实体关系抽取技术研究进展综述*

刘　辉[1], 江千军[2], 桂前进[2], 张　祺[3], 王梓豫[3†], 王　磊[2], 王京景[2]
(1. 国网安徽省电力有限公司, 合肥 230009; 2. 国网安徽省电力有限公司安庆供电公司, 安徽 安庆 246000;
3. 合肥工业大学 电气与自动化工程学院, 合肥 230009)

摘　要:实体关系抽取是指从句子中抽取两个实体之间的关系类别的任务。作为自然语言处理的关键性技术,实体关系抽取在信息检索、知识图谱、自动问答系统等领域具有广阔的应用前景。对于实体关系抽取研究历程作出了详细评述,包括从传统的实体关系抽取到目前基于深度学习的实体关系抽取。重点阐述了基于深度学习的实体关系抽取的主要模型以及流程框架,并对实体关系抽取存在的技术难点加以总结,最后对实体关系抽取的发展进行展望。

关键词:实体关系抽取; 有监督方法; 无监督方法; 开放领域实体关系抽取; 深度学习

0　引言

随着信息时代的到来,互联网中包含着海量的信息,信息数据以指数形式爆炸增长,且模态多样化。多数的信息属于非结构或半结构数据,无法被计算机系统直接利用构建知识图谱。读者在获取信息的过程中存在信息过载、资源迷向等问题,在搜索信息时仍需费时费力筛选才能获得有效信息,因此,如何快速简洁地提取出对读者有效的信息变得愈加重要。

自然语言处理(natural language processing,NLP)是从非结构或半结构化的文本中抽取文本中特定事件,并结构化处理特定事件,进而被计算机识别利用的一种机器技术。自然语言处理过程可粗略分为分词与词性标注、命名实体识别、实体关系抽取。实体关系抽取(relation extracion,RE)作为自然语言处理的子任务之一,目的是在命名实体识别的基础上从文本中抽取实体之间存在的各类关系类别,构成〈实体 1,关系类型,实体 2〉的结构化形式,实现语义关系的自动抽取和查询匹配。例如"首批医护人员分乘 11 架大中型运输机抵达武汉。"中的实体"医护人员"与实体"武汉"拥有"抵达"的关系。实体关系抽取一直是经典而又富有挑战性的任务,在过去二十多年的研究发展下取得了很多阶段性的突破。目前实体关系抽取的研究成果主要应用在知识图谱构建、自动问答系统、机器翻译和海量文本摘要等领域。

从早期基于模式匹配的关系抽取到后来基于机器学习的关系抽取,实体关系抽取得到了广泛的关注。目前随着以深度学习为基础的人工智能潮流席卷全球,自然语言处理也取得了突破进展。深度学习下实体关系抽取有效改善了传统标注工具的自身缺陷,取得了良好的效果,并成为近些年研究的热点与关键。然而实体关系抽取至今仍面临许多挑战,如实体语义关系的复杂性、句与句之间实体关系的模糊性、数据规模不足与模型学习能力的冲突等都制约着实体关系抽取的发展。

目前已有一些综述性文章都对目前的实体关系抽取方法进行了总结归纳[1~5]。例如:文献[1]中对实体关系抽取传统 NLP 标注方法作出了完整的综述;文献[2]介绍了基于深度学习的实体关系抽取方法,并进行比对;文献[3]对基于深度学习的实体关系抽取的基本框架和流程进行了介绍。但是这些综述性文章主要集中对实体关系抽取的部分传统方法所使用的数据集或者只对深度学习模型部分性的讲解,缺乏对于实体关系抽取的历史作出整体综合性的评述,同时一些崭新的实体关系抽取数学模型、流程框架没有被归纳对比。本文将从实体关系抽取主流的分类角度详细梳理现有的实体关系抽取方法,对方法的优缺点进行阐述,着重对深度学习流程框架和数学模型介绍,并对目前方法共有的不足和实体关系抽取未来发展进行了探讨。

1　实体关系抽取的研究现状

实体关系抽取作为信息抽取领域的子任务,得到了国内外学者的广泛关注。在信息抽取领域相关学者不断钻研探索下,实体关系抽取的难题被解决。实体关系抽取任务所遵循的方法可被归纳为有监督实体关系抽取方法、半监督实体关系抽取方法、无监督实体关系抽取方法等。这些方法被广泛应用于知识图谱构建、自动问答系统、机器翻译、海量文本摘要等领域。下面对几种方法进行详细介绍。

1.1　基于模式匹配的实体关系抽取

关系抽取早期的研究是通过模式匹配来完成识别实体的语义关系,实现关系抽取任务。模式匹配的过程主要运用了语言学和自然语言处理学的知识。在关系抽取任务之前,通过人工构造实体关系的特征词典或规则,并将它们存储下来。在实体关系抽取任务中,将规则与预处理后的非结构文本相匹配,提取出了三元关系组。

例如 Miller 等人[6]采用了对实体信息词汇化和概率分布的上下文无关的语法解析器,生成规则用于关系抽取。吴明智等人[7]对生物医学中使用基于模式匹配的方法来关系抽取进行了系统论述。邓擘等人[8]在使用模式匹配技术的基础上引入了词汇语义匹配技术对汉语实体关系进行提取,提出了一种新的汉语实体关系提取技术,经实验表明性能优于单独使用模式匹配的抽取方法。

然而这种方法存在明显的缺陷:要求规则与词典的制定者有本专业较高的语言学基础,同时对自然语言处理学有深入了解与研究;制定规则难度大,耗费大量的时间与人力,可迁移性差,无法直接植入至其他领域。为此学者尝试采用统计机器学习方法,通过对实体间语义关系进行算法建模,替代预定义的语法规则和语义词典。

