第 3 章　开放领域的实体关系分析

针对开放领域的实体关系分析问题，常常使用自然语言处理方法对大量的互联网文本进行实体关系抽取，提取的关系三元组来构建知识图谱，其提供结构化的主题信息，可以帮助用户解决知识查询的问题。当使用在对话系统上，可以为理解用户输入的语义信息打下基础[82]，同时可以转化为结构化的知识数据来完善知识库，使其更加智能。而使用自然语言处理技术处理医学领域文本可以推动数字化医疗的快速发展。医学领域的实体关系抽取是对疾病、病症、药物、蛋白质、基因等重要医学实体之间的语义关系（如治疗关系、诱导关系、突变关系）的揭示，是构建领域知识图谱、本体与知识库、临床决策支持系统的重要基础[83]，对进一步辅助智慧医疗与精准医学具有重要现实意义。

目前，实体关系抽取技术大多局限于封闭的、单一领域下的研究，虽然在固定语料上取得了好的效果，但由于现实中关系种类的动态变化和领域的不确定性，现有的模型难以应用于真实场景。因此，为了处理开放的关系种类，模型需要在识别句子中的实体的基础上，进行已知实体关系的分类、检测未知的实体关系、发现新的关系类型并且可以对新发现的关系种类进行持续学习。因此，研究面向开放域的实体关系抽取方法成为一种新的研究方向。

3.1　开放领域的实体关系抽取

早期开放关系抽取（open relation extraction，OpenRE）方法依赖句法或句法模式来提取表面形式关系[84,85]。然而，这些方法的推广在实际场景中是非常有限的。为了克服这一缺点，最新的方法是通过对关系实例进行聚类来抽取开放关系。此外，Wu 等[86]提出了一种关系孪生网络，通过将有标记数据中的关系知识转移到无标记数据中来识别新的关系。最近，Wang 等[87]提出了一个度量学习框架，该框架利用从已知关系实例中获取的丰富监督信号，直接对未知的关系实例进行聚类。这些方法[86,87]利用从已知关系中学习到的语义知识，有效地提高了聚类性能。在借鉴前人研究成果的基础上，本书利用已知关系来指导未知关系实例的模型聚类。

3.2 相关研究方法综述

3.2.1 自监督学习

自监督学习已经在自然语言处理领域得到了广泛应用，大部分预训练模型采用的是自监督学习范式，通过预训练在下游任务上微调使模型得到巨大提升。Devlin 等[20]采用掩码语言模型和下一句预测为自监督学习损失来训练文本词向量表示模型，同时在不同下游任务上微调，最后结果得到巨大提升。Soares 等[88]通过一种匹配空白的关系预训练方法，通过实体连接的文本构建与任务无关的关系表示，并在多个关系抽取任务上微调实现最佳结果。Wang 等[89]为解决训练数据的瓶颈，用多任务方法组合两个自监督的语言不流利检测任务，并在特定任务上微调，通过少量标注数据获得了性能提升。

3.2.2 开放世界分类

实际上，模型需要将已知（可见类的）数据分类到它们各自的类别中，并拒绝/检测来自未知（不可见类）的实例。这个问题被称为开放世界学习（或开放世界分类）[90,91]。初期，研究人员使用传统的基于机器学习的方法来解决这个问题。Scheirer 等[92]提出了一种 1-vs-set 机器，旨在从二元支持向量机的边缘距离构造一个决策空间。Jain 等[93]估计开集问题的非归一化后验概率，并用统计极值理论拟合概率分布。Bendale 和 Boult[91]扩展了最接近类均值分类器，计算未知类中心与已知类中心之间的距离。然而，这些方法需要真实的负样本来选择决策边界或概率阈值。

近年来，基于深度学习的方法因其强大的表达能力而受到越来越多的关注。Bendale 和 Boult[94]首先提出用 OpenMax 代替 Softmax 层，OpenMax 用 Weibull 分布校准输出概率。DOC[95]使用 Sigmoids 的 1-vs-rest 最后一层，并通过高斯拟合收紧 Sigmoid 函数的决策边界来检测未知类。这些方法需要一个阈值来区分已知和未知，同时面临阈值选择的挑战。Oza 和 Patel[96]提出了一种条件自动编码器（C2AE），它使用统计建模的极值理论对重构误差进行建模，并选择阈值来识别已知/未知类的样本。然而，它们需要使用额外的生成模型生成新的训练实例。Zhou 等[97]提出学习开放集问题的占位符，通过为数据和分类器分配占位符来为未知的类做准备。尽管如此，这些方法仍然需要额外的训练参数来检测未知的类。

