软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(6):1793−1818 [doi: 10.13328/j.cnki.jos.005817]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

深度学习实体关系抽取研究综述*

鄂海红[1,2], 张文静[1,2], 肖思琪[1,2], 程 瑞[1,2], 胡莺夕[1,2], 周筱松[1,2], 牛佩晴[1,2]

[1](北京邮电大学 计算机学院 数据科学与服务中心,北京 100876)

[2](教育部信息网络工程研究中心(北京邮电大学),北京 100876)

通讯作者: 鄂海红, E-mail: ehaihong@bupt.edu.cn

摘 要: 实体关系抽取作为信息抽取、自然语言理解、信息检索等领域的核心任务和重要环节,能够从文本中抽取实体对间的语义关系.近年来,深度学习在联合学习、远程监督等方面上的应用,使关系抽取任务取得了较为丰富的研究成果.目前,基于深度学习的实体关系抽取技术,在特征提取的深度和模型的精确度上已经逐渐超过了传统基于特征和核函数的方法.围绕有监督和远程监督两个领域,系统总结了近几年来中外学者基于深度学习的实体关系抽取研究进展,并对未来可能的研究方向进行了探讨和展望.

关键词: 实体关系抽取;深度学习;联合学习;远程监督;生成对抗网络

中图法分类号: TP183

中文引用格式: 鄂海红,张文静,肖思琪,程瑞,胡莺夕,周筱松,牛佩晴.深度学习实体关系抽取研究综述.软件学报,2019,30(6):1793−1818. http://www.jos.org.cn/1000-9825/5817.htm

英文引用格式: E HH, Zhang WJ, Xiao SQ, Cheng R, Hu YX, Zhou XS, Niu PQ. Survey of entity relationship extraction based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2019,30(6):1793−1818 (in Chinese). http://www.jos.org.cn/1000-9825/5817.htm

Survey of Entity Relationship Extraction Based on Deep Learning

E Hai-Hong[1,2], ZHANG Wen-Jing[1,2], XIAO Si-Qi[1,2], CHENG Rui[1,2], HU Ying-Xi[1,2], ZHOU Xiao-Song[1,2], NIU Pei-Qing[1,2]

[1](Data Science and Service Center, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

[2](Engineering Research Center of Information Networks of Ministry of Education (Beijing University of Posts and Telecommunications), Beijing 100876, China)

Abstract: Entity relation extraction is a core task and an important part in the fields of information extraction, natural language understanding, and information retrieval. It can extract the semantic relationships between entity pairs from the texts. In recent years, the application of deep learning in the fields of joint learning, remote supervision has resulted in relatively abundant research results in relation extraction tasks. At present, entity relationship extraction technology based on deep learning has gradually exceeded the traditional methods which are based on features and kernel functions in terms of the depth of feature extraction and the accuracy. This paper focuses on the two fields of supervision and remote supervision. It systematically summarizes the research progress of Chinese and overseas scholars' deep relationship-based entity relationship extraction in recent years, and discusses and prospects future possible research directions as well.

Key words: entity relationship extraction; deep learning; joint learning; remote supervision; generative adversarial network

* 基金项目: 国家重点研发计划(2018YFB1403501)
Foundation item: National Key R&D Program of China (2018YFB1403501)
收稿时间: 2018-04-25; 修改时间: 2018-07-08, 2018-09-02, 2018-10-13; 采用时间: 2018-12-29; jos 在线出版时间: 2019-03-27
CNKI 网络优先出版: 2019-03-28 07:32:30, http://kns.cnki.net/kcms/detail/11.2560.TP.20190327.2337.016.html

随着互联网技术的发展,人们需要处理的数据量激增,领域交叉现象突出.如何快速高效地从开放领域的文本中抽取出有效信息,成为摆在人们面前的重要问题.实体关系抽取作为文本挖掘和信息抽取[1]的核心任务,其主要通过对文本信息建模,自动抽取出实体对之间的语义关系,提取出有效的语义知识.其研究成果主要应用在文本摘要、自动问答[2]、机器翻译[3]、语义网标注、知识图谱[4]等.随着近年来对信息抽取的兴起,实体关系抽取问题进一步得到广泛关注和深入研究,一些研究成果及时出现在近几年人工智能、自然语言处理等相关领域的国际会议上,如 ACL[5−13]、EMNLP[14−22]、ICLR[23,24]、AAAI[25]、KDD[26]、NAACL[27]、 ECML-PKDD[28]等.

经典的实体关系抽取方法主要分为有监督、半监督、弱监督和无监督这 4 类.有监督的实体关系抽取主要分为基于特征和基于核函数的方法.Zhou[29]和郭喜跃[6]等人利用 SVM 作为分类器,分别研究词汇、句法和语义特征对实体语义关系抽取的影响.有监督方法需要手工标注大量的训练数据,浪费时间精力,因此,人们[30]继而提出了基于半监督[31]、弱监督和无监督的关系抽取方法来解决人工标注语料问题,其中:Brin[32]利用 Bootstrapping 方法对命名实体之间的关系进行抽取;Craven 等人[33]在研究从文本中抽取结构化数据、建立生物学知识库的过程中,首次提出了弱监督机器学习思想;Hasegawa 等人[34]在 ACL 会议上首次提出了一种无监督的命名实体之间关系抽取方法.

经典方法存在特征提取误差传播问题,极大影响实体关系抽取效果.随着近些年深度学习的崛起,学者们逐渐将深度学习应用到实体关系抽取任务中[7].基于数据集标注量级的差异,深度学习的实体关系抽取任务分为有监督和远程监督两类.基于深度学习的有监督实体关系抽取方法是近年来关系抽取的研究热点,该方法能避免经典方法中人工特征选择等步骤,减少并改善特征抽取过程中的误差积累问题.根据实体识别及关系分类两个子任务完成的先后顺序不同,基于深度学习的有监督实体关系抽取方法可以分为流水线(pipeline)方法和联合学习(joint learning)方法.Zeng 等人[20]在 2014 年首次提出使用 CNN 进行关系分类,Katiyar 等人[13]在 2017 年首次将注意力机制 Attention 与递归神经网络 Bi-LSTM 一起用于联合提取实体和分类关系,神经网络模型在有监督领域的拓展皆取得不错效果.同时,基于深度学习的远程监督实体关系抽取方法因具有缓解经典方法中错误标签和特征抽取误差传播问题的能力而成为研究热点,主要基础方法包括 CNN,RNN,LSTM 等网络结构[35,36].近年来,学者们在基础方法之上提出了多种改进,如 PCNN 与多示例学习的融合方法[37]、PCNN 与注意力机制的融合方法[10]等.Ji 等人[38]提出在 PCNN 和 Attention 的基础上添加实体的描述信息来辅助学习实体的表示,Ren 等人[39]提出的 COTYPE 模型、Huang[40]提出的残差网络皆增强了关系提取效果.

为了能够系统综述相关研究成果,我们查阅了近年来的综述论文[30,35,41−43],从中可看出,基于深度学习的实体关系抽取方法与经典抽取方法相比,其主要优势在于深度学习的神经网络模型可以自动学习句子特征,无需复杂的特征工程.所以,本文重点围绕深度学习来深入探讨实体关系抽取方法.

本文首先在第 1 节给出实体关系抽取的问题定义和解决框架.着重在第 2 节、第 3 节介绍基于深度学习的有监督和远程监督领域的实体关系抽取研究进展.之后,在第 4 节介绍基于深度学习的实体关系抽取新模型与新思路.并在第 5 节介绍基于深度学习的实体关系抽取在领域知识图谱构建中的研究进展.最后,在第 6 节、第 7 节给出数据集、评测效果以及对未来研究方向的展望.

1 深度学习实体关系抽取的问题定义和解决框架

1.1 问题定义

预先定义

实体关系抽取作为信息抽取的重要任务,是指在实体识别的基础上,从非结构化文本中抽取出预先定义的实体关系.实体对的关系可被形式化描述为关系三元组$\langle e_1,r,e_2\rangle$,其中,e_1 和 e_2 是实体,r 属于目标关系集 $R\{r_1,r_2,r_3,...,r_i\}$.关系抽取的任务是从自然语言文本中抽取出关系三元组$\langle e_1,r,e_2\rangle$,从而提取文本信息.

基于深度学习实体关系抽取主要分为有监督和远程监督两类.在有监督中,解决实体关系抽取的方法可以分为流水线学习和联合学习两种:流水线学习方法是指在实体识别已经完成的基础上直接进行实体之间关系的抽取;联合学习方法主要是基于神经网络的端到端模型,同时完成实体的识别和实体间关系的抽取.与有监督实体关系抽取相比,远程监督方法缺少人工标注数据集,因此,远程监督方法比有监督多一步远程对齐知识库给

无标签数据打标的过程.而构建关系抽取模型的部分,与有监督领域的流水线方法差别不大.

基于深度学习的实体关系抽取、实体关系识别、实体关系分类是 3 个任务相近、彼此有关联的概念.具体而言,关系抽取[7]在其流水线处理场景中与关系分类处理着相同的任务,此时,关系抽取具体是指在句子中的命名实体对已经被识别的情况下,直接进行实体对的关系分类;而关系抽取在联合学习场景中是将关系分类作为自己的一个子任务,此时,关系抽取具体是指:将实体关系抽取任务分为命名实体识别和关系分类两个子任务,用联合学习模型同时解决这两个子任务.而实体关系识别任务与关系抽取任务相同,在实际处理时也是发现和识别实体间的语义关系[44,45],因此在部分中外综述文献里,实体关系抽取有时也被称为实体关系识别.

1.2 解决问题框架

针对实体关系抽取任务,基于深度学习的抽取框架如图 1 所示.

(1) 获取有标签数据:有监督方法通过人工标记获取有标签数据集,远程监督方法通过自动对齐远程知识库获取有标签数据集;

(2) 构建词语向量表示:将有标签句子分词,将每个词语编码成计算机可以接受的词向量,并求出每个词语与句子中实体对的相对位置,作为这个词语的位置向量,将词向量与位置向量组合作为这个词语的最终向量表示;

(3) 进行特征提取:将句子中每一个词语的向量表示输入神经网络中,利用神经网络模型提取句子特征,进而训练一个特征提取器;

(4) 关系分类:测试时根据预先定义好的关系种类,将特征提取出的向量放入非线性层进行分类,提取最终的实体对关系;

(5) 评估分类性能:最后,对关系分类结果进行评估,评测指标和相关数据集详见第 6 节.

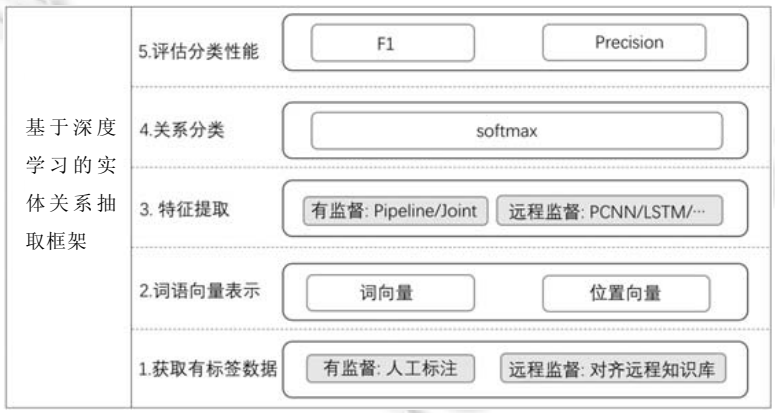


Fig.1 Entity relationship extraction framework based on deep learning

图 1 基于深度学习的实体关系抽取框架

2 基于深度学习的有监督实体关系抽取方法

2.1 有监督实体关系抽取框架演化流程

基于深度学习方法中的有监督方法进行关系抽取,是近年来关系抽取的研究热点,其能解决经典方法中存在的人工特征选择、特征提取误差传播两大主要问题,将低层特征进行组合,形成更加抽象的高层特征,用来寻找数据的分布式特征表示.从基于监督学习的神经网络模型来看,研究主要集中在融合多种自然语言特征来提高识别精确度.有监督的实体关系抽取框架的演化流程如图 2 所示.

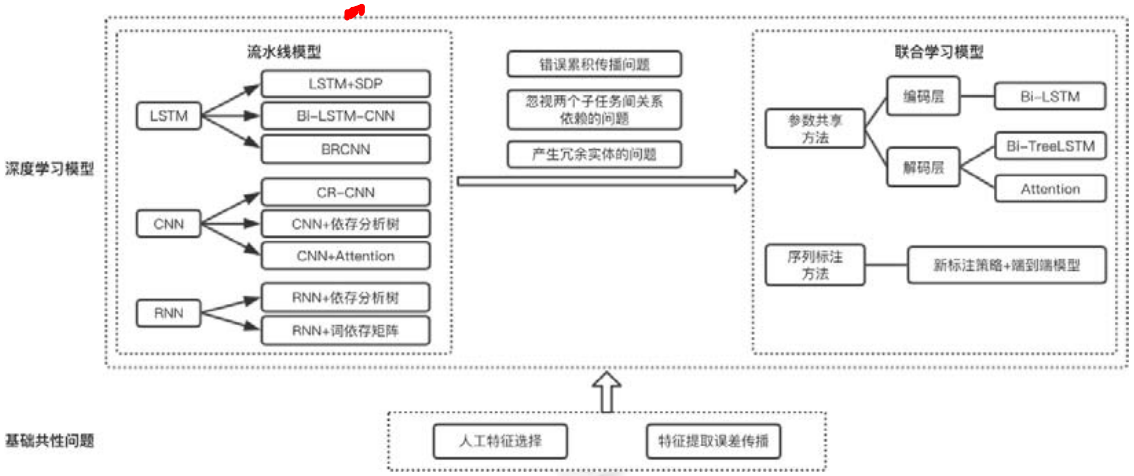


Fig.2 Solution framework based on supervised entity relationship extraction

图 2 基于有监督的实体关系抽取的解决框架

基于深度学习的有监督实体关系抽取可以分为:1) 流水线方法;2) 联合学习方法.这两种方法都基于 CNN, RNN,LSTM 这 3 种框架进行扩展优化.