1.2　基于机器学习的实体关系抽取

机器学习的核心是"使用算法建模并处理数据从中学习,然后对将来某特定事件作出决定或预测"。机器学习可以看做映射,输入为标注语料,输出为期望结果。同时这个映射不仅仅对标注语料有很好的处理能力,而且对任意语料也能够有较好的映射效果,这一过程称为泛化。基于机器学习的实体关系抽取根据对人工语料的依赖程度,可分为有监督的实体关系抽取、半监督的实体关系抽取、无监督的实体关系抽取、面向开放领域的实体关系抽取和远程监督的实体关系抽取。

1.2.1　有监督的实体关系抽取

有监督的抽取方法是实体关系抽取最经典的方法,其核心思想是对机器学习模型投入足量的已标记关系类别的训练语料,然后进行特定关系的匹配识别与抽取任务。有监督的抽取方法包括基于特征向量的方法和基于核函数的方法。

基于特征向量的方法核心思想是从实例句子上提取有关句子的特征如语法信息、词法信息,构造特征向量,从而使用计算特征向量的相似度的方法训练实体关系抽取模型。基于特征向量的实体关系抽取任务分为特征选择、特征权重的选取、分类器的选择三个关键步骤。将训练语料按照语法信息、词法信息等特征项赋予对应的特征值,然后选取恰当的分类器将相似程度高的实体三元组赋予同一语义关系,通过训练后得到的机器学习模型对非结构化文本执行任务。

Kambhatla[9]提出了一种使用最大熵模型结合不同的词汇、句法和语义特征的方式来构建训练模型,在 ACE RDC2003 数据集中实验测试表明能达到较好的拟合效果。车万翔等人[10]对实体关系抽取任务分别采用 Winnow 和支持向量机两种模型,以 2004 年 ACE 数据集为测试语料,Winnow 和 SVM 模型的 F_1 值分别为 73.08% 和 73.27%。实验结果表明,当选择每个实体的左右两个词为特征时,达到最好的抽取效果。黄鑫等人[11]提出采用中文词法、实体、句法、语法基本特征组合的方式进行关系抽取任务,以

收稿日期: 2020-03-29; **修回日期**: 2020-05-28　　**基金项目**:国网安徽省电力有限公司科技资助项目(5212D018008X)
作者简介:刘辉(1978-),男,新疆焉耆人,高级工程师,博士,主要研究方向为实体关系识别;江千军(1972-),男,安徽安庆人,高级工程师,主要研究方向为实体关系抽取;桂前进(1975-),男,安徽潜山人,高级工程师,主要研究方向为实体关系识别;张祺(1999-),男,四川成都人,主要研究方向为命名实体识别;王梓豫(1999-),女(通信作者),四川成都人,主要研究方向为命名实体识别、实体关系抽取(sandoraivy@163.com);王磊(1976-),男,安徽六安人,高级工程师,硕士,主要研究方向为实体关系识别;王京景(1983-),男,安徽淮南人,工程师,博士,主要研究方向为实体关系识别.

ACE2005 的中文语料为测试数据，F_1 值达到 72.77%。甘丽新等人[12]提出一种基于句法语义特征的实体关系抽取方法，融入了句法关系组合特征和句法依赖动词特征特征项，采用 SVM 分类器的方法，实验表明有较高的性能。

上述抽取方法能够较好地实现实体关系抽取，但对句子特征项的选取和判别方面需依靠人的主观经验与判断。Cristianini 等人[13]证明对于指定语义关系无法用有限个特性项表示，形成有限维的特征向量。然而特征项的选择组合是有限的，其性能提升有限，为此提出了基于核函数的实体关系抽取方法。

基于核函数的方法能够充分利用上下句远距离的特征，提高了语义关系识别的能力，同时能够利用特性间的先后顺序和结构等信息，较好地解决了表达词性和语义信息不明的问题。核函数是将原始输入空间映射到一个新的空间，避免了对特征向量的直接计算。Zelenko 等人[14]率先在文本浅层解析描述上定义了核函数及其算法，并将核函数与向量机结合抽取人员隶属关系和组织位置的关系实例，与基于特征的算法进行了比较，证明了基于核函数的方法能够开发挖掘出特征集合进行关系抽取。其他学者在其基础加以改善创新，提出了包括序列核函数抽取方法[15]、依存树核函数方法[16]、最短路径依存树核函数方法[17]、卷积树核函数方法[18]以及它们的组合核函数方法[19]等。表 1 给出了各核函数方法的核心公式、数据集及其评测标准。

表 1 核函数评测标准和评测

抽取方法	评测标准	评测/%	抽取方法	评测标准	评测/%
词序列核函数	F_1	60.8	卷积树核函数	F_1	75.3
依存树核函数	F_1	62.4	组合核函数	F_1	74.6
最短路径依存树核	F_1	68.9			

核函数的核心思想是以树作为基础对象，通过计算树子成分之间的相似程度来进行关系抽取。然而核函数也有它自身的缺陷，随着性能要求的提升，核函数的复合更为复杂，导致了训练语料和测试速度过慢，对于大规模数据处理能力差。

1.2.2 半监督的实体关系抽取

半监督的关系抽取方法的主要思想是根据人为预先设计好的关系类型，通过人工添加合适的少量的实体对作为训练语料，利用模式学习方法进行不断迭代学习，人工进行调整，最终生成关系数据集和序列模式，在一定程度上降低了对于人工标注语料的依赖。目前基于半监督的实体关系抽取任务包裹自举方法(bootstrapping)、标注传播算法(label propagation)、协同训练(co-training)和主动学习(active learning)方法。

