软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):1025−1038 [doi: 10.13328/j.cnki.jos.005929]

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

基于带噪观测的远监督神经网络关系抽取*

叶育鑫[1,2], 薛 环[1], 王 璐[3], 欧阳丹彤[1,2]

[1](吉林大学 计算机科学与技术学院,吉林 长春 130012)

[2](符号计算与知识工程教育部重点实验室(吉林大学),吉林 长春 130012)

[3](北京大学 北京国际数学研究中心,北京 100871)

通讯作者: 欧阳丹彤, E-mail: ouyd@jlu.edu.cn

摘 要: 远监督关系抽取的最大优势是通过知识库和自然语言文本的自动对齐生成标记数据.这种简单的自动对齐机制在将人从繁重的样本标注工作中解放出来的同时,不可避免地会产生各种错误数据标记,进而影响构建高质量的关系抽取模型.针对远监督关系抽取任务中的标记噪声问题,提出"最终句子对齐的标签是基于某些未知因素所生成的带噪观测结果"这一假设.并在此假设的基础上,构建由编码层、基于噪声分布的注意力层、真实标签输出层和带噪观测层的新型关系抽取模型.模型利用自动标记的数据学习真实标签到噪声标签的转移概率,并在测试阶段,通过真实标签输出层得到最终的关系分类.随后,研究带噪观测模型与深度神经网络的结合,重点讨论基于深度神经网络编码的噪声分布注意力机制以及深度神经网络框架下不均衡样本的降噪处理.通过以上研究,进一步提升基于带噪观测远监督关系抽取模型的抽取精度和鲁棒性.最后,在公测数据集和同等参数设置下进行带噪观测远监督关系抽取模型的验证实验,通过分析样本噪声的分布情况,对在各种样本噪声分布下的带噪观测模型进行性能评价,并与现有的主流基线方法进行比较.结果显示,所提出的带噪观测模型具有更高的准确率和召回率.

关键词: 远监督;关系抽取;噪声标签

中图法分类号: TP181

中文引用格式: 叶育鑫,薛环,王璐,欧阳丹彤.基于带噪观测的远监督神经网络关系抽取.软件学报,2020,31(4):1025−1038. http://www.jos.org.cn/1000-9825/5929.htm

英文引用格式: Ye YX, Xue H, Wang L, Ouyang DT. Distant supervision neural network relation extraction base on noisy observation. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):1025−1038 (in Chinese). http://www.jos.org.cn/1000-9825/5929.htm

Distant Supervision Neural Network Relation Extraction Base on Noisy Observation

YE Yu-Xin[1,2], XUE Huan[1], WANG Lu[3], OUYANG Dan-Tong[1,2]

[1](School of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (Jilin University), Changchun 130012, China)

[3](Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China)

Abstract: The great advantage of distant supervision relation extraction is to generate labeled data automatically through knowledge bases and natural language texts. This simple automatic alignment mechanism liberates people from heavy labeling work, but inevitably produces various incorrect labeled data meanwhile, which would have an influential effect on the construction of high-quality relation extraction models. To handle noise labels in the distant supervision relation extraction, here it is assumed that the final label of sentence is

* 基金项目: 国家自然科学基金(61672261, 61872159)
Foundation item: National Natural Science Foundation of China (61672261, 61872159)
本文由"非经典条件下的机器学习方法"专题特约编辑高新波教授、黎铭教授、李天瑞教授推荐.
收稿时间: 2019-05-31; 修改时间: 2019-07-29; 采用时间: 2019-09-20; jos 在线出版时间: 2020-01-10
CNKI 网络优先出版: 2020-01-14 13:22:35, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.1322.027.html

based on noisy observations generated by some unknown factors. Based on this assumption, a new relation extraction model is constructed, which consists of encoder layer, attention based on noise distribution layer, real label output layer, and noisy observation layer. In the training phase, transformation probabilities are learned from real label to noisy label by using automatically labeled data, and in the testing phase, the real label is obtained through the real label output layer. This study proposes to combine the noise observation model with deep neural network. The attention mechanism of noise distribution is focused based on deep neural network, and unbalanced samples are denoised of under the framework of deep neural network, aiming to further improve the performance of distant supervision relation extraction based on noisy observation. To examine its performance, the proposed method is applied to a public dataset. The performance of distant supervision relation extraction model is evaluated under different distribution families. The experimental results illustrate the proposed method is more effective with higher precision and recall, compared to the existing methods.

Key words: distant supervision; relation extraction; noise label

关系抽取是从纯文本生成关系元数据的过程,是自然语言处理中的一项重要任务.全监督关系抽取系统需要大量标注好关系的训练数据,然而,标注数据需人工参与,耗时耗力.Mintz 等人[1]在 2009 年提出远监督的思想,通过对齐知识库和文本自动生成标注数据.他们假设“如果两个实体在知识库中具有某个关系,那么包含这两个实体的所有句子都将表达这个关系”.例如,(苹果,创始人,乔布斯)是知识库中的一个三元组关系实例,那么包含这两个实体的所有句子都会被标注为创始人关系.作为远监督关系抽取任务中的开创性研究,存在数据与关系标签自动对齐的假设过强问题.例如,“乔布斯正在买苹果.”中的乔布斯和苹果并未表达创始人关系,但该句仍被视为一个表达创始人关系的实例.在 Mintz 等人[1]的工作中,生成了诸如此类的大量错误标记,给模型训练带来极大的噪声.Riedel 等人[2]在 2010 年将假设弱化,他们提出“在包含某一关系实体对的句子集合中,存在一个最有可能表达该关系的句子”.选择最有可能表达该关系的句子进行模型训练,保证了最大程度上的噪声筛除,但同时也丢失了大量可用的标记数据,影响了模型训练效果.因此,Hoffmann 等人[3]假设“不止一个句子能够表达该关系”,并采用多实例学习方法定位出这样的标记数据,加入模型训练.Surdeanu 等人[4]进一步假设“包含同一实体对的句子可能表达多种关系”,采用多实例多标签学习方法获取关系间的依赖,进一步扩展了可用的标记数据.Lin 等人[5]的工作则不再区分被自动标记数据的可用性,认为“所有句子和所有关系之间,只存在关联程度的不同”,并采用注意力机制训练不同句子对不同关系的支持度,即权重.

