

$$\begin{aligned} P(T,Y|\boldsymbol{X}) &= P(T|\boldsymbol{X})P(Y|\boldsymbol{X}) \\ &= \prod_{(h,r,t)\in T} P((h,r,t)|\boldsymbol{X}) \prod_{(e,a,v)\in Y} P((e,a,v)|\boldsymbol{X}), \end{aligned} \tag{2-28}$$

其中 $P((h,r,t)|\mathbf{X})$ 表示关系三元组的条件概率，$P((e,a,v)|\mathbf{X})$ 表示属性三元组的条件概率。因此，我们的 KR-EAR 模型包括以下两个部分：

（1）**关系三元组编码器**：用于建模实体和关系之间的相互关联；

（2）**属性三元组编码器**：用于建模实体和属性之间的相互关联，通过一个分类器来实现属性预测。

如图2.9所示，A_1 和 A_2 是两个属性，分别有值域 V_1 和 V_2。在传统的知识表示模型中（左侧），属性 A_1 和 A_2 被试做普通的关系 r_a 和 r_b，而与此不同的是，KR-EAR（右侧）将属性预测视作分类问题。

2.5.1.1　关系三元组编码器

在关系三元组编码器中，我们希望对实体和关系进行编码来建摸两者之间的相互联系。在学习过程中，我们通常通过优化条件概率 $P(h|r,t,\mathbf{X})$, $P(t|h,r,\mathbf{X})$ 和 $P(r|h,t,\mathbf{X})$ 来代替优化 $P((h,r,t)|\mathbf{X})$。

这里，不是一般性，我们采用 TransE 和 TransR 两个经典的知识表示学习模型来建摸关系三元组。以条件概率 $P(h|r,t,\mathbf{X})$ 为例，其具体定义如下：

$$P(h|r,t,\mathbf{X}) = \frac{\exp(g(\mathbf{h},\mathbf{r},\mathbf{t}))}{\sum_{\hat{h}\in E}\exp(g(\hat{\mathbf{h}},\mathbf{r},\mathbf{t}))}, \tag{2-29}$$

其中 $g()$ 表示能量函数，用于衡量关系 r 与实体对 (h,t) 之间的相互联系。这里，我

们可以简单采用 TransE 的能量函数来进行计算：

$$g(\mathbf{h},\mathbf{r},\mathbf{t}) = -||\mathbf{h}+\mathbf{r}-\mathbf{t}||_{L1/L2} + b_1, \tag{2-30}$$

或者采用 TransR 的能量函数来进行计算：

$$g(\mathbf{h},\mathbf{r},\mathbf{t}) = -||\mathbf{h}\mathbf{M_r}+\mathbf{r}-\mathbf{t}\mathbf{M_r}||_{L1/L2} + b_1, \tag{2-31}$$

其中 b_1 是一个常量，$\mathbf{M_r}$ 是 TranR 中的关系 r 对应的映射矩阵。我们将这两个版本的 KR-EAR 模型分别命名为 KR-EAR(TransE) 和 KR-EAR(TransR)。这里，将 TransE 和 TransR 模型表示为概率形式已经在论文[29] 被证明是有效的。

事实上，其他的知识表示学习模型也可以很容易地作为我们的关系三元组编码器，直接将其中能量函数 $g()$ 的计算方式进行替换即可。

2.5.1.2　属性三元组编码器

实体和属性之间的相互联系从直觉上看可以通过一个分类模型来建模。因此，属性三元组编码器对于每一个属性三元组 (e,a,v) 建模其条件概率 $P(v|a,r,\mathbf{X})$，具体定义如下：

$$P((e,a,v)|\boldsymbol{X}) \propto P(v|e,a,\boldsymbol{X}) = \frac{\exp(h(\boldsymbol{e},\boldsymbol{a},\boldsymbol{v}))}{\sum_{\hat{v}\in V_a}\exp(h(\boldsymbol{e},\boldsymbol{a},\hat{\boldsymbol{v}}))}, \tag{2-32}$$

其中 $h()$ 是对于给定实体每一个属性和属性值的评分函数。我们通过如下方式计算 $h()$。首先，我们将实体向量表示通过一个单层神经网络映射到特定属性对应的空间，然后计算映射后实体表示和具体属性值之间的关联系，具体计算如下：

$$h(\mathbf{e},\mathbf{a},\mathbf{v}) = -||f(\mathbf{e}\mathbf{W_a}+\mathbf{b_a}) - \mathbf{V_{av}}||_{L1/L2} + b_2, \tag{2-33}$$

其中 $f()$ 是一个如 tanh 的非线性函数，$\mathbf{V_{av}}$ 是属性值 v 的在属性 a 空间中的低维向量表示，b_2 是一个常量。

众所周知，一个实体的不同属性之间一般会有很强的相互联系。例如，一个国籍是英国的人，一般更有可能会讲英文。而我们的模型将属性和关系进行了显示的切分，更有利于我们对属性之间的相互联系进行建模。

具体地，对于一个实体 e 及其属性三元组 (e,a,v)，假设 $Y(e) = \{(e,\hat{a},\hat{v})|(e,\hat{a},\hat{v}) \in Y\}$ 是实体 e 除了 (e,a,v) 之外的属性三元组集合，那么，我们可以将属性三元组 (e,a,v) 的条件概率 $P((e,a,v)|\boldsymbol{X})$ 进一步近似为：

$$P((e,a,v)|\boldsymbol{X}) \propto P(v|e,a,\boldsymbol{X})P((e,a,v)|Y(e)), \tag{2-34}$$

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，它通过一个 softmax 函数进行定义：

$$P((e,a,v)|Y(e)) = \frac{\exp(z(\boldsymbol{e},\boldsymbol{a},\boldsymbol{v},Y(e))}{\sum_{\hat{v}\in V_a}\exp(z(\boldsymbol{e},\boldsymbol{a},\hat{\boldsymbol{v}},Y(e)))} \tag{2-35}$$

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：

$$z(\boldsymbol{e},\boldsymbol{a},\boldsymbol{v},Y(e)) \propto \sum_{(e,\hat{a},\hat{v})\in Y(e)} P((a,v)|(\hat{a},\hat{v}))(\boldsymbol{A_a}\cdot\boldsymbol{A_{\hat{a}}}), \tag{2-36}$$

其中 $(\boldsymbol{A_a}\cdot\boldsymbol{A_{\hat{a}}})$ 是 $\boldsymbol{A_a}$ 与 $\boldsymbol{A_{\hat{a}}}$ 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直接统计得到的，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。

2.5.1.3　训练方法和实现细节

在这里，我们将 KR-EAR 模型的优化目标定义为公式 (2-28) 的 log 概率函数，具体地：

$$L = \log(P(T,Y|\mathbf{X})) + \alpha C(\mathbf{X}) \tag{2-37}$$

其中 α 是正则因子 $C(\mathbf{X})$ 的权重超参数。$C(\mathbf{X})$ 具体定义如下：

$$\begin{aligned} C(\mathbf{X}) &= \sum_{e\in E}\left[||\mathbf{e}||-1\right]_+ + \sum_{r\in R}\left[||\mathbf{r}||-1\right]_+ \\ &+ \sum_{e\in E}\sum_{i=1}^{|A|}\left[||\mathbf{e}\mathbf{W_i}+\mathbf{b_i}||-1\right]_+ + \sum_{i=1}^{|A|}\left[||\mathbf{V_i}||-1\right]_+, \end{aligned} \tag{2-38}$$

其中 $[x]_+ = \max(0,x)$，用来在 x 和 0 中选取一个最大值。正则因子 $C(\mathbf{X})$ 在训练过程中可以尽量避免模型过拟合。

在训练过程中，KR-EAR 的模型采用随机梯度下降（stochastic gradient descent，SGD）来进行优化。此外，为了加速模型，我们还对模型中的 softmax 函数进行了近似。

Softmax 函数近似　在 KR-EAR 模型中，我们无法直接利用 softmax 函数对条件概率 $P(h|r,t,\mathbf{X})$, $P(t|h,r,\mathbf{X})$, $P(r|h,t,\mathbf{X})$, $P(v|e,a,\mathbf{X})$ 和 $P((e,a,v)|Y(e))$ 进行计算。其原因在于这些条件概率对应的 softmax 函数的计算复杂度和实体总数 $|E|$ 和关系总数 $|R|$ 成正比。而在大规模知识图谱如 Freebase、Wikidata 中，实体和关系的数量通

常都非常多，导致计算量无法承受。因此，我们采用 Mikolov 等人[5] 论文中采用的负采样方法对 softmax 函数进行近似。以公式 (2-29) 中的条件概率 $P(h|r,t,\mathbf{X})$ 为例，其可以通过负采样近似如下：

$$P(h|r,t,\mathbf{X}) = \prod_{(h,r,t)\in T} \left[\sigma(g(\mathbf{h},\mathbf{r},\mathbf{t})) \prod_{i=1}^{c_1} \mathbb{E}_{(h_i,r,t)\in T^-} \sigma(g(\mathbf{h}_i,\mathbf{r},\mathbf{t}))\right] \tag{2-39}$$

其中 $\sigma(x) = 1/(1+\exp(-x))$ 是 sigmoid 函数，c_1 是常量。同样地，我们可以采用负采样对 $P(t|h,r,\mathbf{X})$, $P(r|h,t,\mathbf{X})$, $P(v|e,a,\mathbf{X})$ 和 $P((e,a,v)|Y(e))$ 进行近似。

2.5.2　实验分析

2.5.2.1　数据集与实验设置

表 2.10　数据集 FB24k 的统计数据

数据集	FB24k
实体数量	23,634
关系数量	673
属性数量	314
三元组总数	423,560
训练集 (关系三元组)	205,643
测试集 (关系三元组)	10,766
训练集 (属性三元组)	196,850
测试集 (属性三元组)	10,301

我们在典型的大规模知识图谱 Freebase 上对我们的模型进行了评测。我们以如下方式构建了评测数据集：

（1）过滤掉低频实体和关系，只保留在至少 30 个三元组中出现过的实体和关系；

（2）过滤掉反向关系。在 Freebase 中，每个关系也对应于头尾实体交换后的反向关系。对于每个关系及其反向关系，我们只需要在数据集中保留二者之一。更具体地，对于那些多对一的关系及其一对多的反向关系，由于后者通常与人类的直觉不符，我们会保留前者并舍去后者。例如，对于 Freebase 中的关系 *people.person.nationality*，我们将舍去其反向关系*!people.person.nationality*；

（3）划分属性和关系。在原始 Freebase 中，关系三元组和属性三元组之间没有

表 2.11　实体预测的评测结果

实体	头实体				尾实体				合计			
评测指标	Mean Rank		Hits@10 (%)		Mean Rank		Hits@10 (%)		Mean Rank		Hits@10 (%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
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**

明确的划分。我们人工将原始 Freebase 关系分为两类：属性和关系。最后，我们建立一个名为 FB24k 的数据集，并将数据随机划分为训练集和测试集。

数据集的统计数据在表格2.10中。

我们使用数据集 FB24k 评估了我们的模型和基线模型在链接预测任务中的表现。我们将链接预测任务分为实体预测、关系预测和属性预测三个子任务分别展示和讨论实验结果。在这三个子任务中，我们对 KR-EAR 模型和三个目前效果最好的知识表示模型 TransE、TransH 和 TranR 进行了比较。

对于 KR-EAR 的实验超参数，我们通过网格搜索来寻找最优的超参数。对于 SGD 的学习率 λ，我们搜索了 $\{0.1, 0.01, 0.001\}$；对于正则因子的权重 α，我们搜索了 $\{0.1, 0.01, 0.001\}$；对于实体、关系和属性的表示向量维度，我们搜索了 $\{20, 50, 80, 100\}$；对于所有的偏置常量 b_1, b_2, c_1, c_2，我们搜索了从 -10 到 10 之间的所有整数。最优的超参数是 $\lambda = 0.001$, $\alpha = 0.1$, $k = 100$, $b_1 = 7$, $b_2 = -2$, $c_1 = 10$, $c_2 = 1$ 并采用 L_1 距离度量。在训练中，我们将所有训练三元组（包括关系三元组和属性三元组）上的迭代次数设置为 1000。

2.5.2.2　实体预测

实体预测的目的在于在关系三元组的头尾实体中有一个缺失时对其进行预测。对每一个测试关系三元组 (h, r, t)，我们逐一使用数据集中的实体替换缺失的头尾实体，并计算相应得分 $\sigma(g(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}))$。之后，我们按照得分升序计算数据集中每个候选实体的名次。

在实体预测任务中，我们沿用了[18] 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)。

结果　我们在表格2.11中展示实体预测的评测结果。从表中我们观察到：

（1）KR-EAR 在 Mean Rank 指标下稳定地显著优于包括 TransE、TransH 和

TransR 在内的所有基线方法，包括 TransE、TransH 和 TransR。这表明 KR-EAR 为实体和关系学习更加好的向量表达;

（2）在 Hit@10 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的结果，不会因为特殊考虑了实体的属性值对其造成效果上的影响。

2.5.2.3　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。对每一个测试关系三元组 (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 $\sigma(g(\boldsymbol{h},\hat{\boldsymbol{r}},\boldsymbol{t}))$。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。

在关系预测任务中，同样地，我们使用了 Mean Rank 和 Hits@1 两种评测指标。**关系和属性之间的相关性**　[47] 的结果表明了，通过隐变量模型对关系的头尾实体类型约束进行考虑，可以有效地提升知识表示学习模型的效果。我们认为实体的类型信息本质上是实体属性的一个特例。因此，我们可以进一步利用实体属性对知识表示学习模型进行约束。在 KR-EAR 模型中，由于我们对实体的关系和属性进行了区分，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，我们将其简称为 CRA（Correlation between Relations & Attributes）。

表 2.12　关系预测的评测结果

评测指标	Mean Rank		Hits@1 (%)	
	Raw	Filter	Raw	Filter
TransE	3.1	2.8	65.9	83.8
TransH	3.4	3.1	64.9	84.1
TrasnR	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**

结果　我们在表格2.12中展示了关系预测的评测结果。从表中我们观察到：

（1）KR-EAR 在 Mean Rank 和 Hits@1 下再次优于所有基线方法，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 Hits@1 提高 2.5% 和 1.4%，同时也可以降低 Mean Rank。这证明了在关系预测中考虑实体属性约束的有效性。

2.5.2.4　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做实体预测的一部分[18,21,26]。对每个测试属性三元组 (e,a,v)，我们用每个可能的属性值 $\hat{v}$ 替代 v，并计算相应得得分 $\sigma(h(\boldsymbol{e},\boldsymbol{a},\hat{\boldsymbol{v}}))$。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，我们可以通过将属性值按照 $\sigma(h(\boldsymbol{e},\boldsymbol{a},\hat{\boldsymbol{v}}))\sigma(z(\boldsymbol{e},\boldsymbol{a},\hat{\boldsymbol{v}},Y(e)))$ 排序，进一步在 KR-EAR 模型中加入对属性相关性（AC: Attribute Correlations）的考虑。

在属性预测中，我们同样使用了两种属性预测的评估指标：Mean Rank 和 Hits@1。

表 2.13　属性预测的评测结果

评测指标	Mean Rank		Hits@1 (%)	
	Raw	Filter	Raw	Filter
TransE	10.7	5.6	36.5	55.9
TransH	10.7	5.6	38.5	57.9
TrasnR	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**

结果　我们在表格2.13中给出了属性预测的评测结果。从表中我们可以看出：

（1）KR-EAR 仍然明显优于所有基线方法。这验证了在传统知识表示模型中将属性预测建模为分类模型而不是实体之间的平移向量的必要性；

（2）对于 KR-EAR（TransE）和 KR-EAR（TransR），考虑属性相关性可以分别将 Hits@1 提升 1.4% 和 1.0%。这表明考虑属性之间的相关性在属性预测中是有效的。

表 2.14　属性相关性的例子

属性	关联属性
职业	婚姻状况, 国籍, 性别, 语言, 种族
电影发布地区	电影国家, 电影语言, 电影发布日期, 电影题材
地点时区	国家位置, 地点货币
音乐类型	使用乐器, 所属专辑, 职业, 乐器发声部位
电视剧题材	电视剧国家, 电视剧语言, 电视剧发行网络

属性相关性的个例分析　在表格2.14中，我们给出了一些 KR-EAR 在 FB24k 训练集上得到的属性相关性的例子。我们可以发现，在给定一个属性时，预测得到的相关属性常常反映符合我们认知上的相关性。这表明 KR-EAR 可以有效捕获属性之间的相关性。

2.6　小结

近些年，人们构建了许多大规模知识图谱，以关系事实三元组的形式来表示实体及其关系，用以表示现实世界中的知识。这些知识图谱，为人类学习和理解现实世界提供了有效途径。目前，作为处理大规模图谱的有效手段，知识表示学习模型被广泛应用在很多知识驱动的任务中并取得了显著的效果。然而，尽管现有的知识表示学习模型已经显示出它们对知识图谱进行建模的能力，其仍然存在许多可能的改进方向需要进行探索。在本章节我们主要介绍了我们对于知识表示学习中考虑知识图谱的复杂关系、复杂路径和复杂属性三方面的改进工作。

第 3 章　结构化知识的自动抽取

3.1　引言

近些年来，包括 Freebase、DBpedia、YAGO、Wikidata 在内的不少大型知识图谱逐渐形成体系，并被广泛地应用于包括网页搜索、问答系统、文本检索等诸多自然语言处理任务上，取得了显著的效果提升。虽然现有的大型知识图谱已经包含了海量关系事实，但与现实世界中近乎于无穷无尽的知识量相比，已有知识图谱是远远不够完善的。为了尽可能地完善知识图谱中的关系事实，研究者们花费了很多精力去研究如何自动发现并挖掘世界知识，进而完成知识图谱的构建，其中较为核心的技术就是文本关系抽取。

关系抽取旨在从未经标注的自由文本中抽取实体间的关系事实，进而将实体与关系构成知识并扩展到知识图谱中。传统的文本关系抽取方法主要立足于构建有监督的文本关系抽取系统，其训练极度依赖于大规模的人工标记数据，这带来了巨大的时间与人力耗费。因此，Mintz 等人[31] 提出了基于远程监督的关系抽取方法，通过对齐知识图谱中已有的关系事实和未经标注的自由文本来自动生成训练数据，并能够充分利用大规模的数据训练出可用的关系抽取模型。尽管远程监督技术已经成为当下关系抽取研究中的重要一环，但该方法本身也面临着一些核心难题亟待解决：

1. 数据噪音问题。远程监督通常采用假设极强的启发式规则来自动标注数据，其获取的训练数据中往往存在大量噪音。
2. 数据长尾问题。关系抽取依靠已有的知识图谱来进行远程监督，其自动获取的数据往往呈现幂律分布，大量的长尾实体对的关系是难以通过远程监督来获取数据的。

无论是噪音数据还是长尾数据均为训练抽取系统带来了困难。针对这些挑战性问题，我们立足于文本关系抽取，对如何从文本中获取知识进行了研究与探索，并从以下两个方面上深入与展开：

1. 如何利用远程监督获取大规模训练数据，同时降低其中噪音数据带来的影响，进而训练得到更鲁棒的文本关系抽取系统。
2. 如何引入多语言的额外信息，用以缓解长尾数据带来的影响，进而得到效果更好的文本抽取系统。

上述两点研究方向均有利于获取性能优异的文本关系抽取系统以支持知识图谱的关系事实获取。

在本章节具体介绍我们针对上述问题提出的解决方法之前，我们同样会先回顾文本关系获取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，并着重对关系抽取发展进行梳理。

3.2　相关工作

在本章节的相关工作介绍之前，我们首先在此简要地介绍关系抽取中的常用符号表示。我们将一个知识图谱表示为 $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$，其中 $\mathcal{E}, \mathcal{R}, \mathcal{T}$ 分别表示实体、关系和事实的集合。$(h, r, t) \in \mathcal{T}$ 表示 $h, t \in \mathcal{E}$ 之间存在关系 $r \in \mathcal{R}$。对于给定的实体对 (h, t)，关系抽取旨在从包含 (h, t) 的若干语句中挖掘语义从而最终获取 (h, t) 的关系 r，这些包含 (h, t) 的句子则被聚集为实体对的实例包 $\mathcal{S}_{(h,t)} = \{s^1_{(h,t)}, s^2_{(h,t)}, \ldots\}$。包中的每个实例 $s^i_{(h,t)}$ 被表示为词的序列 $s = \{w_1, w_2, \cdots\}$。

3.2.1　有监督的关系抽取模型

关系抽取是自然语言处理领域中最重要的任务之一，也是世界知识获取的必由之路。之前的研究人员已经在关系抽取方面做出了诸多工作，尤其是基于有监督的关系抽取系统构建。传统有监督的关系抽取模型主要采用统计模型，包括采用特征工程的方法[48-51]，核函数方法[52-58]，图模型方法[59-62]，特征嵌入方法[35,63-64]。尽管上述统计关系抽取模型已被广泛研究，但它们仍面临一些挑战。特征工程方法的核心是设计准确有效的特征，这需要诸多工作来细致观察语言现象并分析它们对提取关系的贡献。与特征工程方法中的复杂特征提取相比，核函数方法的内核设计更加简单，但是，这仍然需要大量的工作来精心设计内核函数。图模型方法和特征嵌入方法能够在一定程度上构建有效特征并预测关系而无需过多的人为干预，然而，这些简单结构的模型具有十分有限的拟合能力。

近来，随着深度学习的快速发展，有监督的神经关系抽取模型得到了广泛的应用和研究。与传统的统计模型相比，这些神经抽取模型可以精准地捕获文本信息，而无需进行明确的语言分析。除了能够摆脱复杂和易错的特征工程之外，神经模型本身的模型性能也是十分强大的，能够很好的拟合数据。有监督的神经关系抽取模型主要集中在引入神经网络以从文本中提取语义特征并对关系进行分类。受计算机视觉任务中卷积神经网络的启发，卷积神经网络首先被用于关系抽取中，并被扩展到诸多变种模型[65-69]。之后，为了更好地建模自然语言这样的序列性数据，循环神经网络也被引入[70-75]。基于依赖关系的神经模型[76-81] 也被提出用于关系抽取，其模型中使用了诸如句法分析树与最短依赖路径之类的信息。

3.2.2　远程监督的关系抽取模型

尽管有监督的关系抽取模型取得了不错的效果，但这些方法大多需要大量的标注数据，构建大规模的标注数据需要耗费大量的时间与人力物力。为了解决这个问题，Mintz 等人[31] 通过远程监督方法将纯文本与知识图谱进行对齐，并自动标注数据。具体来说，远程监督通常以一个极强的启发式条件来进行数据获取：如果两个实体在知识图谱中有关系，那么包含这两个实体的所有句子都将被视为其语义足以表达这种关系。例如，（微软公司，创建者，比尔·盖茨）是世界知识图谱中的关系事实。远程监督将把包含这两个实体的所有句子视为创建者这个关系的训练实例。虽然这是自动标记训练数据的有效策略，但其过强的设定不可避免地会产生错误标注的问题。例如，‘‘比尔 · 盖茨转向慈善事业与微软公司在美国与欧盟遇到的反托拉斯问题有关。’’ 这句话并不表达创建者的关系，但仍将被标注为创建者这个关系的一个训练实例。

为了解决远程监督中不可避免的错误标注问题，不少工作着力于引入多实例学习来规避噪音问题。早期的多实例学习通过考虑每个实例标注的准确性，进而共同作用于最终的结果预测，在预测药物活性等训练数据标注过于模糊的问题上取得了不错的效果。Bunescu 等人[82] 将弱监督学习与多实例学习相结合，并将其扩展到关系抽取上。Riedel 等人[32] 将基于远程监督的关系抽取问题形式化为多实例单标签问题，之后 Hoffmann 与 Surdeanu 等人[33-34] 更进一步地形式化为多实例多标签学习问题。

上述方法均为基于人工特征的方法，并强烈依赖于自然语言处理工具来生成特征，所以在数据标注错误之外也遇到了特征提取错误的叠加与传播的问题。如我们在有监督的关系抽取模型中提到，伴随着深度学习广泛应用于各个领域，许多研究人员着手尝试使用神经网络来提取特征用于关系抽取。尽管这些方法已经取得了很大的成功，他们仍然都是在单个句子上进行抽取，并且深受训练数据匮乏之苦。Zeng 等人[83] 将多实例学习与神经网络模型相结合以进行基于远程监督的关系抽取。尽管基于远程监督的神经关系抽取模型很好地兼顾了鲁棒性与有效性，并因此被广泛运用于知识获取任务之中，但仍然存在诸多问题：

第一，Zeng 等人假设每个实体对所对应的语句集合中只有一个句子是有效的。因此，这样的机制将丢失大量包含在那些被忽略的句子中的丰富信息。针对这个问题我们提出面向多个实例的语句级别注意力机制，这可以充分地利用所有句子的信息。对于信息量丰富的实例，我们的注意力机制可以赋予其更高的权重，而对于信息量较少以及错误标注的实例，我们的注意力机制将赋予其较低的权重以减少其对模型训练的影响。