Brin[20]首先使用了基于 bootstrapping 的方法在 Web 中实现对实体关系的抽取。该方法首先以少量的书名与作者作为初始种子，从文档和语句中抽取新的实例并作为标注样本，根据这样的二元组建立新的抽取模板，利用建立的模板再去更多的文本库中发现新的实体关系，最终建立 Dirpe 系统得到领域关系实例和序列模式。张立邦[21]运用 bootstrapping 模型在中文电子病历的实体关系挖掘中，通过半监督实体关系抽取的方法分析和挖掘电子病历，从中获得大量与患者密切相关的医疗知识。陈锦秀等人[22]在 bootstrapping 的基础上提出了一种基于图的半监督学习算法(标注传递算法)，该方法利用图策略模型替代关系抽取的模型，克服了 bootstrapping 模型中标记数据缺乏时部分实体特征重合带来的弊端，实现全局一致性的目标，提高分类的准确度。Zhang[23]基于协同训练思想的 BootProject 方法在第十三届 ACM 会议上被提出，该方法证实了半监督 SVM 分类器使用不同的词汇和句法特征可以提高分类精度，且提出的基于随机特征投影的 bootproject 算法在只牺牲有限性能的情况下大大减少了对标记训练数据的需求。徐庆伶等人[24]在此基础上进行了改进，结合 co-training 与 tri-training 算法的思想。该方法采用两个不同参数的 SVM 分类器对非结构化样本进行标记，选取置信度高的样本加入到已标记样本集中，进行迭代学习和调整。融合主动学习的半监督学习方法被刘建峰等人[25]提出。该方法采取了融合主动学习并改进了 Bayes 算法，避免了由于被动接受数据而带来的分类效果不理想的问题，得到了很好的分类效果。

半监督式实体关系抽取虽然避免了耗时和繁琐的大量人工标注语料，仅需少量训练语料即可迭代学习构建模型，但对初始标记数据的质量要求较高，并且迭代过程中模板的构建和优化对最后的效果有着至关重要的作用。这种方式普遍存在于迭代学习中噪声引入，进而在不断迭代过程中造成语义漂移的现象；该方法虽然准确率有所提高，但是召回率普遍不高。

1.2.3 无监督的实体关系抽取

无监督的实体关系抽取无须进行人工标准数据，它首先利用某类聚类算法将实体上下句相似程度高的实体对聚成一类，然后选择频率最高的指代词作为该实体对的语义类别。常用的聚类算法有 K-均值、自组织映射聚类算法、遗传算法。

K-均值聚类的算法是由 Kanungo 等人[26]提出的一种十分经典的算法。算法模型能够通过迭代把非结构化文本划分到不同的簇中，使簇内部对象之间的相似度很大，而簇之间对象的相似度很小，实现了先聚类再分类的目的，并选择具有代表性的词语来标记簇集。Kohonen[27]提出了一种自组织映射聚类算法(self-organizing map algorithm)，其核心思想通过网络训练把相类似的输入映射到同一个输出节点上，从而实现对输入数据的聚类。Frey 等人[28]2007 年在 *Science* 上提出了一种仿射传播聚类算法(affinity propagation)，其核心思想是通过迭代过程不断更新每一个点的吸引度和归属度值，直到产生 m 个高质量的聚类中心 exemplar，同时将其余的数据点分配到相应的聚类中。

王娜等人[29]针对传统的欧氏距离测度无法正确描述数据的全局一致性的问题，引入流形距离作为相似度度量，不仅满足了数据的局部一致性假设，而且满足了全局一致性假设，因此对于复杂结构聚类问题更能准确地描述其数据结构。设计同时满足类内相似度大、类间相似度小的聚类准则函数，更好地反映了聚类目标，结果具有确定性，并能取得较好的聚类效果。傅景广等人[30]于 2004 年提出了基于遗传算法的聚类分析方法，遗传算法模拟生物进化的过程，具有很好的自组织、自适应和自学习能力，在无监督实体关系处的抽取任务可以应用。该算法的主要思想是将多维空间中的特征向量按照它们之间的某种距离度量划分为若干个集台，使相同集合中的特征向量之间的距离较为接近。Gong 等人[31]提出了一种基于流形距离的不同度量方法，并将其应用于人工免疫系统的实体关系抽取中。将人工免疫每一抗体编码为代表簇，并利用动态算法从组合优化的角度搜索最优聚类代表。经测试后与 K-均值算法、Maulik 提出的基于遗传算法的聚类和流形距离进化聚类算法相比，该算法具有识别复杂非凸聚类的能力。公茂果等人[32]在文献[31]的基础上提出了一种流形距离作为相似性度量测度算法，经大量人工数据集测试结果显示，与标准的 K-均值算法、基于流形距离的进化聚类算法以及 Maulik 等人提出的基于遗传算法的聚类算法相比有较强的鲁棒性以及较高的准确率。

无监督的实体关系抽取虽然无须人工标注的训练语料，无须预先定义实体关系类型，可迁移性能力强，适合处理大规模的非结构化自然语言文本数据，但无监督式抽取需要事前定义聚类的阈值，同时无监督的实体关系抽取暂无客观的评价标准，且召回率和准确率与有监督抽取方法相比普遍低 10% 左右。

1.2.4 面向开放领域的实体关系抽取

面向开放领域的实体关系抽取既没有对文本类型做要求，也没有人工预先标注语料，同时也不需要知道抽取哪些实体关系，该方法通过前后相邻的短语进行实体关系上的语义构建，借助外部大型实体知识库实现自动实体关系抽取，将非结构化文本数据转换为具有实际意义的事实性命题。面向开放领域的方法包括有句法模式学习技术、自学习技术、句子分解技术。

Corro 等人[33]对于具有复杂关系的自然语言文本提出了 ClausIE的方法。ClausIE 利用有关英语语法的语言知识，首先检测输入句子中的分句，然后根据每个从句成分的语法功能识别每个从句的类型。基于这些信息，ClausIE 能够生成高精度的实体关系三元组。ClausIE 获得了比先前学者的方法更高的召回率和更高的准确率，无论是在高质量的文本上，还是在网络的噪声文本上。姚贤明等人[34]在中文领域的开放领域关系抽取取得了一定进展，提出基于句法分析的抽取方法。该方法对基于句法分析的结果作为根节点，进行迭代逐渐逼近正确的所有谓语的主语、宾语及其定语成分，对最后的结果进行完善，最终获取句子中的多个实体之间的语义关系，结果表明具有一定的参考价值。郭喜跃[35]针对开放领域的中文新闻文本、百度文本和学术期刊文献提出了一种弱监督的实体关系抽取方法，该方法从内容过滤、同义词项合并等角度对其内容进行加工、整合，得到实体关系三元组〈实体 1，关系，实体 2〉，取得了较好的实验结果。李颖等人[36]提出了一种基于依存分析的中文开放式多元实体关系抽取方法。首先对文本依存关系分析；然后将动词视为候选关系词，将有效依存路径的基本名词短语视为实体词，关联两个及两个以上的实体词的关系词可与实体词组成候选多元实体关系组；最后使用分类器对多元实体关系组进行过滤。该方法在百度百科数据集测试中抽取实体关系准确率可达到 81%。