从自动对齐的标记数据可用性角度来看,上述工作都是在识别样本噪声的前提下,尽量保留更多的样本以及丰富样本的特征(Mintz[1]在本领域的开创性基础工作除外,还未涉及到噪声辨识),从而获得性能更高的关系抽取模型.他们的工作在“认为最终句子对齐的标签就是真实关系的反映”的前提下是正确的,然而,真实世界中存在诸多未知因素和未知的噪声扰动,会影响关系抽取模型的真实性.在远监督关系抽取中,简单的对齐和错误标签排查不能保证“最终对齐的标签就是真实关系的反映”,训练所得到的关系抽取模型不具有更好的鲁棒性和泛化能力.因此,本文基于“最终句子对齐的标签是基于某些未知噪声扰动所生成的观测结果”这一假设,提出基于带噪观测的远监督关系抽取模型.通过构造编码层、基于噪声分布的注意力层、真实标签层和带噪观测标签层实现新的远监督关系抽取范式.在现有的远监督关系抽取工作中,比较接近我们的想法的是 Fan 等人[6]提出的矩阵填充方法.他们回避了最终标签关系真实性的问题,直接建立对齐句子与关系标签的对应矩阵,利用矩阵的最小秩分解,为待抽取的句子填充预测关系.该方法在抽取精度上较以往方法有质的飞跃,但无法生成高效的关系抽取器.最小秩分解过程受矩阵的稀疏性和噪声的影响,抽取效率远不如其他现有抽取方法.尽管采用截断核范数等方法可以在一定程度上缓解最小秩分解的计算压力,但无法从根本上解决问题.

远监督关系抽取研究进展中的另一亮点是深度神经网络的引入.鄂海红等人[7]综述了神经网络方法在关系抽取上的发展过程.早期关系抽取中,自然语句的特征提取依靠自然语言处理(natural language processing,简称 NLP)工具加人工经验选择,如徐红艳等人[8]在推荐系统上的应用.一方面,NLP 工具在生成特征的过程中产生的误差也将在后续的关系抽取任务中传播,例如 PosTag(词性标注)等;另一方面,人工经验选择特征不但难以应对大数据样本的指数级增长,而且很难发现不易描述的潜在特征.随着近些年深度学习技术的不断发展,Socher 等人[9]、Zeng 等人[10]、Santos 等人[11]摒弃了使用 NLP 工具提取特征加人工经验特征选择的做法,在关系分类中使用深度神经网络进行自然语言特征的自动提取和选择.他们的工作在基于监督学习的关系抽取任务中,取

得了比以往基于规则、基于支持向量机和基于概率模型等方法更好的效果.Zeng 等人[12]在 2015 年将深度神经网络引入远监督关系抽取中.他们采用多实例学习与神经网络结构相结合的方式构建基于对齐标记数据的关系抽取器,取得了比以往远监督关系抽取模型更好的精度.Qu 等人[13]将 Attention 机制与深度神经网络相结合,给出词级特征权重自动设置的方法,提升远监督关系抽取的特征提取能力.Ji 等人[14]借助 Wikipedia 的额外实体描述信息训练远监督关系抽取模型.它通过深度神经网络模型获得实体描述信息的特征,并加入到实体对所在的句子特征中.欧阳等人[15]利用本体进行关系实例的自动扩充,用于扩充远监督关系抽取任务中待抽取关系的样本数量.Vashishth 等人[16]为对齐的标记数据构造图结构,然后利用图卷积神经网络训练关系分类模型.通过图结构中额外的边信息、实体类型信息来丰富特征,提高模型的关系抽取性能.本文则从基于带噪观测的机器学习角度讨论与深度神经网络的结合,重点研究基于深度神经网络编码的噪声分布注意力机制以及深度神经网络框架下不均衡样本的降噪处理.通过以上研究,进一步提升基于带噪观测远监督关系抽取模型的抽取精度和鲁棒性.

1 相关工作

噪声标签问题是经典机器学习中广泛研究的问题.早期工作(Brodley[17]等人、Rebbapragada[18]等人、Manwani[19]等人)尝试利用 SVM(Natarajan[20]等人)和 Fisher 判别器(Lawrence[21]人)等各种浅层分类器来识别噪声标签.近期,深度神经网络在人工标注的大规模数据集上的噪声处理性能比传统机器学习方法有大幅提升,越来越多的工作关注于利用深度神经网络辨识、缓解或消除噪声对模型学习的影响.这些方法大体上可分为两类:(1) 使用人工标注的数据来帮助缓解噪声标签对模型性能的影响;(2) 直接从含有噪声标签的数据中学习模型.

使用人工标注的数据:这一系列研究都是利用一个小而标注准确的数据集来纠正噪声标签.例如,Li 等人[22]在 2017 年提出有导师学习的神经网络,在损失函数中用软标签重新调整噪声标签的权重.Veit 等人[23]在 2017 年使用标注正确的数据作为标签校正网络.Vahdat 等人[24]在 2017 年使用正确标注的数据作为辅助信息来对潜在的正确标签进行推理.Yao 等人[25]在 2018 年建立了图像噪声标签的可信度以减轻噪声标签对模型的影响.尽管这些方法显示出非常有前景的结果,但它的使用受限于是否存在正确标注数据集这一前提.

直接从含有噪声标签的数据中学习:这类研究通过设计鲁棒的损失函数或通过对潜在的正确标签进行建模,直接从噪声标签中学习.例如,Reed 等人[26]在 2014 年将有放回抽样(booststrapping)方法用于损失函数,以便对类似的图像进行一致的标签预测.Joulin 等人[27]在 2016 年根据样本数对损失函数进行适当加权来缓解噪声标签问题.Jiang 等人[28]在 2017 年提出了一个顺序元学习模型,该模型采用损失值序列并输出标签的权重.Ghosh 等人[29]在 2017 年进一步研究了损失函数的条件,使损失函数可以包容一定的噪声.Mnih 等人[30]在 2012 年提出了一种用于对称标签噪声的噪声适应框架.基于这项工作,Sukhbaatar 等人[31]、Jindal 等人[32]、Patrini 等人[33]、Han 等人[34]通过学习深度神经网络顶部的带噪观测层来解释噪声标签,其中,学习到的转换矩阵表示标签的转移概率.类似地,Xiao 等人[35]提出了图像条件概率噪声模型.Misra 等人[36]构建两个平行分类器,其中一个分类器处理图像识别,而另一个分类器模拟人来报告误差.Bekker 等人[37]提出了基于 EM 迭代算法的神经网络来估计隐藏的真实标签和噪声分布.

由于本文研究的远监督关系抽取任务中无预设精准的人工标注数据集,而通过启发式对齐知识库和自然语句生成的标记数据不可避免地存在大量噪声标签.之前的工作包括:Riedel 等人[2]在包含某对实体的句子中计算选取最可能表达该实体关的句子,并忽略其他的数据(包括噪声数据);Zeng 等人[12]采用类似的方法来进行多实例学习;Lin 等人[5]采用注意力机制,为一个包中的句字和候选关系分配权重,相当于忽略了权重较低的句子;Ji 等人[14]在其基础上增加了实体描述信息;Qu 等人[13]同时在词级别和包级别上采用注意力机制为词和句子分配权重.上述工作都是通过计算噪声数据的概率去忽略这些噪声,而并没有考虑到噪声数据在整体数据中的分布情况,从而丢失了噪声数据的分布信息,因此,本文采用了“直接从含噪声标签数据中学习”的思路去构建关系抽取模型.现有的“直接从含噪声标签数据中学习”的工作主要针对处理图像、处理监督学习任务.本文与现有

工作不同,首次将该方法引入到基于远监督学习的自然语言关系抽取任务中,并且进一步讨论了自然语句的包级特征在噪声分布下的权重学习问题,以及在远监督学习任务中样本不均衡的降噪处理问题.