第二，现有的关系抽取系统只关注从单语言文本中直接抽取实体之间的关系，因而难以应对训练数据中的长尾问题。引入众多多源的额外相关信息，将有助于缓解长尾数据带来的问题。对此，我们研究如何考虑丰富的多语言信息来加强关系抽取效果。我们在此方面着重研究如何结合多语言丰富语料来进行抽取系统的训练与预测。

在本章的剩余部分，我们将详细讨论我们的改进工作，并给出更多的细节。

3.3　基于选择性注意力机制的关系抽取

目前，神经网络模型已经在关系分类任务[66,76,84]中取得了很好的效果。其原因在于神经网络模型采用深度神经网络来抽取每一种关系需要的特征，不再需要人工进行特征设计，更加贴合任务需求。然而，现有的神经网络关系分类模型一般需要基于语句基本的标注数据来构建分类器，由于缺乏大量人工标注的训练数据，因此无法扩展到关系抽取任务中，也无法直接应用于大规模知识图谱的构建。

针对这个问题，研究者希望通过远程监督来构建关系抽取任务的训练数据，从而将神经网络模型的优势扩展到关系抽取任务中。远程监督假设是关系抽取任务中一个重要的前提，其假定：如果两个实体在知识图谱中存在某种关系，那么所有包含这两个实体的句子都会表达这种关系。例如，(比尔 · 盖茨，创立者，微软公司)是知识图谱中的一个关系事实三元组，那么远程监督假设认为所有同时包含 ‘‘比尔 · 盖茨’’ 和 ‘‘微软公司’’ 两个实体的句子都是关系 ‘‘创立者’’ 的正例。基于远程监督假设，我们可以通过将知识图谱中已有关系事实三元组和未经标注的自有文本来自动生成关系抽取任务的训练数据。然而，尽管远程监督假设可以有效地产生大量的标注数据，其过强的假设不可避免地导致了错误标注数据的产生。例如，句子 ‘‘比尔 · 盖茨开始重视慈善事业，其原因与微软公司在美国和欧盟的反托拉斯问题有关。’’ 中同时出现了实体 ‘‘比尔 · 盖茨’’ 和 ‘‘微软公司’’，虽然这个句子和 ‘‘创立者’’ 关系无关，但其仍然被认为是一个有效的训练实体。

针对这个问题，Zeng 等人[83]将多实例学习与神经网络关系分类模型相结合，提出了基于远程监督数据的关系抽取模型。尽管该方法在关系提取方面取得了显著的效果，但仍然远远不能令人满意。该方法假设在所有包含两个实体的句子中，至少有一个句子能够表达这两个实体之间的关系，并且在训练和预测两个阶段仅使用每个实体对所对应的概率最高也就是最有可能正确表达其关系的句子。很明显，这种方法将会丢失大量被忽略的句子中的丰富的有用的信息。

为了综合利用包含同一实体对的所有句子中的有效信息进行关系抽取，本章节将介绍一种基于语句级别选择性注意力机制的神经网络关系抽取模型（NRE），

用于构建基于远程监督的关系抽取系统。如图 3.1所示，NRE 模型使用卷积神经网络（CNN）来提取语句的语义特征并以语义向量的形式来表示语句。为了充分利用包含同一实体对的所有句子的信息，以及解决远程监督带来的错误标注问题，NRE 模型在这些句子的语义向量基础上构建语句级别的选择注意力机制，从而动态地减少噪音句子所对应的权重，同时提升包含有效信息的句子所对应的权重。最后将注意力机制计算的权重与对应的句子向量加权求和作为特征向量来进行关系分类。

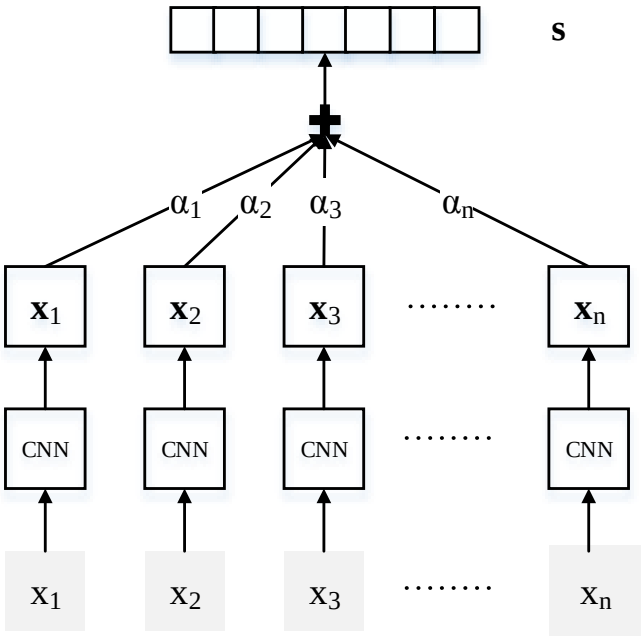

图 3.1　基于语句级别注意力机制的卷积神经网络关系抽取模型，其中 x_i 和 $\boldsymbol{x}_i$ 分别表示其原始的语句字符串和它对应的经过卷积神经网络后得到的输出也就是其向量化的表示，α_i 是由选择注意力机制模型分配给这个句子的权重，$\boldsymbol{s}$ 表示这个实体对所对应的句子集合的向量化表示

3.3.1　算法模型

给定实体对 (h,t) 及包含 (h,t) 的句子集合 $s_{(h,t)} = \{x_1, x_2, \ldots, x_n\}$，本章节介绍的 NRE 模型将预测 (h,t) 与每个关系 $r \in \mathcal{R}$ 形成事实关系三元组的概率。整体模型主要分为以下两个部分：

- **语句编码器**：给定一个句子及其包含的两个目标实体，NRE 模型用一个卷积神经网络来提取句子的分布式表示向量。
- **选择性注意力机制**：当学习到所有句子的分布式表示向量后，NRE 模型使用语句级别的选择性注意力机制来选择那些能够真正表达对应关系的语句。

3.3.1.1　语句编码器

如图 3.2所示, NRE 模型通过卷积神经网络将给定的句子 $x = w_1, w_2, \ldots, w_m$ 转换成它所对应的分布式表示向量 $\boldsymbol{x}$。

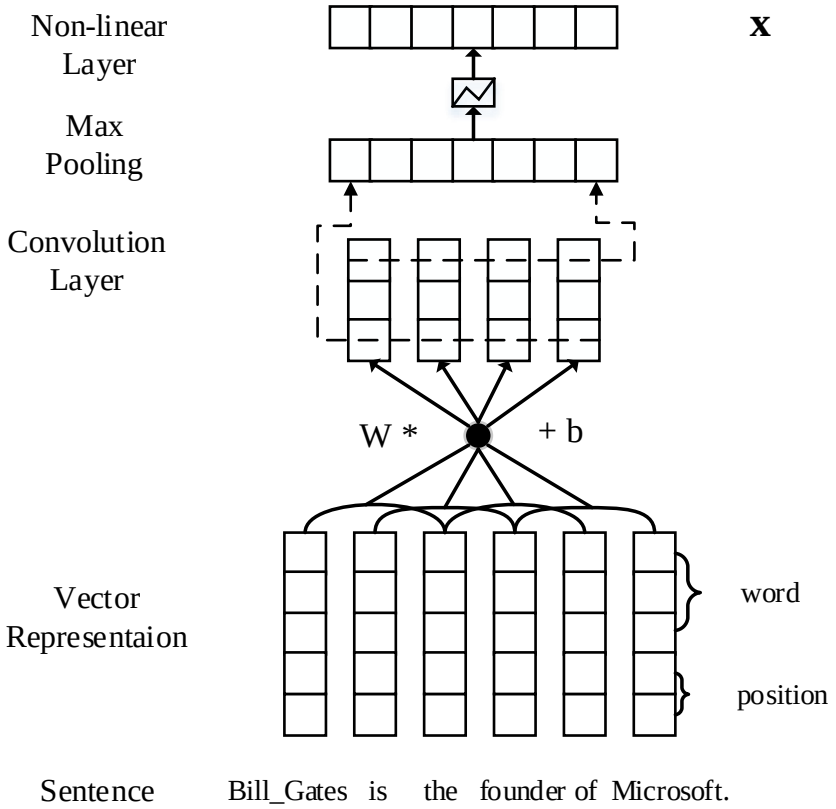

图 3.2　语句编码器 CNN/PCNN 的神经网络结构

输入表示　卷积神经网络输入的是句子 x 的原始单词序列。NRE 模型首先使用词向量矩阵将每个输入单词变换成对应的低维词向量。在词向量之外，为了刻画实体对在句子中所处的位置，NRE 模型还对句子所有单词与头尾实体的相对位置进行了向量化表示。具体地，我们的输入将表示成以下两种形式：

词向量：词向量旨在将离散字符形式的单词转换为连续向量空间中的分布式表示，从而捕捉到单词所对应的语义信息。给定一个句子 x 及其包含的 m 个单词 $x = \{w_1, w_2, \cdots, w_m\}$，每个单词 w_i 将由一个实值向量 $\boldsymbol{w}_i$ 所表示，即其所对应的词向量。

位置向量：在关系抽取任务中，一般情况下越靠近目标实体的单词越具有信息量，并对最终确定目标实体对的关系越具有帮助。类似于 Zeng 等人[66] 的处理方法，NRE 模型将各单词到目标实体对的相对距离向量化，以此来帮助神经网络

去刻画每个单词相对于头尾实体的相对距离。例如，在句子“比尔·盖茨是微软公司的创始人”中，“是”距离头实体比尔·盖茨与尾实体微软公司的相对距离分别为-1 与 1，NRE 模型将-1 与 1 进行了向量化。而一个单词对应的位置向量也相应被定义为该单词相对于头尾实体的距离对应的两个位置向量的组合。

如图3.2所示，其假定单词的词向量维度 d^a 为 3，位置向量的维度 d^b 为 1，最终我们将每个单词的词向量与位置向量拼接起来作为神经网络的输入 $\mathbf{w}$，也就是 $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m]$，其中 $\mathbf{w}_i \in \mathbb{R}^d (d = d^a + d^b \times 2)$。上述词向量与位置向量是关系抽取中的基本设定，在后续的相关工作介绍中，我们将不再一一展开其细节。

编码层　在关系抽取任务中，主要的问题是抽取句子关系对应重要的信息可以出现在句子的任何区域。因此，我们希望利用句子所有局部特征来进行全局关系预测。在这里，NRE 模型使用卷积操作来抽取所有的局部特征。在卷积层中，我们采用一个长度为 l 的滑动窗口在所有句子上提取局部特征。在图3.2所展示的例子中，其滑动窗口的大小 l 为 3。然后，NRE 模型再采用一个最大池化操作，将所有通过卷积操作抽取到的局部特征合并成一个固定长度大小的特征向量，来表示输入的句子。

在这里，卷积操作被定义为一个输入向量序列 $\mathbf{w}$ 和卷积核矩阵 $\mathbf{W} \in \mathbb{R}^{d^c \times (l \times d)}$ 间进行的矩阵乘法操作，其中 d^c 是句子特征向量的维度大小。我们定义 $\mathbf{q}_i \in \mathbb{R}^{l \times d}$ 为第 i 个滑动窗口内的单词输入向量的拼接：

$$\mathbf{q}_i = \mathbf{w}_{i-l+1:i} \quad (1 \leq i \leq m + l - 1). \tag{3-1}$$

由于在滑动窗口从左至右滑动时，可能会超出句子的范围（因为句子的长度是不确定的），所以对于超出的范围，我们统一使用填充向量来进行空白位填充。卷积层的第 i 个卷积输出则为：

$$\mathbf{p}_i = [f(\mathbf{W}\mathbf{q} + \mathbf{b})]_i \tag{3-2}$$

其中 $\boldsymbol{b}$ 是一个偏置向量，$f(\cdot)$ 是激活函数，常用的如双曲正切函数。

句子的最终表示 $\mathbf{x} \in \mathbb{R}^{d^c}$ 的第 i 维通过如下最大池化操作得到：

$$[\mathbf{x}]_i = \max(\mathbf{p}_i), \tag{3-3}$$

此外，Zeng 等人[83] 提出了卷积神经网络的一种变体，采用了分段池化操作来进行关系抽取。卷积层输出结果被头实体和尾实体分成了三部分，最大池化操作

也相应地在三部分上分别进行，即：

$$\begin{aligned}[\boldsymbol{x}_1]_j &= \max_{1\le i\le i_1}[\boldsymbol{p}_i]_j \\ [\boldsymbol{x}_2]_j &= \max_{i_1+1\le i\le i_2}[\boldsymbol{p}_i]_j \\ [\boldsymbol{x}_3]_j &= \max_{i_2+1\le i\le m}[\boldsymbol{p}_i]_j\end{aligned} \tag{3-4}$$

这里 i_1 和 i_2 是头尾实体的句中位置，最后的句子向量为三部分池化结果的拼接，

$$\boldsymbol{x} = [\boldsymbol{x}_1; \boldsymbol{x}_2; \boldsymbol{x}_3] \tag{3-5}$$

由于卷积神经网络及其变种也被广泛用于关系抽取模型中进行编码，在后续的模型介绍中，我们对该部分内容也不过多赘述。

3.3.1.2　面向多实例的注意力机制

假设有一个集合 $s_{(h,t)}$ 包含了 n 个句子，每个句子都包含实体对 (h,t)，即 $s_{(h,t)} = \{x_1, x_2, \ldots, x_n\}$。在预测 h 与 t 之间是否存在关系 r 时，为了充分利用所有句子的信息，我们的模型将集合 $s_{(h,t)}$ 表示成一个统一的特征向量来进行预测。很显然，这个统一特征向量依赖于所有句子的表示 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$，并且每个句子的表示 $\boldsymbol{x}_i$ 都或多或少的含有一些信息有助于判定头尾实体 (h,t) 是否存在关系 r。

对此，一个很直接的想法便是通过句子向量 $\boldsymbol{x}_i$ 的加权平均来计算得到 $s_{(h,t)}$ 的统一表示向量，

$$\boldsymbol{s}_{(h,t)} = \sum_i \alpha_i \boldsymbol{x}_i, \tag{3-6}$$

其中 α_i 表示句子向量 $\boldsymbol{x}_i$ 的权重。在本章介绍的模型中，我们考虑了两种不同的定义 α_i 的方式，即：

平均值　我们假设句子集合 $s_{(h,t)}$ 中的所有句子对集合的表示具有相同的贡献。这意味着集合 $s_{(h,t)}$ 的表示向量是所有句子向量的算术平均值：

$$\boldsymbol{s}_{(h,t)} = \sum_i \frac{1}{n}\mathbf{x}_i, \tag{3-7}$$

这也是本文中的一个基线模型。

选择性注意力机制　然而，在实际情况中，远程监督假设标注的关系抽取数据会不可避免地会遇到错误标注的问题。因此，如果我们在表示句子集合的时候简单

地将每个句子当成同等重要的，那么含有错误标签的句子在训练和测试中会带来大量的噪音，极大地损害模型的效果。因此，我们引入一种语句级别选择性注意力机制来定义 α_i，此时 α_i 也就相应地被定义为：

$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}, \tag{3-8}$$

其中 e_i 是一个能量函数，通过该函数，我们可以刻画输入的语句 x_i 和想要预测的关系类型 r 在多大程度上是匹配的。e_i 越高表示语句 x_i 越能够表述关系 r 的语义。经过大量试验比较，NRE 模型选择了双线性函数作为计算 e_i 的函数：

$$e_i = \boldsymbol{x}_i^{\top} \boldsymbol{A} \boldsymbol{r} \tag{3-9}$$

考虑到远程监督会不可避免地带来错误标注。采用上述的注意力机制可以有效地减少含有噪音的句子所对应的权重值，同时兼顾综合利用所有句子信息的优势。

最后我们通过一个 softmax 层来计算条件概率 $p(r|s_{(h,t)}, \theta)$，即：

$$p(r|s_{(h,t)}, \theta) = \frac{\exp([o]_r)}{\sum_{k=1}^{|\mathcal{R}|} \exp([o]_k)}, \tag{3-10}$$

其中 $|\mathcal{R}|$ 是关系类型的总数量，$\boldsymbol{o}$ 是神经网络的最终输出向量，它表示了对所有关系类型的预测分数，具体定义如下所示：

$$\boldsymbol{o} = \boldsymbol{M} \boldsymbol{s}_{(h,t)} + \boldsymbol{d}, \tag{3-11}$$

其中 $\boldsymbol{d} \in \mathbb{R}^{|\mathcal{R}|}$ 是偏置向量，$\boldsymbol{M}$ 是所有关系类型的表示矩阵（即所有关系类型对应的特征向量所构成的矩阵）。

Zeng 等人[83] 提出的解决远程监督数据噪音问题的多实例学习方法遵循了如下假设：至少有一个包含实体对的句子会反映它们之间的关系，由此为依据，使用且仅使用每个集合中具有最高概率的句子进行训练和预测。因此，他们采用的多实例学习方法可以被视为选择注意力机制一种特殊情况，即，当我们的选择性注意力机制对概率最高的句子的权重设置为 1，其他句子权重设为 0 时，选择注意力机制即退化为多实例学习。

3.3.1.3　优化方法与实现细节

这里我们介绍我们模型的学习过程和优化细节。我们采用交叉熵函数来作为目标函数，即：

$$J(\theta) = \sum_{i=1}^{k} \log p(r_i | s_{(h,t)}, \theta), \tag{3-12}$$

其中 k 表示句子集合的数量，θ 表示我们模型中所有的可学习参数。在训练过程中，我们采用随机梯度下降（SGD）来最小化目标函数。具体训练过程中，我们每次从训练集中随机选择数据进行训练，并不断迭代，直到损失函数收敛。

在具体实现中，我们对于最后的输出层采用了 dropout 操作[85] 从而防止过拟合。dropout 操作可以定义为输入向量和一个遵从伯努利分布的各维度独立的随机向量 $\boldsymbol{h}$ 的逐维度乘积。其中伯努利分布的概率值为 p（即每一维度以 p 的概率取 1,1 – p 的概率取 0）。则公式 (3-11) 可以重新改写为为：

$$\boldsymbol{o} = \boldsymbol{M}(\boldsymbol{s}_{(h,t)} \circ \boldsymbol{h}) + \boldsymbol{d}. \tag{3-13}$$

在进行预测的时候，我们将模型输出集合表示向量长度上扩展到原本的 p 倍，即 $\hat{\mathbf{s}}_{(h,t)} = p\mathbf{s}_{(h,t)}$。经过长度变换的集合表示向量 $\hat{\mathbf{s}}_{(h,t)}$ 也被用于最终的预测关系类型中去。

3.3.2　实验分析

我们在 NYT 数据集上，对本章节介绍的注意力模型以及其他主流的抽取模型进行了实验与比较，下面我们将就实验结果展开具体讨论。

3.3.2.1　数据集和评测指标

关系抽取任务中，Riedel 等人[32] 发布的 NYT10 数据集被全世界研究者广泛应用，所以我们也在这个数据集上验证我们提出的基于选择注意力机制的 NRE 模型。该数据集是通过将 Freebase 知识图谱中的世界知识与纽约时报语料库（New York Times）中的语料进行对齐而生成的。他们使用 Stanford Named Entity Tagger[86] 来进行命名实体标注，并进一步将标注出来的命名实体与 Freebase 知识图谱中实体相链接。之后，他们将 Freebase 中的关系类型分为两部分：一部分用于训练（2005 年到 2006 年的语料库中的句子），一部分用于测试（2007 年后语料库中的句子）。最后，整个数据集合包含 53 种关系类型，包括一种特殊关系 NA 类型，表示头部和尾部实体之间没有关系。训练数据包含 522, 611 个句子，281, 270 个实体对和

18,252 个关系事实。测试集包含 172,448 个句子，96,678 个实体对和 1,950 个关系事实。

与之前的工作类似[31]，我们通过比较我们模型在测试集中挖掘出的世界知识与 Freebase 中的世界知识的重合度来评估我们的模型。而具体的模型性能则是通过精度-召回率曲线和最高置信度预测精度 (P@N) 来体现。

3.3.2.2　实验设置

遵循以前的研究工作，我们使用训练集上的三折交叉验证来确定我们的模型的超参数。我们使用网格搜索来确定最优超参数，并在 $\{0.1, 0.01, 0.001, 0.0001\}$ 中搜索学习率 λ，滑动窗口大小的搜索范围为 $l \in \{1, 2, 3, \cdots, 8\}$，语句表示向量的维度大小在 $\{50, 60, \cdots, 300\}$ 选择，以及在 $\{40, 160, 640, 1280\}$ 中选择训练 batch 大小。对于其他参数，因为它们对结果影响不大，所以我们按照 Zeng 等人[66] 中使用的参数值进行设置。对于训练，我们将所有训练数据的迭代次数设置为 25。在表格3.1中，我们展示了实验中使用的所有超参数。

表 3.1　超参数设置

超参数	数值
卷积窗口大小	3
句子表示大小	230
词向量维度	50
位置向量维度	5
训练批次大小	160
学习率	0.01

3.3.2.3　选择性注意力机制的有效性验证

为了证明语句级别选择性注意力机制的有效性，我们选择 Zeng 等人[66] 中提出的卷积神经网络模型（CNN）及其变种模型（PCNN）作为句子编码器。我们将两种不同类型的卷积神经网络与句子级别注意力机制（ATT），其基线版本（AVE）（该版本中，每个句子集合的向量表示为集合内部句子的平均向量）以及，Zeng 等人[66] 提出的多实例学习（ONE）方法的表现进行了比较。

从图3.3，我们可以得到如下观察结果：

（1）对于 CNN 和 PCNN 模型，加入 ONE 方法可以使得模型相比于原始的 CNN/PCNN 相比具有更好的性能。原因在于原始的基于远程监督得到的训练数据

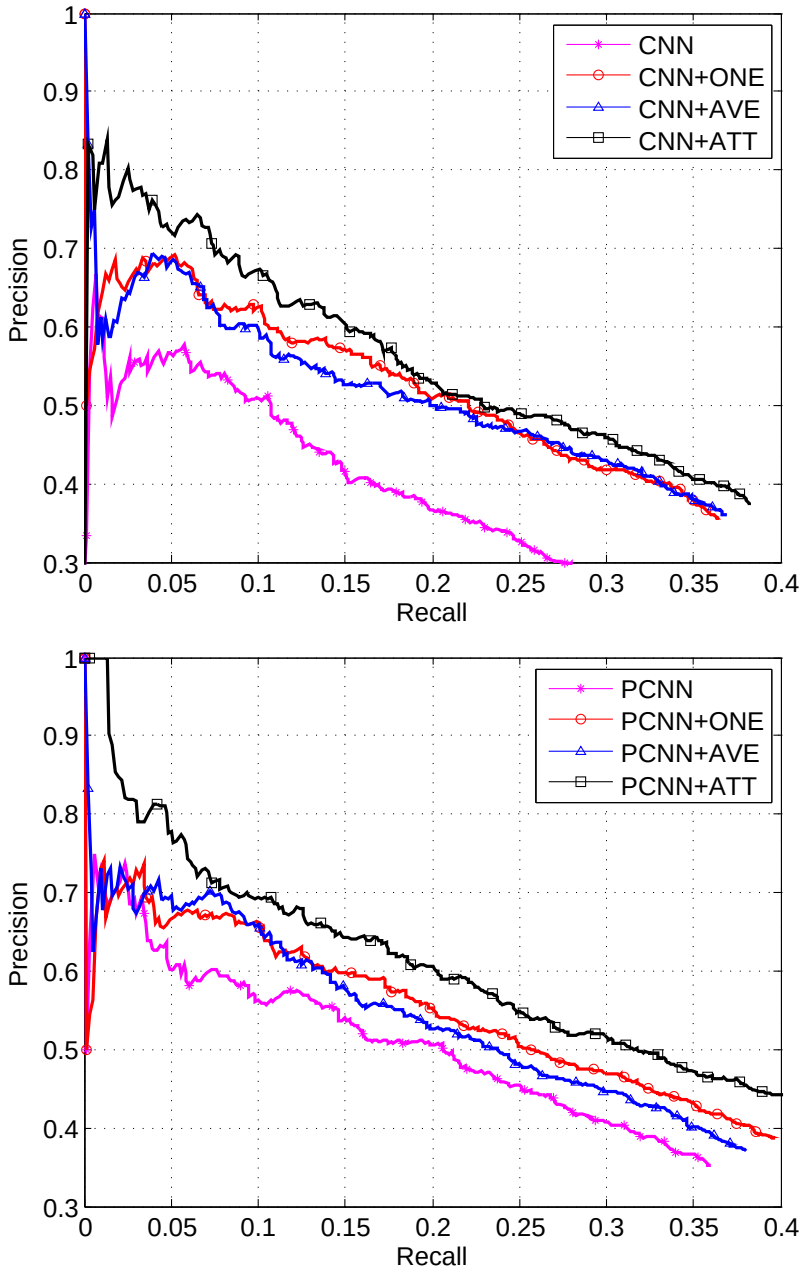

图 3.3　上方：CNN, CNN+ONE, CNN+AVE, CNN+ATT 模型的精度-召回率曲线。下方：PCNN, PCNN+ONE, PCNN+AVE, PCNN+ATT 模型的精度-召回率曲线