王斌等人[37]提出基于远程监督的领域实体属性关系抽取的混合方法，提取词性特征依存关系特征和短语句法树特征，并进行融合训练关系抽取模型，实验表明三种特征融合的 F 值较高，抽取性能有所提高。

虽然面向开放领域的实体关系抽取存在无须预先标注语料、定义语义关系等优点，但由于目前对于面向开放领域的实体关系

抽取的客观评价标准还没有形成,测试数据集、实体知识库的调用呈多样化形式,同时测试的数据都是经过爬虫技术以及数据清洗过后得到的干净数据,如何将其真正用于互联网中杂乱的非结构数据是其日后发展的关键。

1.2.5　远程监督的实体关系抽取

为了降低对人工语料的依赖性,同时提高抽取模型的可迁移能力,相关学者在半监督式实体抽取基础上提出了基于远程监督的实体关系抽取。远程监督式通过知识语料库与未处理的文本自动对齐,即当文本中〈实体 1,关系,实体 2〉实体关系三元组与资料库某个实体对完全一致时,就对这个三元组打上同类标记,以此自动生成大量训练样本,从而生成特征训练的分类器,降低了对人工预料的依赖。

Mintz 等人[38]在 2009 年提出了远程监督的方法。该方法结合了监督 IE(在概率分类器中结合 40 万个含噪模式特征)和无监督 IE(从任意领域的大语料库中提取大量关系)的优点,借用 freebase 大型语义数据库提供远程监督,自动对齐文本并匹配实体关系,其准确率为 67.6%,并且对于歧义关系或表达方式上的词汇距离关系尤其有帮助。余小康[39]针对 DSRE 中训练数据存在大量标注错误的问题,提出一种基于从句识别的去噪算法(NRCI)。实验表明 NRCI 算法可以有效地降低训练数据中的错误标注数据,进而显著地提升远程监督关系抽取的准确率。黄蓓静等人[40]提出了一种利用句子模式聚类及模式评分对远程监督人物关系抽取过程训练集进行去噪的方法。该方法首先利用词向量生成特定关系描述候选词,然后针对关系描述候选词提取句子模式并进行模式聚类,最后对模式聚类结果进行评分。通过筛选评分较低模式对应句子,去掉对关系描述能力不强甚至无法描述关系的句子,得到过滤后的训练集。实验证明利用该方法可以取得 3% ~5% 的准确率。黄杨琛等人[41]提出了一种可以对远程监督自动生成的训练数据去噪的人物实体关系抽取模型。在训练数据生成阶段,通过多示例学习的思想和基于 TF-IDF 的关系指示词发现的方法对远程监督产生的数据进行去噪处理,使训练数据达到人工标注质量。在模型分类器中提出采用词法特征与句法特征相结合的多因子特征作为关系特征向量用于分类器的学习。余小康等人[42]提出一种改进的半监督集成学习算法 ETT(extented-triaining)。ETT 将 DSER 方法中使用的训练数据作为标注数据,未使用的负例数据作为未标注数据,从而可以利用更加丰富的特征来获得更好的分类边界,取得了更高的分类准确率。

远程监督的方法在标记语料时存在着两个问题导致准确率和召回率偏低:a)由于远程监督的假设过于严格,使部分实体之间强加因果导致错误;b)NLP 工具在语料标注、特征提取过程中本身存在错误,导致误差累积。表 2 所示是对上述四种实体关系抽取方法进行比较。

表 2　实体关系抽取方法对比

实体关系抽取方法	优点	缺点	性能提升方法
有监督	召回率和准确率高,在商业化应用中大多仍采用基于监督方式的抽取	严重依赖于人工语料标注	改进特征项、核函数的选择;完善规则的制定
半监督	少量标记语料	对标记预料质量要求高,否则会出现语义飘移现象	挖掘特征项提高性能,降低迭代噪声
无监督	无须人工标注的训练语料,迁移性能力强	普遍准确率偏低	改进特性模式,优化聚类算法
面向开放领域	不需要标注语料和实体关系,适合处理大规模数据	缺乏客观评价标准;测试数据集、实体知识库的调用呈多样化形式	改善文本的噪声过滤算法

2　基于深度学习的实体关系抽取

传统的自然语言处理方法往往依赖于传统标注工具,而传统标注工具(NLP)本身往往存在少量标注错误,错误在后续的迭代过程中会造成严重的影响,降低了算法性能。深度学习下的神经网络模型可通过输入低维、连续的词向量组合形成更抽象的高维向量表示语义类别,能较好地发现数据的最佳特征。同时神经网络模型无须预先特征选择和抽取,降低了特征项选取不适造成的效果不好。近几年深度学习(deep-learning)在语音和图像处理上得到了广泛关注和深度研究,考虑到语音、图像和文本处理机理的相通性,很多学者尝试将深度学习引入实体关系抽取中来。图 1 是深度学习下实体关系抽取的流程框架。首先通过人工标注语料或自动对齐远程知识库获得有标签的文本语料,接着对已标注语料使用 word2vec 模型,用词向量、位置向量、语法关系向量等方式表征词的语义信息,特征向量作为神经网络的输入。特征提取后经 softmax 进一步权重语义特征,最终输出实体关系对。下面对几种主流的基于深度学习的实体关系抽取进行详细介绍。