2 带噪观测的远监督关系抽取模型

本文提出带噪观测的远监督神经网络关系抽取模型来解决噪声标签问题,该模型由编码层、基于噪声分布的注意力层、真实标签输出层和带噪观测层构成.整体结构如图 1 所示.

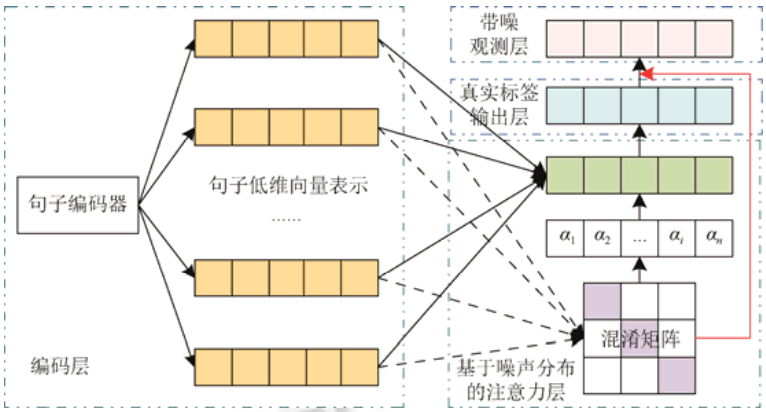

Fig.1 Structure of the proposed model

图 1 模型整体结构

在远监督关系抽取任务中,首先确定同一实体对的一组自然语句,将该组自然语句进行词向量矩阵化,并通过句子编码器得到对应的低维向量表示;然后暂时忽略远监督过程中产生的噪声标签问题,将神经网络生成的标签视为真实标签,在没有带噪观测层的情况下训练神经网络,得到训练集上的混淆矩阵,并把混淆矩阵作为噪声分布的初始化;注意力层根据噪声分布得到每个句子与噪声标签的相关程度,提取自然语句的包级特征;再根据包级特征通过真实标签输出层估计真实标签的概率;最后带噪观测层根据真实标签和混淆矩阵初始化的噪声分布,估计噪声标签的概率,并不断更新混淆矩阵以得到最终的噪声分布.其中,词向量采用 word2vec(https://code.google.com/p/word2vec/)进行特征表示,句子编码器采用卷积神经网络(CNN)的变体分段最大池化卷积神经网络(PCNN),基于噪声分布的注意力层采用混淆矩阵作为噪声分布的表示,真实标签输出层和带噪观测层均使用 softmax 来计算.

3 带噪观测与噪声分布

在基于带噪观测的远监督神经网络关系抽取模型中,带噪观测用于发掘训练样本中的噪声分布信息,并利用噪声分布信息引导模型评估句子与噪声标签间的隐含关系.其中,带噪观测层以混淆矩阵作为噪声分布的初始化,并以真实标签和噪声分布计算噪声标签的概率.通过训练不断更新混淆矩阵,将最终结果作为估计的噪声分布.带噪观测和真实标签输出将在第 3.1 节中给出介绍,噪声分布的初始化将在第 3.2 节中给出介绍.

3.1 带噪观测与真实标签输出

本文假设在远监督关系抽取模型的训练阶段不能直接获得数据的真实标签 y,而只能得到数据的噪声标签 y',并且噪声标签 y' 是由真实标签 y 和未知的噪声分布 w_{noise} 决定的.因此定义噪声标签概率的计算公式如下:

$$p(y'=j\mid x,w,w_{noise})=\sum_{i=1}^{k}p(y'=j\mid y=i;w_{noise})\,p(y=i\mid x,w) \tag{1}$$

在图 1 所示的模型结构中,真实标签输出层是对真实标签的估计,用 $h=h(x)$ 表示以 X 为输入的句子编码器的输出,并通过以下方式计算真实标签 y 的概率:

$$p(y=i\mid x;w)=\frac{\exp(u_i^T h+b_i)}{\sum_{l=1}^{k}\exp(u_l^T h+b_i)} \tag{2}$$

其中,w 是神经网络的参数(包括真实标签输出层参数),$u_1,u_2,\ldots,u_k$ 是真实标签输出层参数,$b_1,b_2,\ldots,b_k$ 是偏置项,接下来根据真实标签和输入特征计算噪声标签 y':

$$p(y'=j\mid y=i,x)=\frac{\exp(u_{ij}^T h+b_{ij})}{\sum_{l=1}^{k}\exp(u_{il}^T h+b_{il})} \tag{3}$$

$$p(y'=j\mid x)=\sum_{l=1}^{k}p(y'=j\mid y=i,x)p(y=i\mid x) \tag{4}$$

其中,$u_{ij}(i,j=1,2,\ldots,k)$是带噪观测层参数.训练阶段模型结构如图 2 所示.

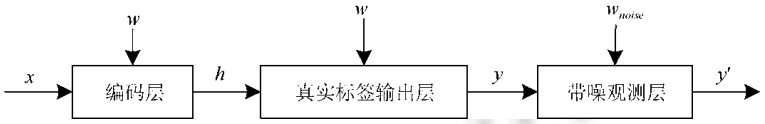

Fig.2 Structure of training phase model

图 2 训练阶段模型结构

图 2 中用 w_{noise} 来代表带噪观测层的所有参数.在训练阶段,给定 n 个特征向量 $x_1,x_2,\ldots,x_n$,并且具有相应的噪声标签 $y'_1,y'_2,\ldots,y'_n$,还有被视为隐藏变量的真实标签 $y_1,y_2,\ldots,y_n$.模型的对数似然函数是

$$S(w,w_{noise})=\sum_t \log p(y'_t\mid x_t)=\sum_t \log\Big(\sum_i p(y'_t\mid y_t=i,x_t;w_{noise})p(y_t=i\mid x_t;w)\Big) \tag{5}$$

其中,w_{noise} 是带噪观测层参数,w 是神经网络参数(不包括带噪观测层参数).这里假设噪声标签只依赖于真实标签,而不依赖输入的句子,因此噪声标签 y'的概率计算就变为

$$p(y'=j\mid y=i)=\frac{\exp(b_{ij})}{\sum_l \exp(b_{il})} \tag{6}$$

$$p(y'=j\mid x)=\sum_i p(y'=j\mid y=i)p(y=i\mid x) \tag{7}$$

模型参数的似然函数就变为

$$S(w,w_{noise})=\sum_t \log p(y'_t\mid x_t)=\sum_t \log\Big(\sum_i p(y'_t\mid y_t=i;w_{noise})p(y_t=i\mid x_t;w)\Big) \tag{8}$$

由于噪声标签是通过向神经网络添加额外的带噪观测层来建模的,因此可以通过设置

$$p(y'=j\mid y=i)=\theta(i,j)=\frac{\exp(b_{ij})}{\sum_l \exp(b_{il})} \tag{9}$$

来使用训练神经网络的标准技术优化 $S(w,w_{noise})$.在训练过程中同时优化带噪观测层参数和分类器,而不是分别优化它们.