- 流水线方法中,基于 RNN 模型的扩展包括在 RNN 基础之上增加依存分析树信息、词依存矩阵信息;基于 CNN 模型的扩展包括在 CNN 基础之上增加类别排名信息、依存分析树、注意力机制;基于 LSTM 模型的扩展包括在 LSTM 基础之上增加最短依存路径(SDP)或将 LSTM 与 CNN 结合.流水线方法存在错误累积传播、忽视子任务间关系依赖、产生冗余实体等问题,因此,联合模型逐渐开始受到重视;
- 联合学习方法根据其建模对象不同,可分为参数共享和序列标注两类子方法:参数共享方法的编码层均使用 Bi-LSTM,解码层则基于 Bi-LSTM、依赖树和注意力机制等方法纷纷进行优化扩展;序列标注方法则用一种新标注策略的端到端模型解决流水线模型中冗余实体的问题.

下面依照流水线方法(基于 RNN 模型的实体关系抽取方法、基于 CNN 模型的实体关系抽取方法、基于 LSTM 模型的实体关系抽取方法)、联合学习方法(基于参数共享的实体关系抽取方法、基于序列标注的实体关系抽取方法)的顺序来介绍有监督领域实体关系抽取方法.

2.2 流水线方法

2.2.1 主要流程

基于流水线的方法进行关系抽取的主要流程可以描述为:针对已经标注好目标实体对的句子进行关系抽取,最后把存在实体关系的三元组作为预测结果输出.一些基于流水线方法的关系抽取模型被陆续提出,其中,采用基于 RNN,CNN,LSTM 及其改进模型的网络结构,因其高精度获得了学术界的大量关注.

2.2.2 主流方法介绍

(1) 基于 RNN 模型的实体关系抽取方法

RNN 在处理单元之间既有内部的反馈连接又有前馈连接,可以利用其内部的记忆来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,已成功应用在多种 NLP 任务中.

基于 RNN 模型进行关系抽取的方法由 Socher 等人[46]于 2012 年首次提出,此方法为分析树中的每个节点分配一个向量和一个矩阵,其中,向量捕获组成部分的固有含义,而矩阵捕捉它如何改变相邻单词或短语的含义.这种矩阵向量 RNN 可以在命题逻辑和自然语言中学习操作符的含义,解决了单词向量空间模型(single-word vector space models)无法捕捉到长短语的构成意义,阻碍了它们更深入地理解语言的问题.

Hashimoto 等人[19]在 2013 年提出了基于句法树的递归神经网络(RNN)模型,与 Socher 等人提出的模型不同的是,Hashimoto 没有使用需要昂贵计算成本的词依存矩阵,而是使用了词性(POS)标签、短语类别和句法头

等附加特征,并向 RNN 模型中引入平均参数,为目标任务的重要短语增加权重,Hashimoto 的模型证明了增加特征及引入平均参数的有效性.

RNN 相比于前馈网络更适合处理序列化输入,但 RNN 也存在着以下两个缺点:(1) 在网络训练时,RNN 容易出现梯度消失、梯度爆炸的问题,因此,传统 RNN 在实际中很难处理长期依赖,这一点在 LSTM 网络中有所改进;(2) 由于 RNN 的内部结构复杂,网络训练周期较长,而 CNN 结构相对简单,主要包括前置的卷积层和后置的全连接层,训练更快速.

(2) 基于 CNN 模型的实体关系抽取方法

CNN 的基本结构包括两层:其一为特征提取层,每个神经元的输入与前一层的局部接受域相连,并提取该局部的特征;其二是特征映射层,网络的每个计算层由多个特征映射组成,每个特征映射是一个平面,平面上所有神经元的权值相等,减少了网络中自由参数的个数.由于同一特征映射面上的神经元权值相同,所以 CNN 网络可以并行学习.

Zeng 等人[20]在 2014 年首次提出了使用 CNN 进行关系抽取,利用卷积深度神经网络(CDNN)来提取词汇和句子层次的特征,将所有的单词标记作为输入,而无需复杂的预处理,解决了从预处理系统中提取的特征可能会导致错误传播并阻碍系统性能的问题.图 3 描述了该论文用于关系分类的神经网络的体系结构.网络对输入句子提取多个级别的特征向量,它主要包括以下 3 个组件:词向量表示、特征提取和输出.图 3 右部分显示了句子级特征向量构建过程:每个词语向量由词特征(WF)和位置特征(PF)共同组成,将词语向量放入卷积层提取句子级特征.图 3 左上部分为提取词汇级和句子级特征的过程,然后直接连接以形成最终的句子特征向量.最后如图 3 左下部分,通过隐藏层和 Softmax 层得到最终的分类结果.

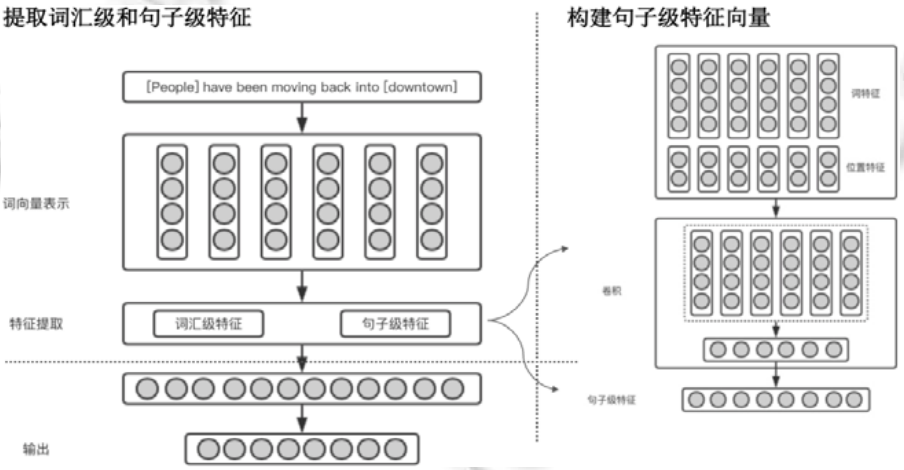


Fig.3 Relationship extraction frameworkbased on convolutional deep neural network

图 3 基于 DNN 的关系抽取框架

Xu 等人[47]于 2015 年在 Zeng 等人工作的基础上提出了基于依存分析树的卷积神经网络的实体关系抽取模型,该模型与 Zeng 等人的 CNN 模型不同的是将输入文本经过了依存分析树,同时提出了一种负采样策略:首先,利用依存路径来学习关系的方向性;然后,使用负采样方法来学习主体和对象的位置分配,采用从对象到主体的最短依存路径作为负样本,并将负样本送到模型中学习,以解决实体对距离较远时,依存分析树引入的无关信息问题.同时,显著提高了关系抽取的性能.

Santos 等人[21]在 2015 年提出了 CR-CNN 模型,与 Zeng 等人的模型相比,CR-CNN 将最后的 Softmax 输出层替换为利用排名进行分类输出:对于给定的输入文本段,网络使用卷积层产生文本的分布向量表示,并将其与文本表示进行比较,以便为每个类生成分数;同时提出了一种新的排名损失函数,能够给予正确的预测类更高的评分、错误的预测类更低的评分.与 Xu 等人的模型相比,本文仅将词向量作为输入特征,而不需要依存分析树等

附加特征,因此可以降低 NLP 工具中提取到错误特征的影响,并提升模型的效果.

Vu 等人[48]在 2016 年提出了一种新的基于 CNN 网络的上下文表示(扩展的中间上下文),与作为 Baseline 的 Zeng 等人的标准 CNN 网络不同的是,Vu 提出的 CNN 模型没有额外的全连接隐藏层;其次,Vu 也尝试使用双向 RNN 进行关系抽取,并为其优化引入 Santos[21]提出的排名损失,改善关系抽取结果.基于两个实体位置可以将上下文分成 3 个不相交的区域:左上下文、中间上下文和右上下文.由于在大多数情况下中间上下文包含关系的最相关信息,因此该文提出了使用两个上下文:(1) 左上下文、左实体和中间上下文的组合;(2) 中间上下文、右实体和右上下文的组合.通过重复中间上下文,迫使网络特别关注它.最后,使用简单的投票机制结合 CNN 和 RNN 网络,并达到了当时的最新技术.

Zeng 等人虽然使用了位置向量来表示指定词与目标实体间的相对距离,但是位置编码不足以完全捕获指定词与目标实体的关系以及它们可能对目标关系的影响.由此,Wang 等人[49]于 2016 年提出的 CNN 架构依赖于一种新颖的多层次注意力机制来捕获对指定实体的注意力(首先是输入层级对于目标实体的注意力)和指定关系的池化注意力(其次是针对目标关系的注意力).这使得模型能够检测更细微的线索,尽管输入的句子异构,但是模型还是能够自动了解句子中的哪些部分与给定的关系类别相关.其次,模型在利用注意力机制来自动识别与关系分类相关的输入句子的部分之后,提出了一种 Attention-based Pooling 的混合方法,认为利用这样的方法会抽取出部分有意义的 *N*-gram 短语,实验证明了在混合层上,能够抽出对关系分类最为显著的 Trigram 字段.最后,论文还引入了一种新的成对的基于边缘的目标函数,并证明其优于标准损失函数.

(3) 基于 LSTM 模型的实体关系抽取方法

由于梯度消失、梯度爆炸的问题,传统的 RNN 在实际中很难处理长期依赖,后面时间的节点对于前面时间的节点感知力下降.而 LSTM 网络通过 3 个门控操作及细胞状态解决了这些问题,能够从语料中学习到长期依赖关系.

Yan 等人[11]在 2015 年提出了基于 LSTM 的融合句法依存分析树的最短路径以及词向量特征、词性特征、WordNet 特征、句法类型特征来进行关系抽取,该论文的模型图如图 4 所示.首先,如图 4 左下部分,利用斯坦福解析器将句子解析为依赖树,并提取最短依赖路径(SDP)作为网络的输入,沿着 SDP,使用 4 种不同类型的信息(称为通道),包括单词、词性标签、语法关系和 WordNet 上位词;在每个通道中(图 4 右部分是每个通道的细节图),词语被映射成向量,捕获输入的基本含义,两个递归神经网络分别沿着 SDP 的左右子路径获取信息,网络中的 LSTM 单元用于有效信息的传播;之后,如图 4 左上部分,最大池化层从每个路径中的 LSTM 节点收集信息,来自不同通道的池化层连接在一起,然后输入到隐藏层;最后,使用 Softmax 输出层用于关系分类.

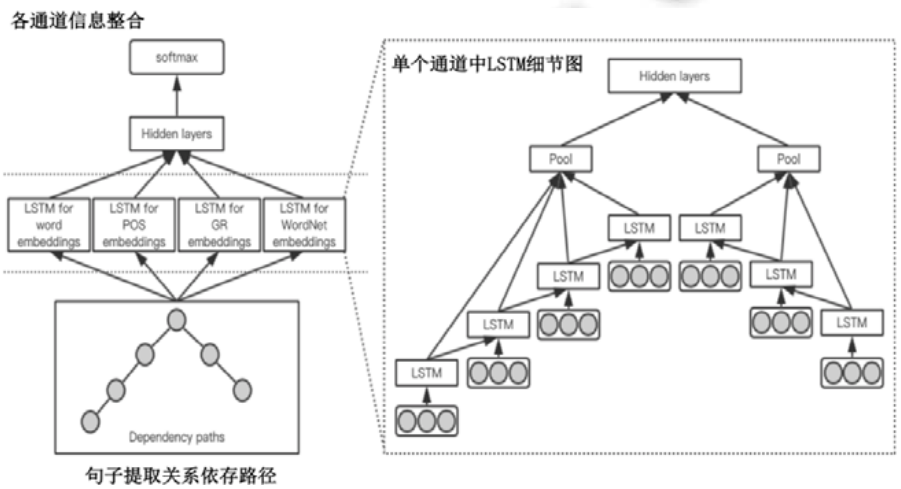


Fig.4 Relationship extraction method based on LSTM and shortest dependency path

图 4 基于 LSTM 及最短依存路径的关系抽取方法

Thien 等人[22]基于已有工作经验,利用传统特征工程并结合 CNN,RNN 网络的优势,在 2015 年提出一种融合传统特征工程和神经网络的方法,首次系统地检测了 RNN 架构以及 RNN 与 CNN 和传统的基于特征的关系抽取方法相结合的工作.本文采用 LSTM 网络的一种变体 GRU(gated recurrent unit)展开实验,同时,首次提出了融合 CNN 和 RNN 网络的 3 种不同的方式:Ensembling(集成)、Stacking(堆叠)、Voting(投票),提高了关系抽取的精确度.