递归神经网络模型(图 2)是最早提出的用于实体识别的深度学习模型。递归神经网络中的基本神经单元存在前向通路和反馈通路。递归神经网络模型在每个时间节点上都有一个词向量输入 xt,然后根据当前节点的状态 ht 计算输出值 yt,而 ht 是根据上一节点状态 $ht-1$ 和当前输入 xt 共同决定的。此过程将一直继续,直到所有时间步骤都被评估完成。Socher 等人[43]首次将矩阵—递归神经网络模型(MV-RNN)应用到自然语言处理中,其核心思想是对测试语料使用句法依存分析使文本的原本顺序转变成解析树结构,在此基础上对解析树的每一个节点分配一个向量和一个矩阵,向量包含单词的内在含义,矩阵确定其如何更改其相邻单词或短语的含义。该模型有效地解决了基于单词向量模型无法捕捉到较长的短语或句子组成意义的问题。其缺点在于解析树每个节点都需要设置训练参数,同时解析树依存于句法依存分析,一旦句法解析原理错误或句子的复杂度过高都会影响后续分类的性能。RNN 可以通过相应变换处理任意时序的序列信息,具有学习任意长度的各种短语和句子的组合向量表示的能力,而利用相应算法训练模型时容易出现梯度爆炸的问题,当输入序列需要长时间标注时问题尤其严重。在连续时间添加实体输入,导致靠前的实体对靠后的感知力下降,分类性能下降。RNN 的一个特例 LSTM 能较好地解决这一问题。实际中并不是一个神经单元的隐藏层而是一个包含多层神经单元的隐藏层,导致训练周期较长,CNN 的提出有效解决了该问题。

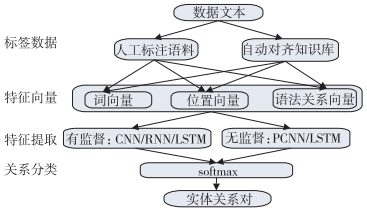

图 1　深度学习实体关系抽取框架

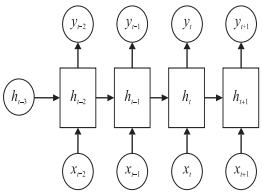

图 2　递归神经网络模型(RNN)中基本神经单元

卷积神经网络模型首先被用在图像方面。卷积神经网络模型首先在卷积层上利用预设的权重滑动矩阵窗口,在训练词向量上滑动窗口,最终输出含有语义特征项的输出矩阵,在提取特征的同时降低了网络总体待训参数的总量;接着在最大池化层(max pooling)上继续选择特征,降低计算难度;最后通过全连接层(fully connected layers)输出期望结果。全连接层相当于卷积神经网络模型的分类器。卷积神经网络在实体关系抽取上有很好的效果,但无法对长句子分析建模。Liu 等人[44]首次提出关系抽取中使用卷积神经网络,该方法结合词汇特征,并通过同义词词典对输入词进行编码,将语义知识集成到神经网络中。在 ACE2005 数据集上的较核函数方法提高了 9% 的 F_1 值。如图 3 所示,Lai 等人[45]首次将目标实体与句子其他词汇特征输入神经网络模型(DNN)得到语句级特征,将两种特征混合为最终向量输入 softmax 层分类。该方法在 SemEval-2010 Task8 关系分类任务取得了最佳效果。Xu 等人[46]在 Zeng 的基础上提出了基于最短依存路径的卷积神经网络,同时采用了负采样策略以解决实体对相对位置过远时依存分析树引入的噪声信息。

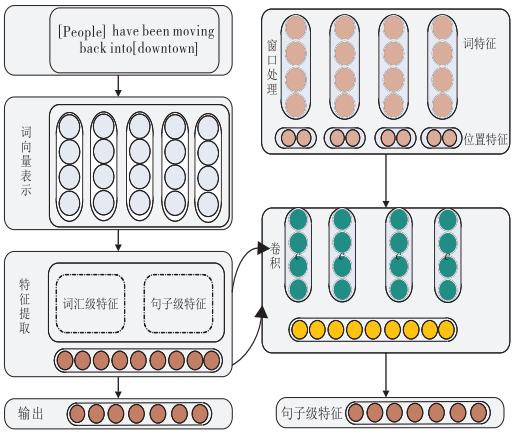

图 3　卷积深度神经网络模型(DNN)实体关系抽取框架

长短期记忆网络模型(LSTM)与递归神经网络模型(RNN)总体框架一致,既有前向通路传递信息,又有自馈通路处理信息,但

是 LSTM 不同点在于允许每个神经单元遗忘或保留信息,在一定程度上解决了 RNN 因输入序列需长时间标注导致的梯度消失梯度爆炸。LSTM 一个神经单元包括遗忘门、输出门、输出门三部分。tanh 函数主要用来调节流经网络的值,使其一直约束在 -1 与 1 之间。sigmoid 将取值约束在 0 与 1 之间,这个特点可以帮助保留和移除信息。如果为 0 则移除;为 1 则保留。当前时刻单元通过接收前一个神经单元的隐藏状态、信息状态和当前输入词向量,最终输出当前单元的信息状态和隐藏状态。遗忘门决定对前一个神经单元信息的遗忘程度,输入门决定对其哪一个神经单元信息的保留程度,输出门决定更新后的信息和隐藏状态。Sundermeyer 等人[47]提出了 RNN 的改进模型长短期记忆网络,通过三个门控操作及细胞状态改善这些问题,使得每个节点的实体都保存了之前的信息,有效地解决了长距离实体对依赖的问题。Sun 等人[48]提出了一种利用长短期记忆网络单元双向结构学习最短依存路径的表示信息,并对 LSTM 的输出使用卷积神经网络训练分类模型,此方法充分地利用两种模型的优点,提高了性能。递归神经网络模型中基本神经单元如图 4 所示。