中包含大量的噪音，而噪音数据会损害关系抽取的性能。

（2）对于 CNN 和 PCNN 模型，加入 AVE 方法可以使得模型相比于原始的 CNN/PCNN 相比具有更好的关系抽取效果。这表明考虑更多的句子有利于关系抽取，因为噪音可以通过信息的互补来减少负面影响。

（3）对于 CNN 和 PCNN 模型，只使用 AVE 方法与只是用 ONE 方法相比具有相似的性能。这说明，尽管 AVE 方法引入了更多的句子信息，但由于它将每个句

子赋予同等的权重，它也会从错误标注的语句中得到负面的噪音信息，从而损害关系抽取的性能。

（4）对于 CNN 和 PCNN 模型，与包括 AVE 方法在内的其他方法相比，ATT 方法在整个召回范围内实现了最高的精度。它表明，我们所提出的选择性注意力机制是有帮助的。它可以有效地滤除对抽取关系没有意义的句子，有效解决基于远程监督的关系抽取中的错误标注问题。

3.3.2.4　语句数量的影响

在原始测试数据集中，有 74,857 个实体对仅对应于一个句子，在所有实体对上占的比例近 3/4。由于我们选择性注意力模型的优势在于处理包含多个句子的实体对的情况，所以我们比较了 CNN/PCNN+ONE，CNN/PCNN+AVE，以及采用了注意力机制的 CNN/PCNN+ATT 在具有多个句子的实体对集合上的表现。然后我们在三个测试设置下检查这三种方法：

- **One**：对于每个测试实体对，我们随机选择其对应的句子集合中的一句话，并将这句话用作关系预测。
- **Two**：对于每个测试实体对，我们随机选择其对应的句子集合中的两句话，并将这两句话用作关系预测。
- **All**：对于每个测试实体对，我们使用其对应的句子集合中的所有句子进行关系预测。

值得注意的是，我们在训练过程中，使用了所有句子。我们将会汇报 P@100, P@200, P@300 这些数值指标以及他们的平均值。

表 3.2　对于拥有不同句子数目的实体对的关系抽取方法的 P@N 指标

实验设置	One				Two				All			
P@N(%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
CNN+ONE	68.3	60.7	53.8	60.9	70.3	62.7	55.8	62.9	67.3	64.7	58.1	63.4
+AVE	75.2	67.2	58.8	67.1	68.3	63.2	60.5	64.0	64.4	60.2	60.1	60.4
+ATT	**76.2**	65.2	**60.8**	67.4	76.2	65.7	62.1	68.0	**76.2**	68.6	59.8	68.2
PCNN+ONE	73.3	64.8	56.8	65.0	70.3	67.2	63.1	66.9	72.3	69.7	64.1	68.7
+AVE	71.3	63.7	57.8	64.3	73.3	65.2	62.1	66.9	73.3	66.7	62.8	67.6
+ATT	73.3	**69.2**	**60.8**	**67.8**	**77.2**	**71.6**	**66.1**	**71.6**	**76.2**	**73.1**	**67.4**	**72.2**

表格3.2展示了所有模型在三种设置下的 P@N 指标。从这个表中，我们可以

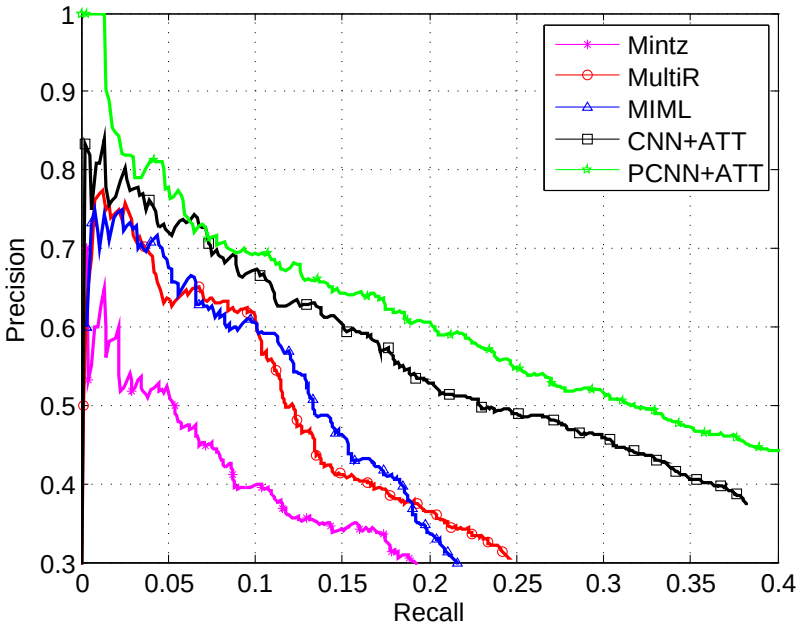

图 3.4　与基于人工特征工程的方法的性能比较

观察到：

（1）对于 CNN 和 PCNN，ATT 方法在所有测试设置中均达到最佳性能。它表明了句子级选择性注意力机制对于多实例学习的有效性。

（2）对于 CNN 和 PCNN，AVE 方法在 One 测试设置下，效果与 ATT 方法相当。然而，当每个实体对的测试句子数量增加时，AVE 方法的性能几乎没有改善。随着句数增加，它甚至在 P@100，P@200 中逐渐下降。原因在于，由于我们将每个句子同等地看待，句子中包含的不表达任何关系的噪音数据对于关系抽取的表现会产生负面影响。

（3）在 ONE 测试设置下，AVE 方法和 ATT 方法相比于 ONE 方法有 5 到 8 个百分点的改进。由于每个实体对在这个测试设置中只有一个句子，这些方法的唯一区别是来自训练。因此，它表明利用所有的句子会带来更多的信息，尽管它也可能带来一些额外的噪音。

（4）对于 CNN 和 PCNN，ATT 方法在 Two 和 All 测试设置下优于其他两个基线 5 到 9 个百分点。它表明，通过考虑更多有用的信息，ATT 方法排名较高的关系事实更可靠，有利于关系抽取。

3.3.2.5　与基于人工特征工程的方法的比较

为了验证我们所提出的方法，我们选择了以下三种基于人工特征的方法来进行性能比较：

- **Mintz**[31] 是一个传统的基于远程监督的模型。
- **MultiR**[33] 提出了一个概率图模型用于多实例学习，它的特点在于可以处理关系类型之间的重合。
- **MIML**[34] 同时考虑了多实例和多关系类型两种情况（即每个实体对可能有多个句子，也可能有多个关系类型）。

我们通过这些作者发布的代码，实现了这些模型。

图3.4中展示了每个方法的精度-召回率曲线。我们可以观察得到：

（1）在整个召回率范围内，CNN/PCNN+ATT 方法显著的优于所有基于人工特征的方法。当召回率大于 0.1 时，基于特征的方法的性能迅速下降。相比之下，在召回率达到约 0.3 之前，我们的模型都具有比较不错的准确率。这表明人工设计的特征不能简洁地表达句子的语义含义，而自然语言处理工具带来的错误必然会损害关系抽取的性能。相比之下，可以自主学习每个句子的向量表示的 CNN/PCNN+ATT 模型可以很好地表达每个句子的语义信息。

2）在整个召回率范围内，PCNN + ATT 与 CNN+ATT 相比表现要好得多。这意味着选择性注意力机制可以很好地考虑所有句子的全局信息，但无法使模型对于单个句子的理解和表示变好。因此，如果我们有更好的句子编码器，我们模型的性能可以进一步提高。

3.3.2.6　案例分析

表格3.3展示了测试集中两个选择性注意力机制的例子。对于每个关系，我们展示了其对应的拥有高注意力权值的句子和拥有低注意力权值的句子。而且我们将每个实体对都进行了加粗显示。

通过这个表格，我们可以发现：

（1）第一个例子是与关系 ‘‘员工’’ 相关的。拥有低注意力权值的句子并没有很好地表达两个实体间的关系，然而高注意力权值对应的句子可以很好地表达 ‘‘梅尔 · 卡尔马津’’ 是 ‘‘天狼星 XM 卫星广播公司’’ 的首席执行官。

（2）第二个例子是与关系 ‘‘出生地’’ 相关。拥有低注意力权值的句子事实上在说的是 ‘‘恩斯特 · 海弗里格’’ 在哪里去世，而不是像高注意力权值句子真的在表达的他出生在哪里。

3.4　基于多语言注意力机制的关系抽取

随着深度学习的发展，Zeng 等人[83] 引入了神经网络模型从样例中自动学习特征来抽取关系。为了解决远程监督数据中的错误标注问题，我们在上一节介绍

表 3.3　NYT 语料上选择性注意力机制的例子

关系	员工
低置信度句子	在霍华德 · 斯特恩正准备跟着他之前的老板**梅尔 · 卡尔马津**给**天狼星 XM 卫星广播公司**制作脱口秀节目的时候，霍兰德尔评价到……
高置信度句子	**天狼星 XM 卫星广播公司**的首席执行官**梅尔 · 卡尔马津**打了一个电话……
关系名	出生地
低置信度句子	瑞士男高音……**恩斯特 · 海弗里格**在一个周六逝世于瑞士的**达沃斯**……
高置信度句子	**恩斯特 · 海弗里格**在 1919 年 7 月 6 日生于**达沃斯**，并在神学院接受了教育……

的 NRE 工作[87] 进一步在神经关系抽取中采用了句子级别的选择注意力机制，获得了不错的效果。

然而，现有的大部分关系抽取系统仅仅关注于从单语言数据中抽取世界知识。事实上，人们使用各种各样的语言来描述世界知识。并且，由于人类经验总结与认知系统的相似性，不同语言之间也是共享着许多知识的。比如，尽管“纽约”和“美国”在英语中分别叫做 *New York* 和 *United States*, 中国人和美国人都认同一个事实：“纽约是美国的一个城市。”因此，充分考虑多语言数据中蕴含的丰富信息也许能够帮助我们建立一个更加有效、更加鲁棒的关系抽取系统。

最简单的利用多语言数据进行关系抽取的方法是为每一种语言构建一个单语关系抽取系统。但是这样难以充分利用到隐藏在各种语言数据中的多样信息。因此，使用多语言数据建立一个统一的系统进行联合关系抽取十分必要，原因如下：

- **一致性**：在实验过程中，我们对远程监督数据进行了随机抽样。我们发现超过一半的中英文句子包含超过 20 个单词，然而这些句子中只有少数几个单词是与关系事实相关的。以表格3.4为例。第一个中文句子有超过 20 个单词，其中只有‘‘纽约’’（New York）和‘‘是美国第一大城市’’（is the biggest city in the United States）是直接反映了关系事实城市的。因此，从这些复杂的句子中找到具体关系对应的表达模式以进行关系抽取并不容易。在各种语言中，同一个关系事实通常有比较相似的表达模式，并且这些表达模式在不同语言中本质上是一致的。我们认为不同语言对于关系事实表达模式上的一致性可以为我们进行关系抽取提供更多的信息。

- **互补性**。在我们随机抽样的远程监督数据中，我们还发现英文数据中 42.2% 的关系事实与中文数据中 41.6% 的关系事实是各自语言独有的。此外，对近一半的关系而言，能够表达这些关系事实的句子数量在不同语言中差异极大。因此，在关系抽取模型训练时，不同语言的文本可以相互补充，特别是对于部分关系，训练语料丰富的语言可以对训练语料匮乏的语言进行补充，从而可以提高模型的整体性能。

表 3.4　关于同一个关系事实（纽约，城市，美国）的中英文句子

关系	城市
英文	1.**New York is a city** in the northeastern **United States**.
中文	1. **纽约**位于美国纽约州东南部大西洋沿岸，**是美国第一大城市**及第一大港. (**New York** is in the United States New York and on the Atlantic coast of the southeast Atlantic, **is the largest city** and largest port **in the United States**) 2. **纽约是美国**人口最多**的城市**. (**New York is** the most populous **city in the United States**)

为了全面地考虑这些问题，本章节介绍一种基于多语言注意力机制的多语言关系抽取（MNRE）模型。在单语上，该模型采用了与以往注意力模型相似的单语注意力机制来筛选每种语言内部信息丰富的句子。在跨语言上，该模型能够充分考虑多语言环境下的信息一致性与互补性，并相应性地采用跨语言注意力机制以进一步利用全局的多语言句子来对关系进行分类。

3.4.1　算法模型

我们提出 MNRE 模型的主要想法是每个关系在不同的语言中有着基本一致的语言表达模式，从而 MNRE 模型可以利用语言间关系表达模式的一致性和互补性来获得更好的关系抽取结果。

形式化地，给定一个实体对，它们在 m 种不同语言中包含该实体对的句子被定义为 $T = \{S_1, S_2, \ldots, S_m\}$，其中 $S_j = \{x_j^1, x_j^2, \ldots, x_j^{n_j}\}$ 对应于第 j 种语言中的 n_j 个句子集合。我们的模型会对每一个关系 r 评估一个分数 $f(T, r)$，当 r 为实体对之间的关系时得分较高，其他时候则得分较低。为了得到该评分函数，MNRE 模型主要包含以下两块模块：

句子编码器。对于一个句子 x 和两个目标实体，我们采用卷积神经网络来将 x 中的句子中与关系相关的表达模式建模为一个表示向量 $\mathbf{x}$。这个句子编码器也可

以用 GRU[88] 或者 LSTM[89] 等循环神经网络结构来实现。从实验效果上看，在关系抽取任务中，卷积神经网络可以在计算效率和效果之间取得更好的平衡。因此，在本文中，我们主要采用卷积神经网络作为我们的句子编码器。

多语言注意力机制。将各种语言中的句子表示成低维向量表示后，我们采用单语言和跨语言的注意力机制来选择那些具有准确的关系表达模式的句子，这些句子往往标注准确且富有更多的信息量。MNRE 模型通过将包含实体对的各句子向量加权求和，并利用注意力机制提高重要句子的权重，从而获得全局表示来进行关系预测。之后我们将详细介绍这两部分模型。

3.4.1.1　句子编码器

句子编码器希望通过卷积神经网络将句子 x 转化为它的低维向量表示 $\mathbf{x}$。首先，模型将输入句子中的单词表示为低维表示向量。接下来，句子编码器使用卷积、最大池化和非线性变换层来进一步获得句子的分布式表示，也就是 $\mathbf{x}$。

输入表示　与 Zeng 等人[66] 的模型类似，我们将每一个输入的单词转换为两种向量的连接作为输入：(1) 词向量，用以刻画单词的句法与语义信息；以及 (2) 位置向量，用以表达该单词相对于两个目标实体的位置信息。通过这种输入方式，我们可以将输入句子表示为一个向量序列 $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$ 其中 $\mathbf{w}_i \in \mathbb{R}^d$，其中 $d = d^a + d^b \times 2$（d^a 和 d^b 分别是词向量和位置向量的维度）。

卷积神经网络　在对输入语句进行编码后，我们使用卷积层来提取局部特征，最大池化操作与非线性层来将所有局部特征合并成整个句子的表示。

首先，卷积层通过在句子上滑动一个长度为 l 的窗口并在每个滑动窗口内进行卷积来抽取局部特征。形式化地，第 i 个滑动窗口的输出通过

$$\mathbf{p}_i = \mathbf{W}\mathbf{w}_{i-l+1:i} + \mathbf{b} \tag{3-14}$$

计算得到，其中 $\mathbf{w}_{i-l+1:i}$ 表示在第 i 个窗口中的连续 l 个词的输入表示的拼接。$\mathbf{W} \in \mathbb{R}^{d^c \times (l \times d)}$ 是卷积矩阵，$\mathbf{b} \in \mathbb{R}^{d^c}$ 是偏置向量（d^c 是卷积层的输出嵌入的维度）。

在这之后，我们通过最大池化操作综合所有的局部特征，并应用双曲正切函数来获得输入句子的固定大小的句子向量。形式化地，输出向量 $\mathbf{x} \in \mathbb{R}^{d^c}$ 的第 i 个元素通过如下公式进行计算：

$$[\mathbf{x}]_j = \tanh\left(\max_i(\mathbf{p}_{ij})\right), \tag{3-15}$$

最终的向量 **x** 即为全句的表示。在这里，除最大池化操作之外，我们也可以采用在分段卷积神经网络模型[83] 中的分段最大池化来获取全句的表示。分段卷积神经网络模型是一个卷积神经网络模型的变体，理论上可以更好地从输入句子中获取到表达关系的语义特征。

3.4.1.2　多语言注意力机制

为了充分利用各种语言的句子信息，我们的模型采用了两种注意力机制以进行多语言关系抽取，包括：(1) 在一种语言中选择信息丰富句子的单语注意力机制；（2）考虑语言间模式一致性和互补性的跨语言注意力机制。

单语注意力机制　为了解决远程监督中的错误标注问题，我们按照 Lin 等人[87] 采用的句子级别选择注意力机制思想，为 MNRE 模型设置了单语注意力机制。直观地说，每种人类语言都有其自身的特点。因此，我们对于每一种不同的语言采用单独的单语注意力机制模块来动态降低每种语言中那些信息不丰富的句子对整体抽取效果的影响。

更具体地来说，对第 j 种语言以及对应的句子集合 S_j，为了关系预测，我们需要将所有句子向量加权求和到一个统一的表示向量 $\mathbf{S}_j$ 中。单语向量 $\mathbf{S}_j$ 就是由句子向量 $\mathbf{x}_j^i$ 的加权求和计算得到:

$$\mathbf{S}_j = \sum_i \alpha_j^i \mathbf{x}_j^i, \tag{3-16}$$

其中 α_j^i 是每一个句子向量 $\mathbf{x}_j^i$ 的单语注意力权重，定义为:

$$\alpha_j^i = \frac{\exp(e_j^i)}{\sum_k \exp(e_j^k)}, \tag{3-17}$$

其中 e_j^i 是句子 x_j^i 反映具体关系 r 的相关性评分。在实际操作中，有许多种方式可以获得 e_j^i，在这里我们简单地用内积来进行计算:

$$e_j^i = \mathbf{x}_j^i \cdot \mathbf{r}_j, \tag{3-18}$$

这里 $\mathbf{r}_j$ 是关系 r 在第 j 种语言中的注意力机制查询向量。

跨语言注意力机制　除了单语注意力机制外，我们的多语言神经网络关系抽取模型 MNRE 模型还提出了跨语言注意力机制，以便更好地利用多语言数据。跨语言注意力机制的关键思想是强调不同语言之间具有较强一致性的句子。在单语注意力机制的基础上，跨语言注意力机制可以进一步利用语言间关系模式一致性来有

效剔除与被预测关系相关性较低的句子，使得模型可以充分考虑那些信息丰富的句子。

跨语言注意力机制的工作方式类似于单语注意力机制。设 j 表示某一种语言，k 是另外一种语言（$k \neq j$）。形式化地，跨语言句子表示 $\mathbf{S}_{jk}$ 被定义为在第 j 种语言的句子 $\mathbf{x}_j^i$ 上的加权和：

$$\mathbf{S}_{jk} = \sum_i \alpha_{jk}^i \mathbf{x}_j^i, \tag{3-19}$$

其中 α_{jk}^i 是每一个语言向量 $\mathbf{x}_j^i$ 相对于第 k 种语言的跨语言注意力权重。α_{jk}^i 定义为：

$$\alpha_{jk}^i = \frac{\exp(e_{jk}^i)}{\sum_k \exp(e_{jk}^k)}, \tag{3-20}$$

其中 e_{jk}^i 表示第 j 种语言句子 x_j^i 与第 k 种语言的关系表达模式的一致性程度评分。与单语注意力机制类似，我们采用如下公式计算 e_{jk}^i：

$$e_{jk}^i = \mathbf{x}_j^i \cdot \mathbf{r}_k, \tag{3-21}$$

其中 $\mathbf{r}_k$ 是关系 r 相对于第 k 种语言的跨语言注意力机制查询向量。为了方便以及统一公式，在下文中我们将单语注意力向量 $\mathbf{S}_j$ 记作 $\mathbf{S}_{jj}$。

3.4.1.3　预测

对于每一个实体对与它们的对应 m 种语言的句子集合 T，我们可以从包含多语言注意力机制的神经网络中得到 $m \times m$ 个向量 $\{\mathbf{S}_{jk} | j, k \in \{1, \ldots, m\}\}$。其中，下标中 $j = k$ 的向量是通过单语言注意力机制获得的句子表示向量，下标中 $j \neq k$ 的是通过跨语言注意力机制获得的句子表示向量。

我们将所有向量 $\{\mathbf{S}_{jk}\}$ 统一进行考虑，并如下定义全局分数函数 $f(T, r)$：

$$f(T, r) = \sum_{j,k \in \{1,\ldots,m\}} \log p(r | \mathbf{S}_{jk}, \theta), \tag{3-22}$$

其中 $p(r|\mathbf{S}_{jk}, \theta)$ 是由句子向量 $\mathbf{S}_{jk}$ 预测关系 r 的条件概率，通过一个 softmax 层如下计算：

$$p(r | \mathbf{S}_{jk}, \theta) = \text{softmax}(\mathbf{M}\mathbf{S}_{jk} + \mathbf{d}), \tag{3-23}$$

其中 $\mathbf{d} \in \mathbb{R}^{|\mathcal{R}|}$ 是一个偏置向量，$|\mathcal{R}|$ 是关系种类的数目，$\mathbf{M} \in \mathbb{R}^{|\mathcal{R}| \times R^c}$ 是一个随机初始化的全局关系矩阵。

为了更好地考虑语言的特殊性，我们进一步引入 $\mathbf{R}_k$ 作为第 k 种语言的关系矩阵。这里我们简单地定义 $\mathbf{R}_k$ 为由等式 (3-21) 中 $\mathbf{r}_k$ 组成的矩阵。因此，公式 (3-23) 可被进一步改写为:

$$p(r|\mathbf{S}_{jk}, \theta) = \text{softmax}[(\mathbf{R}_k + \mathbf{M})\mathbf{S}_{jk} + \mathbf{d}] \tag{3-24}$$

其中 $\mathbf{M}$ 建模了预测关系的全局表达模式，而 $\mathbf{R}_k$ 建模了不同语言独有的表达模型。这里需要注意的是，在训练阶段 $\{\mathbf{S}_{jk}\}$ 通过公式 (3-16) 和 (3-19) 根据标注的关系构造。在预测阶段，由于不能提前得知关系，我们将为每一个可能的关系 r 都构建一个不同的向量 $\{\mathbf{S}_{jk}\}$ 来 $f(T, r)$ 以进行关系预测。

3.4.1.4　优化方法与实现细节

在这里我们介绍 MNRE 模型的学习与优化细节。我们定义目标函数如下:

$$J(\theta) = \sum_{i=1}^{s} f(T_i, r_i), \tag{3-25}$$

其中 s 表示所有实体对的数目，它们每一个在不同语言下都分别对应于一个句子集合，而 θ 表示我们的框架的所有参数。为了解决优化问题，我们采用小批次梯度下降 (SGD) 来最小化这个目标函数。

3.4.2　实验分析

3.4.2.1　数据集和评价指标

我们构建了一个新的多语言关系抽取数据集来评估我们的 MNRE 模型。在实验中，我们关注于在英汉两种语言中进行抽取关系。在此数据集中，中文实例通过对齐中文百度百科与 Wikidata 中的关系事实来生成，而英文实例通过对齐英文维基百科与 Wikidata 中的关系事实来生成。整个数据集合，共包含 176 种关系，其中包括一种特殊的关系 NA（表示实体之间没有关系），我们在表格3.5中列出数据集的统计信息。

3.4.2.2　实验设置

我们通过在验证集上进行网格搜索来确定我们的 MNRE 模型的超参数。在表格3.6中为我们实验中使用的最佳超参数。

表 3.5　数据集的统计信息

数据集		# 关系	# 句子数	# 事实数
英文	训练集	176	1,022,239	47,638
	验证集		80,191	2,192
	测试集		162,018	4,326
中文	训练集	176	940,595	42,536
	验证集		82,699	2,192
	测试集		167,224	4,326

表 3.6　超参数设置

超参数	数值
窗口大小 w	3
句子嵌入大小 d^c	230
词嵌入维度 d^a	50
位置嵌入维度 d^b	5
批次大小 B	160
学习率 λ	0.001
丢弃概率 p	0.5