为避免 Yan 等人提出的模型需要从 NLP 预处理工具中提取附加特征带来的错误传播问题,Li 等人[50]于 2016 年提出一种基于低成本序列特征的 Bi-LSTM-RNN 模型,利用实体对并将它们周围的上下文分段表示来获取更丰富的语义信息,无需词性标注、依存句法树等额外特征.将文本经过 LSTM 网络获得隐藏向量表示后依照两个实体分成五段式的方式输入池化层获得向量表示,再输入分类器进行关系分类,解决了基于句法或依赖性特征等高成本结构特征问题,并证明当不使用依赖解析时,两个目标实体之间的上下文可以用作最短依赖路径的近似替换.

基于 Yan 等人的工作,Cai 等人[51]于 2016 年提出了一种基于最短依赖路径(SDP)的深度学习关系抽取模型:双向递归卷积神经网络模型(BRCNN),通过将卷积神经网络和基于 LSTM 单元的双通道递归神经网络相结合,进一步探索如何充分利用 SDP 中的依赖关系信息.BRCNN 模型结合了 Yan 等人的多通道 LSTM 以及 Zeng 等人的卷积关系抽取的特点,利用基于双向 LSTM 的递归神经网络对最短依存路径中的全局模式进行编码,并利用卷积层捕获依存关系链接的两个相邻词的局部特征,增强了实体对之间关系方向分类的能力.

2.2.3 流水线方法中存在的共性问题

然而,流水线方法存在着以下几个缺点.

1) 错误传播:实体识别模块的错误会影响到接下来的关系分类性能;
2) 忽视了两个子任务之间存在的关系:丢失信息,影响抽取效果;
3) 产生冗余信息:由于对识别出来的实体进行两两配对,然后再进行关系分类,那些没有关系的实体对就会带来多余信息,提升错误率.

2.3 联合学习方法

相比于流水线方法,联合学习[52]方法能够利用实体和关系间紧密的交互信息,同时抽取实体并分类实体对的关系,很好地解决了流水线方法所存在的问题.

2.3.1 主要流程

联合学习方法通过实体识别和关系分类联合模型,直接得到存在关系的实体三元组.因在联合学习方法中建模的对象不同,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.下面分别对这两种方法进行说明.

2.3.2 主流方法介绍

(1) 基于参数共享的实体关系抽取方法

针对流水线方法中存在的错误累积传播问题和忽视两个子任务间关系依赖的问题,基于参数共享的实体关系抽取方法被提出.在此方法中,实体识别子任务和关系抽取子任务通过共享联合模型的编码层来进行联合学习,通过共享编码层,在训练时,两个子任务都会通过后向传播算法更新编码层的共享参数,以此来实现两个子任务之间的相互依赖,最终找到全局任务的最佳参数,实现性能更佳的实体关系抽取系统.在联合学习模型中,输入的句子在通过共享的编码层后,在解码层会首先进行实体识别子任务,再利用实体识别的结果,并对存在关系的实体对进行关系分类,最终输出实体-关系三元组.

Miwa 等人[12]在 2016 年首次将神经网络的方法用于联合表示实体和关系,其模型图如图 5 所示.在该模型中,实体识别子任务和关系分类子任务共享编码层的 LSTM 单元序列表示(编码层包括 LSTM 单元和隐藏层).该方法将实体识别任务当作序列标注任务,使用双向序列 LSTM 输出具有依赖关系的实体标签;之后,通过在双向序列 LSTM 单元上堆叠双向树结构 LSTM 的方法,使关系分类子任务和实体识别子任务共享编码层的 LSTM 单元序列表示,同时,在关系分类子任务中捕获词性标签等依赖特征和实体识别子任务中输出的实体序列,形成

依存树,最终根据依存树中目标实体间的最短路径对文本进行关系抽取.但该模型中的关系分类子任务和实体识别子任务仅共享了编码层的双向序列 LSTM 表示,从严格意义上来说不是真正的联合模型.但是该模型的提出,为之后真正意义上联合学习模型的提出奠定了基础,是基于深度学习方法做联合学习模型的启发者.

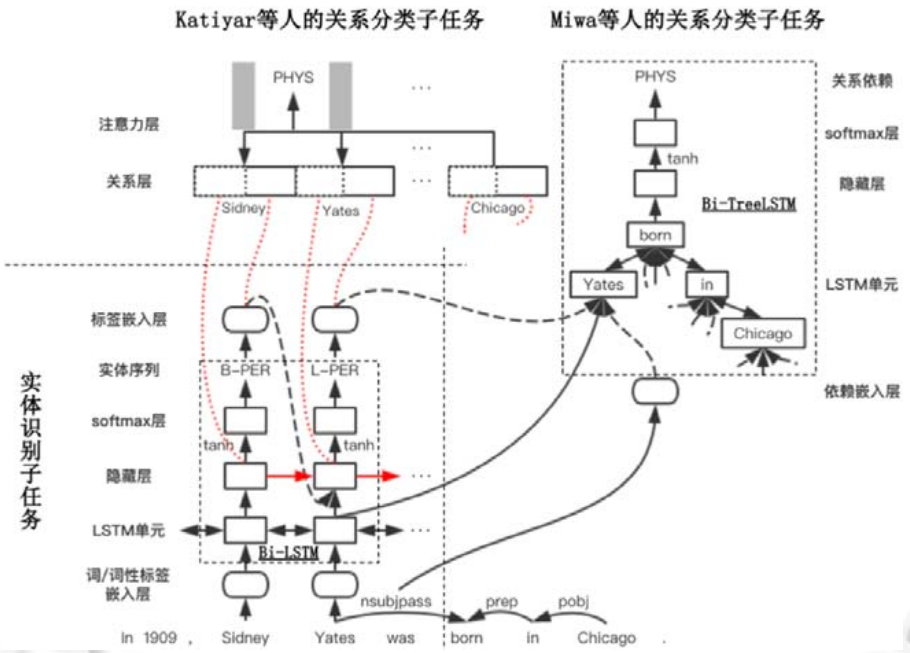


Fig.5 Relational extraction model diagram based on parameter sharing method

图 5 基于参数共享方法的关系抽取模型图

Li 等人[53]在 2017 年将该模型用于提取细菌和细菌位置之间存在的“Live-In”关系,并基于实际应用对 Miwa 模型做出了两点改进:1) 为改善从实体识别子任务到关系分类子任务可能会产生的错误累积传播问题,在关系分类子任务中引入一种新的关系“Invalid_Entity”,对实体识别子任务中产生的实体进行验证,以区分有效实体和无效实体,之后对有效实体再进行“Lives_In”和“not Lives_In”关系的分类;2) 在实体识别子任务中,因贪婪的从左到右逐步预测实体标签的方式可能会在这些标签之间带来错误传播,即先前预测中的错误可能会在随后的预测中引起新的错误,故将模型中原来的贪婪搜索解码换为波束搜索,因波束搜索中的每一步都可以有多个候选预测,在最佳预测不正确的情况下,可以根据全局分数排序来选择候选预测,并在波束搜索中用早期更新技术来训练模型,以缓解实体标签间的错误传播问题.

Katiyar 等人[54]在 2016 年首次将深度双向 LSTM 序列标注的方法用于联合提取观点实体和 IS-FROM,IS-ABOUT 关系,同时还提出了在输出层上添加句子级别的限制和关系级别的优化来提高模型的精确度.但这种方法只能识别观点实体和 IS-FROM,IS-ABOUT 关系,无法提取实体间的关系类型,模型也不能扩展用于抽取其他关系类型.之后,为改进模型无法扩展应用的问题,Katiyar 等人[13]在自己 2016 年模型的基础上,于 2017 年首次将注意力机制与双向 LSTM 一起用于联合提取实体和分类关系.该方法的模型图如图 5 所示,实体识别子任务和关系分类子任务共享编码层表示(编码层包括 LSTM 单元和隐藏层).该模型在实体识别子任务中和 Miwa 等人[12]的模型一致,将实体识别子任务当作序列标注任务,使用多层双向 LSTM 网络来进行实体检测;在关系分类子任务上,该方法改善了 Miwa 等人[12]依赖于词性标签、依赖树等特征的缺点,基于实体识别子任务输出的实体序列表示和共享的编码层表示,使用注意力模型进行关系分类;同时,该模型还可以扩展提取各种定义好的关系类型,是真正意义上的第一个神经网络联合抽取模型.

其中,Miwa 等人[12]和 Katiyar 等人[13]的模型图如图 5 所示.二者在实体识别子任务上的模型图基本相同,

如图左下部分所示,均使用 Bi-LSTM 来进行实体识别子任务(其中,红色箭头部分仅为 Katiyar 等人[13]的模型图所有).图左上部分为 Katiyar 等人[13]的关系分类子任务示意图,基于注意力机制来进行关系分类;图右上部分为 Miwa 等人[12]的关系分类子任务示意图,基于 Bi-TreeLSTM 来进行关系分类.

(2) 基于序列标注的实体关系抽取方法

基于参数共享的实体关系抽取方法,改善了传统流水线方法中存在的错误累积传播问题和忽视两个子任务间关系依赖的问题.但因其在训练时还是需要先进行命名实体识别子任务,再根据实体预测信息对实体进行两两匹配,最后进行关系分类子任务,因其在模型实现过程中分开完成了命名实体识别和关系分类这两个子任务,仍然会产生没有关系的实体这种冗余信息.为了解决这个问题,基于新序列标注方法的实体、关系联合抽取方法被提出.

Zheng 等人[55]在 2017 年提出了基于新的标注策略的实体关系抽取方法,把原来涉及到命名实体识别和关系分类两个子任务的联合学习模型完全变成了一个序列标注问题.在该方法中,共包含 3 种标注信息:(1) 实体中词的位置信息{*B*,*I*,*E*,*S*,*O*},分别表示{实体开始,实体内部,实体结束,单个实体,无关词};(2) 实体关系类型信息,需根据实际需要自定义关系类型并编码,如{*CF*,*CP*,…};(3) 实体角色信息{1,2},分别表示{实体 1,实体 2}.该方法能使用序列标注的方法同时识别出实体和关系,避免了复杂的特征工程,通过一个端到端的神经网络模型直接得到实体-关系三元组,解决了基于参数共享的实体关系抽取方法可能会带来的实体冗余的问题.新序列标注方法的模型图如图 6 所示.在该端到端的神经网络模型中,对输入的句子,首先,编码层使用 Bi-LSTM 来进行编码;之后,解码层再使用 LSTM 进行解码;最终,输出模型标注好的实体-关系三元组.另外,Zheng 等人[55]在这篇论文中还对该端到端模型增加了偏置损失函数,该函数增强了相关实体对之间的联系,削弱了无效实体标签的影响力,提高了关系分类的准确率;并基于这种新的标注方法,该论文中还学习用不同的端到端模型来解决关系抽取问题.

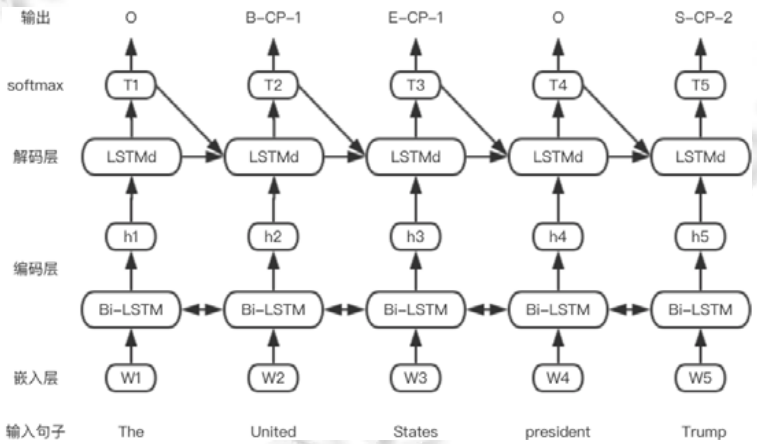


Fig.6 New sequence annotation method model diagram

图 6 新序列标注方法模型图

2.3.3 联合学习方法中存在的共性问题

联合学习方法包括基于参数共享的实体关系抽取方法和基于新序列标注的实体关系抽取方法:前者很好地改善了流水线方法中存在的错误累积传播问题和忽视两个子任务间关系依赖的问题;而后者不仅解决了这两个问题,还解决了流水线方法中存在的冗余实体的问题.但这两种方法对于现今有监督领域存在的重叠实体关系识别问题,并未能给出相关的解决方案.

2.4 基于深度学习的有监督领域关系抽取方法与经典方法的对比

基于有监督学习的经典方法严重依赖于词性标注、句法解析等自然语言处理标注工具中提供的分类特征,

而自然语言处理标注工具中往往存在大量错误,这些错误会在关系抽取系统中不断传播放大,最终影响关系抽取的效果.而基于深度学习的有监督方法可以在神经网络模型中自动学习特征,将低层特征进行组合,形成更加抽象的高层特征,用来寻找数据的分布式特征表示,能够避免人工特征选择等步骤,减少并改善特征抽取过程中的误差积累问题.

2.5 有监督领域实体关系抽取核心公式

流水线和联合方法是有监督实体关系抽取领域主流的两个派系,这两个派系的实体关系抽取现今衍生出多种不同的抽取方法,其抽取方法的核心公式见表 1.