欲在测试阶段预测真实的标签.因此,删除了旨在消除训练集中的噪声的带噪观测层,只计算真实标签的概率 $p(y=i|x;w)$.在测试阶段,模型结构如图 3 所示.

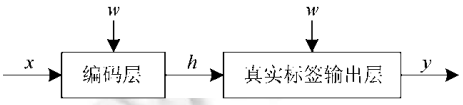

Fig.3 Structure of testing phase model

图 3 测试阶段模型结构

3.2 噪声分布的初始化

在我们的带噪观测模型中,噪声分布信息用来引导模型评估句子的真实标签和噪声标签的概率转移关系.如果随机初始化噪声分布,那么这个噪声分布会包含不准确的噪声信息,模型将不容易估计到与真实情况接近

的噪声分布,而包含一定噪声分布信息的初始化可以引导模型估计到更接近真实情况的噪声分布.根据训练数据的标签和不带噪声观测层模型计算的标签得到的混淆矩阵,比随机初始化的更容易接近带噪观测估计的噪声分布.因此本文中采用混淆矩阵初始化噪声分布,使模型更容易在训练的过程中收敛到接近真实情况的噪声分布.用于初始化噪声分布的混淆矩阵是利用训练数据和不含带噪观测层的关系抽取模型生成的.在生成的过程中,我们将神经网络生成的标签视为真实标签,将训练数据集的标签视为噪声标签.其计算方式如下:

$$b_{ij}=\sum_t 1_{\{y_t'=j\}}p(y_t=i\mid x_t) \tag{10}$$

其中,混淆矩阵 b 的行对应真实标签,列对应噪声标签,在 b 的行上计算真实标签 i 连接 j 的概率,并取其对数.计算公式如下:

$$b_{ij}'=\log\left(\frac{b_{ij}}{\sum_t p(y_t=i\mid x_t)}\right) \tag{11}$$

然后在学习噪声分布和估计真实标签的训练过程中,把混淆矩阵 b'作为带噪观测层的初始化参数.

4 基于噪声分布的注意力机制

在已有的神经网络关系抽取中,注意力机制是用来动态调节句子和候选标签的权重,以缓解噪声标签问题.在本文的带噪观测模型中,注意力机制则关注于自然语句和噪声标签间权重的自动调节,以更好地估计噪声分布和计算真实标签的概率.本文基于噪声分布的注意力机制,采用混淆矩阵作为参数初始化,并在训练过程中不断更新.编码层将在第 4.1 节中加以介绍,基于噪声分布的注意力机制将在第 4.2 节中加以介绍.

4.1 编码层

对于训练数据中的所有句子,首先把包含同一实体对的句子分配到同一组中,再通过词向量方法把句子向量矩阵化.

本文使用的预训练词向量是从大型文本语料库中训练得到的,并且可以表达单词上下文的句法和语义信息.除了词向量方法之外,还使用单词的位置信息.位置信息表明每个单词在句子中相对实体对的距离.位置信息如图 4 所示.每个单词相对实体对的位置信息在开始时被随机化为位置向量,并在训练期间不断更新.最后把词向量和位置向量拼接在一起,一个句子就转换为向量矩阵 $X=\{x_1,x_2,\ldots,x_{|X|}\}$,其中,$x_i\in R^d(d=d^w+d^p\times 2)$$d^w$ 是词向量长度,d^p 是位置向量长度.

Fig.4 Location information of words relative to entities

图 4 单词相对实体的位置信息

与 Zeng 等人[12]、Lin 等人[5]的工作相同,把向量矩阵化的句子用 PCNN 编码到低维向量表示.卷积是两个维度相同矩阵间的运算操作,定义如公式(1)所示,其中,矩阵 $A=(a_{ij})_{m\times n}$,矩阵 $B=(b_{ij})_{m\times n}$.

$$A\otimes B=\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}b_{ij} \tag{12}$$

把卷积作用在矩阵化的句子 $X=\{x_1,x_2,\ldots,x_{|X|}\}$上,其中,$x_i\in R^d$ 是句子中第 i 个单词的向量表示,取 $X_{i:j}=[x_i,x_{i+1},\ldots,x_j]$.给定一组卷积核 $W=\{w_1,w_2,\ldots,w_{|m|}\}$,其中,$w_t\in R^{l\times d}$,$l$ 是卷积核大小.对句子 $X_{i:j}$ 和卷积核 w_t 进行卷积操作得到 $c_t\in R^{|X|-l+1}$.经过 m 个卷积核后得到 $C=\{c_1,c_2,\ldots,c_m\}$.

然后把卷积编码后的句子做分段最大池化操作,分段是把卷积后的结果根据实体对在句子中的位置分割成 3 段,$c_i=\{c_{i1},c_{i2},c_{i3}\}$,最大池化是分别在每段句子上取最大值,$p_{i1}=\max(c_{i1})$,$p_{i2}=\max(c_{i2})$,$p_{i3}=\max(c_{i3})$.对于每个卷积核得到的结果做分段最大池化,得到 $p_i=\{pi_1,p_{i2},p_{i3}\}$,最后把分段最大池化的结果进行拼接,得到句子低维向量编码 $P\in R^{3m}$.由于单池化操作会过快降低编码后句子的维度,无法提取句子的细粒度特征,因此,本文采用分段最大池化操作.整个句子编码过程如图 5 所示.

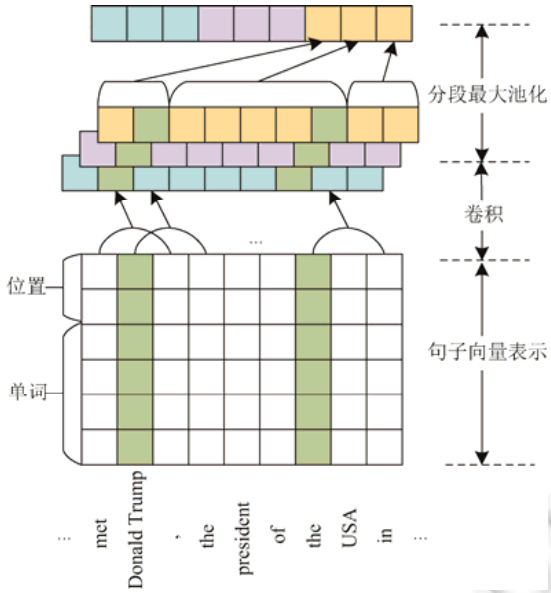

Fig.5 Encoder layer

图 5 编码层

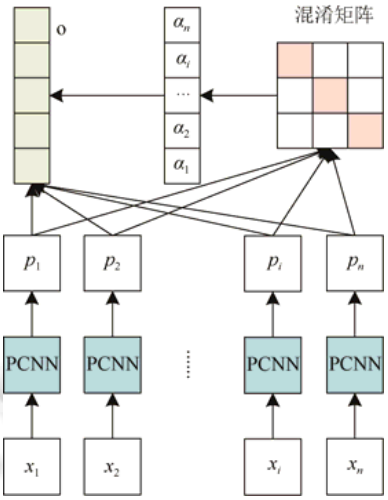

Fig.6 Attention with noise distribution

图 6 基于噪声分布的注意力机制

4.2 基于噪声分布的注意力机制

在远监督关系抽取任务中,已有的工作都是通过注意力机制来动态调节句子和候选标签间的权重来判断句子和候选标签的相关程度,达到缓解噪声标签的效果.不同于以往工作,本文从含有噪声标签的数据去估计噪声分布和真实标签,并利用注意力机制和噪声分布信息来估计当前句子的真实标签的置信度.