3.4.2.3　考虑语言一致性的有效性

为了验证不同语言之间的关系表达模式具有一致性，且对多语言关系抽取具有帮助作用，我们在构建的数据集上进行了实验。我们将我们的模型与仅使用英语数据训练 (PCNN-En、CNN-En) 和仅使用中文数据训练的 (PCNN-Zh、CNN-Zh) 的模型进行了对比，同时我们也实现了几个简单版本的多语关系抽取模型，包括一个使用 PCNN-En 和 PCNN-Zh 联合预测的联合模型 (PCNN+joint) 以及一个用公共的关系表示矩阵训练的联合模型 (PCNN+share) 做了对比。我们在 CNN 上也进行了同样的实验（即 CNN+joint 与 CNN+share）。

从图3.5中，我们得到了以下观察结果:

(1) [P]CNN+joint 模型和 [P]CNN+share 模型相比于 [P]CNN-En 模型和 [P]CNN-Zh 模型取得了更好的表现。这表明从多种语言信息中共同得到验证的关系事实更加可靠，也意味着联合利用中英文句子的信息有利于更好地提取关系事实。

(2) CNN+share 模型与 CNN+joint 模型相比仅仅取得了差不多的效果，甚至在召回率在 0.1 到 0.2 之间时 CNN+share 模型表现更差。此外，从总体上看，PCNN+share 模型相比于 PCNN+joint 模型几乎在整个召回率范围中都表现更差。

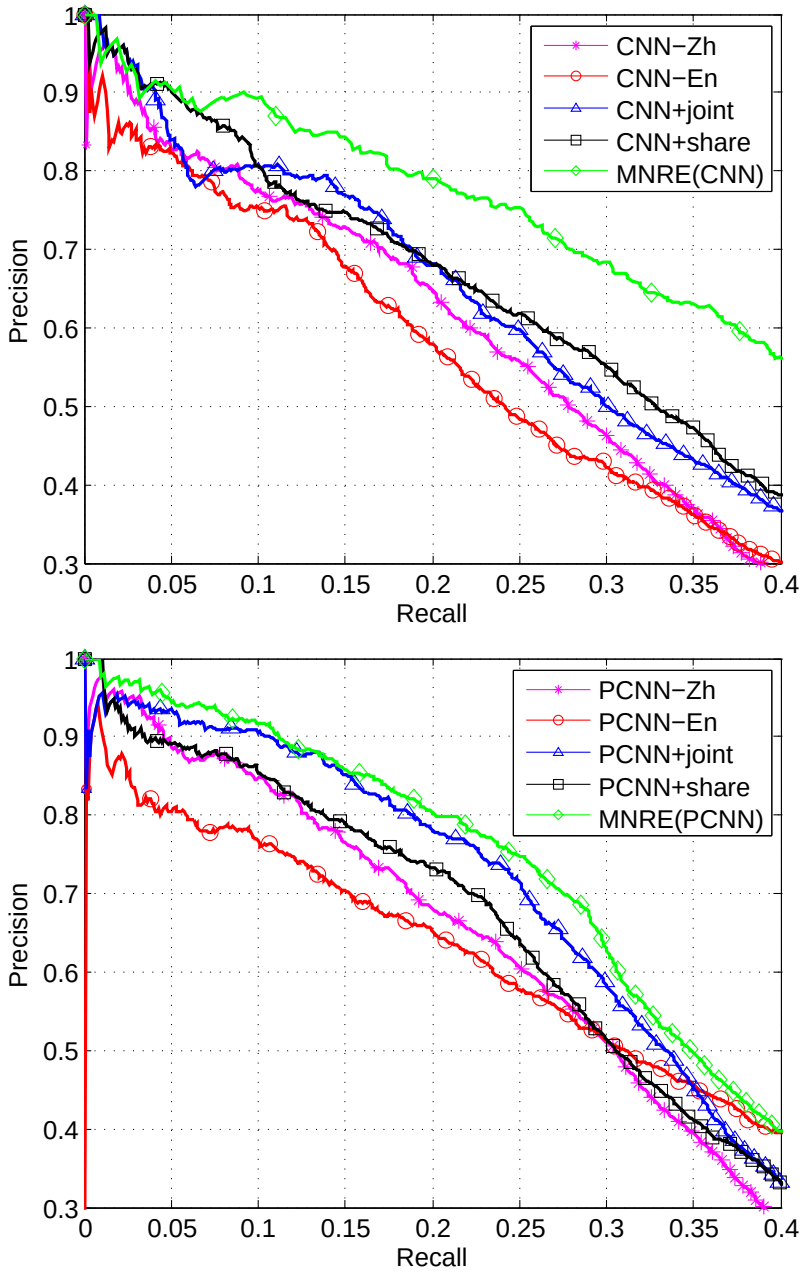

图 3.5　上方: CNN-En, CNN-Zh, CNN+joint, CNN+share 和 MNRE(CNN) 的精确率/召回率曲线；下方: PCNN-En, PCNN-Zh, PCNN+joint, PCNN+share, 和 MNRE(PCNN) 的精确率/召回率曲线

这表明，通过共享关系表示矩阵的简单组合方法并不能进一步刻画各种语言之间的隐式相关性，并进一步帮助多语言上的关系抽取。

(3) 相比于包括 [P]CNN+joint 和 [P]CNN+share 在内的其他基准方法，我们的 MNRE 模型在整个召回率范围中都取得了最高的精确率。通过对这些基线模型的参数进行网格搜索，我们可以观察到 [P]CNN+joint 和 [P]CNN+share 完全不能获得与 MNRE 模型接近的效果，甚至在增加了输出层的大小后也是如此。这表明简单

表 3.7　多语言注意力机制的一个例子。低、中、高表示注意力权重

CNN+Zh	CNN+En	MNRE	Sentence
—	中	低	1. **Barzun** is a commune in the Pyrénées-Atlantiques department in the Nouvelle-Aquitaine region of south-western **France**.
—	中	高	2. **Barzun** was born in Créteil , France
中	—	低	3. 作为从**法国**移民到美国来的顶尖知识分子，**巴尔赞**与莱昂内尔·特里林、德怀特·麦克唐纳等人一道，在冷战时期积极参与美国的公共知识生活 ...(As a top intellectual immigrating from **France** to the United States, **Barzun**, together with Lionel Trilling and Dwight Macdonald, actively participated in public knowledge life in the United States during the cold war . . .)
中	—	高	4. **巴尔赞**于 1907 年出生于法国一个知识分子家庭，1920 年赴美。(**Barzun** was born in a **French** intellectual family in 1907 and went to America in 1920.)

地增加模型大小并不能够从数据从学习到更多有用的信息。与之不同的是，我们提出的 MNRE 模型可以成功地通过考虑不同语言间的关系表达模式的一致性来提高多语言关系抽取的表现。

为了更加直观地展示 MNRE 模型的跨语言注意力机制如何刻画关系表达的一致性，我们进一步地在表格3.7中给出了一个关于跨语言注意力机制的例子。它展示了 MNRE 模型训练后关系 ‘‘出生地’’ 的句子集合中，中对英与英对中跨语言选择注意力权重最高和最低的四个句子。我们用粗体突出了实体对。为了做对比，我们同时展示它们在 CNN+Zh 和 CNN+En 模型中的选择注意力权重。从表格中我们可以看到，尽管这四个句子都表达出了巴尔赞出生在法国的关系事实。第一个和第三个句子显然包含了更多的可能迷惑关系抽取系统的噪声信息。由于采用跨语言注意力机制考虑了两种语言间句子的结构一致性，MNRE 模型可以凭借相比 CNN+Zh 模型和 CNN+En 模型更高的注意力权重更多地在预测关系时考虑第二个和第四个句子，并更清晰地表达关系 ‘‘出生地’’。

表 3.8　对一些特定的关系的详细结果 (precision@1)。#Sent-En 和 #Sent-Zh 分别表示被标注为对应关系的中英文句子个数

关系	#Sent-En	#Sent-Zh	CNN-En	CNN-Zh	MNRE-En	MNRE-Zh
包含	993	6984	17.95	69.87	73.72	75.00
上级	1949	210	43.04	0.00	41.77	50.63
父亲	1833	983	64.71	77.12	86.27	83.01
公民	25322	15805	95.22	93.23	98.41	98.21

3.4.2.4　考虑语言互补性的有效性

为了验证语言之间关系模式的互补性，我们比较了下列几种基线方法：CNN-En、CNN-Zh、PCNN-En、PCNN-Zh，以及多语言训练后仅仅使用单语向量来预测关系的模型 MNRE-En 和 MNRE-Zh。

图3.6展示了 CNN 和 PCNN 的四种模型的精确率/召回率曲线，从中我们可以看到:

(1) MNRE-En 和 MNRE-Zh 模型几乎在整个召回率范围内的效果都优于 CNN-En、CNN-Zh、PCNN-En 以及 PCNN-Zh 模型。这说明通过带有多语言注意力机制的联合训练，中英文关系抽取子系统都可以从另一个语言学习到有用的语义信息。

(2) 尽管 PCNN-En 模型的效果差于 PCNN-Zh 模型，CNN-En 模型的效果差于 CNN-Zh 模型，借助于带有多语言注意力机制的联合训练，MNRE-En 模型达到了和 MNRE-Zh 模型接近的效果。这说明通过我们提出的多语言注意力机制，中英文的关系抽取子系统都可以充分利用两种语言的文本，从而提升单语模型上的效果。

表格3.8展示了对于某些中英文训练样例数不平衡的关系的详细结果，从该表我们可以看出:

（1）对于关系 ‘‘包含’’ 来说，英文训练样例数仅仅是中文的 7 分之一。由于缺乏训练数据，CNN-En 与 CNN-Zh 相比效果差距非常大。然而，通过利用多语言注意力机制进行联合训练，MNRE-En 模型与 MNRE-Zh 模型的效果就较为相近，且均有提升。

（2）对于关系 ‘‘上级’’ 来说，中文训练样例数仅仅是英文 9 分之一，CNN-Zh 甚至预测不出任何正确的关系事实。原因可能是由于中文训练样例仅仅有 210 个，CNN-Zh 模型没有得到充分的训练。然而，通过利用多语言注意力机制进行联合训练，MNRE-En 模型和 MNRE-Zh 都能达到相对满意的效果。

（3）对于中英文句子数较为平衡的关系 ‘‘父亲’’ 和 ‘‘公民’’，我们的 MNRE 模型仍然可以在中英文关系抽取的表现上获得提升。

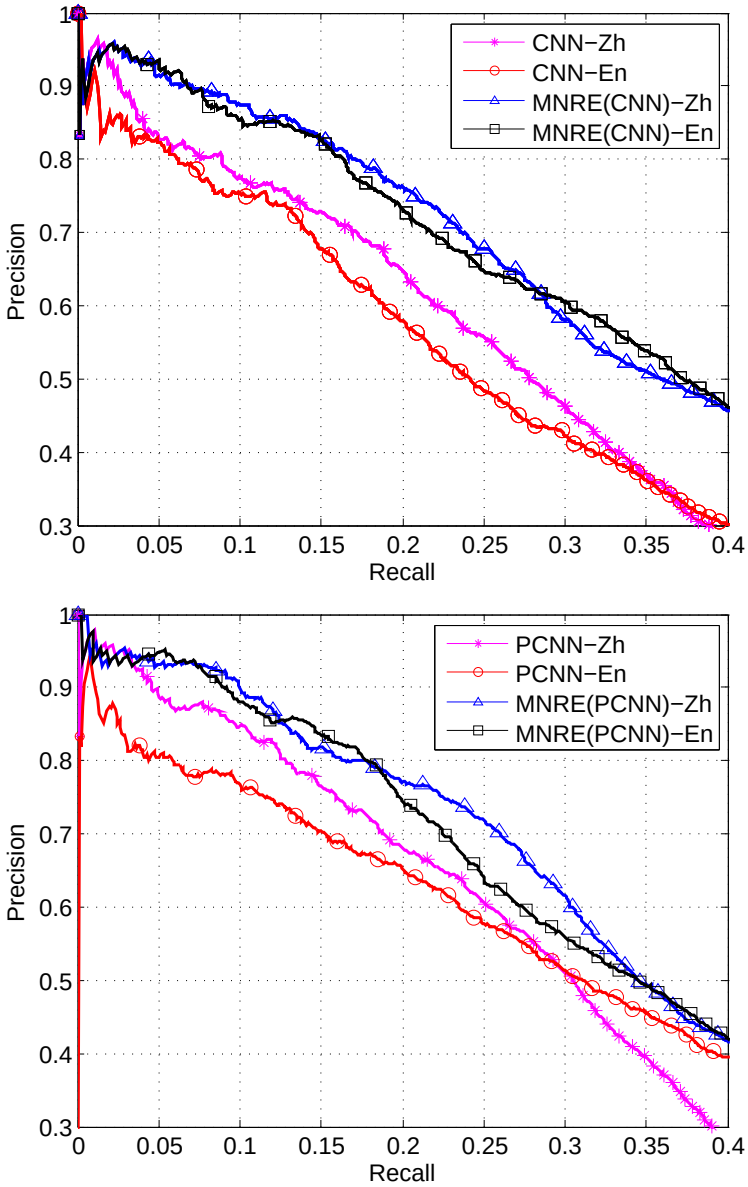

图 3.6　上方: CNN-En、CNN-Zh、MNRE(CNN)-En 和 MNRE(CNN)-Zh 的精确率/召回率曲线；下方: PCNN-En、PCNN-Zh、MNRE(PCNN)-En 和 MNRE(PCNN)-Zh 的精确率/召回率曲线

3.4.2.5　关系矩阵的有效性

对于关系预测，我们采用了两种关系矩阵，包括: 用一个统一的矩阵 **M** 来考虑关系的全局一致性，和用语言相关的矩阵 **R** 来考虑关系在每种语言上的特性。为了验证这两种关系矩阵的效果，我们将采用了两种关系矩阵的 MNRE 模型与仅采用 **M** 矩阵的 MNRE-M 模型和仅采用 **R** 矩阵的 MNRE-R 模型做了对比。

图3.7展示了每种方法的精确率/召回率曲线。从图中我们可以观察到:

(1) MNRE-M 的效果与 MNRE-R 和 MNRE 相比差距很大。这说明我们不能只

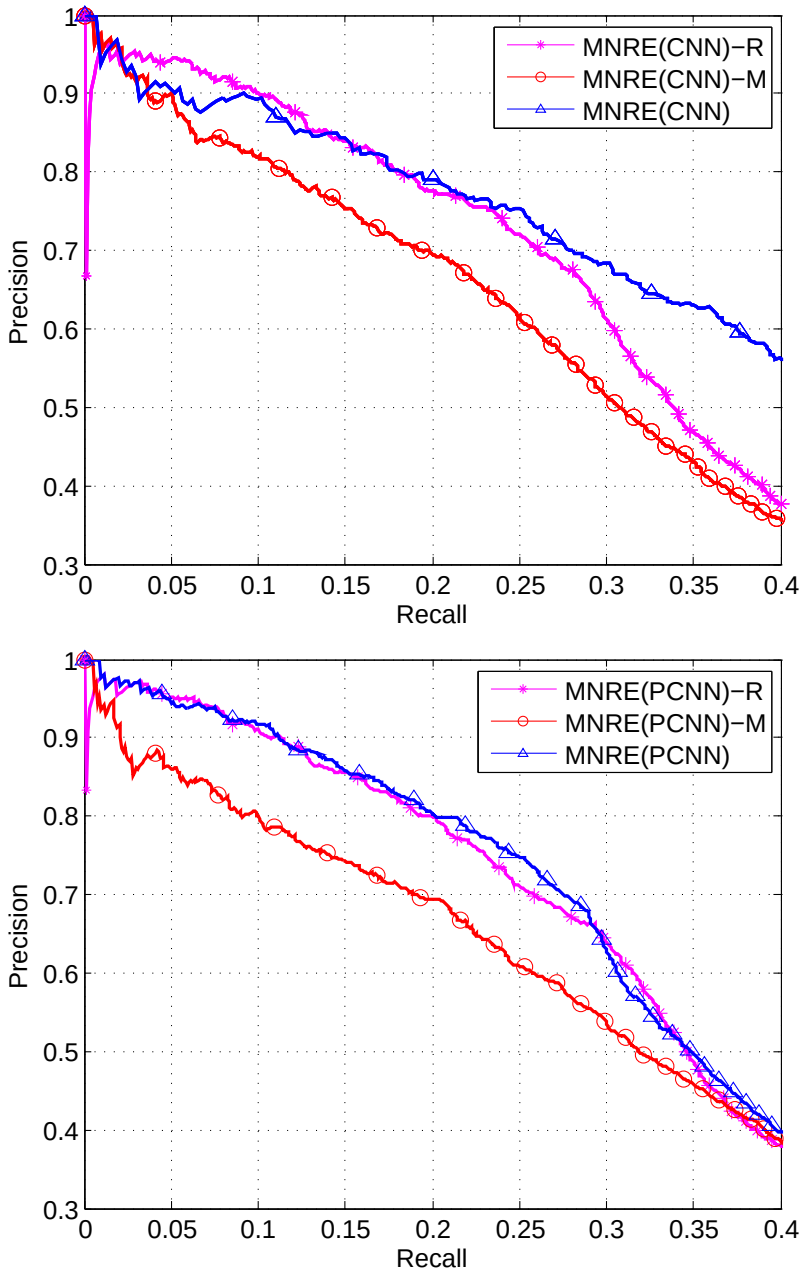

图 3.7　上方: MNRE(CNN)-M、MNRE(CNN)-R 和 MNRE 的精确率/召回率曲线；下方: MNRE(PCNN)-M、MNRE(PCNN)-R 和 MNRE(PCNN) 的精确率/召回率曲线

用全局关系矩阵进行关系预测。究其原因是因为每种语言都有其特有的表达关系模式的特点，不能很好地整合到一个单一的关系矩阵中。

(2) 当召回率较低时，MNRE-R 与 MNRE 有相似的表现。然而在召回率到达 0.25 时 MNRE-R 模型的效果急剧下降。这表明语言之间的关系模式也存在全局一致性，这一点在多语言关系抽取中也是不可忽视的。因此，我们应当在多语言关系抽取上联合使用 **M** 和 **R** 来刻画关系的表达模式。

3.5　小结

互联网文本是人类知识的主要载体。知识图谱构建需要利用文本关系抽取技术从文本中自动获取知识。这方面的著名项目包括华盛顿大学的 KnowItAll[90] 和 TextRunner[91]、卡内基梅隆大学的 Never-Ending Language Learner（NELL）[92] 以及微软亚洲研究院的 Probase[93]。然而，尽管现有的文本关系抽取技术已经成为补全知识图谱信息的一个重要手段，其仍然存在许多潜在的改进方向需要进行探索。在本章节中我们主要介绍了我们对文本关系抽取技术的改进包括（1）如何解决远程监督数据中的噪音问题；（2）如何充分利用好多语言数据信息进行联合关系抽取。

第 4 章　结构化知识的计算应用

4.1　引言

近些年来，伴随着 Freebase、DBpedia、YAGO、Wikidata 等大型知识图谱的构建，研究者们将其应用到很多知识相关的下游任务如网页搜索、问答系统等。知识图谱信息的引入有效地提高了知识相关下游任务的效果。然而，知识图谱原始的存储方式为离散符号组成的结构化信息，无法直接融入到目前大多数下游任务中效果最好的深度学习模型中。为了解决这个问题，研究者们提出了知识表示学习用于将知识图谱中原始的离散结构化信息转换为低维连续空间中的向量表示，大大降低了对知识图谱信息进行处理的复杂度，增强了知识图谱的可用性。因此，由知识表示学习处理得到的知识图谱实体和关系的表示可以被广泛地应用于各种需要引入知识信息的下游任务中。

本章将介绍我们在应用知识表示学习方面的相关工作。我们将详细介绍知识表示学习在细粒度实体分类和开放域问答任务中能够发挥的作用。

细粒度实体分类较一般的实体分类任务，预设了更加细致多样的实体类型，对实体分类模型的分辨能力具有更高的要求。传统的实体分类模型仅使用了实体在文本中的上下文信息，却忽略了现有知识图谱这一包含丰富信息的资源。本章将会展示如何通过引入基于知识表示学习的注意力机制，使知识图谱中的实体以及实体之间的关系信息能够在细粒度实体分类任务中发挥出重要的作用。

开放域问答旨在从大规模无标注语料找到用户问题的答案。开放域问答任务中问题和文档往往会涉及到许多实体以及实体之间的关系。传统的开放域问答系统仅仅从语义层面对问题和文档进行理解，缺乏了必要的背景知识。本章将会展示我们如何通过引入知识表示学习得到的知识图谱表示以及与关系抽取任务的多任务学习，使得开放域问答系统可以更好地在知识层面对问题和文档进行理解。

在本章节，在介绍我们的工作前，我们将首先对知识表示学习最典型的几项应用进行简要的介绍。

4.2　相关工作

知识表示学习最典型的应用就是知识图谱链接预测。现有的知识表示学习模型已经在链接预测任务中验证了其有效性。除了知识图谱链接预测以外，知识表示学习也被作为外部信息广泛应用于各项不同的自然语言处理任务中，并取得了

显著的提升效果。在本节中，我们将介绍几个知识表示学习的典型应用：语言模型、自动问答、信息检索、推荐系统、等等。

4.2.1　语言模型

语言模型旨在学习给定单词序列（一般为句子，也可以是段落、文章）的概率分布，是一项非常重要经典的自然语言处理任务。近些年，神经网络语言模型如循环神经网络语言模型已经在很多需要语言模型的任务中取得了不错的效果。然而，现有的神经网络语言模型往往面临可解析性差和缺乏背景知识的问题。其原因在于简单地对词共现进行考虑无法解决文本生成过程中需要涉及到不同类型知识的问题，这导致对于低频单词特别是低频实体的生成效果特别差。

为了解决这个问题，Ahn 等人[94] 提出了基于知识的神经网络语言模型 NKLM。NKLM 模型可以在循环神经网络生成自然语言序列的时候考虑来自知识图谱的背景知识。NKLM 模型采用两种不同的模型来生成下一个单词：（1）一种是根据循环神经网络计算的分数从 ‘‘单词词表’’ 里选择下一个要生成的单词；（2）一种是根据外部知识图谱从 ‘‘知识词表’’ 里选择下一个要生成的单词。

NKLM 模型探究了如何在神经网络模型中同时结合以往循环神经网络语言模型和外部知识图谱的符号信息。然而，由于 NKLM 模型需要根据上下文来确定当前位置的主题，然后再进一步确定需要使用的外部知识，这使得其无法在很多主题无关的文本中进一步进行扩展。尽管如此，我们还是坚信将外部知识的信息在语言模型的建模中进行考虑是非常有必要的。

4.2.2　自动问答

自动问答旨在自动回答用户给定的问题。解决自动问答需要同时解决问题的理解以及答案的推理。而自动问答这两个关键的步骤都需要考虑外部知识才能做好。因此，自动问答也是知识表示学习的一个非常直接的应用场景。传统的自动问答系统一般把知识图谱作为内部数据库进行使用。这些自动问答系统首先把用户的问题转化为一般查询语句，然后在知识图谱中查找正确的答案。然而，这种做法忽略了知识图谱中实体和关系相互之间的联系。最近，随着深度学习的发展，人们开始探索如何用神经网络模型来帮助理解问题和回答问题。

Serban 等人[95] 结合了知识表示学习模型和神经网络模型，根据知识图谱进行事实性问题的生成。除此之外，Yin 等人[96] 将知识图谱中的关系事实信息与神经网络产生式问答模型相结合，用于回答事实性的问题。更进一步地，He 等人[97] 提出了一个端到端的神经网络生成模型，同时结合复制和检索机制来利用知识表示

学习技术帮助生成更加自然的问答回复。

4.2.3　信息检索

信息检索旨在对用户的查询返回与之相关的文档。和自动问答类似，如何正确地理解用户的意图对于信息检索非常重要。在信息检索中，用户的查询往往会涉及到很多实体信息或者背景信息，要精确地理解用户的意图就需要理解这些实体信息和背景信息。因此，知识图谱的信息对做好信息检索非常重要。传统的信息检索系统一般把用户的查询和要检索的文档当做文本序列，然后通过人工构造的特征（如词袋特征）来计算它们之间的相似度。然而，现有的系统仅仅通过简单地字符串匹配还远远无法真正地理解用户的意图。

近些年来，人们开始研究如何利用知识表示学习技术来增强现有信息检索系统的效果。其中，最为简单有效的方法就是通过知识表示学习学习到的实体表示来增强原本的词表示。Hasibi 等人[98] 提出了一个基于实体的检索模型，用于帮助信息检索中对于用户查询的理解。该模型可以和以往的基于词的检索模型相结合，进一步提升检索效果。与此同时，他们还提出可以在用于文档排序的深度神经网络中利用知识表示学习融入知识图谱的信息；Xiong 等人[99] 也提出了一个基于实体的模型用于对信息检索系统中用户的查询和相关文档进行表示。此外，Nguyen 等人[100] 还提出可以在用于文档排序的深度神经网络中利用知识表示学习融入知识图谱的信息；Xiong 等人[101] 还提出利用知识表示学习将信息检索系统中的用户查询和相关文档表示到实体空间中，然后计算它们在知识图谱中的相似度。