目前已有的许多关系抽取研究都有一个假设，那就是模型运行在没有开放关系的封闭世界中。Gao 等[98]在小样本关系分类任务中添加了对以上都不是（none-of-the-above，NOTA）关系的检测，这表明查询实例不表达任何给定的关系。但是，NOTA 与开放分类

不同，因为任务设置不同，在测试阶段检测查询实例时它包含多个支持实例。所以本书将研究在开放领域下，使模型既能正确分类已知关系又要检测未知关系。

3.2.3　无监督聚类

许多经典的聚类算法已被应用到关系聚类中，如基于划分的算法[99]、基于密度的算法[100]和基于图的算法[101]。然而，对于高维数据，聚类性能很差，无法在聚类前学习到关系语义特征。最近，一些研究集中在基于深度神经网络的聚类[102-106]，它可以融合聚类和特征学习。Caron 等[103]提出了 DeepCluster 方法，该方法使用 k-means 对特征进行分组，并分配标签作为监督信号，迭代更新网络权值。Caron 等[107]融合了自我监督和聚类，从大规模数据中获取互补信息，并验证了非策划数据的表示学习能力。Zhan 等[104]提出了一种在线聚类框架，该框架同时进行聚类和网络更新，而不是交替进行。本书的重点是学习数据聚类中的良好特征表示，以发现新的关系。

3.2.4　深度度量学习

正如在计算机视觉领域中广泛使用的那样，深度度量学习（deep metric learning，DML）通常用于通过特定的损失函数来学习样本到特征的映射 $\| f(\cdot) \|$。损失函数在 DML 框架中起着至关重要的作用。最近已经提出了各种损失函数。Wang 等[108]提出了一种新颖的损失函数，即大边距余弦损失（LMCL），其引入了一个余弦边际余量来进一步最大化所学习的特征在角度空间的决策边界。Wang 等[109]提出了一种基于集合的排序动机结构化损失来学习判别嵌入。Kim 等[110]提出了一种新的基于代理的损失，它使用基于样本对和基于代理样本点的方法来提高模型的收敛速度。Sun 等[111]提出了圆形损失，它利用重新加权每个相似性来突出未优化的相似性分数，获得圆形决策边界以实现更好的收敛。计算机视觉的成功证明了 DML 在学习特征表示方面的强大潜力。在本书将使用 DML 来学习数据中的语义知识，使模型有效地发现新的关系。

3.2.5　持续学习

现有的持续学习模型主要集中在 3 个领域：①基于正则化的方法[112-113]，通过施加约束项来更新历史任务中重要的神经网络权重从而来缓解灾难性遗忘；②动态架构方法[114-115]，动态扩展模型架构以学习新任务，并有效防止忘记旧任务，然而这些方法不适合 NLP 应用问题，因为模型大小会随着任务的增加而急剧增加；③基于记忆的方法[116-119]，从旧任务中保存一些样本，并在新任务中不断学习它们，以减轻灾难性遗忘。Dong 等[120]提出了一个简单的关系蒸馏增量学习框架，以平衡保留旧知识和适应新知识。Yan 等[121]提出了一种新的两阶段学习方法，该方法使用动态可扩展表示来进行更

有效的增量概念建模。

在这些方法中，基于记忆的方法在 NLP 任务[122-124]中最为有效。受基于记忆的方法在 NLP 领域的成功，启发本书使用记忆重放的框架来学习不断出现的新关系。

3.2.6 对比学习

对比学习(contrastive learning，CL)旨在使相似样本的表示在特征空间中彼此更接近，而不同样本的特征表示应该更远[125]。近年来，CL 的兴起在自监督表示学习方面取得了长足的进步[126-129]。这些方法的共同点是数据集没有可用的类型标签。因此，正负对是通过数据增强形成的。最近，有监督的对比学习[130]受到了很多关注，它使用标签信息来扩展对比学习。Hendrycks 等[131]将监督对比损失与 ImageNet-C 数据集上的交叉熵损失进行比较，并验证监督对比损失对优化器的超参数设置或数据增强不敏感。Chen 等[132]提出了一种用于视觉表示的对比学习框架，不需要特殊的架构或记忆库。Khosla 等[130]将自监督批量对比方法扩展到完全监督设置，它使用监督对比损失学习更好地表示。Liu 和 Pieter [133]提出了一种基于能量模型的混合判别生成训练方法。在本书中，对比学习应用于持续关系抽取，以提取更好的关系表示。