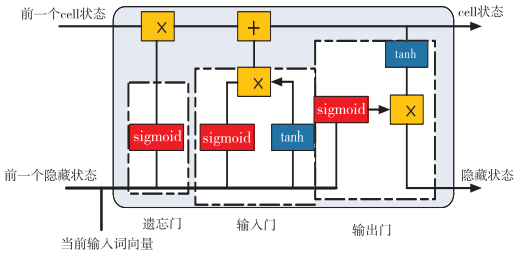

图 4　递归神经网络模型中基本神经单元

流水线方式的关系抽取任务可能出现以下的错误迭代:实体识别模块中实体关系对错误提取导致关系分类的性能下降;忽视了命名实体识别(NER)与实体关系抽取(RE)之间存在的关系,导致原数据信息丢失,影响最终的抽取效果;识别与抽取之间都会标记语料,多次迭代得到 SVM 分类器,会产生冗余信息,同时实体关系抽取进行两对实体对匹配生成〈实体 1,关系,实体 2〉,没有关系的实体也会带来冗余信息因而降低效率。为了缓解这些缺陷,提出了联合学习方法也称为端到端的抽取方式。该方法能实现 NER 与 RE 融合,直接得到实体三元组,实现标记语料,特征选取共享,迭代调整模型共存,降低了流水线模式可能带来的错误累积、信息冗余。

联合学习方法又可以分为参数共享方法和序列标注方法。参数共享方法分别对实体和关系进行建模,而序列标注方法则是直接对实体关系三元组进行建模。Miwa 等人[49]提出了一种新的端到端的神经网络模型用来获取文本数据的实体和实体关系。使用了双向 LSTM 结合树型结构 LSTM(tree-LSTM)获取单词序列和依赖树子结构信息。双向 LSTM 用来实体识别,tree-LSTM 用来实体抽取。前者的 word embeddings 和 label embeddings 可作为后者的输入,实现参数的共享。通过实体预处理和预定抽样进一步提高抽取性能。Zheng 等人[50]提出了新的标注方法来解决联合实体关系抽取,很好地解决了参数共享方法带来的实体关系冗余问题,同时在神经网络模型中加入了偏置损失函数,该函数的利用有效增强了有效实体对的关系,强制忽略无效实体标签的作用。该方法在公开的数据集上得到了很好的测试效果。

远程监督的实体关系抽取需要事先预定特征项,在多次迭代后导致性能下降。随着深度学习在有监督领域中的应用,利用字向量、词向量代替实体特征向量,利用神经网络模型提取句子向量进行分类,能很好地解决该问题。Zeng 等人[51]首次提出了将深度学习与远程监督联合用于实体关系抽取任务,在卷积神经网络模型的基础上提出了 PCNN 模型。该模型将实体对所在句子切分三段分段池化,避免了对隐藏层节点过度忽略,从而能够获得更多的实体对上下文信息,提高模型的准确率和召回率。同时在输入方面构建词嵌入特征向量时采用了词向量、位置向量多种特征向量。PCNN 这一模型被后续的学者广泛采用。针对远程监督中可能存在噪声引入的问题,Ji 等人[52]使用了多实例(MIT)的方法,将实体对看做包,在包含同一实体对的所有句子中选择语义关系概率最大的指示词作为该实体对的语义指示词,在降低噪声的同时丢失了大量有效信息。He 等人[53]在 Zeng 的基础上结合了注意力机制,对包内句子分配权重能够表示对于指定实体关系该句子的重要程度,最终正确标签语义关系分配权大,贡献率高;错误标签语义关系分配权小,贡献率低。该模型在保证充分利用包内信息的同时降低了噪声的影响。

表 3 给出了上述几种典型的深度学习方法对比。除了以上几种典型的深度学习方法,随着置信网络、生成对抗网络、强化学习在深度学习中的不断应用,目前已有学者探索基于深度置信网络、生成式对抗网络、深度强化学习的实体关系识别技术,这也是未来实体关系识别技术的研究热点与方向。

表 3　基于深度学习的实体关系识别模型对比

对比项		模型	特点
基于深度学习的有监督实体关系抽取	流水线方法	MV-RNN	使文本的原本顺序转变成解析树结构，解决了基于单词向量模型无法捕捉到较长的短语,但参数计算量大,难以解决大规模数据
		CNN+词汇特征	结合词汇特征,并通过同义词词典对输入词进行编码,将语义知识集成到神经网络中
		CNN+SDP	采用基于最短依存路径的卷积神经网络,并利用负采样策略降低噪声
		LSTM	通过采用 LSTM 使神经单元保存上一时刻的信息,有效地解决了长距离实体对依赖的问题
		LSTM+SDP+CNN	利用长短期记忆网络单元双向结构学习最短依存路径的表示信息,并对 LSTM 的输出使用卷积神经网络训练分类模型
	联合学习方法	参数共享	实体识别与关系抽取共享编码层的 LSTM 单元序列
		序列标注	将联合学习模型转换为序列标注问题,有效降低冗余信息,避免复杂工程
基于深度学习的远程监督实体关系抽取		PCNN+多示例	将实例句子三段池化以得到更多上下文信息,通过多实例学习以词袋形式进一步权重实体对的表示词
		PCNN+attention	将实例句子三段池化以得到更多上下文信息,结合注意理机制进一步降低噪声,提高准确率

3　实体关系抽取的挑战和趋势

随着深度学习的引入,实体关系抽取任务取得了突破性的进展,同时也带来了相应的挑战:a)在不同神经网络模型的测试任务中,输入词向量大多都是较为简单的例句且每类语义关系分配合理均匀,然而在实际情况中往往要复杂得多;b)对于实际任务存在着测试语料规模与模型学习能力互相制约的难题,面对实际文本中存在的成千上万的语义关系以及数以千计的实体对,人工标注的测试语料是十分有限的,同时神经网络模型又需要大规模的数据进行学习,无法做到“举一反三”;c)在面对实体关系对存在于同一篇文章不同段落不同句子时,神经网络模型无法有效解决该情况。接下来的研究可从如何解决在更加复杂的语境下抽取实体关系对入手。