4.2.4　推荐系统

伴随着海量互联网信息的出现，推荐系统已经成为很多互联网应用中的重要组成部分。推荐系统旨在预测用户对于给定物品的兴趣度。由于知识图谱可以提供很多用户和物品的信息，推荐系统一般将知识图谱作为一个外部资源，用于帮助在推荐过程中对于用户和物品进行精确的理解。

Cheekula 等人[102] 探索了如何在推荐系统利用 DBpedia 中的目录结构信息来刻画物品之间的联系，然后利用一个传播算法来为用户推荐可能感兴趣的物品。除此之外，Passant[103] 也尝试利用知识图谱中艺术实体之间的语义关联来构建一个音乐推荐系统。然而，这些系统仍然停留在简单利用知识图谱的结构的程度。最近，Zhang 等人[104] 提出了在推荐系统和知识图谱的协同过滤过程中同时学习不同实体的表示。

4.2.4.1　链接预测

通常情况下，知识图谱中显式记录的关系事实信息并不完整，有大量客观存在的关系事实信息并没有以关系三元组的形式收录进知识图谱中。这一问题主要是两部分原因造成的：

1. 现有的基于文本的知识图谱构建方法尚不成熟，关系级别的事实抽取任务召回率还很低。准确判断相关文本资源稀缺的实体对之间的关系对于现有的知识图谱构建方法而言仍是一项挑战。
2. 包含完整关系信息的知识图谱规模将会非常庞大。考虑下面的简单情形：假设一知识图谱中有两种关系 ‘‘毕业于’’ 和 ‘‘校友’’，对于一所学校 A，如果该知识图谱中包含 N 个人物实体与其存在 ‘‘毕业于’’ 的关系，则该知识图谱中蕴含的完整的与 ‘‘校友’’ 相关的关系三元组数量将达到 N^2 级别。对于包含数十万实体的大规模知识图谱而言，这种数量的关系数据是难以直接存储的。

不过，正如我们在第二点中指出的那样，实体之间的关系事实并不是相互独立的。我们可以通过推理的方式，从知识图谱中已知的关系推断出新的关系。换言之，我们可以从知识图谱中现有的关系得到实体的相对语义，再根据实体语义信息推测新关系。这就是利用知识表示学习进行知识图谱链接预测的基本思想。

由于知识图谱链接预测的对象为关系三元组，而表示学习的基本训练数据也为关系三元组，知识图谱补全在最近的表示学习研究中被广泛用作测试任务，几乎已成为事实标准。在本书第2章介绍的表示学习方法的实验部分，我们已经展示了这些表示学习方法如何被应用于知识图谱链接预测，因此我们在这里就不再赘述。

当然，在上面提到的这些任务以外，知识表示学习技术还被用于很多其他的自然语言处理任务中，用以帮助对其中的文本中涉及到知识图谱信息的理解，如对话系统[105-106]，实体消歧[107-108]，知识图谱对齐[109-110]，依存分析[111]，等等。此外，知识表示学习技术还被用于很多其他领域中，如视觉关系抽取[112-113]，社交网络关系抽取[114]，等等。然而，目前对于知识表示学习技术的应用探索还非常初步。如何在自然语言处理任务中灵活有效地结合知识表示学习技术仍需进一步探索。

4.3　基于知识的实体分类

实体分类旨在判断出一段文本中命名实体的语义类型。该任务对于许多自然语言处理的后续任务都非常重要，如命名实体识别[115]，关系抽取[116]，问答系统[117]等，可以帮助这些后续任务理解文本中涉及的相关实体。目前，神经网络实体分

类模型得益于其强大的文本建模能力，已经在实体分类任务中取得了最好的效果。然而，神经网络实体分类模型仍然存在着以下问题：

- **实体和单词表示的割离**：现有的神经网络实体分类模型在对命名实体和其上下文进行特征表示提取时是相互独立的。而事实上，不同上下文的单词的重要性在对不同实体的类别进行判断时是有很大区别的。例如，在句子 ‘‘比尔 · 盖茨和保罗 · 艾伦一起创立了世界上最大的互联网公司之一的微软公司。’’ 中，当我们在考虑命名实体 ‘‘微软公司’’ 的语义类型的时候，单词 ‘‘公司’’ 非常重要，而当我们在考虑命名实体 ‘‘比尔 · 盖茨’’ 的语义类型的时候，单词 ‘‘创立’’ 反而更加重要。因此，我们需要在不同的实体的语义类型的时候动态地调整不同单词的重要性。
- **没有考虑知识图谱信息**：知识图谱中的关系事实可以为判断一个命名实体的语义类型提供很多额外的补充信息。文本中命名实体之间的关系可以有效地帮助我们判断命名实体的语义类型。例如，当一句话中出现了命名实体 ‘‘美国’’ 和 ‘‘加拿大’’ 的时候，如果我们知道关系事实 (美国，接壤，加拿大)，我们可以根据关系事实两端实体的特性来判断 ‘‘美国’’ 和 ‘‘加拿大’’ 可能都是国家。然而，现有的神经网络实体分类模型并没有在预测命名实体语义类型的时候考虑知识图谱的信息。

为了解决神经网络实体分类模型的上述问题，我们提出了基于知识注意力机制的神经细粒度实体分类模型 KNET（Knowledge-Attention Neural Fine-grained Entity Typing）[①]。如图4.1所示，KNET 模型包括两个部分。KNET 模型首先通过一个神经网络来产生文本单词和命名实体的上下文表示。然后 KNET 模型利用知识注意力机制来增强命名实体的上下文表示。通过在上下文表示中引入受知识图谱实体表示指导的注意力机制，KNET 模型在实体分类任务中借用了知识图谱这一外部资源。

4.3.1　算法模型

对于给定的句子和该句子中的实体，我们将句子表示为单词序列 $s = \{\cdots, l_2, l_1, m_1, m_2, \cdots, r_1, r_2, \cdots\}$，其中 m_i 为组成命名实体的词，l_i 为出现在实体前的词，而 r_i 为出现在实体后的词。对于每个命名实体，KNET 模型首先生成一个特征向量 $\boldsymbol{x}$，然后根据特征向量来推断实体类型的概率分布。

KNET 模型由两部分组成：(1) 句子编码器，它将句子 s 编码成句子向量 $\boldsymbol{x}$；(2) 类型预测器，它通过 $\boldsymbol{x}$ 计算出命名实体类型分布向量 $\boldsymbol{y}$，从而来推断命名实体的

① 该工作为与本科生辛极的合作研究工作

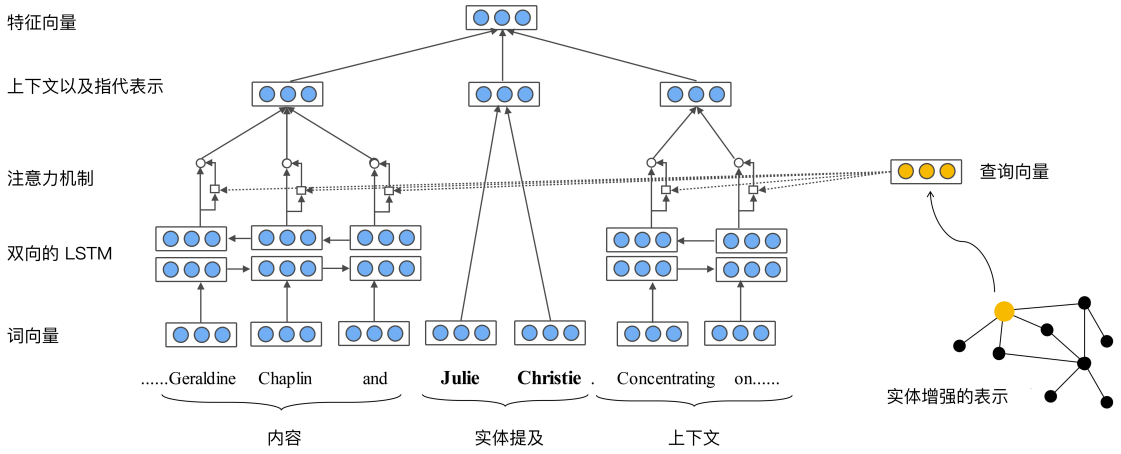

图 4.1　基于知识注意力机制的神经细粒度实体分类的模型框架

语义类型。

4.3.1.1　句子编码器

KNET 模型首先在词向量的基础上得到命名实体与上下文单词的表示。由于命名实体包含的词的数量通常较少，如果采用复杂的模型（如 CNN 或 RNN）倾向于过拟合，KNET 模型采用了一种简单高效的方法计算命名实体的表示，即直接将命名实体中 n_m 个词的词向量进行平均求和：

$$\boldsymbol{m} = \frac{1}{n_m}\sum_{i=1}^{n_m}\boldsymbol{m}_i. \tag{4-1}$$

至于上下文文本的表示，我们希望模型能够在其中体现出不同词的重要程度，故在 KNET 模型设计了注意力机制，为上下文中不同的词计算不同的注意力分值。具体地，上下文 $\{..., l_3, l_2, l_1\}$ 和 $\{r_1, r_2, r_3, ...\}$ 的词向量被分别输入到 LSTM 网络中，上下文表示 $\boldsymbol{c}$ 则是 LSTM 网络的输出 $\overrightarrow{\boldsymbol{h}_i^l}, \overleftarrow{\boldsymbol{h}_i^l}, \overrightarrow{\boldsymbol{h}_i^r}, \overleftarrow{\boldsymbol{h}_i^r}$ 的加权总和：

$$\boldsymbol{c} = \frac{\sum_{i=1}^{L}\left(a_i^l\begin{bmatrix}\overrightarrow{\boldsymbol{h}_i^l}\\ \overleftarrow{\boldsymbol{h}_i^l}\end{bmatrix} + a_i^r\begin{bmatrix}\overleftarrow{\boldsymbol{h}_i^r}\\ \overrightarrow{\boldsymbol{h}_i^r}\end{bmatrix}\right)}{\sum_{i=1}^{L}a_i^l + a_i^r}, \tag{4-2}$$

其中 a_i^l, a_i^r 分别为 l_i, r_i 的注意力分值，而 L 为模型的一个超参数，表示考虑的上下文的窗口大小。

KNET 模型将由上述过程得到的命名实体和上下文的表示进行拼接，得到输

入样例的特征表示向量：

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{c} \end{bmatrix}. \tag{4-3}$$

4.3.1.2　类别预测器

KNET 模型将句子表示向量 $\boldsymbol{x}$ 通过两层多层全连接网络（MLP）计算得到实体的类型向量 $\boldsymbol{y}$。$\boldsymbol{y}$ 上的每一项都表示命名实体在该类型上的预测概率:

$$\boldsymbol{y} = \sigma\left(\boldsymbol{W}_{y1} \tanh\left(\boldsymbol{W}_{y2}\boldsymbol{x}\right)\right) \tag{4-4}$$

$$\boldsymbol{y}^{(i)} = p(t^{(i)}|s,\theta) \tag{4-5}$$

其中 θ 表示模型的所有参数, $\sigma(\cdot)$ 表示 `sigmoid` 函数，$\boldsymbol{W}_{y1}$ 和 $\boldsymbol{W}_{y2}$ 是全连接网络的参数矩阵，$t^{(i)}$ 表示该实体预测为第 i 个类型。如果一个类型的预测概率大于 0.5，则该类型为最终预测结果；如果所有类型中没有一个大于 0.5，则具有最大预测概率的类型被视为最终的预测结果。

KNET 模型的训练目标函数定义为所有训练数据的交叉熵之和：

$$L = -\sum_{i,j} \boldsymbol{y}_i^{*(j)} \log \boldsymbol{y}_i^{(j)} + (1 - \boldsymbol{y}_i^{*(j)}) \log(1 - \boldsymbol{y}_i^{(j)}), \tag{4-6}$$

其中 $\boldsymbol{y}^*$ 表示命名实体的正确类型，$\boldsymbol{y}_i^{(j)}$ 是向量 $\boldsymbol{y}_i$ 的第 j 个分量，我们通过最大化损失函数 L 来学习模型的所有参数 θ。

4.3.1.3　注意力机制

注意力机制在 KNE 模型中起着重要的作用。在本节中，我们介绍如何在上下文表示中计算公式4-2中的注意力权重 a_i^l、a_i^r。我们期望，注意力机制能够考虑命名实体与上下文的关联、命名实体与知识图谱的关联。在这里，我们探讨了三种形式的注意力机制:

(1) **语义注意力机制**。将上下文的表示本身作为注意力的查询向量来计算注意力权重，这是由 Shimaoka 等人[118] 提出的。

(2) **实体注意力机制**。采用命名实体表示向量 $\boldsymbol{m}$ 作为注意力的查询向量，它将捕捉命名实体和上下文信息之间的语义关联。

(3) **知识注意力机制**。将从外部知识图谱中获取实体表示特征作为注意力的查询向量，它将捕捉实体与上下文、实体与知识图谱之间的语义关联。

语义注意力机制（Semantic Attention，SA）　Shimaoka 等人[118] 使用一个多层全连接网络（MLP）来计算语义注意力：

$$a_i^{\mathrm{SA}} = \sigma\big(\boldsymbol{W}_{S1} \tanh\left(\boldsymbol{W}_{S2} \begin{bmatrix} \overrightarrow{\boldsymbol{h}_i} \\ \overleftarrow{\boldsymbol{h}_i} \end{bmatrix}\right)\big), \tag{4-7}$$

其中 $\boldsymbol{W}_{S1}$ 和 $\boldsymbol{W}_{S2}$ 是多层全连接网络的参数矩阵。在后文中，我们将省略掉上标 l 和 r，因为它们对于注意力权重的计算方式是完全一样的。在语义注意力机制中，所有实体实际上共享用于计算注意力权重的的参数，也就是说针对上下文词汇计算的注意力数值与实体是无关的。因此，语义注意力机制很难将提高与相应实体高度相关的上下文词语的重要性。

实体注意力机制（Mention Attention，MA）　为了考虑实体与上下文的相关性，在计算注意力权重时，我们可以很自然地想到用命名实体的表示向量来当做注意力机制中查询向量。具体地，给定公式4-1中提到的命名实体的表示向量 $\boldsymbol{m}$，我们用如下的方式计算注意力权重：

$$a_i^{\mathrm{MA}} = f\left(\boldsymbol{m}\boldsymbol{W}_{\mathrm{MA}} \begin{bmatrix} \overrightarrow{\boldsymbol{h}_i} \\ \overleftarrow{\boldsymbol{h}_i} \end{bmatrix}\right), \tag{4-8}$$

其中 $\boldsymbol{W}_{\mathrm{MA}}$ 是一个双线性参数矩阵。$f()$ 是一个非线性函数，这里我们采用最简单的二次函数 $f(x) = x^2$，它是正定的，且易于求导（同样的设定也适用于方程4-9)。

知识注意机制（Knowledge Attention，KA）　知识图谱提供了实体之间大量的关系事实，这些信息可以有效地帮助实体分类。我们使用最经典的知识表示学习方法 TransE 模型[119] 来对知识图谱中实体和关系进行编码。在训练过程中，我们已知命名实体对应的知识图谱中的实体 e。因此，与公式4-8相似，我们可以直接计算注意力权重如下：

$$a_i^{\mathrm{KA}} = f\left(\boldsymbol{e}\boldsymbol{W}_{\mathrm{KA}} \begin{bmatrix} \overrightarrow{\boldsymbol{h}_i} \\ \overleftarrow{\boldsymbol{h}_i} \end{bmatrix}\right), \tag{4-9}$$

其中 $\boldsymbol{e}$ 是命名实体 m 对应知识图谱实体通过知识表示学习获得的表示向量，而 $\boldsymbol{W}_{\mathrm{KA}}$ 是双线性参数矩阵。

与训练时不同，在测试过程采用知识注意力机制存在一个问题：我们并不知道在知识图谱中哪个实体与我们在句子中的命名实体相对应，甚至有可能句子中的命名实体在知识图谱中没有对应实体。一个简单的解决方式是使用实体链接算

法先将句子中的命名实体链接到知识图谱中的对应实体。但是，实体链接本身就是一个难度非常大的任务，不可避免地会引入额外的错误。此外，实体链接也无法解决解决存在知识图谱以外的实体问题。

为了解决这个问题，我们提出利用文本信息来重构实体表示向量。具体地说，对于一个实体 e 及其上下文句子 s，我们使用 LSTM 网络将其上下文文本分别编码为 $\boldsymbol{c}_l$ 与 $\boldsymbol{c}_r$，并进一步学习基于文本的表示 $\boldsymbol{e}$：

$$\hat{\boldsymbol{e}} = \tanh\left(\boldsymbol{W}\begin{bmatrix}\boldsymbol{m}\\ \boldsymbol{c}_l\\ \boldsymbol{c}_r\end{bmatrix}\right), \tag{4-10}$$

其中 $\boldsymbol{W}$ 是参数矩阵，$\boldsymbol{m}$ 是在公式 (4-1) 中提到的命名实体表示向量。注意，这里使用的 LSTM 网络的参数不同于公式 (4-2) 中的 LSTM 网络。为了将文本重构的实体表示向量与和基于知识图谱的实体表示向量联系起来，在训练过程中，我们通过在公式 (4-6) 中的目标函数 J 中添加额外的一项 J_{KB}，来学习 $\hat{\boldsymbol{e}}$。

$$J_{\mathrm{KB}} = -\sum ||\boldsymbol{e} - \hat{\boldsymbol{e}}||^2 \tag{4-11}$$

其中求和是针对在训练集中的所有实体而进行的。这样，在测试过程中，我们就可以直接使用公式 (4-10) 来获得近似的实体表示向量，并使用公式 (4-9) 来计算知识注意力权重。

进一步地，如果我们可以通过使用命名实体涉及的单词来缩小候选实体类型的范围，我们就可以获得更准确的信息。这些信息可以作为对方程4-10利用文本重构出的实体表示向量的补充。因此，我们首先对实体进行消歧：（1）我们通过匹配知识图谱中的实体和文本中的命名实体，来构建候选实体的列表；（2）计算文本重构表示向量 $\hat{\boldsymbol{e}}$ 与知识图谱中的候选实体表示向量之间的 L_2 距离，并选择距离最小的候选实体作为我们最终选择的实体。图4.2即为上述算法的一个示例。

为了减小由于实体消歧效果可能带来的错误信息，我们对消歧实体的 L_2 距离设置了一个阈值 α。对于所选择的实体 e 和它的 L_2 距离 d，如果 d 小于 α，这意味着我们实体消岐的结果正确的可能性较高，我们就选择 $\boldsymbol{e}$ 来计算知识注意力权重。而如果 d 大于 α，这意味着实体消除的结果可能是错误的，或者这个实体在知识图谱中没有任何类似的实体，此时我们就直接使用 $\hat{\boldsymbol{e}}$ 来计算知识注意力权重。

4.3.2　实验分析

4.3.2.1　数据集与实验设置

`Figer` 是一个广泛使用的数据集，它在 Ling 等人[120] 的工作中被提出，用于实体分类。但是，`Figer` 的训练集不包括 KNET 模型所需要的实体链接信息。此

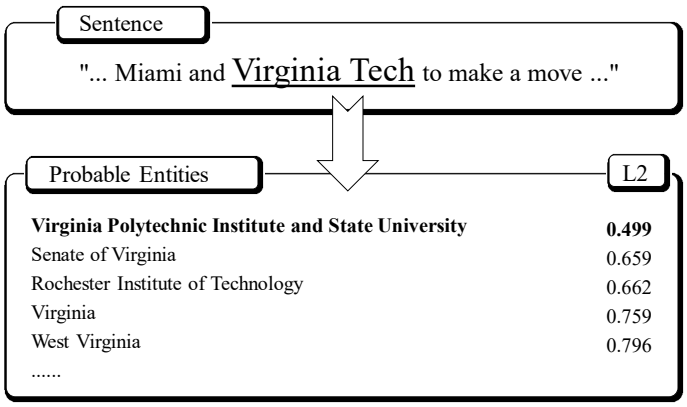

图 4.2　命名实体通过实体消歧链接到正确实体上的过程

外，该测试集的粒度还不够细 (例如，超过 38% 的实体只有人物标注，没有更细粒度的标签)。因此，我们构建了一个新的数据集用于 KNET 模型的评测，该数据集包含自动标注数据和手动标注数据两部分。

自动标注数据集（Wiki-auto）。类似于 Ling 等人[120] 的做法，我们使用维基百科和 Freebase 来生成训练、验证与测试三部分数据集，并采用远程监督技术[31] 来进行数据自动标注。具体来说，我们在维基百科中检索包含链接到另一个维基页面的锚文本的句子，该锚文本可以进一步链接到一个 Freebase 实体。我们将 Freebase 中的实体类型信息标注来为句子中的命名实体标注类型。在数据构建过程中，我们在维基百科中主要检索 FB15K 的实体，FB15K 是 Bordes 等人[119] 构造的 Freebase 子集。

Freebase 包含数千种类型，通常比较混乱且带有噪声。例如，实体*纽约市*有 85 种类型，包括*城镇*、*有狗的城市*和*获奖者*。为了避免这种混淆，我们只保留在 FB15K 中至少有 50 个对应实体的类型，然后手工地将它们映射为一个包含 74 个类型的双层类别集合。

手动标注数据集（Wiki-manu）。远程监督不可避免地会将噪声引入自动标记数据集[121-122]。因此，我们从维基百科中随机抽取了 100 个实体和它们所在的句子，然后用与自动标记数据集相同的类别集合来对其进行手工标注。这个手工标注数据集仅用于模型评测。

自动和手动数据集各自有其优缺点。人工标注数据集的缺点是规模小，自动标注数据集的缺点是远程监督的假设过强从而引入噪音。但是，从人工标记数据集的观察结果来看，这种噪音现象在自动标注数据集中并不严重：只有很小一部分实体在不同的上下文中有不同的标签（例如，在 `Figer` 的测试集中只有 3.9%）。我们在两个数据集上分别进行了实验和结果分析，结论基本上是一致的。我们在

表格4.1中对数据集 Figer、Wiki-auto 和 Wiki-man 进行了详细比较。

表 4.1　不同数据集合的比较

数据集	Wiki-auto	Wiki-manu	Figer
实体总数	100,000	100	562
实体平均类型数	3.07	2.32	1.38
人物	22.47%	16.00%	43.42%
组织	14.76%	11.00%	28.11%
地点	39.90%	52.00%	18.15%
其他类型	22.87%	21.00%	12.81%

注意，在我们的实验中，Freebase 有两种作用：（1）为知识表示学习模型 TransE 模型提供学习实体表示向量需要的关系事实三元组；（2）为标注数据集提供类型信息。这两方面不一定需要用同一个知识图谱来完成，相反，它们可以由两个独立的知识图谱来实现。在我们的实验中，我们确保（1）中的三元组和（2）中的实体类型信息相互独立，用以验证模型能否被推广到使用不同数据图谱的情况。

参考 Ling 等人[120] 的设定，我们使用宏平均 f1、微平均 f1 和准确率来评估模型的效果。一般来说，我们认为微平均 f1 是最能代表细粒度实体分类性能的度量指标，进一步的细节可以在之前的相关工作中找到。参考 Shimaoka 等人[118] 的设定，我们使用 GloVe[123] 预训练的词向量来初始化模型的词向量。在超参数选择上，我们搜索了不同的超参数设置：学习率 λ 在 $\{0.01, 0.005, 0.01\}$ 中选择，LSTM 的隐状态大小在 $\{100, 150, 200\}$ 中选择，词向量的大小 $\{50, 100, 300\}$ 中选择，窗口大小 L 在 $\{5, 10, 15\}$ 中选择，每次训练集合大小 B 在 $\{100, 500, 1000\}$ 中选择。基于模型在验证集上的表现，我们选取的最优超参数可见表格4.2。

表 4.2　超参数设定

超参数	值
Learning rate	0.005
LSTM hidden-size	100
Word vector size	300
Window size	15
Batch size	1,000

表 4.3　实体分类的表现 (%)

数据集	Wiki-auto						
评价指标	宏平均				微平均		
	Acc	Pre	Rec	F1	Pre	Rec	F1
AFET	20.32	67.00	45.82	54.75	69.29	42.40	52.61
KB-ONLY	35.12	69.65	71.35	70.49	54.85	74.99	63.36
HNM	34.88	68.09	61.03	64.37	72.80	64,48	68.39
SA	42.77	75.33	69.69	72.40	77.35	72.63	74.91
MA	41.58	73.64	71.71	72.66	75.94	75.52	75.72
KA	45.49	74.82	72.46	73.62	76.96	75.49	76.22
KA+D	**47.20**	**75.72**	**74.03**	**74.87**	**77.96**	**77.87**	**77.92**
数据集	Wiki-man						
评价指标	Strict	宏平均			微平均		
	Acc	Pre	Rec	F1	Pre	Rec	F1
AFET	18.00	64.50	50.00	56.33	64.29	50.43	56.52
KB-ONLY	17.00	55.50	72.83	63.00	27.81	74.57	40.52
HNM	15.00	61.80	68.00	64.75	62.35	68.53	65.30
SA	18.00	66.67	73.67	69.44	65.54	75.43	70.14
MA	26.00	65.13	78.50	71.19	64.09	82.33	72.08
KA	23.00	64.69	78.92	71.10	63.25	82.68	71.67
KA+D	**34.00**	**68.41**	**82.83**	**74.94**	**66.12**	**87.50**	**75.32**