给定一组句子$\{x_1,x_2,\ldots,x_n\}$,然后通过 PCNN 把句子编码到低维向量表示$\{p_1,p_2,\ldots,p_n\}$,每个句子的权重α_i计算方式如下:

$$\alpha_i=\frac{\exp(e_i)}{\sum_{j=1}^{k}\exp(e_j)} \tag{13}$$

其中,k是关系个数,$e_i=p_i w_{noise} r$,利用噪声分布w_{noise}来计算句子p_i与噪声标签r的置信度,r是噪声标签的向量表示,然后得到这组句子的加权求和表示:

$$s=\sum_{i=1}^{n}\alpha_i p_i \tag{14}$$

模型的输出为

$$o=Ms+d \tag{15}$$

其中,M是关系的矩阵表示,d是偏置项.最后经过真实标签输出层来计算真实标签的概率:

$$p(y\mid s,w)=\frac{\exp(o_i)}{\sum_{j=1}^{k}\exp(o_j)} \tag{16}$$

其中,w是神经网络参数.

5 不均衡样本的降噪处理

远监督关系抽取任务中数据标记的启发式对齐,会导致所生成样本的不均衡.以 Freebase 与《纽约时报》语料库(NYT)对齐而生成的标记数据样本为例,其中,负例(标签为“NA”的样本)的数量远远超过正例,占总样本的 72.5%左右.Surdeanu 等人[4]针对该问题,通过在负例中随机采样 10%来缓解样本分布的不均衡,避免正例在训练过程中选择“NA”作为正确标签.

在我们的带噪观测模型中,同样存在样本负例过多的问题,自然而然地,噪声会更多地分布在负例中.如果

直接使用混淆矩阵 b'初始化带噪观测层参数,则会得不到接近真实情况的噪声分布,那么模型在估计所有句子真实标签的概率时会更趋向于“NA”,导致错误分类.为了减少样本分布不均衡对带噪观测模型性能的影响,本文对用来初始化带噪观测层的混淆矩阵做了局部降噪处理.

在第 3.2 节中,矩阵 b 的行对应真实标签,列对应噪声标签.当真实标签为 i 时,在第 i 行中除第 i 个元素值之外最大的 l 是最有可能的噪声标签.因此,将真实标签输出层中的第 i 个元素连接到其最可能的噪声标签 l 上.如果将第 1 个 softmax 层中的第 i 个标签连接到带噪观测层中的第 i 个标签,那么本文提出的方法将不能获得噪声分布信息.此外,由于训练数据中标签为“NA”数据过多,非“NA”标签数据在训练过程中都有选择“NA”作为标签的倾向,因此矩阵的对角线不一定全都非零,为了消除矩阵对角线为 0 时对带噪观测层估计噪声分布的影响,本文对混淆矩阵作了一定的限制,如果第 i 行对角线上的值为 0,那么第 i 行对角线上的那个数据固定为 1.此外,噪声大多会分布在“NA”所在混淆矩阵的行和列,为了减少噪声分布不均衡对模型性能的影响,本文对“NA”所在混淆矩阵的行和列上做局部降噪处理.对于公式(11)作了如下判断:

$$b_{ij}=\begin{cases}1, & b_{ij}=0 \text{ and } i=j\\ 0, & i=0, j\neq 0 \text{ or } i\neq 0, j=0\\ \sum_t 1_{\{y_t'=j\}} p(y_t=i\mid x_t), & \text{otherwise}\end{cases} \tag{17}$$

6 实 验

本节拟通过实验验证在远监督关系抽取任务中,基于带噪观测假设所构建的深度神经网络关系抽取模型较以往模型在关系抽取性能上的提升.首先在第 6.1 节描述实验设计所使用的数据集和评估指标;然后在第 6.2 节给出模型的参数设置,进一步地,根据计算混淆矩阵的实验结果在第 6.3 节分析了样本噪声的分布情况,对在各种样本噪声分布下的带噪观测模型进行性能评价,并在第 6.4 节中与几种基线方法进行比较.

6.1 数据集和评价指标

对于远监督关系抽取数据集,本文使用了两个数据集.第 1 个数据集是 2010 年由 Riedel 等人[2]通过将 Freebase 中的关系与《纽约时报》语料库(NYT)对齐而生成的.其中,训练集是对齐 Freebase 和 NYT 中 2005 年～2006 年的句子产生的数据,测试集是对齐 Freebase 和 NYT 中 2007 年的句子产生的数据.该数据集中包含 53 类关系,包括特殊关系类型“NA”,表示两个实体之间没有关系.得出的训练和测试数据分别包含 570 088 和 172 448 个句子.Hoffmann 等人[3]、Surdeanu 等人[4]、Zeng 等人[12]、Lin 等人[5]也使用该数据集作了相关研究.第 2 个数据集是 2012 年 Ling 等人[38]通过维基百科文章和 Freebase 生成的数据集 Wiki-KBP,其中,训练集包含 23 111 个句子,测试集包含 15 847 个句子.该数据集共 7 类关系,包括特殊关系类型“NA”.

与 Mintz[1]、Lin[5]的工作类似,使用 held-out 评估本文提出的方法,并在 held-out 评估中比较所有基线的准确率和召回率,对比本文提出的模型的效果.

6.2 参数设置

本文通过调整句子最大长度、学习率、权重衰减率和批量大小等参数来测试模型在验证数据集上的性能.对于其他参数,本文使用与 Lin 等人[5]相同的参数.表 1 为本文实验中使用的主要参数.

Table 1 Parameter settings

表 1 参数设置

卷积核大小	3
卷积核个数	230
词向量维度	50
位置向量维度	5
批大小	160
Dropout	0.5

6.3 噪声分布情况分析

本文提出的方法主要是为了估计实体对隐藏真实标签的概率和噪声分布.为了初始化带噪观测层,本文首先在没有带噪观测层的情况下训练神经网络,忽略错误标签问题,计算训练集上的混淆矩阵来初始化噪声适应层参数,混淆矩阵如图 7 所示.