实体关系抽取的部分方法已被商业化利用,但仍有大量最新成果停留在理论阶段,对于实体关系抽取未来的发展方向与趋势,下文进行了详细梳理。

1)基于深度学习的实体关系抽取的持续研究

目前实体关系抽取的飞速发展是以深度学习为基础的。从 2014 年至 2017 年,学者们对递归神经网络模型、卷积神经网络模型、长短期记忆网络模型进行了广泛的深度研究。从 2017 年至今,随着注意力机制、迁移学习、强化学习等其他方法的引入结合,使得实体关系抽取的正确率、召回率、F_1 值进一步突破,性能提升。在基于深度学习的框架下,如何迁移其他领域理论方法至实体关系抽取任务成为今后发展的主要方向。

2)面向开放领域的关系抽取

实体关系抽取目前存在着测试语料规模和模型学习能力相互制约难题,但增大数据集的规模与减少神经网络对测试预料的依赖性并不能从根本程度上解决目前的困境。互联网每日可产生 1 TB 的数据,实体关系类别以各种方式出现且关系类别与日俱增,封闭的数据库无法满足目前现状。如何实现没有对文本类型做要求也没有人工预先标注语料,同时也不需要知道抽取哪些实体关系,实现自动实体关系抽取,即面向开放领域的实体关系抽取成为日后研究的热点。

3)面向复杂语境的关系抽取

主流的关系抽取主要针对二元关系抽取任务。面对跨句子的关系抽取、跨段落的关系类别,神经网络模型无法有效解决该问题。同时在具有一定语义环境、情感基础的基础下,人类对关系类别有了更深的理解,可以从文章中挖掘出更多的语义关系,但在基于深度学习的实体关系抽取属于空白领域,没有较好的理论方法。

在对各种理论方法性能评测时,采用 ACE 关系抽取任务数据集、SemEval2010 Task8 数据集等,而数据集往往是经过数据清洗后的“干净”测试语料,对于网络环境中颗粒度大以及复杂的语境下的语料,效果往往不佳。

4 结束语

实体关系抽取任务一直是自然语言处理中的热门领域。近二十年的发展中,从最早的基于模式匹配和规则的实体关系抽取,发展到机器学习模式的关系抽取,再到如今的基于深度学习的实体关系抽取,以及深度学习中神经网络与注意力机制、多示例模式、强化学习等方法结合使用,使得实体关系抽取任务处理方式更加多样化。

有监督的实体关系抽取严重依赖于人工语料标注,召回率和准确率高,在商业化应用中,大多仍采用基于监督方式的抽取。半监督式实体关系抽取只需少量标记语料,但对标记预料质量要求高,否则会出现语义飘移现象。无监督的实体关系抽取不需要人工标注的训练语料,迁移性能力强,但普遍准确率偏低。面向开放领域的实体关系抽取不需要标注语料和实体关系。通过前后相邻的短语进行实体关系上的语义构建,借助外部大型实体知识库,实现自动实体关系抽取,具有广泛的应用前景。深度学习的方法在最近几年发展迅速,一度赶超传统模式的实体关系抽取,但基于深度学习的实体关系抽取缺乏标记数据集和测试数据集。从目前的研究成果来看,实体关系任务已经取得了极大的成功,但仍存在着重叠实体关系识别难、标签错误传播、迭代过程噪声过大等共性问题,仍需学者不断探索努力。

参考文献:

[1] 谢德鹏,常青. 关系抽取综述[J]. 计算机应用研究,2020,37(7):1921-1924,1930.

[2] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述[J]. 软件学报,2019,30(6):1793-1818.

[3] 李枫林,柯佳. 基于深度学习框架的实体关系抽取研究进展[J]. 情报科学,2018,36(3):169-176.

[4] 武文雅,陈钰枫,徐金安,等. 中文实体关系抽取研究综述[J]. 计算机与现代化,2018,276(8):25-31,38.

[5] 刘绍毓,李弼程,郭志刚,等. 实体关系抽取研究综述[J]. 信息工程大学学报,2016,17(5):541-547.

[6] Miller S,Fox H,Ramshaw L,*et al*. A novel use of statistical parsing to extract information from text[M]//North American Chapter of the Association for Computational Linguistics. 2000:226-233.

[7] 吴明智,崔雷. 生物医学实体关系抽取的研究[J]. 中华医学图书情报杂志,2010,19(5):5-10.

[8] 邓擘,樊孝忠,杨立公. 用语义模式提取实体关系的方法[J]. 计算机工程,2007,33(10):218-220.

[9] Kambhatla N. Combining lexical,syntactic,and semantic features with maximum entropy models for extracting relations[C]//Proc of Meeting of the Association for Computational Linguistics. 2004.

[10] 车万翔,刘挺,李生. 实体关系自动抽取[J]. 中文信息学报,2005,19(2):1-6.

[11] 黄鑫,朱巧明,钱龙华,等. 基于特征组合的中文实体关系抽取[J]. 微电子学与计算机,2010,27(4):198-200,204.

[12] 甘丽新,万常选,刘德喜,等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展,2016,53(2):284-302.

[13] Cristianini N,Shawetaylor J. An introduction to support vector machines and other kernel-based learning methods[J]. Kybernetes,2001,30(1):103-115.

[14] Zelenko D,Aone C,Richardella A,*et al*. Kernel methods for relation extraction[J]. Journal of Machine Learning Research,2003,3(3):1083-1106.

[15] 刘克彬. 基于特征选择和语义扩展的词序列核函数研究[C]//中国中文信息学会. 第三届学生计算语言学研讨会论文集. 2006.

[16] Culotta A,Sorensen J. Dependency tree kernels for relation extraction[C]//Proc of Meeting of the Association for Computational Linguistics. 2004:423-429.

[17] Bunescu R C,Mooney R J. A shortest path dependency kernel for relation extraction[C]//Proc of Conference on Human Language Technology & Empirical Methods in Natural Language Processing. [S. l.]:Association for Computational Linguistics,2005.

[18] Zhang Min,Zhang Jie,Su Jian. Exploring syntactic features for relation extraction using a convolution tree kernel[C]//Proc of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. 2006.