4.3.2.2　实验结果

最近的许多工作[120,124-125] 已经证明神经网络实体分类模型的效果好于绝大多数基于特征的模型。因此，我们选取以下两个经典的神经网络实体分类模型作为我们的基线模型:

(1) **带语义关注的神经模型（SA）**。根据我们的了解，这个工作[118] 是当前最优的模型。由于他们的代码还没有公开，我们自己实现了他们的模型，并取得了与作者报告相当的效果。

(2) **混合神经模型（HNM）**。我们也实现了混合神经模型[126]，这是一个包含全连接层和循环层的神经网络实体分类模型，但是没有采用注意力机制。

我们还比较了一个效果最好的基于特征的模型:

(3) **AFET 模型**[127]，它也使用了来自知识图谱的辅助信息，但没有考虑将实体

之间的关系知识结合到模型中以获得更好地实体表示。

考虑到引入了知识图谱这样的外部信息，我们进一步考虑了一个基线:

(4) **KB-ONLY 模型**，它只使用知识图谱的表示向量来进行实体分类（以阈值 α 来控制，用 $\boldsymbol{e}$ 或 $\hat{\boldsymbol{e}}$ 来替代方程4-4中的 $\boldsymbol{x}$）。

我们将这四个基线模型与我们的神经分类模型进行比较，其中包括实体注意力机制 (MA)、知识注意力机制 (KA)、基于消歧的知识注意力机制 (KA+D)。表格4.3中报告了实验结果。从表中我们可以看到:

（1）所有的神经网络实体分类模型的效果都比基于特征的 AFET 模型更好。这展示了神经网络模型能够有效利用大规模训练数据学习到实体分类的对应模式。

（2）与 SA 模型相比，MA 模型的表现稍好一些，这是因为 MA 模型采用了一种简单的与实体相关的注意力机制。这表明了考虑实体和上下文相互关联性在实体分类中的重要性。

（3）KA 和 KA+D 模型在所有方法中取得最佳效果。原因是 KA 和 KA+D 模型都从知识图谱中引入了丰富的实体和关系信息，并比其他方法能更准确地关注上下文单词。它表明将通过知识表示学习将知识图谱信息应用到实体分类模型上的有效性。

（4）KA+D 模型在所有评价指标下的性能都优于 KA 模型。该模型通过对命名实体与知识图谱实体的相似程度进行消歧操作，从而可以从知识图谱中得到更精确的信息。

（5）KB-ONLY 的性能比 KA 和 KA+D 差很多。它表明，虽然知识图谱信息对实体类型分类有好处，但它并不能单独进行实体分类。相反，它必须以更复杂的方式与文本信息一起考虑，并最终对实体分类任务产生作用。

4.3.2.3　不同实体上的模型有效性

为了研究模型的细节，我们进一步将它们在测试集的不同子集中与基线模型进行比较。测试集基于实体的粗类型或消除歧义操作的难度来划分。

实体粗粒度类型　我们比较了模型在三种粗粒度实体类型上的效果：人物、组织和地点。为了更好的比较模型的效果，我们还将它们与一个简单的基线模型比较：M-ONLY，只使用命名实体来做分类（用 $\boldsymbol{m}$ 替换公式4-4中的 $\boldsymbol{x}$）。实验结果显示在表格4.4中。从表中我们可以看到：KA 和 KA+D 在更 ‘‘有难度’’ 的粗类型上获得了更大的改进，比如人物与组织。原因是，简单地根据命名实体，就可以较容易地确定一个地点实体的细粒度类型，因为它经常包含像河流或大道这样的信息性词汇。但是，对于人物和组织类型，我们必须更多地依赖上下文信息。在这种情况

表 4.4　在不同粗粒度类型上的效果比较（%）

数据集	Wiki-auto		
类型	人物	组织	组织
M-ONLY	58.64	63.95	87.65
HNM	63.79	66.85	86.26
SA	68.47	71.85	90.74
KA	70.77	74.18	91.23
KA+D	**74.87**	**75.16**	**91.75**
数据集	Wiki-man		
类型	人物	组织	组织
M-ONLY	52.63	71.19	75.54
HNM	54.00	50.00	76.69
SA	55.77	81.36	79.26
KA	67.72	75.41	79.29
KA+D	**67.14**	**90.32**	**81.62**

下，KA 和 KA+D 模型显示了它们在建模上下文信息上的优势。M-ONLY 的性能高低则一定程度上显示了每个粗粒度类型的判断 ‘‘难度’’。

消岐难度的影响　在 KA+D 中，在对知识图谱中的实体进行消歧时，需要考虑上下文的信息。上下文可以提供关于实体属性的丰富、有用的信息，也可能几乎不包含任何有用的提示信息。我们根据消除歧义操作的结果是否正确，将测试集划分为两个子集，分别名为正确集和错误集，并探讨各种模型在其中的效果。结果展示在表格4.5中。

从表格中我们可以看到：

（1）KA 模型在两个子集中始终优于所有基线模型。它表明使用知识图谱信息可以有效地帮助实体分类。

（2）所有方法在正确子集中的效果比在错误子集中更好。结果是符合直觉的，因为在正确子集中的实体的上下文能够提供更准确的信息，并使得类型分类结果更好。

（3）在正确子集上，KA+D 可以通过消除歧义从知识图谱中获得精确的实体信息，从而显著优于其他所有方法。在错误子集中，KA+D 的优势较小，原因在于其无法有效地对命名实体进行消歧，但是它仍然优于基线。原因是，在阈值 α 的控制下，在这个子集中的一个命名实体将被对应到一个类似的实体（其表示向量也

表 4.5　在正确/错误子集上的模型结果（%）

数据子集	正确集		错误集	
评估指标	准确率	微平均 F1	准确率	微平均 F1
Wiki-auto	80.53		19.47	
HNM	37.60	68.39	23.60	52.15
SA	46.66	78.63	26.64	57.61
MA	44.32	79.29	28.26	59.05
KA	49.24	79.83	**29.99**	**59.42**
KA+D	**51.77**	**82.33**	28.27	57.56
数据子集	正确集		错误集	
评估指标	准确率	微平均 F1	准确率	微平均 F1
Wiki-man	83.00		17.00	
HNM	15.66	67.80	11.76	51.95
SA	20.48	75.05	5.88	47.37
MA	28.92	75.22	11.76	53.85
KA	24.10	75.23	**17.65**	53.93
KA+D	**34.94**	**78.32**	12.50	**54.77**

会有用），又或者保持原始文本的表示向量，从而在一定程度上减轻错误对模型效果的影响。

我们进一步在图4.3中展示阈值 α 对 KA+D 模型的影响。结果显示，当 α 增加（即模型对消除歧义结果更有把握）时，KA+D 的效果在正确子集中得到了显著地提升，但在错误子集中变得更差。因此，在现实世界的应用中，我们必须根据正确/错误子集的比率 (数据集的消歧困难度)，调整 α 以来实现效果上的权衡。考虑到当 α 从 0.55 上升到 0.7 时，错误集中模型效果的急剧降低和正确的效果相对增长缓慢，我们将 α 设置为 0.55，如表格4.2中最优超参数所示。

4.3.2.4　案例分析

我们在图4.4中给出一个例子，对由 SA 和 KA+D 模型给出的注意力权重进行可视化。从这个示例中，我们可以看到：SA 模型未能将注意力权重集中在那些对实体分类的有用词汇上。而 KA+D 模型，通过正确地考虑知识图谱中的实体信息，可以将注意力权重更多地放在那些有意义的单词上，例如 `starred`、`the film`、`Omar Sharif` 和 `Geraldin Chaplin`。由 KA+D 所预测的类型是人物、艺术家和演员，它们与带标注的标签相同。而除了这三种类型，SA 模型还预测了三种

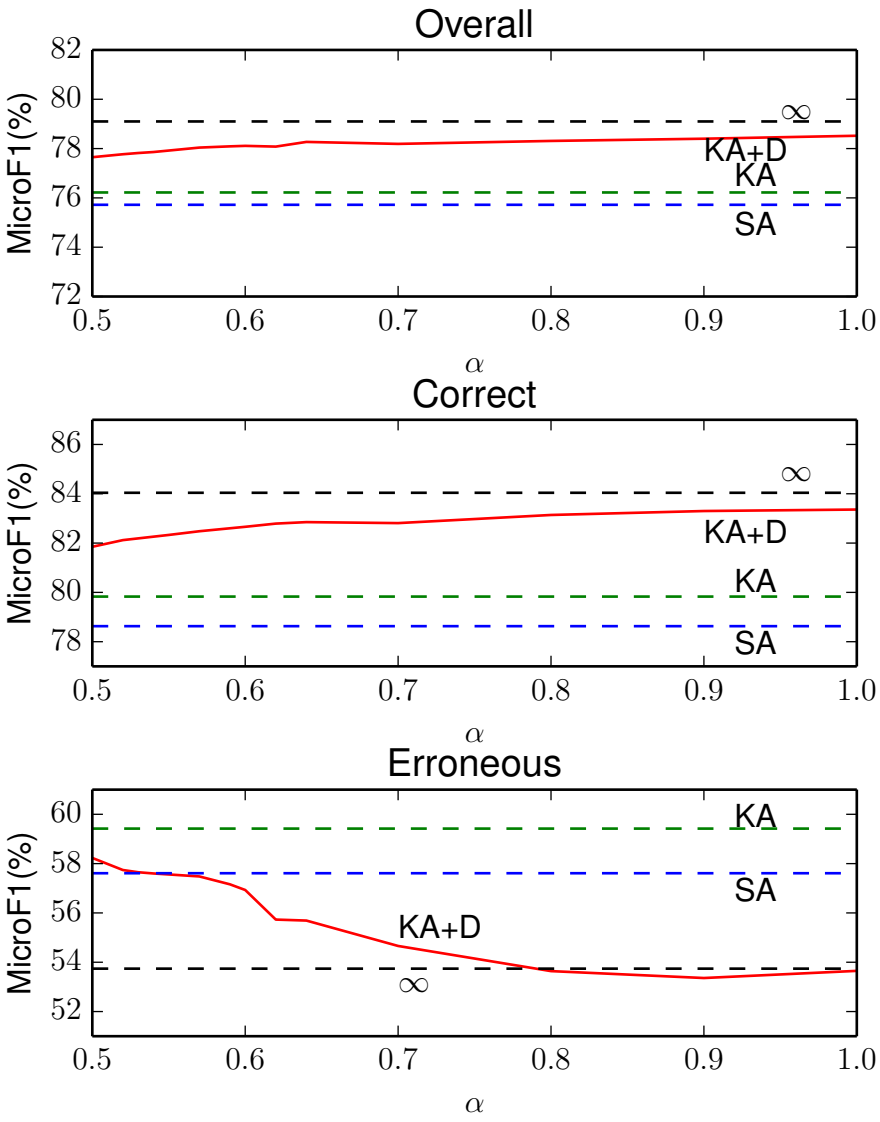

图 4.3　α 对 KA+D 的影响。∞ 表示不使用 α 的 KA+D 模型

多余的错误类型。

4.4　基于知识的开放域问答

近些年，阅读理解作为问答领域的一个子任务，其旨在根据指定文档的信息来回答用户的问题，已经成为了自然语言处理的热门方向。现有的神经网络阅读理解系统[128-132] 得益于其多层网络结构和注意力机制强大的建模能力，已经能够在一定程度上做到对问题的答案进行推理。可以说，阅读理解系统已经展示了现有神经网络模型在阅读、处理和理解自然语言文本方面的能力。

尽管现有阅读理解系统已经取得了不错的效果，但是这些系统在回答问题时

模型	句子和注意力权重
SA	...tradition starred Omar Sharif, Geraldine Chaplin and **Julie Christie**. Concentrating on the love triangle aspects of the novel, the film ...
KA+D	...tradition starred Omar Sharif, Geraldine Chaplin and **Julie Christie**. Concentrating on the love triangle aspects of the novel, the film ...

图 4.4　案例分析

仍然依赖于用户预先指定好的文档，因而无法在实际问答场景中广泛使用。和阅读理解任务不同，开放域问答旨在通过大量无标注文档的信息来回答用户的问题，更加符合现实应用场景的需求。因此，Chen 等人[133] 提出了结合信息检索和阅读理解两者的技术来实现开放域问答：先利用信息检索技术检索出和问题相关的文档，然后利用阅读理解技术从中抽取出问题的答案。

Chen 等人提出的开放域问答模型为人们解决开放域问答提出了一种新的思路，然而，其提出的开放域问答模型仍然远远不能应用于实际任务。其原因在于其无法解决以下两个问题：

- **背景知识缺乏**：现有的开放域问答模型缺乏背景知识的信息，导致其无法很好地理解对于问题和相关文档中涉及到背景知识的部分（如实体）。在现实场景中，人们回答一个问题一般不仅仅根据查询到的相关文档，也需要使用自身的背景知识。例如，在表格4.6中，对于问题 “Do you know that <Query> has been made in Texas from wild mustang grapes since the Antebellum Era’’，我们必须拥有相关背景知识才能理解文档中 “Prior to the American Civil War of 1861-1865’’ 指的是问题中的 “Antebellum Era’’，才能正确地回答出问题。因此，我们需要在神经网络开放域问答模型中对背景知识进行建模。
- **噪音数据**：由于现有开放域问答系统一般通过远程监督进行数据标注，其不可避免地会面临噪音问题。例如，对于问题 ‘‘都柏林是哪个国家的首都？’’，我们可能会面临以下问题：（1）检索到的文档 ‘‘都柏林是爱尔兰最大的城市...’’ 事实上和问题无关；（2）对检索到文档 ‘‘作为爱尔兰的首都，都柏林同时也是爱尔兰最大的旅游城市之一...’’，其中第二个 ‘‘爱尔兰’’ 实际上不是问题的真正答案，因为其上下文并没有在描述问题。如何处理这些噪音数据是开放域问答模型在训练和预测时最关键的问题之一。

针对现有开放域问答系统的上述问题，我们借鉴了人类回答开放域问题的模

表 4.6　需要背景知识的开放域问答的例子

问题	文档	背景知识	答案
Do you know that <Query> has been made in Texas from wild mustang grapes **since the Antebellum Era**?	Prior to the American Civil War of 1861-1865, Thomas Affleck ... made **mustang wine** on his Glenbly the Plantation located in Gay Hill, Washington County, Texas ... mustang grapes would first have to be crushed and left to ferment for twelve to fifteen hours ...	"Prior to the American Civil War of 1861-1865" 指的是 "Antebellum Era".	Mustang Wine

式，提出了一个基于 ''粗读-精读-总结'' 模式的开放域问答模型 DS-QA。进一步地，我们通过加入知识表示得到的实体表示和采用关系抽取任务进行多任务学习，在我们的 DS-QA 模型中引入了来自知识图谱的背景知识，极大地提升了模型在回答需要背景知识问题时的效果。

4.4.1　算法模型

我们的 DS-QA 模型旨在对于一个给定的问题在大规模无标注文本中找到问题的答案。我们首先采用信息检索技术从开放域文本库中检索出和问题相关的段落，然后从这些检索到的段落中抽取出问题的答案。

具体地，给定一个问题 $q = (q^1, q^2, \cdots, q^{|q|})$，我们首先检索 m 个相关段落 $P = \{p_1, p_2, \cdots, p_m\}$，其中 $p_i = (p_i^1, p_i^2, \cdots, p_i^{|p_i|})$（$p_i^j$ 表示一个单词或者实体，也就是说，我们把每一个实体合并成一个单词）表示第 i 个检索到的段落。我们的模型对给定问题 q 和相关段落集合 P，建模抽取答案 a 的条件概率。如图4.5所示，DS-QA 模型主要包含以下两个部分：

段落选择器：　给定问题 q 和相关段落集合 P，段落选择器旨在判断哪些段落是真正包含问题答案的。具体地，段落选择器对于检索到的所有段落输出一个条件概率分布 $\Pr(p_i|q, P)$。

段落阅读器：　给定问题 q 和特定段落 p_i，段落阅读器采用一个多层 LSTM 网络来计算抽取出答案 a 的概率 $\Pr(a|q, p_i)$。

最终，对给定问题 q 和相关段落集合 P，我们定义抽取答案 a 的条件概率如下：

$$\Pr(a|q, P) = \sum_{p_i \in P} \Pr(a|q, p_i) \Pr(p_i|q, P). \tag{4-12}$$

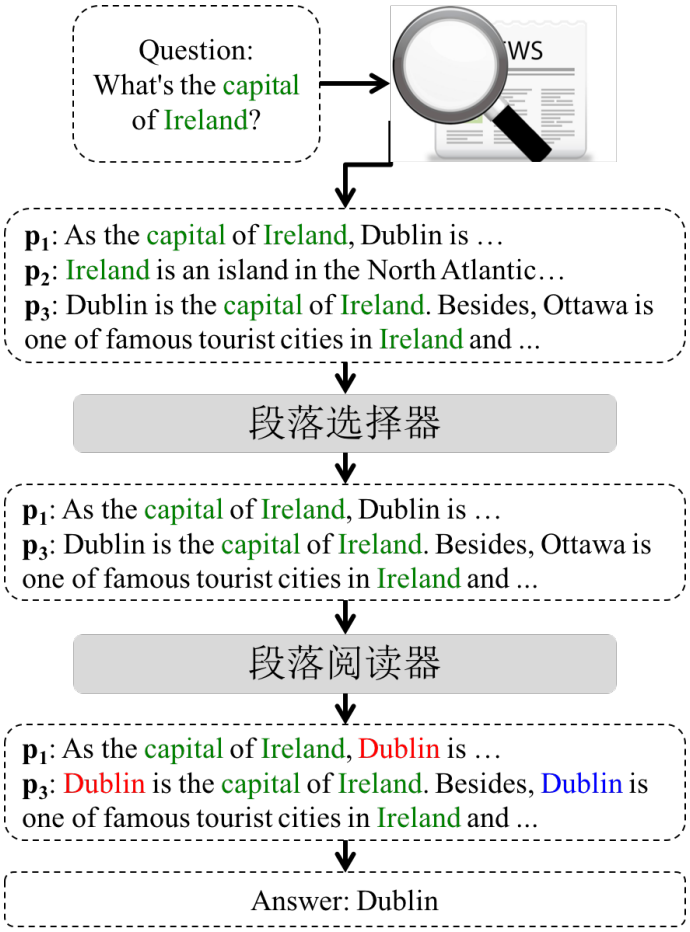

图 4.5　DS-QA 模型的示意图：包括段落选择器和段落阅读器两个主要部分

除此之外，为了在 DS-QA 模型中引入外部知识图谱的信息，我们在模型的输入表示中加入了知识表示学习得到的实体表示。另外，为了显式地考虑实体之间的关系信息，我们还加入了关系抽取任务进行多任务学习。

4.4.1.1　段落选择器

由于开放域问答系统通过远程监督进行数据标注，其不可避免地会面临错误标注问题。为了有效地利用所有检索到的段落中包含的信息，我们需要对其中的噪音数据进行过滤。因此，我们采用一个段落选择器来评估各个检索到的段落包含答案的可能性。

输入表示：　我们首先将段落 p_i 中的每一个单词 p_i^j 转化为向量表示，其向量表示包含以下两个部分 $\mathbf{p}_i^j$：（1）**词向量：**词向量旨在将离散字符形式的单词转换为连续向量空间中的分布式表示，从而捕捉到单词所对应的语义信息。（2）**实体向量：**实体向量旨在将通过知识表示学习将知识图谱中的实体、关系以及关系事实等知

识信息转化为分布式表示，从而将知识图谱中的信息融入到开放域问答系统中。

段落表示：　我们将段落中单词的输入表示传到一个神经网络中获得其隐式表示 $\hat{\mathbf{p}}_i^j$。这里我们采用了两种不同的神经网络结构包括：（1）多层感知机模型（MLP）

$$\hat{\mathbf{p}}_i^j = \text{MLP}(\mathbf{p}_i^j), \tag{4-13}$$

以及（2）循环神经网络模型（RNN）

$$\{\hat{\mathbf{p}}_i^1, \hat{\mathbf{p}}_i^2, \cdots, \hat{\mathbf{p}}_i^{|p_i|}\} = \text{RNN}(\{\mathbf{p}_i^1, \mathbf{p}_i^2, \cdots, \mathbf{p}_i^{|p_i|}\}). \tag{4-14}$$

我们希望 $\hat{\mathbf{p}}_i^j$ 可以有效地刻画单词 p_i^j 与其上下文之间的语义联系。对于循环神经网络，我们采用一个单层双向 LSTM 网络来进行实现，并将两层的隐状态拼接得到最终的隐式表示 $\hat{\mathbf{p}}_i^j$。

问题表示：　和段落表示类似，我们首先将问题中的每一个词 q^i 表示为一个由词向量和实体向量拼接而成的输入表示 $\mathbf{q}^i$，然后传入多层感知机（MLP）：

$$\hat{\mathbf{q}}_i^j = \text{MLP}(\mathbf{q}_i^j), \tag{4-15}$$

或者循环神经网络（RNN）：

$$\{\hat{\mathbf{q}}^1, \hat{\mathbf{q}}^2, \cdots, \hat{\mathbf{q}}^{|q|}\} = \text{RNN}(\{\mathbf{q}^1, \mathbf{q}^2, \cdots, \mathbf{q}^{|q|}\}). \tag{4-16}$$

之后，我们采用一个自选择注意力机制来获得最终的问题表示 $\mathbf{q}$：

$$\hat{\mathbf{q}} = \sum_j \alpha^j \hat{\mathbf{q}}^j, \tag{4-17}$$

其中，α_j 表示问题中每一个单词的重要性，具体计算如下：

$$\alpha_i = \frac{\exp(\mathbf{w}\mathbf{q}_i)}{\sum_j \exp(\mathbf{w}\mathbf{q}_j)}, \tag{4-18}$$

其中 $\mathbf{w}_b$ 是权重向量。

最后，我们将采用最大池化层和 softmax 层获得每一个段落的概率：

$$\Pr(p_i|q, P) = \text{softmax}\left(\max_j(\hat{\mathbf{p}}_i^j \mathbf{W}\mathbf{q})\right), \tag{4-19}$$

其中 $\mathbf{W}$ 为权重矩阵。

4.4.1.2　段落阅读器

段落阅读器旨在对于给定段落 p_i 抽取出问题的答案。和段落选择器类似，我们首先将段落 p_i 通过一个多层双向 LSTM 网络表示为隐式向量 $\{\bar{\mathbf{p}}_i^1, \bar{\mathbf{p}}_i^2, \cdots, \bar{\mathbf{p}}_i^{|p_i|}\}$。同时，我们也采用一个基于自选择注意力机制的多层双向 LSTM 网络来获取问题表示 $\bar{\mathbf{q}}$。

段落选择器旨在从给定段落中抽取出最可能是正确答案的片段。这里，我们把问题转换为预测正确答案的开始位置和结束位置。因此，对于给定问题 q 和段落 p_i，抽取出答案 a 的概率可以表示为：

$$\Pr(a|q, p_i) = P_s(a_s)P_e(a_e), \tag{4-20}$$

其中 a_s 和 a_e 表示答案在段落中的开始和结束位置，而 $P_s(a_s)$ 和 $P_e(a_e)$ 分别是 a_s 和 a_e 作为答案开始位置和结束位置的概率，具体计算如下：

$$P_s(j) = \text{softmax}(\bar{\mathbf{p}}_i^j \mathbf{W}_s \bar{\mathbf{q}}), \tag{4-21}$$

$$P_e(j) = \text{softmax}(\bar{\mathbf{p}}_i^j \mathbf{W}_e \bar{\mathbf{q}}), \tag{4-22}$$

其中 $\mathbf{W}_s$ 和 $\mathbf{W}_e$ 是双线性网络中的可学习权重矩阵。在开放域问答任务中，由于我们事先并不知道答案的具体位置，而段落中可能有多处片段可以匹配上答案。令 $\{(a_s^1, a_e^1), (a_s^2, a_e^2), \cdots, (a_s^{|a|}, a_e^{|a|})\}$ 表示段落中多处匹配上答案的片段，公式 (4-20) 可以进一步通过如下方式进行定义：(1) 最大值，也就是，我们假设段落中只有一处匹配上答案的片段是问题的答案：

$$\Pr(a|q, p_i) = \max_j \Pr_s(a_s^j) \Pr_e(a_e^j) \tag{4-23}$$

(2) 加和值，也就是，我们假设中所有匹配上答案的片段是同一事物：

$$\Pr(a|q, p_i) = \sum_j \Pr_s(a_s^j) \Pr_e(a_e^j). \tag{4-24}$$

4.4.1.3　多任务学习

对于 DS-QA 模型，我们可以定义以下损失函数进行优化：

$$L = -\sum_{(\bar{a}, q, P) \in T} \log \Pr(\bar{a}|q, P) - \alpha R(P), \tag{4-25}$$