Table 1 Supervised entity relationship extraction core formula

表 1 有监督实体关系抽取核心公式

类别	序号	方法名称	核心公式	公式类型
Pipeline	1	Hashimoto, 2013[19]	$d(x)=soft\max\left(W^{label}x+\sum_i W_i^{add}x_i'+b^{label}\right),\overline{\theta}=\frac{1}{T+1}\sum_{t=0}^{T}\theta_t$ x_i' 是同一棵句法树中的任何其他节点的特征向量; $\overline{\theta}$ 是模型平均参数,θ_t 是 t 次优化迭代之后的模型参数向量	分类
	2	CR-CNN[21]	$S_\theta(x)_c=r_x^T[W^c]_c$ $L=\log(1+\exp(\gamma(m^+-s_\theta(x)_{y^+})))+\log(1+\exp(\gamma(m^++s_\theta(x)_{c^-})))$ $S_\theta(x)_c$ 是类标签 $c\in C$ 的分数;γ是缩放因子,对预测误差进行更多惩罚, y^+为正确的类标签,而 c^-是错误的类标签	目标函数
	3	SDP-LSTM[11]	$J=-\sum_{i=1}^{n_c}t_i\log y_i+\lambda\left(\sum_{i=1}^{w}\|W_i\|_F^2+\sum_{i=1}^{v}\|U_i\|_F^2\right)$ t 是 groundtruth 的 one-hot 编码表示,λ用于指定权重惩罚的大小	目标函数
	4	Bi-LSTM-RNN[50]	$x_{penul}=r_{before}\oplus r_{former}\oplus r_{middle}\oplus r_{latter}\oplus r_{after}$ $p(R_i)=soft\max(R_i)=\frac{e^{w_{2i}\times x_{penul}}}{\sum_{j=1}^{\|R\|}e^{w_{2i}\times x_{penul}}}$ $r_{before},r_{former},r_{middle},r_{latter},r_{after}$ 是实体上下文的 5 个表示, x_{penul} 是实体最终表示,$p(R_i)$是各关系概率	分类
Joint	1	Bi-LSTM+ Bi-TreeLSTM[54]	$h_p^{(r)}=\tanh(W^{(r_h)}d_p+b^{(r_h)}),y_p=soft\max(W^{(r_y)}h_t^{(r)}+b^{(r_y)})$ d_p表示关系候选中的词对 p 之间的路径的依赖层关系	分类
	2	Bi-LSTM+ Attention[12]	$u_{st}^t=v^T\tanh(W_1[z_{st};b_{st}]+W_2[z_t;b_t]),p_{st}^t=soft\max(u_{st}^t)$ $z_{\leqslant t}$ 是前一时间步的顶部隐藏层表示,$b_{\leqslant t}$ 表示对应的标签嵌入	分类
	3	Bi-LSTM+ CNN[13]	$y_r=W_R\cdot(R_S\circ r)+b_R,p_r^i=\frac{\exp(y_r^i)}{\sum_{j=1}^{nc}\exp(y_r^j)}$ r 是由伯努利的概率 p 表示的二进制掩码矢量 nc 表示关系量的总数量,R_S表示关系特征	分类
	4	End-to-End+ 序列标注[56]	{B,I,E,O,S}表示单词在实体中的位置,分别表示 {实体开始,实体内部,实体结束,单个实体,其他的无关词} {$CF,CP,\ldots$}表示自定义的关系类型,{1,2}表示关系中的实体信息角色	序列标注方法

3 基于深度学习的远程监督实体关系抽取方法

3.1 远程监督实体关系抽取框架演化流程

面临大量无标签数据时,有监督的关系抽取消耗大量人力,显得力不从心.因此,远程监督实体关系抽取应运而生.Mintz[14]于 2009 年首次提出将远程监督应用到关系抽取任务中,其通过数据自动对齐远程知识库来解决开放域中大量无标签数据自动标注的问题.远程监督标注数据时主要有两个问题:噪声和特征提取误差传播.噪声问题是由于远程监督的强假设条件,导致大量数据的关系被错误标记,使得训练数据存在大量噪声;而特征

提取中的误差传播问题是由于传统的特征提取主要是利用 NLP 工具进行数据集的特征提取,因此会引入大量的传播误差.针对错误标签问题,Surdeanu[8]于 2010 年提出的多示例多标签学习方法、Lin[10]于 2016 年提出的 Attention 机制,都有效减弱了远程监督错误标签对抽取性能的影响.而自从深度学习的崛起和其在有监督领域取得良好的关系抽取效果后,用深度学习提取特征的思路来替代特征工程是一个非常自然的想法:用词向量、位置向量来表示句子中的实体和其他词语;用深度模型对句子建模,构建句子向量;最后进行关系分类.深度学习模型及其特点有:CNN 的扩展模型 PCNN+MIL[37]、PCNN+ATT[10](Attention 机制作为多示例机制的一种泛化)弱化错误标签问题;LSTM[57]获取实体对方向性信息;COTYPE[39]联合抽取实体和关系信息;深度残差网络[40]防止错误标签噪声的逐层累积.基于远程监督实体关系抽取框架的演化流程如图 7 所示.下面按照 PCNN 及其扩展模型、LSTM、COTYPE、深度残差网络的顺序来进行远程监督领域实体关系抽取的主流方法介绍.

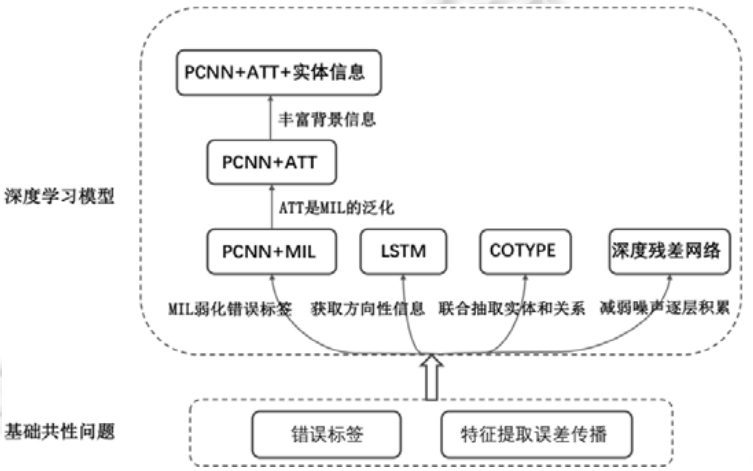


Fig.7 Evolutionary process of entity relationship extraction framework based on distantsupervision

图 7 基于远程监督的实体关系抽取框架的演化流程

3.2 基于深度学习的远程监督领域实体关系抽取主流方法介绍

3.2.1 基于 PCNN 及其扩展模型的实体关系抽取

经典的实体关系抽取在提取特征时使用 NLP 工具,会导致误差逐层传播,影响关系抽取效果.深度学习中的 PCNN 方法有效解决了特征提取误差传播的问题.而对于远程监督中错误标签引入噪声的问题,本模块采用多示例和注意力两种机制来缓解噪声问题.以下是基于 PCNN 及其扩展模型的实体关系抽取过程.

(1) 基于 PCNN 和多示例(MIL)的实体关系抽取

Zeng[20]提出了 PCNN 结合多示例的方法进行远程监督实体关系抽取,与 CNN 不同的是,PCNN 根据实体所在位置将句子切分成 3 段进行池化,从而得到更多和实体相关的上下文信息.而多示例学习是将实体对看成包,基于 At-least-one 假设,在包含实体对的所有句子中,选择使得关系概率最大的示例语句作为实体对的表示.关系抽取的具体流程为:

a) 示例语句编码:词向量、位置向量共同组成词语表示向量;

b) 卷积层:卷积部分是采用了常见的针对文本的卷积核设计,单向滑动;

c) 三段池化与最终关系分类:在池化层,是按照分段进行 Max Pooling 的,而 PCNN 的 P 是 Piecewise,将句子按照两个实体进行分割,分割得到 3 段,将这 3 段分别进行 Max Pooling.最后,使用一个 Softmax 分类器进行类别判断.

PCNN 结合多实例的方法虽然优化了传统远程监督的效果,但多实例实际上是给包打标签而不是给语句打标签,即从包含实体对的所有语句中只选择了一个语句,这必然导致丢失大量有用的句子信息.

(2) 基于 PCNN 和注意力机制(ATT)的实体关系抽取

Zeng 的多示例方法只用了包中一条语句信息,这就在一定程度上丢失了很多信息.针对此问题,Lin[10]在 Zeng 的基础上采用 Attention 机制,充分利用包内的信息,进一步减弱错误打标的示例语句产生的噪声.最终,标签正确分类的示例语句贡献较大,分配权重较高;标签错误分类的示例语句贡献较小,分配权重较低.从而提高分类的准确率.具体流程主要分为:

a) 包中示例分类:将实体对作为包,含实体对的句子作为包中示例;

b) 示例语句编码(句子特征提取):句子分词,将句子词语和实体转化为稠密实数向量,然后利用卷积、池化和非线性转换等操作构建起对应的句向量.句向量编码过程如图 8 所示;

c) 给句子加入注意力机制:给不同的句子赋予不同的权重$\alpha_1,\alpha_2,\alpha_3,\ldots,\alpha_n$,隐式地摒弃一些噪音语料,以此提升分类器的性能.这样使得网络的输出数目和关系数目相等,方便后续 Softmax 层进行分类.图 9 为原始句子包生成句子包向量的过程,原始句子通过 CNN 提取句子特征,构建句子向量,给包中不同句子添加不同的权重,构建出一个句子包向量.

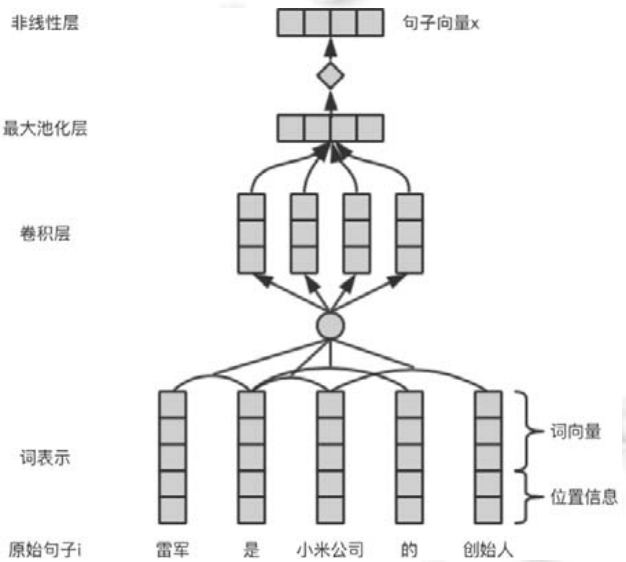


Fig.8 Construction process of sentence vector

图 8 句向量构建过程

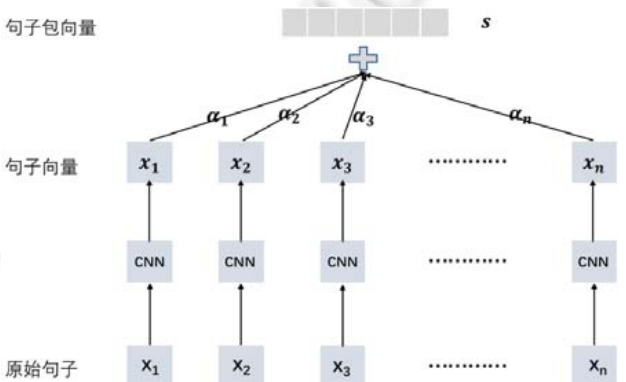


Fig.9 Generation process ofsentence package vector added attention mechanism

图 9 添加注意力机制的句子包向量生成过程

Attention 机制虽与多示例方法都是减弱错误标签带来的噪声问题,但多示例只用了包中一条语句信息,而 Attention 机制综合利用了包中所有示例语句信息,更好地提升了远程监督中关系抽取的效果.

(3) 基于 PCNN、注意力机制和实体表示信息的实体关系抽取

目前的远程监督关系抽取都集中在探索句子的语义信息层次上,忽略了实体本身的描述信息对关系抽取效果的影响.对此,Ji 在文献[38]中提出加入实体表示信息的深度学习实体关系抽取模型.此模型是在 PCNN 和 Attention 的基础上添加了实体的描述信息来辅助学习实体的表示,从而提高准确率.其提取关系流程主要为:

a) PCNN 模块:用 PCNN 提取句子特征,每个实体对对应一个包,用句子级别注意力机制给包中每个句子分配一个权重,综合利用包中所有句子的信息;

b) 提取实体信息:从 Freebase 和 Wikipedia 页面中提取实体描述以补充实体关系提取的背景知识,用一个传统的 CNN 模型(一个卷积层和一个最大池化层)从实体描述中提取特征.背景知识不仅为预测关系提供了更多信息,而且为注意力机制模块带来了更好的实体表示;

c) 特征融合:用交叉熵最小化目标函数,目标函数由句子级别注意力机制和实体信息共同决定.

本文实际检测到:当前远程监督关系抽取模型如果在没有实体背景信息的情况下,其在抽取某些实体对关系时效果不佳.针对此问题,作者提出使用实体表示信息丰富其背景知识,以便更好地预测关系.实验表明在前人模型的基础上加入此创新点,均明显地提升了当前模型的效果.

3.2.2 基于 LSTM 的实体关系抽取方法

传统的远程监督方法在提取特征时采用 NLP 工具包,加重了错误传播、错误积累的问题,所以 He 等人[57]提出一种 SE-LSTM 结合多示例学习的方法来解决远程监督中错误传播、错误积累问题,其模型如图 10 所示.

a) LSTM 网络抽取实体对方向性信息(图 10 左上部分):HE 等人首先将句子的最短依存路径(SDP)分割成两个子路径作为 LSTM 结构的输入,自动地抽取特征,以此来抽取实体对的方向性信息;

b) CNN 网络提取句子整体信息(图 10 右部分):尽管 SDP 对关系抽取非常有效,但是这并不能捕捉到句子的全部特征.针对此问题,作者将全部句子放进 CNN 网络,进而抽取句子的全部信息(sentence embedding);

c) 特征融合(图 10 左下部分):最后,将 LSTM 隐藏层单元以及 CNN 的非线性单元相融合,通过 Softmax 层来标注实体对对应的关系.