[19] 郭剑毅,陈鹏,余正涛,等. 基于多核融合的中文领域实体关系抽取[J]. 中文信息学报,2016,30(1):24-29.

[20] Brin S. Extracting patterns and relations from the World Wide Web[C]//Proc of International Workshop on the Web and Databases. 1998:172-183.

[21] 张立邦. 基于半监督学习的中文电子病历分词和名实体挖掘[D]. 哈尔滨:哈尔滨工业大学,2014.

[22] 陈锦秀,姬东鸿. 基于图的半监督关系抽取[J]. 软件学报,2008,19(11):2843-2852.

[23] Zhang Zhu. Weakly-supervised relation classification for information extraction[C]//Proc of Conference on Information and Knowledge Management. 2004:581-588.

[24] 徐庆伶,汪西莉. 一种基于支持向量机的半监督分类方法[J]. 计算机技术与发展,2010,20(10):115-117,121.

[25] 刘建峰,吕佳. 融合主动学习的改进贝叶斯半监督分类算法研究[J]. 计算机测量与控制,2014,22(6):1938-1940.

[26] Kanungo T,Mount D M,Netanyahu N S,*et al*. An efficient K-means clustering algorithm:analysis and implementation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2002,24(7):881-892.

[27] Kohonen T. The self-organizing map[J]. Neurocomputing,1998,21(1):1-6.

[28] Frey B J,Dueck D. Clustering by passing messages between data points[J]. Science,2007,315(5814):972-976.

[29] 王娜,杜海峰,王孙安. 一种基于流形距离的迭代优化聚类算法[J]. 西安交通大学学报,2009,43(5):76-79.

[30] 傅景广,许刚,王裕国. 基于遗传算法的聚类分析[J]. 计算机工程,2004,30(4):122-124.

[31] Gong Maoguo,Jiao Licheng,Liu Fang,*et al*. The quaternion model of artificial immune response[C]//Proc of Artificial Immune Systems Proceedings. Berlin:Springer,2005:207-219.

[32] 公茂果,焦李成,马文萍,等. 基于流形距离的人工免疫无监督分类与识别算法[J]. 自动化学报,2008,34(3):367-375.

[33] Corro L D,Gemulla R. ClausIE:clause-based open information extraction[C]//Proc of Web Conference. 2013:355-366.

[34] 姚贤明,甘健侯,徐坚. 面向中文开放领域的多元实体关系抽取研究[J]. 智能系统学报,2019,14(3):597-604.

[35] 郭喜跃. 面向开放领域文本的实体关系抽取[D]. 武汉:华中师范大学,2016.

[36] 李颖,郝晓燕,王勇. 中文开放式多元实体关系抽取[J]. 计算机科学,2017,44(S1):80-83.

[37] 王斌,郭剑毅,线岩团,等. 融合多特征的基于远程监督的中文领域实体关系抽取[J]. 模式识别与人工智能,2019,32(2):133-143.

[38] Mintz M,Bills S,Snow R,*et al*. Distant supervision for relation extraction without labeled data[C]//Proc of International Joint Conference on Natural Language Processing. 2009:1003-1011.

[39] 余小康. 结合从句识别和半监督集成学习的远程监督关系抽取[D]. 杭州:浙江大学,2016.

[40] 黄蓓静,贺樑,杨静. 远程监督人物关系抽取中的去噪研究[J]. 计算机应用与软件,2017,34(7):11-18,31.

[41] 黄杨琛,贾焰,甘亮,等. 基于远程监督的多因子人物关系抽取模型[J]. 通信学报,2018,39(7):103-112.

[42] 余小康,陈岭,郭敬,等. 结合从句级远程监督与半监督集成学习的关系抽取方法[J]. 模式识别与人工智能,2017,30(1):54-63.

[43] Socher R,Huval B,Manning C D,*et al*. Semantic compositionality through recursive matrix-vector spaces[C]//Proc of Empirical Methods in Natural Language Processing. 2012:1201-1211.

[44] Liu Chunyang,Sun Wenbo,Chao Wenlan,*et al*. Convolution neural network for relation extraction[C]//Advanced Data Mining and Applications. 2013:231-242.

[45] Zeng Daojian,Liu Kang,Lai Siwei,*et al*. Relation classification via convolutional deep neural network[C]//Proc of the 25th International Conference on Compulational Linguistics. 2014.

[46] Xu Kun,Feng Yansong,Huang Songfang,*et al*. Semantic relation classification via convolutional neural networks with simple negative sampling[C]//Proc of Empirical Methods in Natural Language Processing. 2015:536-540.

[47] Sundermeyer M,Schluter R,Ney H,*et al*. LSTM neural networks for language modeling[C]//Proc of Conference of the International Speech Communication Association. 2012:194-197.

[48] Sun Ziyang,Gu Junzhong,Yang Jing. Chinese entity relation extraction method based on deep learning[J]. Computer Engineering,2018,44(9):164-170.

[49] Miwa M,Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures[C]//Proc of Meeting of the Association for Computational Linguistics. 2016:1105-1116.

[50] Zheng Suncong,Wang Feng,Bao Hongyun,*et al*. Joint extraction of entities and relations based on a novel tagging scheme[C]//Proc of Meeting of the Association for Computational Linguistics. 2017:1227-1236.

[51] Zeng Daojian,Liu Kang,Lai Siwei,*et al*. Relation classification via convolutional deep neural network[C]//Proc of International Conference on Computational Linguistics. 2014:2335-2344.

[52] Ji Guoliang,Liu Kang,He Shizhu,*et al*. Distant supervision for relation extraction with sentence-level attention and entity descriptions[C]//Proc of National Conference on Artificial Intelligence. 2017:3060-3066.

[53] He Dengchao,Zhang Hongjun,Hao Wenining,*et al*. A customized attention-based long short-term memory network for distant supervised relation extraction[J]. Neural Computation,2017,29(7):1964-1985.