其中 $\bar{a}$ 表示正确答案，T 为训练集，$R(P)$ 为防止段落选择器过拟合的正则项。这里，$R(P)$ 定义为 $\Pr(p_i|q, P)$ 和概率分布 $\mathcal{X}$ 之间的 KL 距离。其中，对于正确答案

片段的段落，$\mathcal{X}_i = \frac{1}{c_P}$（$c_P$ 是包含正确答案片段的段落数目），反之，$\mathcal{X}_i = 0$。具体地，$R(P)$ 定义如下：

$$R(P) = \sum_{p_i \in P} \mathcal{X}_i \log \frac{\mathcal{X}_i}{\Pr(p_i|q,P)}. \tag{4-26}$$

为了对知识图谱中实体间的关系事实信息进行考虑，我们这里进一步加入了关系抽取任务进行多任务学习。假设检索到的段落 p_i 包含实体 $\{e_1, e_2, \cdots, e_m\}$，那么 DS-QA 模型将对其两两进行关系预测，假设 e_j 与 e_k 在知识图谱中的关系为 r_{jk}，那么我们将最大化 e_j 与 e_k 在段落中预测出关系 r_{jk} 的概率 $\Pr(r_{jk}|p_i, e_j, e_k)$。具体地，概率 $\Pr(r_{jk}|p_i, e_j, e_k)$ 计算如下：

$$\Pr(r_{jk}|p_i, e_j, e_k) = NRE(p_i, e_j, e_k) \tag{4-27}$$

其中 NRE 模型为我们在上一章节介绍的基于选择注意力机制的神经网络关系抽取模型，且 NRE 模型与 DS-QA 中的段落阅读器共享隐式表示。因此，对于 DS-QA 模型，其损失函数可以进一步定义为：

$$L = -\sum_{(\bar{a},q,P)\in T} \log \Pr(\bar{a}|q,P) - \beta \sum_{p_i \in P} \sum_{jk} \Pr(r_{jk}|p_i, e_j, e_k) - \alpha R(P). \tag{4-28}$$

4.4.1.4　模型预测

在进行预测是，我们抽取概率值最高的答案 $\hat{a}$ 如下：

$$\begin{aligned} \hat{a} &= \arg\max_a \Pr(a|q,P) \\ &= \arg\max_a \sum_{p_i \in P} \Pr(a|q,p_i) \Pr(p_i|q,P). \end{aligned} \tag{4-29}$$

这里，段落选择器可以看做是对所有段落的 ‘‘粗读’’，段落阅读器可以看做是对具体段落的 ‘‘精读’’。在经过 ‘‘粗读’’ 和 ‘‘精读’’ 之后，DS-QA 模型总结了所有段落的有效信息来进行答案预测。

4.4.2　实验分析

4.4.2.1　数据集与实验设置

我们在以下五个公开的开放域问答评测数据集上对 DS-QA 模型进行了评测。

Quasar-T[①] [134] 包含了 43,000 个开放域日常问题，其问题对应的 50 个相关段落是从 ClueWeb09 中采用 LUCENE 检索得到。

① https://github.com/bdhingra/quasar

SearchQA[①][135] 是一个大型的开放域问题数据集，其问题-答案对由从 J! Archive 中抽取得到，而其问题对应的 50 个文档是采用 Google Search API 搜索得到。

TriviaQA[②][136] 包含了 95,000 从日常琐事对话中获取的问题-答案对，其问题对应的 50 个文档是采用 Bing Web Search API 搜索得到。

CuratedTREC[③][137] 是一个基于 TREC QA 任务的数据集，包含了从 TREC1999、2000、2001 和 2002 中抽取出的 2,180 个问题。

WebQuestions[④][138] 是一个知识库问答数据集，其问题的答案都是知识图谱 Freebase 中的实体，我们采用 LUCENE 从英文维基百科中检索出了问题的相关段落。

此外，为了评测我们模型通过知识表示学习和多任务学习考虑知识图谱信息的效果，我们还在 HotpotQA 数据集上对我们模型进行了进一步分析。上述数据集的详细数据展示在表格4.7中。

表 4.7　开放域问答数据集

数据集	训练集	验证集	测试集
Quasar-T	37,012	3,000	3,000
SearchQA	99,811	13,893	27,247
TriviaQA	87,291	11,274	10,790
CuratedTREC	1,486	-	694
WebQuestions	3,778	-	2,032
HotpotQA	90,564	7405	7405

参考 Chen 等人的工作[133] 中的评测方式，我们采用了两个不同的评测指标包括精确匹配 (EM) 和 F1 来评价 DS-QA 模型的效果。

我们在开发集上对我们的模型参数进行了搜索（我们加粗高亮了所有最优参数）：对于 LSTM 模型的隐状态大小，我们搜索了 $\{32, 64, \mathbf{128}, \cdots, 512\}$；段落阅读器中的 LSTM 模型的层数，我们搜索了 $\{1, 2, \mathbf{3}, 4\}$；对于正则化参数 α，我们搜索了 $\{0.1, \mathbf{0.5}, 1.0, 2.0\}$；对于 batch 大小，我们搜索了 $\{4, 8, 16, \mathbf{32}, 64, 128\}$。对于其余参数，由于对模型效果没有特别大的影响，我们采用了[133] 中的参数。

① https://github.com/nyu-dl/SearchQA
② http://nlp.cs.washington.edu/triviaqa/
③ https://github.com/brmson/dataset-factoid-curated/tree/master/trec
④ https://github.com/brmson/dataset-factoid-webquestions

表 4.8　在五个不同开放域问答数据集上的评测效果

数据集	Quasar-T		SearchQA		TriviaQA		CuratedTREC	WebQuestions	
模型	EM	F1	EM	F1	EM	F1	REM	EM	F1
GA[129]	26.4	26.4	-	-	-	-	-	-	-
BiDAF[139]	25.9	28.5	28.6	34.6	-	-	-	-	-
AQA[140]	-	-	40.5	47.4	-	-	-	-	-
R^3[141]	35.3	41.7	49.0	55.3	47.3	53.7	28.4	17.1	24.6
Our + AVG	38.5	45.7	55.6	61.0	42.6	48.2	28.6	17.8	24.5
+ FULL	**42.2**	**49.3**	**58.8**	**64.5**	**48.7**	**56.3**	**29.1**	**18.5**	**25.6**

4.4.2.2　基线模型

为了评测我们 DS-QA 模型的效果，我们选取了目前效果最好的模型作为我们的基线模型：（1）**GA** 模型[129]，一个利用门机制进行多步逻辑推理的阅读理解模型；（2）**BiDAF** 模型[139]，一个采用了双向选择注意力机制网络的阅读理解模型；（3）**AQA** 模型[140]，一个利用强化学习对问题进行改写，然后综合所有改写问题抽取出的答案的问答模型；（4）$\mathbf{R}^3$ 模型[141]，一个利用强化学习选择最有信息量的段落来回答问题的开放域问答模型。

除了上述模型之外，我们还比较了我们模型的简化版本（认为所有相关段落包含正确答案的概率相同）。在实验中，我们把 DS-QA 模型命名为 “Our+FULL’’，将其简化版本命名为 “Our+AVG’’。

4.4.2.3　实验结果

DS-QA 模型和基线模型效果比较的结果展示在表格4.8中。从表示中的结果，我们可以看到：

（1）我们的两个模型（Our+AVG 和 Our+FULL）几乎在所有数据集上的效果都好于基线模型。其原因在于我们的模型可以有效地综合利用不同段落中的信息来回答问题，而大多数基线模型只能考虑其中信息量最大的段落。这证明了我们认为在过滤掉数据中噪音的基础上综合利用所有段落信息可以更好地回答开放域问题的设想。

（2）在所有数据集上，Our+FULL 模型的效果都要显著地好于 Our+AVG 模型。这说明了我们的段落选择器可以有效地过滤掉和问题无关的段落并降低噪音数据对于模型的影响。

表 4.9　段落选择器和传统信息检索模型在段落选择任务上的效果比较

数据集	Quasar-T				SearchQA			
任务	段落选择		总体		段落选择		总体	
Models	Hits@1	Hits@3	EM	F1	Hits@1	Hits@3	EM	F1
IR	6.3	10.9	-	-	13.7	24.1	-	-
Our + INDEP	26.8	36.3	40.6	46.9	59.2	70.0	57.0	62.3
Our + FULL	27.7	36.8	41.1	48.0	58.9	69.8	58.8	64.5

4.4.2.4　段落选择效果分析

为了分析我们提出的段落选择器是否能有效地过滤掉与问题无关的段落，我们比较了我们的段落选择器和传统信息检索模型[①]。此外，我们还比较了我们段落选择器的简化版 Our+INDEP（不与段落阅读器联合训练而是只用段落是否包含正确答案片段进行训练）。这里，我们采用远程监督数据对段落选择的效果进行评测，即我们认为一个段落只要包含正确答案的片段就被当成正确的。我们采用 Hit@N 进行评价。

结果展示在表格4.9中。从表格中的实验结果，我们可以看到：

（1）Our+INDEP 和 Our+FULL 模型在段落选择任务中都显著优于传统信息检索模型。这说明了我们提出的段落选择器能够有效地发现问题和段落之间的相互联系。

（2）Our+FULL 和 Our+INDEP 模型在段落选择任务上效果非常接近，但是在开放域问答任务上的总体效果却显著地优于 Our+INDEP 模型。这说明了通过段落选择器与段落阅读器的联合训练，段落选择器可以更加有效地在匹配上正确答案的文本片段中分辨出哪些片段是真正在回答问题。

4.4.2.5　段落数量的影响

我们的段落选择器的功能可以看做是在精读段落之前对段落做一次粗读来进行信息过滤，加速回答问题的速度。为了研究我们的段落选择器是否能否有效地加速整个开放域问答系统，我们比较了我们的模型只使用最具有信息的前几个段落来回答问题的效果（分别通过段落选择器和传统信息检索系统进行段落选择）。

实验结果展示在图4.6中。我们可以看到，我们的 Our+FULL 模型只需要原来大概一半的段落数量即可达到与使用所有段落接近的总体效果。与之相反，采用传统信息检索模型进行段落选择的 Our+IR 模型在段落数量减少时效果下降很多。这说明了我们的段落选择器可以很好地帮助我们加速开放域问答模型。

① 我们使用 LUENCE 实现了基于 BM25 的信息检索模型。

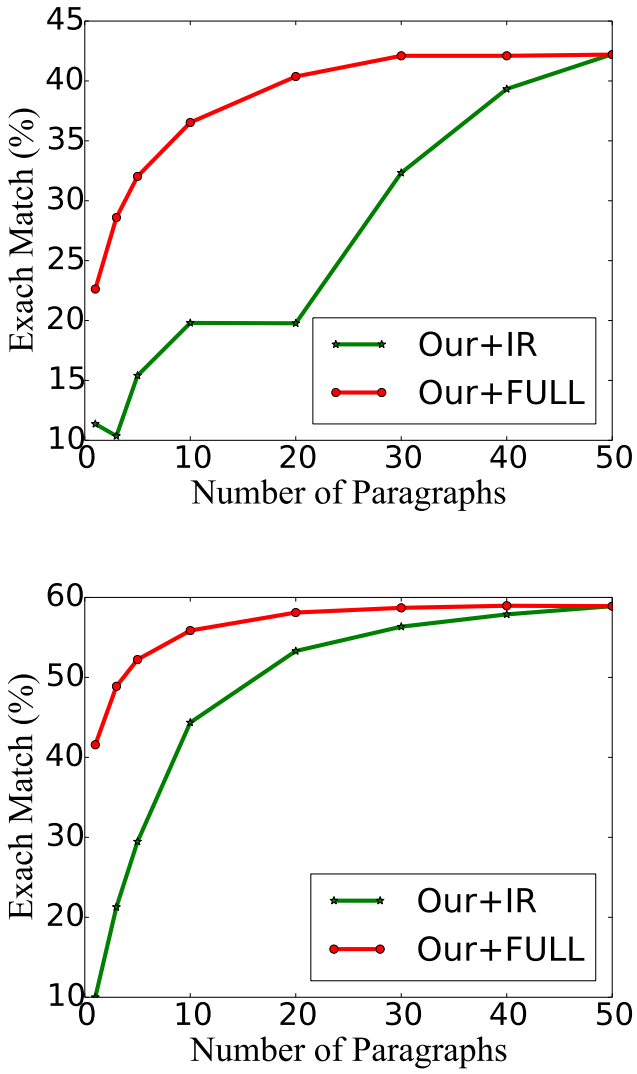

图 4.6　在数据集 Quasar-T（上方）和 SearchQA（下方）上模型采用不同数量段落的效果

4.4.3　考虑知识图谱信息的效果

为了分析考虑知识图谱的信息能否帮助我们的开放域问答系统更好地理解问题及其相关文档，我们在 HotpotQA 数据集上对比了是否加入知识图谱信息（知识表示学习获得的实体向量以及多任务学习）对于我们模型的效果的影响。选择 HotpotQA 数据集进行比较的原因在于该数据集涉及到的实体以及实体之间的关系相比于其他数据集更多，更能展现知识图谱信息的影响。我们将考虑知识图谱信息的模型命名为 “Our+Knowledge”, 将不考虑知识图谱信息的模型命名为 “Our-Knowledge”

实验结果展示在图4.10中。从实验结果中我们可以看到，Our+Knowledge 模型的效果在 EM 和 F1 上都显著地好于 Our-Knowledge 模型。这说明了通过使用知识

表 4.10　知识图谱信息对于模型的影响

模型	EM	F1
Our+Knowledge	44.44	58.26
Our-Knowledge	42.17	56.37

表 4.11　Our+Knowledge 模型做对而 Our-Knowledge 模型做错的例子

问题	文档	背景知识	答案
The Album Against the Wind was the 11th Album of a Rock singer Robert C Seger born may 6 1945. What was the Rock singers stage name?	Against the Wind is the eleventh studio album by American rock singer Bob Seger and his fourth with the Silver Bullet Band. It was released in February 1980. It is Seger's only number-one album to date, spending six weeks at the top of the Billboard Top LPs chart, knocking Pink Floyd's The Wall from the top spot.	"Robert C Seger" 指的是 "Robert Seger".	Robert Seger

表示学习获得的实体向量进行输入表示和与关系抽取任务进行多任务学习，我们的模型可以有效地在开放域问答系统中考虑知识图谱的信息，提高了模型的总体效果。

为了直观地展示加入知识图谱信息的效果，我们给出了一个在验证集中 Our+Knowledge 模型做对而 Our-Knowledge 模型做错的例子。可以看到，在这个例子中，模型需要知道 "Robert C Seger" 指的是 "Robert Seger'' 这一背景知识才能正确地将两者联系在一起从而回答出问题。Our+Knowledge 模型通过考虑知识图谱的信息引入了这一背景知识，从而能够正确地回答出问题的答案。

4.5　小结

目前，知识图谱及其表示学习已经成为许多自然语言处理下游任务如网页搜索、知识推理、问答系统等的重要信息来源。然而，由于现实场景的复杂性，现有的基于知识的自然语言处理模型还远远不能有效地对知识图谱的信息进行使用。在本章节中，我们主要介绍了结构化知识在实体分类和开放域问答中的应用。

第 5 章　总结与展望

知识图谱将人类知识组织成结构化知识库系统。知识图谱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，国内外互联网公司纷纷推出知识图谱产品，如谷歌 Knowledge Graph、微软 Bing Satori、百度知心以及搜狗知立方等。在著名的 IBM Watson 问答系统和苹果 Siri 语音助手的背后，知识图谱也扮演着重要的角色。如谷歌在介绍知识图谱时所说的：“Things，Not Strings.’’ 可以说，知识图谱的兴起拉开了自然语言理解从字符串匹配跃迁至智能理解的序幕。基于知识的自然语言处理技术已经成为目前最热门的研究方向之一。

为此，我们需要探索如何更好地将结构化知识图谱融入目前基于深度学习的自然语言处理模型中。然而，实现自然语言处理与知识图谱的有机结合，做到真正的自然语言理解并非简单的工作，需要解决以下几个关键问题：

（1）**知识表示**。如何在自然语言处理模型中充分利用大规模知识图谱，需要首先解决知识表示的问题。在第二章中，我们介绍了我们针对知识表示学习的三个重要问题做出的改进，分别提出了（a）考虑知识图谱复杂关系的 TransR 模型，使得不同实体在不同关系下可以有不同的表示，可以有效解决一对多、多对一、多对多关系的表示问题；(b) 考虑知识图谱复杂路径的 PTrasnE 模型，研究如何在知识表示学习模型中选择合适的关系路径以及如何表示关系路径；（c）考虑知识图谱复杂属性的 KR-EAR 模型，提出了一个同时考虑知识图谱中实体、关系、属性的知识表示学习框架。

（2）**知识获取**。如何从互联网大规模的结构化、半结构和无结构数据中自动获取知识，是构建大规模知识图谱的关键问题，因此我们需要探索知识自动获取技术。在第三章中，我们介绍了我们针对知识自动获取中的文本关系抽取任务作出的两个重要改进，分别是（a）基于选择注意力机制的 NRE 模型，解决了远程监督文本关系抽取中的噪音问题；（b）基于多语言注意力机制的 MNRE 模型，利用多语言数据构建了统一的多语言文本关系抽取模型，使得各个语言之间的数据可以相互验证和补充。

（3）**知识应用**。我们需要系统探索如何面向不同自然语言处理任务，将知识图谱信息合理地融合到该任务的模型中，实现知识指导的自然语言理解。在第四章中，我们介绍了我们在实体分类和开放域问答上的研究，分别是（a）基于知识注意力机制的实体分类 KNET 模型，引入知识图谱中的实体和关系信息用于帮助

发现命名实体和上下文之间的联系；（b）基于人类 ‘‘粗读-精读-总结’’ 问答模式的 DS-QA 模型，引入知识图谱实体和关系信息用于帮助问题和文档理解。

我们提出的方法有效地解决了基于知识的自然语言处理技术中面向知识图谱的知识表示、知识获取、知识应用中的关键问题，为迈向真正的自然语言理解打下了坚实的基础。接下来，我们将探讨面向大规模结构化知识的表示学习、自动获取与计算应用的未来挑战。

面向不同知识类型　目前的研究工作通常根据其映射属性将知识图谱中的关系划分为四种类型，包括一对一，一对多，多对一和多对多关系。在处理这四种不同的关系时，不同的知识表示学习方法表现差异极大。这表明我们需要针对不同类型的知识或关系专门设计不同的知识表示学习框架。然而，目前的关系划分方式过于简单，不能有效直观地解释知识的本质类型特点。根据知识的认知和计算特征，现有知识可以分为几种类型：（1）层次知识（例如 ‘‘整体-部分’’ 关系和 ‘‘上下位’’ 关系），表示实体之间的从属关系。（2）属性知识（例如 ‘‘国籍’’ 属性），表示实体的属性信息。许多实体可以共享相同的属性，尤其是对于诸如性别，年龄等的枚举属性。（3）关联关系（例如 ‘‘朋友’’ 关系），其指示实体之间的关系。直观的是，这些不同类型的关系应该以不同的方式建模。

知识的动态性　现有的知识表示学习方法通常在关系事实三元组进行建模的时候往往忽略了知识图谱中中包含的时间信息，将整个知识统一表示到唯一的语义空间中。事实上，知识不是一成不变的，而是随着时间而变化。对于任何时间点，应该有一个具有相应时间戳的唯一知识图谱。例如，‘‘乔治 · 布什’’ 在 1995-2000 年是 ‘‘美国’’ 的 ‘‘总统’’，而在其从其位置离任后不应该继续当做 ‘‘总统’’ 进行建模。对事实三元组的时间信息进行考虑有助于知识表示学习模型更加准确地理解知识图谱中的实体及其关系。更重要的是，对知识图谱变迁的研究不仅对研究知识图谱的理论和应用十分重要，也有助于对人类历史和认知的研究。目前，有一些研究工作[142-144] 已经在尝试将时间信息考虑进知识表示学习对知识图谱的建模过程中，但是这些工作仍然是对知识动态性的初步探讨，其仍然需要进一步探索。

多语言表示学习　Mikolov 等人[5] 观察到不同语言的向量空间之间相应概念的几何分布具有很强的相似性。在此基础在，他们认为在跨语言的不同向量空间之间进行语义上的映射在技术上是可行的，而且提出一个基于联合语义空间的跨语言表示模型十分有价值。原因在于统一的跨语言表示模型其中关于语言不变的表示特征可以帮助很多已有模型进行跨语言的迁移。此外，目前已经有了很多项目如

DBpedia，YAGO，Freebase，Wikidata 等，通过从维基百科中提取结构化信息来构建多语言的知识图谱。多语言知识图谱的构建对于全球化下的知识共享非常重要，并且已经在很多跨语言任务如跨语言检索，机器翻译和跨语言问答中发挥了重要的作用。然而，目前对于多语言知识表示学习的研究工作还几乎没有。因此，多语言知识表示学习是一项有意义但具有挑战性的工作，可以通过富语言中的知识图谱的信息来帮助进行其他语言的知识表示学习且进一步帮助其后续任务。

多模态表示学习　随着高速网络的快速发展，来自世界各地的数十亿人可以即时自由地上传和共享多媒体内容。正如我们所看到的，现在互联网不仅包含纯文本页面和超链接，还包含了海量的音频，图片和视频。如何有效和高效地利用从文本到视频的多模态信息正成为知识表示学习中一个关键且具有挑战性的问题。目前，在考虑多模态信息的知识表示学习的相关工作还比较有限，主要是简单地对实体的描述信息、结构化信息和图像信息进行考虑。这些模型信息来源比较有限，且对多源信息的使用手段还非常初步。我们仍需研究如何设计更合理更有效的知识表示学习模型以便更好地利用知识图谱的多模态信息。此外，社交网络等其他形式的多源信息仍然与知识表示学习相隔离，需要进一步探讨是否可以将各种不同来源的信息在统一的知识表示模型中进行建模。

文档/跨文档级别关系抽取　目前，神经网络关系抽取模型已经展示了其极强的能力，能够很好地从文本中找到每一个关系对应的文本表达模式。然而，现有的神经网络关系抽取模型往往只关注如何从一个单一的句子中抽取世界知识。事实上，很多世界知识通常没有被直接用一个句子进行表达，而是用多个句子、多个段落甚至多篇文章进行讲述。我们需要进一步研究如何从文档/跨文档级别的文本中进行关系抽取，以提高现有关系抽取系统能够获取的世界知识的覆盖率。然而，目前对于多文档/跨文档级别关系抽取的研究工作还几乎没有。因此，文档/跨文档级别关系抽取是一项有意义但具有挑战性的工作。相比于目前句子级别的关系抽取，文档/跨文档级别关系抽取需要考虑更多地因素包括（1）如何考虑文档内和文档间实体的相互指代问题；（2）如何考虑文档内和文档间的各个出现的关系事实之间存在逻辑推理关系；（3）如何将现有神经网络模型适配到更大规模的文档级别的文本上。

少次/零次关系抽取　最近，如何处理少次/零次的类别已经成为自然处理领域的一个重要问题，如文本分类、机器翻译等。少次/零次学习旨在学习如何为一个没有样例或者只有少数几个样例的类别进行分类。在关系抽取中，一个最为关键的问

题是现有的关系抽取模型由于受到训练语料类别不平衡以及部分类别训练样例较少的影响，只能抽取常见类别的关系事实。而往往在现实场景中，常见类别的信息在知识图谱中已经非常完善，我们更希望能够抽取出那些训练样例较少的关系对应的世界知识。因此，如何利用好丰富的外部信息帮助进行少次/零次关系抽取，以及设计适合少次/零次关系抽取的模型是目前关系抽取任务重一个非常重要的研究问题。

开放域关系抽取　现有的神经网络关系抽取技术还仅仅停留在研究如何进行封闭域 (close-domain) 关系抽取。封闭域关系抽取是指目标进行抽取的关系是来自于预先定义好的固定的集合。与之相反，开放域 (open-domain) 关系抽取旨在在不预先给定关系集合的情况下进行关系抽取，一般使用文本中的一些词来表示两个实体之间的关系。与封闭域关系抽取相比，开放域关系抽取任务无需我们事先指定可能存在的关系，适用范围更广。但是，与之同时，开放域关系抽取相比于封闭域关系抽取难度更大，推测新的关系类型难度非常大。原因在于有新的关系类型没有固定的形式且种类非常多。开放域关系抽取任务的主要难点在于：（1）如何将神经网络关系抽取模型从现有的分类模型转化为更加符合开放域关系抽取的具体模式；（2）如何在训练数据非常难以获得的情况下学习好神经网络关系抽取模型。

知识图谱质量　在现实任务中应用知识图谱的一个最主要的难点在于知识图谱本身的质量还远远没有达到要求。现有的大规模知识图谱如 Freebase、DBpedia、Yago、Wikidata 等中的关系事实有很大一部分是通过自动构建的方式从互联网信息中抽取得到。因此，由于缺乏人工标注，这些知识图谱本身不可避免地存在着噪音。这些噪音的存在很大程度地对其应用到下游任务中的效果产生了影响。因此，如何自动地检测出知识图谱中存在的矛盾或者错误是目前在下游任务中应用知识图谱信息的一个重要问题。

知识图谱规模　现有的大型知识图谱规模巨大，无法很高效地应用于现实场景的具体任务中。例如，截止至目前为止，Freebase 拥有大约 2300 万的实体和 19 亿的关系事实。由于知识图谱规模宏大，现有的很多基于知识的模型事实上仅仅在知识图谱中的一个子集验证了模型的有效性，而这些模型由于计算效率的问题往往无法直接应用于现实场景。因此，如何改进现有的基于知识表示的模型，使得其可以在现实场景中实现效果和效率上的平衡仍然是我们需要解

参考文献

[1] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10):78-85.