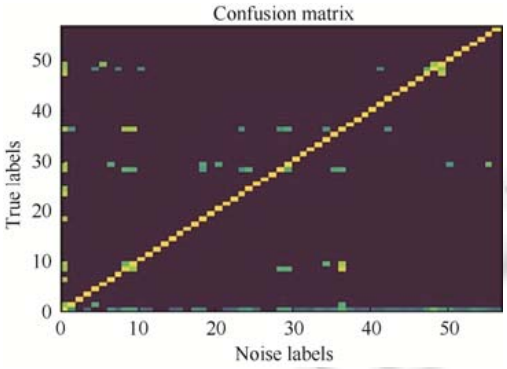

Fig.7 Confusion matrix

图 7 混淆矩阵

从图 7 可以看出,训练数据中绝大多数样本的标签分布在矩阵对角线上,是真实标签,没有分布在矩阵对角线上的属于可能的噪声标签.在训练过程中,标签“NA”的编号为 0,因此矩阵的第 1 行表示真实标签为“NA”,第 1 列表示噪声标签为“NA”.由于训练数据中句子标签为“NA”的数据过多,所以噪声标签大多分布在矩阵的第 1 列和第 1 行.

6.4 对比基线

在 NYT 数据集上,为了评估本文提出的方法,我们选择了几种经典的基线方法,通过 held-out 评估进行比较,Mintz 等人[1]2009 年提出的基于包含某一实体对的所有句子都表达某个关系的假设方法;MIMLRE 是 Riedel 等人[2]2010 年提出的多实例多标签学习方法;cnn_att 是 Lin 等人[5]2016 年提出的采用 CNN 编码基于句子级注意力机制的远监督关系抽取方法;PCNN+ATT 是 Lin 等人[5]2016 年提出的采用 PCNN 编码基于句子级注意力机制的远监督关系抽取方法;RESIDE 是 Vashishth 等人[16]2018 年提出引入额外的边信息和实体类型信息,并采用图卷积神经网络作为句子编码器.在 Wiki-KBP 数据上,本文与 PCNN+ATT 模型进行了比较.

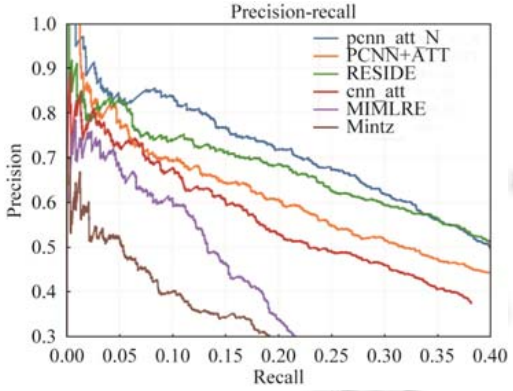

Fig.8 NYT experimental results

图 8 NYT 实验结果

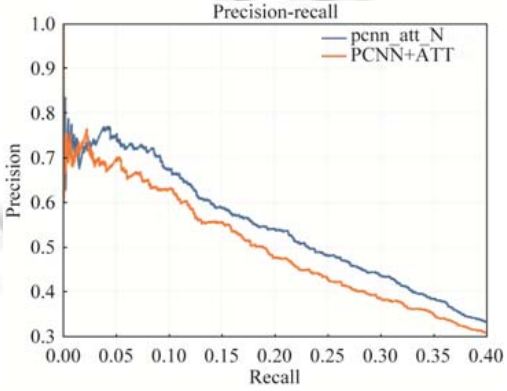

Fig.9 Wiki-KBP experimental results

图 9 Wiki-KBP 实验结果

为了进行公平比较,在 NYT 数据集上,对于上述基线方法,本文的模型和 PCNN+ATT 使用相同的超参数,其他方法直接使用作者发布的结果.对于本文提出的方法和 PCNN+ATT,总共运行 45 轮,在训练期间保存每一轮

的训练模型.最后把在测试集上效果表现最好的模型作为本文模型的最终结果.图 8 显示了所有方法的准确率/召回率曲线,包括本文提出的模型(标记为 pcnn_att_N).对于所有基线方法可以看到,RESIDE 显示出比其他方法更好的性能,证明图卷积神经网络和边信息、实体类型对模型的性能提升有很大帮助.虽然 RESIDE 显示出比其他基线方法有显著提高,但本文方法仍然比 RESIDE 有一定的提高,说明带噪观测层学得了一定的噪声分布信息,并有效缓解了噪声标签问题.PCNN+ATT 相对其他早先的基线性能也有很高的提升,说明句子级注意力机的有效性.从图 9 可以看出,在 Wiki-KBP 数据集上,本文提出的方法效果也要好于 PCNN+ATT.

表 2 比较了本文提出的模型和所有其他基线方法在 NYT 数据集上的 *P@N*.对于 RESIDE,本文采用了作者原始论文中的 *P@N* 值.本文提出的模型达到了 *P*@100、*P*@200、*P*@300 的最高值,平均 *P@N*(%)值比 RESIDE 的平均值高 3.7,比 PCNN+ATT 的平均值高 9.2.

Table 2 Evaluation of performance

表 2 评估结果

P@N (%)	100	200	300	Mean
pcnn_att_N	**84.2**	**84.6**	**80.7**	**83.1**
PCNN+ATT	81.1	71.1	69.4	73.9
cnn_att	76.2	70.6	66.4	71.1
RESIDE	84.0	78.5	75.6	79.4
MIMLRE	70.9	62.9	60.9	64.9
Mintz	51.8	50.0	44.8	48.9

同时,在其他参数与 Lin[5]相同的条件下,本文分析了不同参数对模型性能的影响,表 3 为不同参数下模型在 NYT 数据集上模型性能的对比.

Table 3 Model parameters

表 3 模型参数

P@N (%)	100	200	300
卷积核大小{3,4,5}	{84.2,84.0,**84.3**}	{**84.6**,80.2,79.8}	{**80.7**,78.9,79.7}
卷积核个数{100,230,300}	{82.1,84.2,**84.4**}	{79.6,**84.6**,82.3}	{74.3,80.7,**80.8**}
句子长度{50,120,200}	{80.2,**84.2**,83.5}	{78.3,**84.6**,80.4}	{70.5,**80.7**,78.1}
学习率{0.01,0.1,0.5}	{**84.2**,80.8,79.3}	{**84.6**,78.6,76.3}	{**80.7**,76.1,73.4}

从表 3 可以看出,卷积核大小对模型性能的影响不是很大,适当增加卷积核的个数会提升模型的性能,过短或过长的句子对模型有一定的影响,学习率对模型的性能影响也比较大.

为了证明本文提出方法的扩展性,本文在 cnn 模型上做了同样的实验,对于 cnn_att 和 cnn_att_N,本实验采用了相同的超参数,实验结果如图 10 所示,可以看到,本文提出的基于带噪观测的 cnn_att_N 方法对比 Lin 等人[5]提出的 cnn_att 方法在性能上有显著提升.

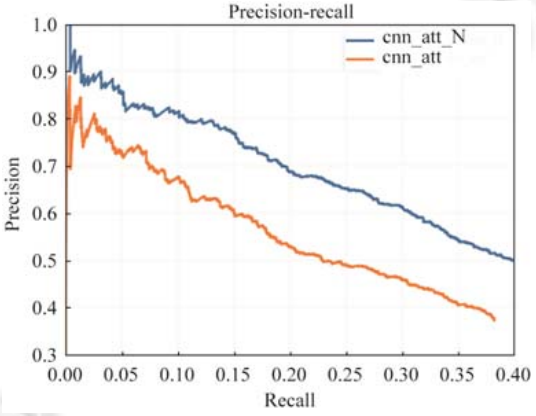

Fig.10 Experimental results of cnn_att

图 10 cnn_att 实验结果

从表 4 可以看出,本文提出的 cnn_att_N 对比 cnn_att 在 *P*@100、*P*@200、*P*@300 上都有 10%左右的提高,证明了带噪观测模型的易扩展性和普适性.