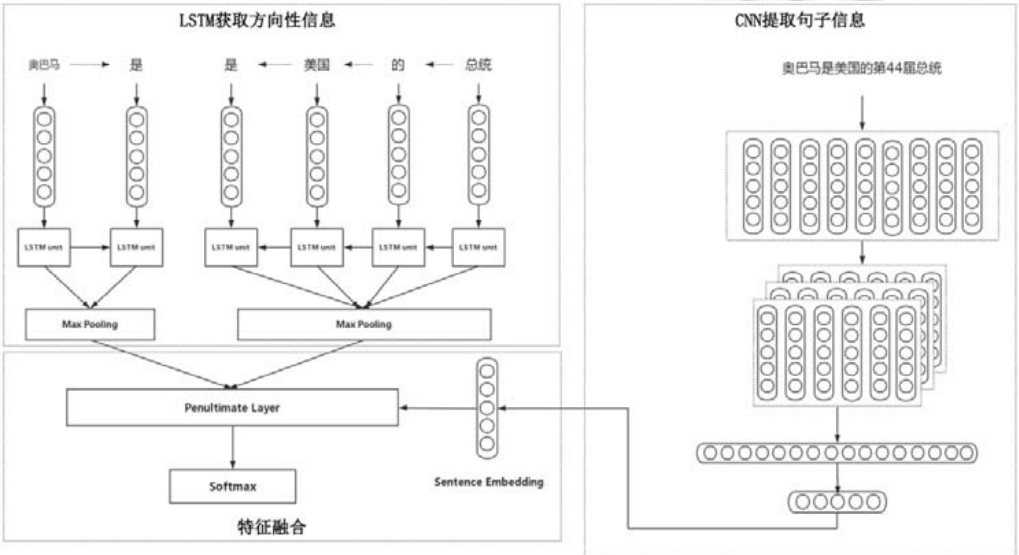


Fig.10 Distantsupervision entity relationship extraction framework based on LSTM

图 10 基于 LSTM 的远程监督实体关系抽取框架

本文提出的 SE-LSTM 网络结合多示例的方法,其可以在不需要任何 NLP 工具包的帮助下自动地抽取特征,并且通过两个 LSTM 提取实体对的方向性信息.实验表明,该方法大大地提升了关系抽取的准确率.

3.2.3 基于 COTYPE 联合抽取模型的实体关系抽取方法

现有的远程监督关系抽取模型通常只能在某一特定领域进行关系抽取工作,并且将实体抽取和关系抽取两项工作分开进行,分开进行的方式会导致错误的累积传播,不易优化扩展模型.针对此问题,Ren 在文献[39]中提出了联合抽取模型 COTYPE,此模型的提出,主要解决在远程监督关系抽取过程中面临的 3 大挑战:1) 事先训练好的命名实体识别器限制了领域之间的扩展;2) 将实体抽取和关系抽取分开导致错误的累积传播;3) 在远程监督中标签噪声问题.COTYPE 的框架主要分为 3 个部分.

a) 数据预处理:在训练语料上运行文本分割算法,得到候选实体;给同一句话的两个候选实体构建关系,用三元组表示;最后分析文本,抽取文本特征;
b) 联合训练实体和关系向量空间:将候选实体、候选关系、文本特征等嵌入到关系空间以及实体空间,并对两者进行联合建模;
c) 对实体类型和关系类型进行推理预测.

COTYPE 模型与 PCNN 等单模型相比不仅可以扩展到不同领域,而且通过把实体抽取和关系抽取两个任务结合,较好地减弱了错误的累积传播.实验结果表示,其明显提升了当时 State-of-the-art 的效果.

3.2.4 基于深度残差网络的实体关系抽取方法

一般来说,深层神经网络能抽取更深的语义特征,所以 Huang[40]实验了 9 层 CNN 的实体关系抽取模型.但事实发现,9 层 CNN 抽取效果不如单层.Huang 猜测可能是由于远程监督的数据里有太多错误标签的数据,错误标签带来的噪声随着神经网络层次的加深逐渐被放大,导致 9 层效果比单层的差.因此,提出一种深度残差网络模型来解决深层网络增大噪声的问题,其采用残差网络设法使浅层网络的特征跳跃传递至深层网络,让网络可以选择较不被噪声影响的那层网络特征来进行关系分类.在性能上,9 层的残差网络可达到 State-of-the-art (PCNN+ATT)模型相似的效果.

3.3 基于深度学习的远程监督关系抽取方法与经典方法的对比

经典的远程监督方法是在解决远程监督中强假设条件造成大量错误标签的问题,而深度学习方法主要是是在解决特征提取中误差传播问题.

远程监督的提出,是因为在开放域中存在大量无规则非结构化数据,人工标注虽能使标注的准确率较高,但是时间和人力消耗巨大,在面对大量数据集时显得不切实际.因此,远程监督实现一种数据集自动对齐远程知识库进行关系提取的方法,可进行自动标注数据.但由于其强假设条件造成大量错误标签问题,之后,经典的远程监督的改进都是在改进处理错误标签的算法.

深度学习的提出,是因数据特征构造过程依赖于 NER 等 NLP 工具,中间过程出错会造成错误传播问题.且现今基于深度学习的远程监督实体关系抽取框架已包含经典方法中对错误标签的探讨解决,因此可以认为现今的远程监督关系抽取框架是基于传统方法的扩展优化.

3.4 基于深度学习的远程监督关系抽取方法与有监督方法的对比

有监督的实体关系抽取依靠人工标注的方法得到数据集,数据集准确率、纯度较高,训练出的关系抽取模型效果较好,具有很好的实验价值.但其人工标注数据集的方法耗费大量人力成本,且标注数据的数量有限、扩展性差、领域性强,导致构造的关系抽取模型对人工标注的数据具有依赖性,不利于模型的跨领域泛化能力,领域迁移性较差.

远程监督在面对大量无标签数据时,相较于有监督实体关系抽取具有明显优势.人力标注大量无标签数据显得不切实际,因此远程监督采用对齐远程知识库的方式自动标注数据,极大地减少了人力的损耗且领域迁移性较强.但远程监督自动标注得到的数据准确度较低,因此在训练模型时,错误标签的误差会逐层传播,最终影响整个模型的效果.因此,现今的远程监督实体关系抽取模型的效果普遍比有监督模型抽取效果效果差.基于深

度学习的有监督和远程监督实体关系抽取效果对比可见表 2.

Table 2 Comparison of supervised and remotely supervised entity relationships based on deep learning

表 2 基于深度学习的有监督和远程监督实体关系抽取对比

	有监督	远程监督
数据集标注方法	人工标注	远程对齐知识库
数据集特点	准确度高,噪声小	准确度低,噪声大
数据集规模	较小(通常情况)	较大
成本	较高	较低
迁移性	较差	较好
领域性	较强	较低
抽取效果	较好	较差

3.5 远程监督领域实体关系抽取方法核心公式

现今,基于深度学习的远程监督实体关系抽取研究点主要集中在远程监督的噪声问题和特征提取的误差传播两方面,远程监督部分实体关系抽取核心公式为表 3.

Table 3 Distant-supervised entity relationship extraction core formula

表 3 远程监督实体关系抽取核心公式

方法名称	创新方法	核心公式	公式类别
远程监督+多示例[58]	多示例单标签	$F_m^{once}(z,y_{truth})=\begin{cases}1, & y_{truth}\neq NA\wedge\|z\|\geqslant 1\\ -1, & y_{truth}\neq NA\wedge\|z\|=0\\ -\|z\|, & \text{otherwise}\end{cases}$ $\|z\|$作为实体对存在某关系的句子个数,F 为含此实体对的包标签	多示例的包表示
PCNNs+MIL[37]	三段池化层	$p_{ij}=\max(c_{ij}),1\leqslant i\leqslant n,1\leqslant j\leqslant 3$ p_{ij} 为池化层结果,将一个句子向量按实体位置分割为三段池化	三段池化层取值
PCNN+ATT[10]	多注意力	$\alpha_i=\dfrac{\exp(e_i)}{\sum_k\exp(e_k)}$, $\vec{s}=\sum_i\alpha_i x_i$,α_i 为句子权重,$\vec{s}$ 为句子集向量	注意力机制
APCNNs+D[38]	添加实体描述信息	$L_e=\sum_{i=1}^{\|D\|}\|e_i-d_i\|^2,\min L=L_A+\lambda L_e$ L_e 为实体描述信息的目标函数,L_a 为 APCNNs 部分目标函数,L 为模型最终目标函数 LA	目标函数
COTYPE[39]	实体与关系信息联合抽取	$\min_{\{z_i\},\{c_j\},\{r_k\},\{m_i\},\{c_j'\},\{y_k\}}=M+Z+ZM$ 最小化联合期望函数,M 为实体期望函数,Z 为关系期望函数,ZM 为关系、实体相互影响魔性的期望函数	期望函数
深度残差网络[40]	9 层残差网络	$\tilde{c}_l=f(w_1+c_{i:i+h-1}+b_1),c_l'=f(w_2+c_{i:i+h-1}+b_2),c=c+c'$ $\tilde{c}_l$ 和 c_l' 分别是两个卷积层,c 是最终的残差结果	卷积层

4 基于深度学习的实体关系抽取新模型与新思路

4.1 融合深度增强学习的实体关系抽取

近期,随着增强学习方法的兴起,给予实体关系抽取又一种新的思路.有学者试图将增强学习[59]的方法与深度学习的方法融合起来,进行实体和关系的联合抽取.Feng 等人[60]在 2017 年提出了基于增强学习和深度学习的联合学习方法抽取实体和关系的模型.该模型中,增强学习将任务建模为两步决策过程,如图 11 所示:第 1 个决策根据实体抽取的初步结果,判断包含目标实体对的句子是否是一个关系;第 2 个决策将关系进行分类.通过设计每步的奖励函数,可以将实体提取的信息传递给关系提取并获得反馈,以便同时提取实体和关系.深度学习用于自动捕获非结构化文本中最重要的信息,这些信息代表决策过程中的状态,首先使用 Bi-LSTM 来模拟上下文信息,将实体抽取任务视为序列标注任务,实现初步的实体提取;在提取结果的基础上,基于注意力机制的方法可以表示包括目标实体对的句子,以在决策过程中生成初始状态;接着使用 Tree-LSTM 来表示关系,在决策过

程中生成过渡状态;最后采用Q-Learning 算法,在两步决策过程中得到控制策略π.该方法解决了在增强学习与深度学习的联合模型中,如何将实体信息传递给关系抽取,使实体和关系信息能够交互并获得反馈的问题.在ACE2005 数据集上的实验结果,比现有技术的方法获得更好的性能,并且召回率评分提高了 2.4%.

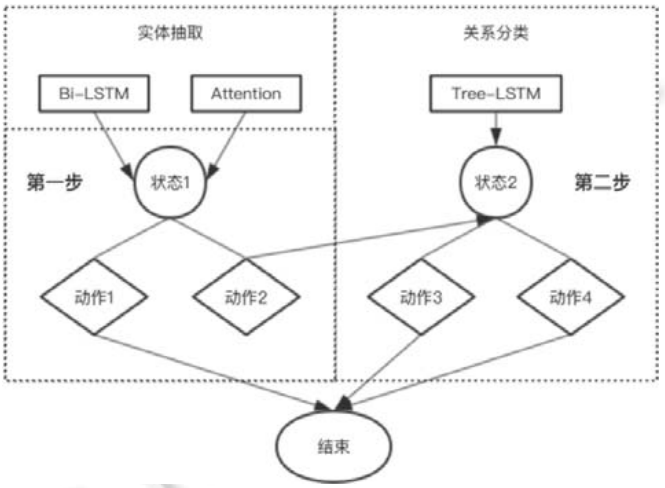


Fig.11 Two-step decision process

图 11 两步决策过程

Qin[61]于 2018 年 ACL 会议上提出一种深度增强学习的远程监督实体关系抽取方法,认为多示例和注意力机制并非最理想的降噪方法,那些被错误打标的数据依旧作为模型的训练数据,影响着关系抽取的效果.因此,Qin 用深度增强学习方法训练一个正例、负例数据识别器.不同于之前研究中将负例移除的方式,Qin 是将不存在目标关系的示例语句放入负例集中,将正例数据和负例数据正确分类,并充分利用了正例数据和负例数据的信息.

4.2 融合生成对抗网络的实体关系抽取

生成对抗网络是实体关系提取中的新兴方法,其通过在词向量表示阶段引入对抗性噪声并给出新的损失函数来增加模型的准确率.其主要思路是:生成器和判别器为博弈对方,生成器拟合数据的产生过程生成模型样本,判别器通过增加噪声样本增强模型准确率和鲁棒性,优化目标是达到纳什均衡.

生成对抗网络是由 GoodFellow 等人[62]在 2014 年提出的一种生成模型,在图像和视觉领域取得广泛的研究和应用.从 2016 年开始,Miyato[23,63]逐渐将对抗训练引入文本分类任务中.Wu[24]于 2017 年将生成对抗网络引入弱监督实体关系抽取中,证明词向量加入对抗性噪声之后,其进入 CNN 或 RNN 等深度模型中的提取效果比直接进入深度模型提取关系的准确率高.Qin 在文献[17]将对抗的思路加入模型中来对隐含话语的关系进行分类,通过隐式网络和竞争特征鉴别器之间的竞争来实现自适应模仿方案,在 PDTB 基准测试中实现了最先进的性能.Qin[64]于 2018 年将生成对抗网络引入到远程监督关系抽取中,用于筛选错误标签,最终达到降噪的效果.实验结果表明,此模型优于现今效果最好的远程监督实体关系抽取模型.