[2] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of SIGMOD. 2008: 1247-1250.

[3] Auer S, Bizer C, Kobilarov G, et al. Dbpedia: A nucleus for a web of open data[J]. The semantic web, 2007:722-735.

[4] Hoffart J, Suchanek F M, Berberich K, et al. Yago2: A spatially and temporally enhanced knowledge base from wikipedia[J]. Artificial Intelligence, 2013, 194:28-61.

[5] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of NIPS. 2013: 3111-3119.

[6] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[7] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society: volume 1. Amherst, MA, 1986: 12.

[8] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases [C]//Proceedings of AAAI. 2011: 301-306.

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

[10] 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.

[11] Sutskever I, Tenenbaum J B, Salakhutdinov R R. Modelling relational data using bayesian clustered tensor factorization[C]//Proceedings of NIPS. 2009: 1821-1828.

[12] Jenatton R, Roux N L, Bordes A, et al. A latent factor model for highly multi-relational data [C]//Proceedings of NIPS. 2012: 3167-3175.

[13] Yang B, Yih W t, He X, et al. Embedding entities and relations for learning and inference in knowledge bases[C]//Proceedings of ICLR. 2015.

[14] Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data[C]//Proceedings of ICML. 2011: 809-816.

[15] Nickel M, Tresp V, Kriegel H P. Factorizing yago: scalable machine learning for linked data [C]//Proceedings of WWW. 2012: 271-280.

[16] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs[C]//Proceedings of AAAI. 2016.

[17] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//Proceedings of NIPS. 2013: 926-934.

[18] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of NIPS. 2013: 2787-2795.

[19] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of ICLR. 2013.

[20] Mikolov T, Yih W t, Zweig G. Linguistic regularities in continuous space word representations [C]//Proceedings of NAACL: volume 13. 2013: 746-751.

[21] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of AAAI. 2014: 1112-1119.

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

[23] Xiao H, Huang M, Zhu X. Transg : A generative model for knowledge graph embedding[C]// Proceedings of ACL. 2016: 2316-2325.

[24] He S, Liu K, Ji G, et al. Learning to represent knowledge graphs with gaussian embedding[C]// Proceedings of CIKM. 2015: 623-632.

[25] Xiao H, Huang M, Zhu X. From one point to a manifold: Knowledge graph embedding for precise link prediction[C]//Proceedings of IJCAI. 2016: 1315-1321.

[26] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of AAAI. 2015.

[27] Guo S, Wang Q, Wang L, et al. Jointly embedding knowledge graphs and logical rules[C]// Proceedings of EMNLP. 2016: 1488-1498.

[28] Lin Y, Liu Z, Luan H, et al. Modeling relation paths for representation learning of knowledge bases[C]//Proceedings of EMNLP. 2015: 705-714.

[29] Wang Z, Zhang J, Feng J, et al. Knowledge graph and text jointly embedding.[C]//Proceedings of EMNLP: volume 14. 2014: 1591-1601.

[30] Miller G A. Wordnet: a lexical database for english[J]. Communications of the ACM, 1995, 38 (11):39-41.

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

[32] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[C]// Proceedings of ECML-PKDD. 2010: 148-163.

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

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

[35] Weston J, Bordes A, Yakhnenko O, et al. Connecting language and knowledge bases with embedding models for relation extraction[C]//Proceedings of EMNLP. 2013: 1366-1371.

[36] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Proceedings of JMLR, 2003, 3:1137-1155.

[37] Tong H, Faloutsos C, Pan J Y. Fast random walk with restart and its applications[C]//Proceedings of ICDM. 2006: 613-622.

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

[39] Lao N, Subramanya A, Pereira F, et al. Reading the web with learned syntactic-semantic inference rules[C]//Proceedings of EMNLP-CoNLL. 2012: 1017-1026.

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

[41] Gardner M, Talukdar P P, Kisiel B, et al. Improving learning and inference in a large knowledge-base using latent syntactic cues.[C]//Proceedings of EMNLP. 2013: 833-838.

[42] Neelakantan A, Roth B, McCallum A. Compositional vector space models for knowledge base inference[C]//2015 AAAI Spring Symposium Series. 2015.

[43] Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4):046115.

[44] Lü L, Zhou T. Link prediction in complex networks: A survey[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6):1150-1170.

[45] Mitchell J, Lapata M. Vector-based models of semantic composition[C]//Proceedings of ACL. 2008: 236-244.

[46] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]// Proceedings of Interspeech. 2010: 1045-1048.

[47] Krompaß D, Baier S, Tresp V. Type-constrained representation learning in knowledge graphs [C]//Proceedings of ISWC. 2015: 640-655.

[48] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of ACL. 2004: 22.

[49] GuoDong Z, Jian S, Jie Z, et al. Exploring various knowledge in relation extraction[C]// Proceedings of ACL. 2005: 427-434.

[50] Jiang J, Zhai C. A systematic exploration of the feature space for relation extraction[C]// Proceedings of NAACL. 2007: 113-120.

[51] Nguyen D P, Matsuo Y, Ishizuka M. Relation extraction from wikipedia using subtree mining [C]//Proceedings of AAAI. 2007: 1414-1420.

[52] Culotta A, Sorensen J. Dependency tree kernels for relation extraction[C]//Proceedings of ACL. 2004: 423.

[53] Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction[C]// Proceedings of EMNLP. 2005: 724-731.

[54] Zhao S, Grishman R. Extracting relations with integrated information using kernel methods [C]//Proceedings of ACL. 2005: 419-426.

[55] Zhang M, Zhang J, Su J, et al. A composite kernel to extract relations between entities with both flat and structured features[C]//Proceedings of ACL. 2006: 825-832.

[56] Mooney R J, Bunescu R C. Subsequence kernels for relation extraction[C]//Proceedings of NIPS. 2006: 171-178.

[57] Zhang M, Zhang J, Su J. Exploring syntactic features for relation extraction using a convolution tree kernel[C]//Proceedings of NAACL. 2006: 288-295.

[58] Wang M. A re-examination of dependency path kernels for relation extraction[C]//Proceedings of IJCNLP. 2008: 841-846.

[59] Roth D, Yih W t. Probabilistic reasoning for entity & relation recognition[C]//Proceedings of ACL. 2002: 1-7.

[60] Roth D, Yih W t. A linear programming formulation for global inference in natural language tasks[C]//Proceedings of CoNLL. 2004.

[61] Sarawagi S, Cohen W W. Semi-markov conditional random fields for information extraction [C]//Proceedings of NIPS. 2005: 1185-1192.

[62] Yu X, Lam W. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach[C]//Proceedings of ACL. 2010: 1399-1407.

[63] Riedel S, Yao L, McCallum A, et al. Relation extraction with matrix factorization and universal schemas.[C]//HLT-NAACL. 2013.

[64] Gormley M R, Yu M, Dredze M. Improved relation extraction with feature-rich compositional embedding models[C]//Proceedings of EMNLP. 2015: 1774-1784.

[65] Liu C, Sun W, Chao W, et al. Convolution neural network for relation extraction[C]//Proceedings of ICDM. 2013: 231-242.

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

[67] Nguyen T H, Grishman R. Relation extraction: Perspective from convolutional neural networks[C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015: 39-48.

[68] Santos C N d, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks[C]//Proceedings of ACL-IJCNLP. 2015: 626-634.

[69] Huang Y Y, Wang W Y. Deep residual learning for weakly-supervised relation extraction[C]// Proceedings of EMNLP. 2017: 1803-1807.

[70] Zhang D, Wang D. Relation classification via recurrent neural network[J]. arXiv preprint arXiv:1508.01006, 2015.

[71] Nguyen T H, Grishman R. Combining neural networks and log-linear models to improve relation extraction[J]. arXiv preprint arXiv:1511.05926, 2015.

[72] Vu N T, Adel H, Gupta P, et al. Combining recurrent and convolutional neural networks for relation classification[C]//Proceedings of NAACL. 2016: 534-539.

[73] Zhang S, Zheng D, Hu X, et al. Bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. 2015: 73-78.

[74] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of ACL. 2016: 207-212.

[75] Xiao M, Liu C. Semantic relation classification via hierarchical recurrent neural network with attention[C]//Proceedings of COLING. 2016: 1254-1263.

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

[77] Xu K, Feng Y, Huang S, et al. Semantic relation classification via convolutional neural networks with simple negative sampling[C]//Proceedings of EMNLP. 2015: 536-540.

[78] Xu Y, Mou L, Li G, et al. Classifying relations via long short term memory networks along shortest dependency paths.[C]//Proceedings of EMNLP. 2015.

[79] Liu Y, Wei F, Li S, et al. A dependency-based neural network for relation classification[C]// Proceedings of ACL-IJCNLP. 2015: 285-290.

[80] Xu Y, Jia R, Mou L, et al. Improved relation classification by deep recurrent neural networks with data augmentation[C]//Proceedings of COLING. 2016: 1461-1470.

[81] Cai R, Zhang X, Wang H. Bidirectional recurrent convolutional neural network for relation classification[C]//Proceedings of ACL. 2016: 756-765.

[82] Bunescu R, Mooney R. Learning to extract relations from the web using minimal supervision [C]//Proceedings of ACL: volume 45. 2007: 576.

[83] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of EMNLP. 2015.

[84] dos Santos C N, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks[C]//Proceedings of ACL: volume 1. 2015: 626-634.

[85] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. JMLR, 2014, 15(1):1929-1958.

[86] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling[C]//Proceedings of ACL. Association for Computational Linguistics, 2005: 363-370.

[87] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances [C]//Proceedings of ACL: volume 1. 2016: 2124-2133.

[88] Cho K, Van Merriënboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv preprint arXiv:1409.1259, 2014.

[89] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

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

[91] Yates A, Cafarella M, Banko M, et al. Textrunner: open information extraction on the web[C]// Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2007: 25-26.

[92] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning[C]//Proceedings of AAAI. 2010.

[93] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding[C]// Proceedings of SIGMOD. ACM, 2012: 481-492.

[94] Ahn S, Choi H, Pärnamaa T, et al. A neural knowledge language model[J]. arXiv preprint arXiv:1608.00318, 2016.

[95] Serban I V, García-Durán A, Gulcehre C, et al. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus[C]//Proceedings of ACL. 2016: 588-598.

[96] Yin J, Jiang X, Lu Z, et al. Neural generative question answering[C]//Proceedings of IJCAI. 2016.

[97] He S, Liu C, Liu K, et al. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning[C]//Proceedings of ACL. 2017: 199-208.

[98] Hasibi F, Balog K, Bratsberg S E. Entity linking in queries: Tasks and evaluation[C]// Proceedings of ICTIR. 2015: 171-180.

[99] Xiong C, Callan J, Liu T Y. Bag-of-entity representation for ranking[C]//Proceedings of ICTIR. 2016: 181-184.

[100] Nguyen G H, Tamine L, Soulier L, et al. Toward a deep neural approach for knowledge-based ir[J]. arXiv preprint arXiv:1606.07211, 2016.

[101] Xiong C, Power R, Callan J. Explicit semantic ranking for academic search via knowledge graph embedding[C]//Proceedings of WWW. 2017: 1271-1279.

[102] Cheekula S K, Kapanipathi P, Doran D, et al. Entity recommendations using hierarchical knowledge bases[J]. 2015.

[103] Passant A. dbrec–music recommendations using dbpedia[C]//Proceedings of ISWC. Springer, 2010: 209-224.

[104] Zhang F, Yuan N J, Lian D, et al. Collaborative knowledge base embedding for recommender systems[C]//Proceedings of SIGKDD. 2016: 353-362.

[105] Le P, Dymetman M, Renders J M. Lstm-based mixture-of-experts for knowledge-aware dialogues[C]//Proceedings of ACL workshop. 2016.

[106] Zhu W, Mo K, Zhang Y, et al. Flexible end-to-end dialogue system for knowledge grounded conversation[J]. arXiv preprint arXiv:1709.04264, 2017.

[107] Huang H, Heck L, Ji H. Leveraging deep neural networks and knowledge graphs for entity disambiguation[J]. arXiv preprint arXiv:1504.07678, 2015.

[108] Fang W, Zhang J, Wang D, et al. Entity disambiguation by knowledge and text jointly embedding. [C]//Proceedings of CoNLL. 2016: 260-269.

[109] Chen M, Tian Y, Yang M, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment[C]//Proceedings of IJCAI. 2017: 1511-1517.

[110] Zhu H, Xie R, Liu Z, et al. Iterative entity alignment via joint knowledge embeddings[C]// Proceedings of IJCAI. 2017: 4258-4264.

[111] Kim A Y, Song H J, Park S B, et al. A re-ranking model for dependency parsing with knowledge graph embeddings[C]//Proceedings of IALP. IEEE, 2015: 177-180.

[112] Zhang H, Kyaw Z, Chang S F, et al. Visual translation embedding network for visual relation detection[C]//Proceedings of CVPR. 2017: 5532-5540.

[113] Baier S, Ma Y, Tresp V. Improving visual relationship detection using semantic modeling of scene descriptions[C]//Proceedings of ISWC. Springer, 2017: 53-68.

[114] Tu C, Zhang Z, Liu Z, et al. Transnet: Translation-based network representation learning for social relation extraction[C]//Proceedings of IJCAI. 2017: 2864-2870.

[115] Chabchoub M, Gagnon M, Zouaq A. Collective disambiguation and semantic annotation for entity linking and typing[C]//Proceedings of SWEC. 2016.

[116] Liu Y, Liu K, Xu L, et al. Exploring fine-grained entity type constraints for distantly supervised relation extraction.[C]//Proceedings of COLING. 2014.

[117] Yahya M, Berberich K, Elbassuoni S, et al. Robust question answering over the web of linked data[C]//Proceedings of CIKM. 2013.

[118] Shimaoka S, Stenetorp P, Inui K, et al. Neural architectures for fine-grained entity type classification[C]//Proceedings of EACL. 2017.

[119] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of NIPS. 2013.

[120] Ling X, Weld D S. Fine-grained entity recognition[C]//Proceedings of AAAI. 2012.

[121] Ren X, He W, Qu M, et al. Label noise reduction in entity typing by heterogeneous partial-label embedding[C]//Proceedings of KDD. 2016.

[122] Yaghoobzadeh Y, Adel H, Schütze H. Noise mitigation for neural entity typing and relation extraction[C]//Proceedings of EACL. 2017.

[123] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]// Proceedings of EMNLP: volume 14. 2014.

[124] Yosef M A, Bauer S, Hoffart J, et al. HYENA: Hierarchical type classification for entity names [C]//Proceedings of COLING. 2012.

[125] Yogatama D, Gillick D, Lazic N. Embedding methods for fine grained entity type classification. [C]//Proceedings of ACL. 2015.

[126] Dong L, Wei F, Sun H, et al. A hybrid neural model for type classification of entity mentions [C]//Proceedings of IJCAI. 2015.

[127] Ren X, He W, Qu M, et al. Afet: Automatic fine-grained entity typing by hierarchical partial-label embedding[C]//Proceedings of EMNLP. 2016.

[128] Chen D, Bolton J, Manning C D. A thorough examination of the cnn/daily mail reading comprehension task[C]//Proceedings of ACL. 2016: 2358-2367.

[129] Dhingra B, Liu H, Yang Z, et al. Gated-attention readers for text comprehension[C]//Proceedings of ACL. 2017: 1832-1846.

[130] Cui Y, Chen Z, Wei S, et al. Attention-over-attention neural networks for reading comprehension [C]//Proceedings of ACL. 2017: 593-602.

[131] Shen Y, Huang P S, Gao J, et al. Reasonet: Learning to stop reading in machine comprehension [C]//Proceedings of SIGKDD. ACM, 2017: 1047-1055.

[132] Wang W, Yang N, Wei F, et al. Gated self-matching networks for reading comprehension and question answering[C]//Proceedings of ACL. 2017: 189-198.

[133] Chen D, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions[C]// Proceedings of the ACL. 2017: 1870-1879.

[134] Dhingra B, Mazaitis K, Cohen W W. Quasar: Datasets for question answering by search and reading[J]. arXiv preprint arXiv:1707.03904, 2017.

[135] Dunn M, Sagun L, Higgins M, et al. Searchqa: A new q&a dataset augmented with context from a search engine[J]. arXiv preprint arXiv:1704.05179, 2017.

[136] Joshi M, Choi E, Weld D, et al. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of ACL. 2017: 1601-1611.

[137] Voorhees E M, et al. The trec-8 question answering track report.[C]//Proceedings of TREC. 1999: 77-82.

[138] Berant J, Chou A, Frostig R, et al. Semantic parsing on Freebase from question-answer pairs [C]//Proceedings of EMNLP. 2013: 1533-1544.

[139] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension [C]//Proceedings of ICLR. 2017.

[140] Buck C, Bulian J, Ciaramita M, et al. Ask the right questions: Active question reformulation with reinforcement learning[J]. arXiv preprint arXiv:1705.07830, 2017.

[141] Wang S, Yu M, Guo X, et al. R3: Reinforced ranker-reader for open-domain question answering [C]//Proceedings of AAAI. 2018.

[142] Jiang T, Liu T, Ge T, et al. Encoding temporal information for time-aware link prediction.[C]// Proceedings ofEMNLP. 2016: 2350-2354.

[143] Esteban C, Tresp V, Yang Y, et al. Predicting the co-evolution of event and knowledge graphs [C]//Proceedings of FUSION. IEEE, 2016: 98-105.

[144] Trivedi R, Dai H, Wang Y, et al. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs[C]//Proceedings of ICML. 2017: 3462-3471.

致　谢

衷心感谢我的导师孙茂松教授。在博士期间，您对我悉心指导，并创造良好的学术氛围、优越的科研环境。您的严谨认真的科研态度一直是我的学习榜样。

衷心感谢我的副指导老师刘知远副教授，在过去 7 年的本科和博士生涯中，您一直是我的良师益友。您在科研生活和日常生活中都给予了我许多的关怀和指导，同时在未来职业发展规划上为我提供了巨大的指导和帮助。

在腾讯微信合作研究期间，感谢周杰、李鹏博士的指导和帮助，同时感谢腾讯微信实习期间共事过的同事们。

感谢刘洋老师、栾焕博老师和其他在科研和生活中曾经给予我帮助的各位老师。感谢我的各位论文合作者沈世奇、曾文远、岂凡超、辛极、姚远、刘家骅、计昊哲，和你们宝贵的合作机会让我感受到了优秀的清华人在科研上的不断追求。感谢自动问答、知识计算研究小组的各位研究生和本科生同学，是你们的共同努力才能让我们研究小组一直向前发展。感谢实验室的秘书王宇星、尹向荣老师，感谢你们在我博士期间在财务报销、日常生活、科研服务器等方面对我提供的帮助。此外，我要感谢实验室所有同学们：赵宇、陈新雄、李莉、沈世奇、柳春洋、阿雅娜、张檬、涂存超、杨成、谢若冰、陈翱、张嘉成、陈慧敏、张菡、刘正皓、丁延卓、梁健楠、矣晓沅、武彬、韩旭、郭志芃、叶德铭、杜家驹、黄轩成、王硕、周界、张凯韬等，感谢你们在我博士生涯对我的帮助和鼓励。

特别感谢我的女朋友郝雪，感谢你在日常对我无微不至的照顾，在我情绪低落时对我的关心和开导，你是我前进的最大动力。

最后感谢我的父亲和母亲，你们是我最踏实的后盾，让我有勇气面对任何艰难险阻。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1991 年 4 月 23 日出生于广东省汕头市。

2010 年 9 月考入清华大学计算机科学与技术系计算机科学与技术专业，2014 年 7 月本科毕业并获得工学学士学位。

2014 年 9 月免试进入清华大学计算机科学与技术系攻读博士学位至今。

发表的学术论文

[1] **Yankai Lin**, Haozhe Ji, Zhiyuan Liu, Maosong Sun. Denoising Distantly Supervised Open-Domain Question Answering. The 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). (**CCF A**)

[2] **Yankai Lin**, Zhiyuan Liu, Maosong Sun. Neural Relation Extraction with Multilingual Attention. The 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). (**CCF A**)

[3] **Yankai Lin**, Shiqi Shen, Zhiyuan Liu, Maosong Sun. Neural Relation Extraction with Selective Attention over Instances. The 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). (**CCF A**)

[4] **Yankai Lin**, Zhiyuan Liu, Maosong Sun. Knowledge Representation Learning with Entities, Attributes and Relations. International Joint Conference on Artificial Intelligence (IJCAI 2016). (**CCF A**)

[5] **Yankai Lin**, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, Song Liu. Modeling Relation Paths for Representation Learning of Knowledge Bases. The Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). (**CCF B**)

[6] **Yankai Lin**, Zhiyuan Liu, Maosong Sun, Yang Liu, Xuan Zhu. Learning Entity and Relation Embeddings for Knowledge Graph Completion. Long paper. The 29th AAAI Conference on Artificial Intelligence (AAAI 2015). (**CCF A**)

[7] Yuan Yao, Deming Ye, Peng Li, Xu Han, **Yankai Lin**, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, Maosong Sun. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). (**CCF A**)

[8] Hao Zhu, **Yankai Lin**, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, Maosong Sun. Graph

Neural Networks with Generated Parameters for Relation Extraction. The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). (**CCF A**)

[9] Shun Zheng, Xu Han, **Yankai Lin**, Peilin Yu, Lu Chen, Ling Huang, Zhiyuan Liu, Wei Xu. DIAG-NRE: A Neural Pattern Diagnosis Framework for Distantly Supervised Neural Relation Extraction. The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). (**CCF A**)

[10] Jiahua Liu, **Yankai Lin**, Zhiyuan Liu, Maosong Sun. XQA: A Cross-lingual Open-domain Question Answering Dataset. The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). (**CCF A**)

[11] Ji Xin, **Yankai Lin**, Zhiyuan Liu, Maosong Sun. Improving Neural Fine-Grained Entity Typing with Knowledge Attention. The 32th AAAI Conference on Artificial Intelligence (AAAI 2018). (**CCF A**)

[12] Zihao Fu, **Yankai Lin**, Zhiyuan Liu, Wai Lam. Fact Discovery from Knowledge Base via Facet Decomposition. The 2019 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT 2019). (**CCF C**)

[13] Fanchao Qi, **Yankai Lin**, Maosong Sun, Hao Zhu, Ruobing Xie, Zhiyuan Liu. Cross-lingual Lexical Sememe Prediction. The Conference on Empirical Methods in Natural Language Processing (EMNLP 2018). (**CCF B**)

[14] Xiaozhi Wang, Xu Han, **Yankai Lin**, Zhiyuan Liu, Maosong Sun. Adversarial Multi-lingual Neural Relation Extraction. The 27th International Conference on Computational Linguistics (COLING 2018). (**CCF B**)

[15] Wenyuan Zeng, **Yankai Lin**, Zhiyuan Liu, Maosong Sun. Incorporating Relation Paths in Neural Relation Extraction. The Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). (**CCF B**)

[16] Ayana, Shiqi Shen, **Yankai Lin**, Cunchao Tu, Yu Zhao, Zhiyuan Liu, Maosong Sun. Recent Advances on Neural Headline Generation. Journal of Computer Science and Technology (JCST 2017). (**CCF B**)

[17] Huimin Chen, Maosong Sun, Cunchao Tu, **Yankai Lin** and Zhiyuan Liu. Neural Sentiment Classification with User and Product Attention. The Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). (**CCF B**)

[18] Stefan Heinrich, Cornelius Weber, Stefan Wermter, Ruobing Xie, **Yankai Lin**, Zhiyuan Liu. Crossmodal language grounding, learning, and teaching. The 30th Conference on Neural Information Processing Systems (CoCo@NIPS'16). Workshop.

[19] Zhiyuan Liu, Maosong Sun, **Yankai Lin**, Ruobing Xie. Knowledge Representation Learning: A Review. Journal of Computer Research and Development, 2016. (In

Chinese)