Table 4 Evaluation of performance

表 4 评估结果

P@*N* (%)	100	200	300	Mean
cnn_att_N	**86.1**	**80.6**	**79.1**	**81.9**
cnn_att	76.2	70.6	66.4	71.1

由于 NYT 数据集训练数据中句子标签为“NA”的占了 72.5%左右,从而使样本噪声分布极不均衡,为了分析样本噪声的不同分布对模型性能的影响,本文对用来初始化带噪观测层的混淆矩阵作了 4 种局部降噪处理:混淆矩阵 *N* 表示删除了分布在矩阵第 1 行和第 1 列中的噪声,混淆矩阵 *R* 表示删除了分布在第 1 行中的噪声,混淆矩阵 *C* 表示删除了分布在第 1 列中的噪声,混淆矩阵 *O* 表示不对矩阵作任何操作,图 11 显示了对矩阵降噪后的结果.

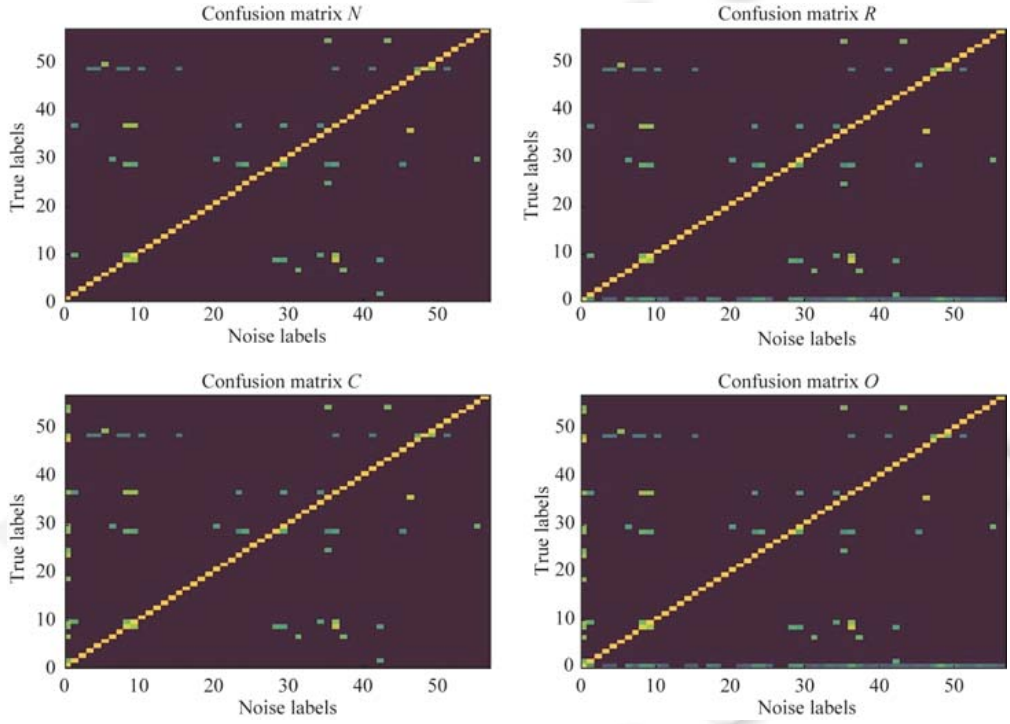

Fig.11 Confusion matrix

图 11 混淆矩阵

然后,本文把对混淆矩阵做的 4 种操作分别用来初始化带噪观测层参数,并做了对比实验,从图 12 可以看出,pcnn_att_*N* 的整体效果最好,pcnn_att_*R* 的效果与 pcnn_att_*N* 的效果接近,因此可以判断,过多的负例会对模型性能造成一定的影响.pcnn_att_*C* 和 pcnn_att_*O* 都是没有对第 1 列的噪声做处理,效果相比前两者比较差,我们判断模型在训练的过程中非“NA”标签数据选择“NA”标签当作正确标签的概率比较高,因此导致后两者模型的性能较差.同样,为了判断带噪观测层和基于噪声分布的注意力机制对模型整体性能的影响,本文做了对比实验,pcnn_att_*A* 为去掉带噪观测层,pcnn_att_*M* 为去掉基于噪声分布的注意力机制,从图 13 可以看出,带噪观测层对模型性能的提升较大,基于带噪观测的注意力机制对模型的性能也有一定的提升.

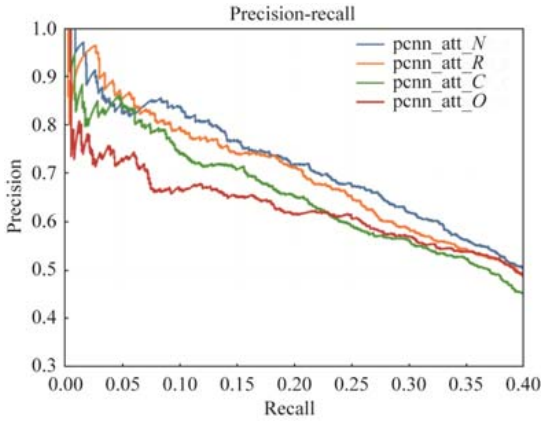

Fig.12 Experimental results

图 12 实验结果图

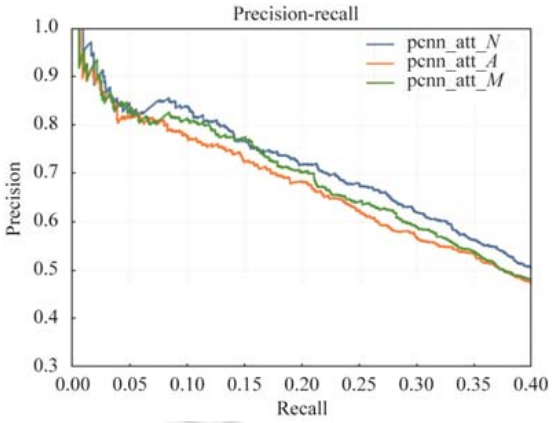

Fig.13 Comparative experimental results

图 13 对比实验结果

7 结束语

本文在基于被标记数据是带有噪声的观测结果这一前提假设下,构建了基于带噪观测的远监督神经网络关系抽取模型来学习数据样本中未知的噪声分布,提出了基于噪声分布的注意力机制来计算句子与噪声标签间的相关程度.进一步地,探讨了样本噪声分布不平衡对该模型性能的影响,实验结果表明,负例中的样本噪声对模型的性能有一定的影响,但影响不大.而在正例中的样本噪声对模型性能影响较大,对它们都做局部降噪处理后模型性能达到最好,表明该模型能够可靠地从含有噪声标签的数据中学到噪声分布信息并计算出真实标签的概率.最后,将基于带噪观测的模型在当前主流的一些神经结构网络上进行测试,性能均有提高,平均提升在 8%左右.从而也证明了本文构建模型的易扩展性和普适性.