生成对抗网络筛选错误标签数据的流程如图 12 所示.

a) 预训练:对生成器和鉴别器进行预训练,得到生成器和鉴别器的参数θ_G 和θ_D.由于在良好初始化参数的情况下对抗训练很容易趋于收敛,因此预训练具有很好的优化效果.本文生成器和鉴别器都用简单的卷积神经网络,相比于循环神经网络,卷积神经网络具有更少的参数;

b) 数据划分:一次迭代(epoch)扫描远程监督训练集中所有正例集 $P=\{s_1,s_2,\ldots,s_j,\ldots\}$,将其划分为 N 个包:$B=\{Bag_1,Bag_2,\ldots,Bag_k,\ldots\}$,一次处理一个包中的全部数据;

c) 生成器训练:生成器计算包中正样本的概率分布,其产生的高置信样本被认为是真实的正例样本,然

后根据这个概率分布进行抽样;

d) 对抗器训练:对抗器接收这些高置信度样本,但将其视为负样本;相反,低置信度的样本仍被视为正样本.在这个过程中,模型会以预训练的参数进行初始化;

e) 交替训练:对于生成的样本,生成器使真正的概率最大;相反,对抗器使这个概率最小.两个网络交替进行训练,更新θ_G和θ_D.

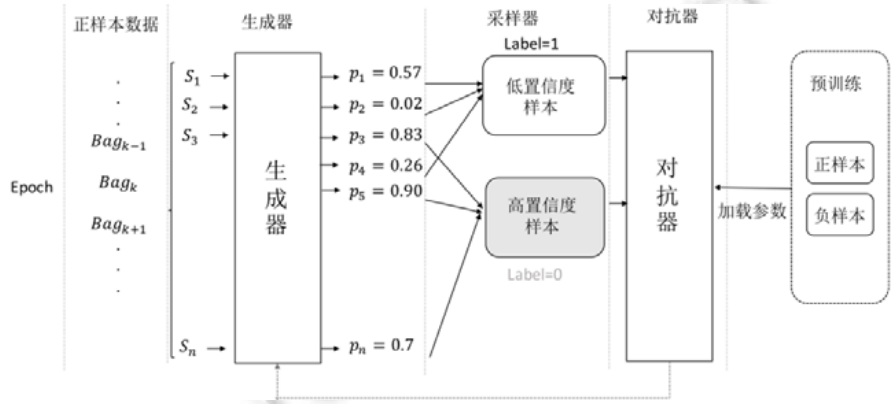


Fig.12 Process of filtering error tags by generative adversarial networks

图 12 生成对抗网络筛选错误标签数据的流程

对比实验结果表明,PCNN+ATT+DSGAN 模型较 PCNN+ATT 而言,AUC 和 *p*-values 均有明显的改善.用生成对抗网络进行训练集噪声数据筛选,会提高远程监督领域实体关系抽取效果.

5 基于深度学习的实体关系抽取在生物医药领域中的最新应用进展

实体关系抽取是信息抽取的核心任务[65,66],其主要通过对文本信息建模,自动抽取出实体对之间的语义关系,提取出有效的语义知识.目前,基于深度学习的实体关系抽取已逐渐应用到垂直领域并取得了不错的效果,其中,实体关系抽取在生物医药领域的应用尤为广泛.深度学习实体关系抽取可以发掘生物医学中药品实体与疾病间深层次的特征,在毒理学研究、药物发现和药物安全监测方面有着广泛的应用.下面依次从 CNN, LSTM 模型的角度简要介绍深度学习实体关系抽取在医药领域的最新应用.表 4 为深度学习模型在生物医药领域中的应用.

Table 4 Deep learning entity relationship extraction usedin biomedicine field

表 4 深度学习实体关系抽取在生物医药领域中的应用

领域	深度学习方法	提出年份	解决问题
生物医药	CNN[67]	2016	首次应用深度模型,对临床文本等碎片化内容进行关系抽取
	最大熵+CNN[68]	2017	抽取化学药物与疾病之间的关系
	Bi-LSTM-RNN[53]	2017	抽取药物与疾病实体之间的关系、细菌与位置实体之间的关系
	Bi-LSTM+ATT[69]	2018	实体识别和不良药物事件提取
	SVM+CNN+RNN[70]	2018	抽取生物医学文献中化学品和蛋白质之间的关系
	Bi-LSTM[26]	2018	抽取疾病与治疗药品间关系
	CNN+LSTM[71]	2018	抽取化学药物与疾病之间的关系

从文本中提取生物医学实体及其关系,对生物医学研究具有重要的应用价值.以前的工作主要是利用基于特征的流水线模型来处理这个任务,当采用基于特征的模型时,需要进行大量特征工程工作,耗费时间且抽取效果参差不齐.因此,学者们试图将深度学习的方法引入生物医药领域的关系抽取中来提升效果.

从生物医疗领域的科研文章、医疗报告、电子医疗记录抽取相关信息,已经成为了当前生物医药领域的研究热点.2016 年 6 月,Sahu 等人[67]首次提出基于卷积神经网络(CNN)的临床文本关系提取新框架,临床文本相较

于科研文章而言,内容更具碎片化和不完整性,因此关系抽取的过程更具挑战性.Sahu 将每个句子用词级向量、位置向量、词性特征、词干特征、实体类型信息来共同表示,丰富句子表示信息;并且用 CNN 网络进行关系抽取,减少了对专家特征知识定义质量的依赖,模型在 i2b2-2010 临床关系提取挑战数据集超过了当前 state-of-the-art 的效果.

2017 年 1 月,Gu 等人[68]的论文用最大熵改进了 Sahu 的 CNN 模型,对化学药物与疾病之间的关系进行抽取.通过 CNN 网络抽取了文本句子的上下文特征以及依存特征,获得了更加精确、有效的句子信息.模型在 BioCreative-V CDR 语料库(包括 1 500 篇美国国立医学图书馆生物医学数据库论文(MEDLINE),所有论文都被手工标注了化学与疾病)上达到了当前 State-of-the-art 的效果.

Peng 等人在 2018 年 BioCreative VI Workshop 上发表的文献[70]结合了 SVM,CNN 和 RNN 模型,联合挖掘生物医学文献中化学品和蛋白质之间的关系,从而证明了生物医学文献自动关系提取方法的有效性.Peng 将句子向量、位置向量、词干特征、句子的依存特征作为 SVM,CNN 以及 RNN 模型的输入,最终将 3 种模型预测的结果进行投票,获得最终的关系预测.结果表明,在 BioCreative VI 的 CHEMPROT 系统精确度为 0.726 6,召回率为 0.573 5,*F* 值为 0.641 0.此模型在 2017 年挑战期间取得了最高效果.

Li[53]在 2017 年 BMC Bioinformatics 会议上提出将 CNN 和 Bi-LSTM-RNN 应用于生物医药领域的关系抽取任务中,在药物与疾病实体之间的关系抽取、细菌与位置实体之间的关系抽取这两个任务中分别比最新技术提高了 8.0%和 9.2%.本文所用模型对应上文的有监督领域的联合模型,同时进行命名实体识别与实体间关系抽取两个任务.使用 CNN 提取字符级信息,用 Bi-LSTM 识别生物医学实体,再结合 Bi-LSTM-RNN 沿着两个目标实体的最短依存路径(SDP)方向学习实体间关系表示.这些表示用于确定实体间最后的关系类别.此模型在实际应用中取得了杰出的效果,这也表明了深度学习实体关系抽取在生物医学文本挖掘中研究的重要性.

药物引起的不良反应是一个潜在的危险问题,可能导致患者死亡和发病.提取药物不良事件以及挖掘药物与疾病间关系,是生物医学研究中的重要问题.2018 年 1 月,Ramamoorthy 等人[69]采用 Bi-LSTM 结合注意力机制的序列模型进行实体识别和不良药物事件提取,利用临床文本中的当地语言实现序列内相互作用,以便对药物和疾病实体间关系进行共同学习,从而抽取到最合适的关系.模型证明,用此种方式进行事件和关系抽取的性能优于先前工作中使用的基于最短依存路径(SDP)方法.

Chikka 等人[26]在 2018KDD 上提出一种结合深度学习和规则的关系抽取模型,解决如何抽取疾病与治疗药品间关系的问题.文中利用深度学习的词级和句子级表示信息来提取治疗方案与医疗问题之间的关系,使用基于规则的方法处理数据集中可用的样本数量较少的关系,最终通过 Bi-LSTM 和基于规则的模型联合得出最终关系.最终结果在 I2b2 2010 关系提取任务的关系类上取得了良好的性能.结合深度学习和基于规则的模型可以深入挖掘疾病与药品之间关系,在决策支持系统、安全监视和新的药品发现中有着广泛应用.

Nguyen[71]在 2018 年 BioNLP 上提出通过 CNN+CNNchar 和 CNN+LSTMchar 模型来抽取生物医学文本中化学药品与疾病之间的关系.不同于之前模型中只用 CNN 提取词语与字符级信息,Nguyen 提出 CNN 和 LSTM 共同训练字符级别的词向量,解决生物医药领域专有名词没有特定的词向量这一问题,将字符集别词向量和词级别词向量拼接作为 CNN 关系抽取网络的输入.将模型应用于 BioCreative-V CDR 语料库中的任务数据,其结果表明:利用基于 CNN 和 LSTM 的字符级单词表示模型改进了不使用此信息模型的关系抽取效果,更好地抽取化学药品与疾病之间的关系.

6 基于深度学习的实体关系抽取的数据集及其评测方法

6.1 数据集介绍

近年来,用作深度学习关系抽取实验评估的标准数据集主要有 SemEval-2010 Task 8 公开数据集、ACE2004 实验语料、NYT-FB 数据集等.

(一) 有监督领域

有监督领域的实体关系抽取主要采用 MUC 关系抽取任务数据集、ACE04、ACE05、SemEval-2010 Task 8

公开数据集,部分论文采用 MPQA 2.0 语料库和 BioNLP-ST 2016 的 BB 任务数据集.有监督方面评测标准主要以 *F*1 值来统计.

- MUC 关系抽取任务数据集:MUC-7 包含五大评测任务:命名实体识别、指代消解、模版元素填充、模版关系确定和场景模版填充.其中,关系抽取首次作为一个独立的评测任务被提出来.MUC-7 的数据语料主要是取自新闻语料,主要是飞机失事事件报道和航天器发射事件报道.MUC 会议停开以后, ACE 会议也将关系抽取任务作为会议的一个子任务;
- ACE 关系抽取任务数据集:ACE 会议从 2002 年~2007 年一直将关系抽取任务作为一个子任务,其中获得广泛认可的是 ACE04/ACE05.其中,ACE04 语料库来源于语言数据联盟(linguistic data consortium,简称 LDC),分成广播新闻(BNEWS)和新闻专线(NWIRE)两个部分,总共包含 451 个文档和 5 702 个关系实例.ACE04 提供了丰富的标注信息,从而为信息抽取中的实体识别、指代消解和关系抽取凳子任务提供基准(benchmark)的训练和测试语料库.而 ACE05 作为 ACE04 的扩充,对 ACE04 数据集进行了适当的修改与完善;
- SemEval-2010 Task 8 数据集:SemEval 是由 Senseval 演变而来的语义评测.Senseval 是由 ACL-SIGLEX 组织的国际权威的词义消歧评测,但由于 Senseval 中除词义消歧外有关语义分析的任务越来越多,之后,Senseval 委员会决定把评测名称改为国际语义评测(SemEval).SemEval-2010 Task 8 数据集是 2010 年 SemEval 语义评测的子任务,构建于 2009 年,此任务用于名词间多种语义关系的分类.数据集根据预设定的 9 种互不相容关系从各大数据源收集而来,数据源包括 WordNet,Wikipedia data,Google *n*-grams 等.数据集共包含 10 717 条数据,其中,训练集有 8 000 条,测试集有 2 717 条.数据集中 9 种关系,分别为:Cause-Effect(因果关系),Instrument-Agency(操作、使用关系),Product-Producer(产品-生产者关系),Content-Container(空间包含关系),Entity-Origin(起源关系),Entity-Destination(导向关系),Component-Whole(组件-整体关系),Member-Collection(成员-集合关系),Message-Topic(主题关系).每条数据是一个包含实体对的句子,类别标签为实体对在该句中表现出的关系;
- MPQA 2.0 语料库:包含来自各种新闻源的新闻文章和社论,数据集中共有 482 个文档,包含 9 471 个带有短语级别注释的句子.数据集中包含观点实体的黄金标准注释,如观点表达、观点目标和观点持有者;还包含观点关系的注释,如观点持有者和观点表达之间的 IS-FROM 关系、观点目标和观点表达之间的 IS-ABOUT 关系;
- BioNLP-ST 2016 的 BB 任务:此任务是针对细菌/位置实体抽取和两者间 Lives_In 关系抽取而设立的一个标准竞赛,数据集由来自 PubMed 的 161 个科学论文摘要组成,数据集中包含 3 种类型的实体:细菌、栖息地和地理位置;包含一种关系:Lives_In,指由细菌-栖息地构成的 Lives_In 关系或由细菌-地理位置构成的 Lives_In 关系.