3.3 本章小结

在开放关系检测方面，当前的实体关系抽取方法还未将开放关系检测作为一个具体任务，但在真实的开放域场景下，所获得的语料往往是一些有标注的数据和大量的无标注数据，如何对已知关系进行分类的同时，使模型检测出没有先验知识的未知关系是值得研究的。在开放关系发现方面，现有大部分关系抽取模型假设具有封闭的关系集合，但开放域自然语言语料中包含大量开放的实体关系，而且新关系的数量仍在不断增长中，如何利用深度学习模型自动发现不同领域中实体间的新关系并实现开放关系抽取，实现开放域下的知识发现，是值得深入研究的问题。另外，在实际场景中可能具有某些非专业领域的标注数据，但缺乏对齐的标注数据，而无监督实体关系发现方法所能学习到的信息量是有限的，如何有效利用标注数据中的关系语义知识，帮助模型发现大量无标注数据中的新关系，是实现开放领域关系抽取的关键问题。在实体关系持续学习方面，随着新的实体关系不断出现，现有的模型总是假定一组预定义的关系并在固定的数据集上进行训练，无法很好地处理现实生活中不断增长的关系类型。如何帮助模型学习新关系的同时，保持对旧关系的准确分类，实现模型对实体关系的持续学习，是值得深入研究的问题。另外，目前存在的一些实体关系持续学习方法大多是基于记忆重放的，但随着学习的实体关系的增加，在记忆重放的过程中利用到的语义知识是不充分的，如何更加有效地利用记忆中的样

本，使模型能够保持稳定的学习能力，是实现实体关系持续学习的关键问题。

本书将结合图神经网络、自监督学习、开放世界分类方法、无监督聚类、深度度量学习、持续学习和对比学习来克服以上挑战，进而实现开放域下的实体关系抽取。

本书主要内容安排：本书内容总体论述上，分别从命名实体识别、面向垂直领域的关系抽取和面向开放领域的关系抽取 3 个层次上依次系统化地论述人机对话信息中的命名实体识别与关系抽取问题。本书的第 2 篇从识别文本中的命名实体最基本的问题求解方法开始讨论，分别重点介绍了基于 S-LSTM 构建英文 NER 新的上下文词状态与句子状态表示模型、基于句子语义与 Self-Attention 机制的中文和英文命名实体识别模型，以及针对中文融合了拼音特征与五笔特征的 NER 模型，并深入对比了不同实体识别方法的优劣。在本书的第 3 篇，在实体识别的基础上，重点介绍面向垂直领域的实体关系抽取问题，进一步挖掘实体之间所存在的关系。第 4 篇在垂直领域关系抽取的基础上，进一步探讨了如何针对开放领域所进行实体关系抽取。同时，第 4 篇在上述研究工作的基础上呈现了笔者通过开放共享的方式提供的开放域文本关系抽取实验演示平台，为开展本方面工作的相关人员提供重要的平台支撑。

本 篇 小 结

近年来，对话系统和信息抽取的研究受到学术界和工业界的广泛关注。本篇主要围绕其中关键的命名实体识别、实体关系抽取任务两个方面对当前的命名实体识别方法和实体关系分析方法进行了系统化的介绍。

首先，在命名实体识别方面，分别综述了命名实体识别的相关工作，以及用于实体识别任务的相关深度学习方法。通过对命名实体识别算法进行基本概述，将本书的工作方法与已有的研究工作进行对比，并指出其存在的问题，针对相应的问题提出了相应的解决方法。

其次，在将句子或文档进行命名实体识别的基础上，紧接着就是进行实体关系抽取，即将文本中包含的知识三元组进行抽取。根据垂直领域关系抽取可以分为远监督关系抽取、小样本关系抽取、文档关系抽取及实体和关系的联合抽取。针对当前方法的不足，本书融合了卷积神经网络、图神经网络、对抗训练等策略，提升了模型的鲁棒性和抽取性能。

最后，在真实的开放领域，关系种类和数量在不断增长。因此，本篇面向开放领域的关系抽取进行综述性介绍，并详细分析了每种方法的适用场景。针对当前面向开放领域下关系抽取方法存在的不足，本书将探究如何有效利用自监督学习、持续学习、对比学习机制，引导模型可以对开放关系进行检测、发现及持续学习的方法。