References:

[1] Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009. 1003–1011.

[2] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg Springer-Verlag, 2010. 148–163.

[3] Hoffmann R, Zhang C, Ling X, Zettlemoyer L, Daniel S. Knowledge-based weak supervision for information extraction of overlapping relations. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011. 541–550.

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

[5] Lin Y, Shen S, Liu Z, Luan H, Sun M. Neural relation extraction with selective attention over instances. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. 2016,1:2124–2133.

[6] Fan M, Zhao D, Zhou Q, Liu Z, Zheng TF, Chang EY. Errata: Distant supervision for relation extraction with matrix completion. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014,1:839–849.

[7] E HH, Zhang WJ, Xiao SQ, Cheng R, Hu YX, Zhou XS, Niu PQ. A survey of entity relationship extraction based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2019,30(6):1793–1818 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5817.htm [doi: 10.13328/j.cnki.jos.005817]

[8] Xu HY, Zhao H, Wang RB, Fu HC, Liu YL. Research on the movie recommendation system based on user similarity. Journal of Liaoning University, 2018,45(3):193–200 (in Chinese with English abstract).

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

[10] 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. 2014. 2335–2344.

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

[12] 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.

[13] Qu J, Ouyang D, Hua W, Ye Y, Li X. Distant supervision for neural relation extraction integrated with word attention and property features. Neural Networks, 2018,100:59–69.

[14] Ji G, Liu K, He S, Zhao J. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. 2017. 3060–3066.

[15] Ouyang DT, Qu JF, Ye YX. Extending training set in distant supervision by ontology for relation extraction. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):2088−2101 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4638.htm [doi: 10.13328/j.cnki.jos.004638]

[16] Vashishth S, Joshi R, Prayaga SS, Bhattacharyya C, Talukdar P. Reside: Improving distantly-supervised neural relation extraction using side information. arXiv Preprint arXiv:1812.04361, 2018.

[17] Brodley CE, Friedl MA. Identifying mislabeled training data. Journal of Artificial Intelligence Research, 1999,11:131–167.

[18] Rebbapragada U, Brodley CE. Class noise mitigation through instance weighting. In: Proc. of the European Conf. on Machine Learning. Berlin, Heidelberg: Springer-Verlag, 2007. 708–715.

[19] Manwani N, Sastry PS. Noise tolerance under risk minimization. IEEE Trans. on Cybernetics, 2013,43(3):1146–1151.

[20] Natarajan N, Dhillon IS, Ravikumar PK, Tewari A. Learning with noisy labels. In: Advances in Neural Information Processing Systems. 2013. 1196–1204.

[21] Lawrence ND, Schölkopf B. Estimating a kernel Fisher discriminant in the presence of label noise. ICML. 2001,1:306–313.

[22] Li Y, Yang J, Song Y, Cao L, Luo J, Li L. Learning from noisy labels with distillation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1910–1918.

[23] Veit A, Alldrin N, Chechik G, Krasin I, Gupta A, Belongie SJ. Learning from noisy large-scale datasets with minimal supervision. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 839–847.

[24] Vahdat A. Toward robustness against label noise in training deep discriminative neural networks. In: Advances in Neural Information Processing Systems. 2017. 5596–5605.

[25] Yao J, Wang J, Tsang IW, Zhang Y, Sun J, Zhang C, Zhang R. Deep learning from noisy image labels with quality embedding. IEEE Trans. on Image Processing, 2019,28(4):1909–1922.

[26] Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, Rabinovich A. Training deep neural networks on noisy labels with bootstrapping. arXiv Preprint arXiv:1412.6596, 2014.

[27] Joulin A, van der Maaten L, Jabri A, Vasilache A. Learning visual features from large weakly supervised data. In: Proc. of the European Conf. on Computer Vision. Cham: Springer-Verlag, 2016. 67–84.

[28] Jiang L, Zhou Z, Leung T, Li LJ, Li FF. Mentornet: Regularizing very deep neural networks on corrupted labels. arXiv Preprint arXiv:1712.05055, 2017,4.

[29] Ghosh A, Kumar H, Sastry PS. Robust loss functions under label noise for deep neural networks. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. 2017.

[30] Mnih V, Hinton GE. Learning to label aerial images from noisy data. In: Proc. of the 29th Int'l Conf. on Machine Learning. 2012. 567–574.

[31] Sukhbaatar S, Bruna J, Paluri M, Bourdev L, Fergus R. Training convolutional networks with noisy labels. arXiv Preprint arXiv:1406.2080, 2014.

[32] Jindal I, Nokleby M, Chen X. Learning deep networks from noisy labels with dropout regularization. In: Proc. of the 16th IEEE Int'l Conf. on Data Mining. IEEE, 2016. 967–972.

[33] Patrini G, Rozza A, Krishna MA, Nock R, Qu L. Making deep neural networks robust to label noise: A loss correction approach. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1944–1952.

[34] Han B, Yao J, Niu G, Zhou M, Tsang IW, Zhang Y, Sugiyama M. Masking: A new perspective of noisy supervision. In: Advances in Neural Information Processing Systems. 2018. 5836–5846.

[35] Xiao T, Xia T, Yang Y, Huang C, Wang X. Learning from massive noisy labeled data for image classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2691–2699.

[36] Misra I, Lawrence ZC, Mitchell M, Girshick RB. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 2930–2939.

[37] Bekker AJ, Goldberger J. Training deep neural-networks based on unreliable labels. In: Proc. of the 2016 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. IEEE, 2016. 2682–2686.

[38] Ling X, Weld DS. Fine-grained entity recognition. In Proc. of the 26th AAAI Conf. on Artificial Intelligence. 2012. 94–100.

附中文参考文献:

[7] 鄂海红,张文静,肖思琪,程瑞,胡莺夕,周筱松,牛佩晴.基于深度学习的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 [doi: 10.13328/j.cnki.jos.005817]

[8] 徐红艳,赵宏,王嵘冰,付瀚臣,刘逸伦.融合用户相似度的影视推荐系统研究.辽宁大学学报(自然科学版),2018,45(3):193–200.

[15] 欧阳丹彤,瞿剑峰,叶育鑫.关系抽取中基于本体的远监督样本扩充.软件学报,2014,25(9):2088−2101. http://www.jos.org.cn/1000-9825/4638.htm [doi: 10.13328/j.cnki.jos.004638]

叶育鑫(1981－),男,吉林长春人,博士,教授,博士生导师,CCF 专业会员,主要研究领域为人工智能,信息抽取.

王璐(1988－),女,博士后,主要研究领域为非参数模型,生物统计.

薛环(1992－),男,硕士生,主要研究领域为远监督关系抽取.

欧阳丹彤(1968－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自动推理,模型诊断.