(二) 远程监督领域

远程监督领域的实体关系抽取主要采用 NYT-FB 数据集.这个数据集是由 Freebase 知识库对其纽约时报的文本获得的数据集.训练数据为知识库对其 2005 年、2006 年文本获得的,测试库数据为知识库对其 2007 年文本获得的.NYT-FB 数据集中共有 53 种关系,共计 695 059 条数据(其中训练集包含 522 611 条训练语句,训练数据中有近 80%的句子的标签为 NA,测试集包含 172 448 条测试语句),通过结合 FreeBase 对 NYT 语料做实体链接、关系对齐等操作进行标注,最终得到一个被广泛使用的关系抽取数据集.

6.2 评测方法介绍

关系抽取领域有 3 项基本评价指标:准确率(precision)、召回率(recall)和 *F* 值(*F* measure).

(一) 准确率

准确率是从查准率的角度对实体关系抽取效果进行评估,其计算公式为

$$Precision_R = \frac{\text{被正确抽取的属于关系}R\text{的实体对个数}}{\text{所有被抽取为关系}R\text{的实体对个数}} \tag{6-1}$$

(二) 召回率

召回率是从查全率的角度对抽取效果进行评估,其计算公式为

$$Recall_R = \frac{\text{被正确抽取的属于关系}R\text{的实体对个数}}{\text{实际应被抽取的属于关系}R\text{的实体对个数}} \tag{6-2}$$

(三) F 值

对与关系抽取来说,准确率和召回率是相互影响的,二者存在互补关系,因此,F 值综合了准确率和召回率的信息,其计算公式为

$$F_\beta = \frac{(\beta^2+1)\cdot Precision \cdot Recall}{Precision + Recall} \tag{6-3}$$

β是一个调节准确率与召回率比重的参数,实际测试中,一般认为准确率与召回率同等重要,因此,β值一般设置成 1.因此,上式可以表示为

$$F_1 = \frac{2\cdot Precision \cdot Recall}{Precision + Recall} \tag{6-4}$$

6.3 深度学习实体关系抽取典型论文的数据集与评测标准

不同模型的数据集及其评测标准见表 5.

Table 5 Different models of data sets and their evaluation criteria

表 5 不同模型的数据集及其评测标准

关系抽取方法	序号	Model	数据集	评测指标	评测值	发表年份	发表会议
流水线	1	MV-RNN(POS, WordNet,NER)[46]	SemEval-2010 Task 8	*F*1	82.4	2012	EMNLP
	2	RNN[19]	SemEval-2010 Task 8	*F*1	79.4	2013	ACL
	3	Convolutional DNN[20]	SemEval-2010 Task 8	*F*1	82.7	2014	COLING
	4	SDT-LSTM[11]	SemEval-2010 Task 8	*F*1	83.7	2015	EMNLP
	5	CR-CNN[21]	SemEval-2010 Task 8	*F*1	84.1	2015	ACL
	6	Vote-BIDIRECT[22]	ACE05	*F*1	84.1	2015	Computer Science
	7	Dependency paths from the object to subject[47]	SemEval-2010 Task 8	*F*1	85.4	2015	Computer Science
	8	ER-CNN + R-RNN[48]	SemEval-2010 Task 8	*F*1	84.9	2016	NAACL
	9	Multi-Level attention CNNs[49]	SemEval-2010Task 8	*F*1	88.0	2016	ACL
	10	Bi-LSTM-RNN[50]	SemEval-2010Task 8	*F*1	83.1	2016	ACL
联合学习	11	Bi-LSTM+Bi-TreeLSTM[12]	ACE05	*F*1	55.6	2016	ACL
	12	LSTM[54]	MPQA 2.0 语料库	*F*1	54.98(IS_ABOUT) 58.22(IS_FROM)	2016	ACL
	13	Bi-LSTM+ Bi-TreeLSTM[53]	BioNLP-ST 2016 的 BB 任务数据集	*F*1	28.5	2017	PAKDD
	14	Bi-LSTM+Attention[13]	ACE 2005	*F*1	55.9	2017	ACL
	15	Novel tagging scheme[55]	NYT	*F*1	52.0	2017	ACL
远程监督	16	PCNNs+MIL[37]	NYT-FB	Precision (Top100)	86.0	2015	EMNLP
	17	APCNNs[38]	NYT-FB	Precision (Top100)	87.0	2016	ACL
	18	APCNNs+D[38]	NYT-FB	Precision (Top100)	87.0	2017	AAAI
	19	DMN[72]	NYT-FB	Precision (Top100)	89.0	2017	IJCAI
	20	APCNN+soft_label[73]	NYT-FB	Precision (Top100)	84.0	2017	ACL
	21	JointD+KATT[74]	NYT-FB	Precision (Top100)	80.6	2018	AAAI
	22	CNN+RL[75]	NYT-FB	*F*1	42.0	2018	AAAI
	23	MIMLCNN[76]	NYT-FB	Precision (Top100)	69.0	2016	COLING
	24	RNN-Adv[24]	NYT-FB	*F*1	38.2	2017	ACL
	25	ResCNN-9[40]	NYT-FB	Precision (Top50)	88.0	2017	ACL

表 5 中,序号 1~序号 15 是有监督领域实体关系抽取的典型模型与其相关信息介绍,序号 16~序号 25 是远程监督领域实体关系抽取的典型模型与其相关信息介绍.其中,1~10 是有监督领域中流水线类别的模型,序号 11~序号 15 是有监督领域的联合学习类别的模型.

参考常耀成[77]在《软件学报》中的数据集整理的方式,本文数据集描述与下载链接见表 6.

Table 6 Dataset description and download link

表 6 数据集描述与下载链接

数据集	简述	URL
SemEval-2010 Task 8	包含 10 717 条数据(训练集 8 000 条,测试集 2 717 条);共包含 9 种互不相容的关系,如因果关系、包含关系等	https://www.researchgate.net/publication/271452073_SemEval-2010_task_8
MPQA 2.0 corpus	来自各种新闻源的新闻文章和社论,数据集中共有 482 篇文章,9 471 条句子	http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/
ACE05	7 种实体类型和 6 种实体关系类型	https://www.nist.gov/speech/tests/ace/ace05
openIE	50 亿网页数据,提取开放域关系三元组	https://github.com/dair-iitd/OpenIE-standalone
ACE04	7 种实体类型和 7 种实体关系类型	https://www.nist.gov/speech/tests/ace/ace04
NYT	53 种关系,共计 695 059 条数据(其中训练集包含 522 611 条训练语句,测试集包含 172 448 条测试语句)	https://github.com/shanzhenren/CoType

7 未来研究方向和总结

目前,基于深度学习的实体关系抽取已经取得了极大成功,但依旧值得学者们不断探索.通过对现有实体关系抽取研究工作进行总结,未来可从以下几个方面展开相关研究.

(1) 重叠实体关系识别

目前,就重叠实体关系识别这一问题,已有的实体关系识别模型还没有给出相应的解决方法.尽管 Zheng[55]提出的新标注策略解决了参数共享方法存在冗余实体的问题,真正做到了将两个子任务合并成一个序列标注问题,但该方法仍然没有解决重叠实体关系问题.故未来重叠实体关系仍会是学者研究和攻克的一大难题.此外,因 Zheng[55]新标注策略的提出,未来在这套标注策略上也可以进行更多的改进和发展,来进一步完善端到端的关系抽取任务.

(2) 跨句子级别关系抽取

现今,关系抽取任务集中在对一句话内识别出的实体对进行关系分类,而按照自然语言的习惯,实体对分别位于不同句子中的情况也十分常见.现有的指代消解任务可以通过指代对象识别和指代对象中心词抽取有效影响多种自然语言处理任务系统的性能,但其存在依赖人工特征强、精确度不够高的问题.因此,融合并改进指代消解和关系抽取模型,是未来解决跨句子级别关系抽取任务中可以研究探讨的一种方案.

此外,Peng 等人[78]于 2017 年提出了基于图的 LSTM 网络(graph LSTM)的一般关系提取框架,可以很容易地扩展到跨句子 N 元关系提取.图公式提供了一种探索不同 LSTM 方法的统一方法,它能结合各种句内和句间的依赖关系,如顺序、句法和语篇关系等;能学习实体的上下文表示,以用作关系分类器的输入,简化与任意元关系的处理,并且能够利用相关关系进行多任务学习.通过在两个重要的精确医学环境中评估该框架,证明了其在传统监督学习和远程监督方面的有效性.因此,基于图结构进行实体关系抽取也可作为解决跨句子级别关系抽取问题的一种方案.

(3) 关系类型 OOV 问题

现今,完成关系抽取任务的主流方法中,均没有有效地解决关系类型 OOV(out of vocabulary)问题.对于没出现在训练集中的关系类型,已有的模型框架无法准确地预测出实体对所属的正确关系类型.在 SemEval-2010 的评测任务 8 中,因考虑到句子实例中实体对的先后顺序问题,引入了 Other 类对不属于已有关系类型的实例进行描述,然而这只是减少了存在关系的实体对的损失,提升了模型判断关系提及的能力,对 Other 类中实体对的关系却难以定义,关系模糊,需要人工干预和判断.因此,关系类型 OOV 问题也是未来亟待解决的问题之一.

(4) 解决远程监督的错误标签问题

远程监督中的假设过于肯定,难免引入大量的噪声数据.为缓解错误标注的问题,目前主流的方式是:(a) 利用多示例学习方法对测试包打标签;(b) 采用 Attention 机制对不同置信度的句子赋予不同的权值.但这两种方法都不可避免地会将一些不具有某个关系的句子作为这个关系的训练语句:在多示例学习方法的情况下,若一个包中全是负例(包中没有一个句子的关系是实体对对齐知识库得到的关系),即使取出概率最大的语句作为这个包的训练语句,其仍是噪声语句;而在 Attention 机制下,虽将并不代表实体对关系的语句给予较小的权重,但本质上仍是将其作为正例放入训练集中,仍是会引入噪声.Qin[61]将深度增强学习引入远程监督领域,将不存在目标关系的示例语句放入负例集中,是远程监督领域解决噪声问题的一个新兴方法.但解决噪声的方法远不止这 3 种,如何采用有效的方式来解决远程监督的错误标签问题,是实体关系提取发展过程中研究的重要问题.

(5) 远程监督领域错误传播问题

现今,实体关系抽取的典型模型是 PCNN+ATT,但其主要利用的是句子的语义信息.虽已有论文利用句子的语法信息[79]将依存句法树用于实体关系抽取,但效果并不惊人.因此,如何将语义与语法信息有效融合来抽取实体关系,也是今后优化深度模型的主要方向之一.

References:

[1] Golshan PN, Dashti HR, Azizi S. A study of recent contributions on information extraction. arXiv preprint arXiv:1803.05667, 2018.

[2] Xu J, Zhang ZX, Wu ZX. Review on techniques of entity relation extraction. New Technology of Library and Information Service, 2008,24(8):18−23 (in Chinese with English abstract).

[3] Gan LX, Wan CX, Liu DX, Zhong Q, Jiang TJ. Chinese entity relationship extraction based on syntactic and semantic features. Journal of Computer Research and Development, 2016,53(2):284−302 (in Chinese with English abstract).

[4] Liu Q, Li Y, Duan H, Liu Y, Qin ZG. A survey of knowledge mapping construction techniques. Journal of Computer Research and Development, 2016,53(3):582−600 (in Chinese with English abstract).

[5] Ye H, Chao W, Luo Z, Li Z. Jointly extracting relations with class ties via effective deep ranking. arXiv:preprint arXiv:1612.07602, 2016.

[6] Guo XY, He TT, Hu XH, Chen QJ. Chinese entity relationship extraction based on syntactic and semantic features. Journal of Chinese Information Processing, 2014,28(6):183−189 (in Chinese with English abstract).

[7] Kumar S. A survey of deep learning methods for relation extraction. arXiv:arXiv preprint arXiv:1705.03645, 2017.

[8] Surdeanu M, Tibshirani J, Nallapati R, Manning CD. Multi-instance multi-label learning for relation extraction. In: Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012. 455−465.

[9] Zheng S, Hao Y, Lu D, Bao H, Xu J, Hao H, Xu B. Joint entity and relation extraction based on a hybrid neural network. Neurocomputing, 2017,257:1−8.

[10] Lin Y, Shen S, Liu Z, Luan H, Sun M. Neural relation extraction with selective attention over instances. In: Proc. of the Meeting of the Association for Computational Linguistics. 2016. 2124−2133.

[11] Xu Y, Mou LL, Li G, Chen YC, Peng H, Jin Z. Classifying relation via long short term memory networks along shortest dependency paths. Conf. on Empirical Methods in Natural Language Processing, 2015,42(1):56−61.

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

[13] Katiyar A, Cardie C. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In: Proc. of the Meeting of the Association for Computational Linguistics. 2017. 917−928.

[14] Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proc. of the Joint Conf. of the Meeting of the ACL and the Int'l Joint Conf. on Natural Language Processing of the Afnlp. 2009. 1003−1011.

[15] Yu XK, Chen L, Guo J, Cai YY, Wu Y, Wang JC. Relationship extraction method combining clause-level remote supervision and semi-supervised integration learning. Pattern Recognition and Artificial Intelligence, 2017,30(1):54−63 (in Chinese with English abstract).

[16] Yao L, Riedel S, Mccallum A. Unsupervised relation discovery with sense disambiguation. In: Proc. of the Annual Meeting of the Association for Computational Linguistics. 2012.

[17] Qin L, Zhang Z, Zhao H, Hu Z, Xing EP. Adversarial connective-exploiting networks for implicit discourse relation classification. 2017. 1006−1017. https://arxiv.org/abs/1704.00217

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

[19] Hashimoto K, Miwa M, Tsuruoka Y, Chikayama T. Simple customization of recursive neural networks for semantic relation classification. In: Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing. 2013. 18−21.

[20] Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation classification via convolutional deep neural network. In: Proc. of the 25th Int'l Conf. on Computational Linguistics: Technical Papers (COLING 2014). 2014. 2335−2344.

[21] Santos CND, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks. Computer Science, 2015,86: 132−137.

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

[23] Miyato T, Dai AM, Goodfellow I. Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605. 07725, 2016.

[24] Wu Y, Bamman D, Russell S. Adversarial training for relation extraction. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 2017. 1778−1783.

[25] He Z, Chen W, Li Z, Zhang M, Zhang W, Zhang M. SEE: Syntax-aware entity embedding for neural relation extraction. arXiv preprint arXiv:1801.03603, 2018.

[26] Chikka VR, Karlapalem K. A hybrid deep learning approach for medical relation extraction. arXiv preprint arXiv:1806.11189, 2018.

[27] Wen J, Sun X, Ren X, Su Q. Structure regularized neural network for entity relation classification for chinese literature text. arXiv preprint arXiv:1803.05662, 2018.

[28] Adilova L, Giesselbach S, Rüping S. Making efficient use of a domain expert's time in relation extraction. arXiv preprint arXiv: 180704687, 2018.

[29] Zhou GD, Su J, Zhang J, Zhang M. Exploring various knowledge in relation extraction. In: Proc. of the Conf. on Meeting of the Association for Computational Linguistics (ACL 2005). University of Michigan, 2002. 419−444.

[30] Huang X, You HL, Yu Y. A Survey of Research on Relationship Extraction Technology. New Technology of Library and Information Service, 2013,29(11):30−39 (in Chinese with English abstract).

[31] Liu JW, Liu Y, Luo XL. Semi-supervised Learning Method. Chinese Journal of Computers, 2015,38(8):1592−1617 (in Chinese with English abstract).

[32] Brin S. Extracting patterns and relations from the World Wide Web. In: Proc. of the Int'l Workshop on the World Wide Web and Databases. 1998. 172−183.

[33] Kumlien MCJ. Constructing viological knowledge bases by extraction information from text sources. In: Proc. of the 7th Int'l Conf. on Intelligent Systems for Molecular Biology. AAAI Press, 1999. 77−86.

[34] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora. In: Proc. of the Meeting on Association for Computational Linguistics. 2004. 415.

[35] Jiao LC, Yang SY, Liu F, Wang SG, Feng ZX. Neural Network Seventy Years: Retrospect and Prospect. Chinese Journal of Computers, 2016,39(8):1697−1716 (in Chinese with English abstract).

[36] Zhou FY, Jin LP, Dong J. A survey of Convolutional Neural Networks. Chinese Journal of Computers, 2017,40(6):1229−1251 (in Chinese with English abstract).

[37] Zeng D, Liu K, Chen Y, Zhao J. Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 2015. 1753−1762.

[38] Ji GL, Liu K, He SZ, Zhao J. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proc. of the AAAI. 2017. 3060−3066.

[39] Ren X, Wu Z, He W, Qu M, Voss CR, Ji H, Abdelzaher TF, Han JW. CoType: Joint extraction of typed entities and relations with knowledge bases. 2016. 1015−1024. https://arxiv.org/abs/1610.08763

[40] Huang YY, Wang WY. Deep residual learning for weakly-supervised relation extraction. arXiv preprint arXiv:1707.08866, 2017.

[41] Golshan PN, Dashti HAR, Azizi S, Safari L. A study of recent contributions on information extraction. arXiv preprint arXiv:1803.05667, 2018.

[42] Wang LY. Entity relationship extraction based on deep convolutional neural network [MS. Thesis]. Taiyuan: Taiyuan University of Technology, 2017 (in Chinese).

[43] Yang JF, Yu QB, Guan Y, Jiang ZP. A survey of research on electronic medical record named entity recognition and entity relationship extraction. Acta Automatica Sinica, 2014,40(8):1537−1562 (in Chinese with English abstract).

[44] Qin B, Liu AA, Liu T. Unguided Chinese open entity relationship extraction. Journal of Computer Research and Development, 2015,52(5):1029−1035 (in Chinese with English abstract).

[45] Chinchor N, Marsch E. MUC-7 information extraction task definition. In: Proc. of the 7th Message Understanding Conf. Appendices, 1998. 359−367.

[46] Socher R, Huval B, Manning CD, Ng AY. Semantic compositionality through recursive matrix-vector spaces. In: Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012. 1201−1211.

[47] Xu K, Feng Y, Huang S, Zhao D. Semantic relation classification via convolutional neural networks with simple negative sampling. Computer Science, 2015,71:941−949.

[48] Vu NT, Adel H, Gupta P, Schütze H. Combining recurrent and convolutional neural networks for relation classification. arXiv preprint arXiv:1605.07333, 2016.

[49] Wang L, Cao Z, Melo GD, Liu Z. Relation classification via multi-level attention CNNs. In: Proc. of the Meeting of the Association for Computational Linguistics. 2016. 1298−1307.

[50] Li F, Zhang M, Fu G, Qian T, Ji D. A Bi-LSTM-RNN model for relation classification using low-cost sequence features. arXiv preprint arXiv:1608.07720, 2016.

[51] Cai R, Zhang X, Wang H. Bidirectional recurrent convolutional neural network for relation classification. In: Proc. of the Meeting of the Association for Computational Linguistics. 2016. 756−765.

[52] Zheng S, Xu J, Bao H, Qi Z, Zhang J, Hao H, Xu B. Joint learning of entity semantics and relation pattern for relation extraction. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Cham: Springer-Verlag, 2016. 443−458.

[53] Li F, Zhang M, Fu G, Ji D. A neural joint model for extracting bacteria and their locations. In: Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Cham: Springer-Verlag, 2017. 15−26.

[54] Katiyar A, Cardie C. Investigating LSTMs for joint extraction of opinion entities and relations. In: Proc. of the Meeting of the Association for Computational Linguistics. 2016. 919−929.

[55] Zheng S, Wang F, Bao H, Hao Y, Zhou P, Xu B. Joint extraction of entities and relations based on a novel tagging scheme. 2017. 1227−1236. https://arxiv.org/abs/1706.05075

[56] Fei L, Zhang M, Fu G, Ji D. A neural joint model for entity and relation extraction from biomedical text. Bmc Bioinformatics. 2017, 18:198.

[57] He D, Zhang H, Hao W, Zhang R, Chen G, Jin D, Cheng K. Distant supervised relation extraction via long short term memory networks with sentence embedding. Intelligent Data Analysis, 2017,21:1213−1231.

[58] Riedel S, Yao L, Mccallum A. Modeling relations and their mentions without labeled text. In: Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases. 2010. 148−163.

[59] Liu Q, Zhai JW, Zhang ZC, Zhong S, Zhou Q, Zhang P, Xu J. An overview of deep reinforcement learning. Chinese Journal of Computers, 2018,41(1):1−27 (in Chinese with English abstract).

[60] Feng Y, Zhang H, Hao W, Chen G. Joint extraction of entities and relations using reinforcement learning and deep learning. In: Proc. of the Comput Intell Neurosci. 2017. 1−11.

[61] Qin P, Xu W, Wang WY. Robust distant supervision relation extraction via deep reinforcement learning. arXiv preprint arXiv:1805.09927, 2018.

[62] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Advances in Neural Information Processing Systems, 2014,3:2672−2680.

[63] Miyato T, Dai AM, Goodfellow I. Adversarial training methods for semi-supervised text classification. 2016. https://arxiv.org/abs/1605.07725

[64] Qin P, Xu W, Wang WY. DSGAN: Generative adversarial training for distant supervision relation extraction. arXiv preprint arXiv: 1805.09929, 2018.

[65] Nickel M, Murphy K, Tresp V, Gabrilovich E. A review of relational machine learning for knowledge graphs. Proc. of the IEEE, 2015,104:11−33.

[66] Grainger T, Aljadda K, Korayem M, Smith A. The semantic knowledge graph: A compact, auto-generated model for real-time traversal and ranking of any relationship within a domain. 2016. 420−429. https://arxiv.org/abs/1609.00464

[67] Sahu SK, Anand A, Oruganty K, Gattu M. Relation extraction from clinical texts using domain invariant convolutional neural network. In: Proc. of the 15th Workshop on Biomedical Natural Language Processing, 2016. 206−215.

[68] Gu J, Sun F, Qian L, Zhou G. Chemical-induced disease relation extraction via convolutional neural network. In: Proc. of the Database 2017. [doi: 10.1093/database/bax024]

[69] Ramamoorthy S, Murugan S. An attentive sequence model for adverse drug event extraction from biomedical text. arXiv preprint arXiv:1801.00625, 2018.

[70] Peng Y, Rios A, Kavuluru R, Lu Z. Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models. arXiv preprint arXiv:1802.01255, 2018.

[71] Nguyen DQ, Verspoor K. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. arXiv preprint arXiv:1805.10586, 2018.

[72] Feng X, Guo J, Qin B, Liu T, Liu Y. Effective deep memory networks for distant supervised relation extraction. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. 2017. 4002−4008.

[73] Liu T, Wang K, Chang B, Sui Z. A soft-label method for noise-tolerant distantly supervised relation extraction. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 2017. 1790−1795.

[74] Han X, Liu ZY, Sun M. Neural knowledge acquisition via mutual attention between knowledge graph and text. In: Proc. of the AAAI. 2018.

[75] Feng J, Huang M, Zhao L, Yang Y, Zhu XY. Reinforcement learning for relation classification from noisy data. In: Proc. of the AAAI. 2018.

[76] Jiang X, Wang Q, Li P, Wang B. Relation extraction with multi-instance multi-label convolutional neural networks. In: Proc. of the COLING. 2016. 1471−1480.

[77] Chang YC, Zhang YX, Wang H, Wan HY, Xiao CJ. Features oriented survey of state-of-the-art keyphrase extraction algorithms. Ruan Jian Xue Bao/Journal of Software, 2018,29(7):2046−2070 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5538.htm [doi: 10.13328/j.cnki.jos.005538]

[78] Peng N, Poon H, Quirk C, Toutanova K, Yih WT. Cross-sentence *N*-ary relation extraction with graph LSTMs. arXiv preprint arXiv:1708.03743, 2017.

[79] Li MY, Yang J. Open chinese entity relationship extraction method based on dependency parsing. Computer Engineering, 2016,42(6):201−207 (in Chinese with English abstract).

附中文参考文献:

[2] 徐健,张智雄,吴振新.实体关系抽取的技术方法综述.现代图书情报技术,2008,24(8):18−23.

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

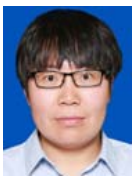
[4] 刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述.计算机研究与发展,2016,53(3):582−600.

[6] 郭喜跃,何婷婷,胡小华,陈前军.基于句法语义特征的中文实体关系抽取.中文信息学报,2014,28(6):183−189.

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

[30] 黄勋,游宏梁,于洋.关系抽取技术研究综述.现代图书情报技术,2013,29(11):30−39.

[31] 刘建伟,刘媛,罗雄麟.半监督学习方法.计算机学报,2015,38(8):1592−1617.
[35] 焦李成,杨淑媛,刘芳,王士刚,冯志玺.神经网络七十年:回顾与展望.计算机学报,2016,39(8):1697−1716.
[36] 周飞燕,金林鹏,董军.卷积神经网络研究综述.计算机学报,2017,40(6):1229−1251.
[42] 王林玉.基于深度卷积神经网络的实体关系抽取[硕士学位论文].太原:太原理工大学,2017.
[43] 杨锦锋,于秋滨,关毅,蒋志鹏.电子病历命名实体识别和实体关系抽取研究综述.自动化学报,2014,40(8):1537−1562.
[44] 秦兵,刘安安,刘挺.无指导的中文开放式实体关系抽取.计算机研究与发展,2015,52(5):1029−1035.
[59] 刘全,翟建伟,章宗长,钟珊,周倩,章鹏,徐进.深度强化学习综述.计算机学报,2018,41(1):1−27.
[77] 常耀成,张宇翔,王红,万怀宇,肖春景.特征驱动的关键词提取算法综述.软件学报, 2018,29(7):2046−2070. http://www.jos.org.cn/1000-9825/5538.htm [doi: 10.13328/j.cnki.jos.005538]
[79] 李明耀,杨静.基于依存分析的开放式中文实体关系抽取方法.计算机工程,2016,42(6):201−207.



鄂海红(1982－),女,辽宁锦州人,博士,副教授,CCF 专业会员,主要研究领域为深度学习知识图谱,自然语言处理,大数据及人工智能技术在交叉领域应用研究.

张文静(1995－),女,硕士生,CCF 学生会员,主要研究领域为自然语言处理,远程监督关系抽取,多轮对话.

肖思琪(1994－),女,硕士生,主要研究领域为自然语言处理,关系抽取.

程瑞(1995－),女,硕士生,CCF 学生会员,主要研究领域为知识图谱,知识抽取,知识融合.

胡莺夕(1993－),女,硕士生,主要研究领域为自然语言处理,深度学习,信息抽取,文本摘要.

周筱松(1995－),女,硕士生,CCF 学生会员,主要研究领域为自然语言处理,知识图谱.

牛佩晴(1994－),女,硕士生,主要研究领域为自然语言处理.