计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(5):253−261 [doi: 10.15888/j.cnki.csa.009131]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于外部知识增强的远程监督关系抽取模型①

曾碧卿, 李砚龙, 蔡　剑

(华南师范大学 软件学院, 佛山 528225)
通信作者: 李砚龙, E-mail: 2020023855@m.scnu.edu.cn

摘　要: 远程监督关系抽取方法旨在高效的构建大规模的监督语料并应用在关系抽取的任务上. 但是由于远程监督构建语料的方式, 带来了噪声标签和长尾分布两大问题. 本文提出了一种新颖的远程监督关系抽取模型架构, 与以往的基于管道的训练形式不同, 除了句子编码器模块, 新添加了外部知识增强模块. 通过对知识库中已存在的实体类型与关系进行预处理和编码, 为模型提供句包文本所没有的外部知识. 有利于缓解数据集中存在部分长尾关系示例不足所导致的信息不足的问题, 以及提升了模型对噪声示例的判别能力. 通过在基准数据集 NYT 和 GDS 上的大量实验, 相较于主流最优模型在 AUC 值上分别提升了 0.9% 和 5.7%, 证明了外部知识增强模块的有效性.

关键词: 远程监督; 关系抽取; 图卷积神经网络; 外部知识

引用格式: 曾碧卿,李砚龙,蔡剑.基于外部知识增强的远程监督关系抽取模型.计算机系统应用,2023,32(5):253–261. http://www.c-s-a.org.cn/1003-3254/9131.html

Distantly-supervised Relation Extraction Model via External Knowledge Enhancement

ZENG Bi-Qing, LI Yan-Long, CAI Jian

(School of Software, South China Normal University, Foshan 528225, China)

Abstract: The distantly-supervised relation extraction method aims to efficiently construct a large-scale supervised corpus and apply it to the task of relation extraction. However, constructing the corpus by distant supervision brings two major problems: noise labels and long tail distribution. In this study, a novel distantly-supervised relation extraction model is proposed. Unlike the previous pipeline-based training, an external knowledge enhancement module is added in addition to the sentence encoder module. By preprocessing and coding the existing entity types and relations in the knowledge base, the external knowledge that the sentence package text does not have is provided for the model. It is conducive to alleviating the problem of insufficient information caused by insufficient long tail relation instances in the data set and improving the discrimination ability of the model to noise instances. Through a large number of experiments on the benchmark data sets NYT and GDS, the AUC value has increased by 0.9% and 5.7% respectively, compared with the mainstream optimal model, which proves the effectiveness of the external knowledge enhancement module.

Key words: distant supervision; relation extraction (RE); graph convolutional network (GCN); external knowledge

随着人工智能技术的蓬勃发展和互联网数据的日益递增, 如何从海量的非结构化数据中提取出结构化信息逐渐成为自然语言处理 (natural language processing, NLP) 领域的研究热点. 信息抽取 (information extraction, IE) 作为自然语言处理的 3 大任务之一, 其目标就是从非结构化的文本中提取结构化信息, 这种

① 基金项目: 国家自然科学基金面上项目 (62076103); 广东省基础与应用基础研究基金 (2021A1515011171); 广东省普通高校人工智能重点领域专项 (2019KZDZX1033); 广州市基础研究计划基础与应用基础研究项目 (202102080282)
收稿时间: 2022-11-05; 修改时间: 2023-01-06; 采用时间: 2023-01-19; csa 在线出版时间: 2023-03-24
CNKI 网络首发时间: 2023-03-27

方式是构建和扩充新的知识图谱以及结构化知识库的重要前置技术. 关系抽取 (relation extraction, RE) 属于信息抽取的一个重要步骤, 旨在基于两个给定实体的相关上下文中提取实体之间的对应关系. 因为具有提取文本信息的能力, 所以有利于 NLP 领域的很多应用 (如信息检索、对话生成和问题回答等).

传统监督模型在关系抽取任务中得到了广泛的探索. 然而, 他们模型上的表现在很大程度上取决于训练数据的规模和质量. 为了构建大规模的数据, Mintz 等人[1] 提出了一种新的远程监督机制, 通过将现有的知识库与文本对齐来自动标注训练示例构建数据集, 而这属于一种弱监督的学习方式. 因为远程监督构建语料效率高、人工成本低的缘故吸引了许多国内外研究者的关注. 远程监督关系抽取的第 1 个挑战是噪声标签, 为了解决这一问题, 一些研究者试图通过整合外部信息来丰富模型的背景知识. 一般来说, 外部信息, 如实体描述[2]、实体类型[3] 和知识图谱[4], 将被编码为向量形式, 然后通过简单的连接或注意力机制集成到远程监督关系抽取模型 (distantly-supervised relation extraction, DSRE) 中, 除了上述隐式知识, 知识库 (knowledge base) 中存在一些显式的信息没有充分利用到, 而这种信息可以有效地增强模型在有噪声示例影响下的判别能力. 远程监督关系抽取第 2 个挑战是长尾关系, 长尾关系是指存在一些关系类别的训练示例数目过少的问题, 该问题在实际应用场景中经常可见. 在最广泛使用的 NYT-10 数据集[5] 中有着大约 70% 的长尾关系数据, 这就意味着远程监督关系抽取模型在限定数量的训练示例中学习能力十分重要, 因此, 本文探索了一种基于双通道的结合外部知识增强的关系抽取方法, 通过丰富的外部知识信息与文本信息结合, 提高 DSRE 模型的关系判别能力.

1 相关工作

多示例学习 (multi-instance learning, MIL) 方法[6] 是目前远程监督关系抽取的主流方法之一, 其基本思想是将通过远程监督构建的具有相同的实体对的句子文本组成一个包 (或称为句包), 并以包为单位进行预测. Zeng 等人[7] 提出了一种分段卷积神经网络 (piecewise convolutional neural networks, PCNN), 该工作使用了预训练的词向量将单词映射到低维度向量空间, 然后使用卷积神经网络对句包文本进行特征提取, 从而能够自动捕捉上下文信息. 注意力机制 (attention mechanism) 也是目前远程监督关系抽取的热门方法之一, Lin 等人[8] 提出了一种句子级别的注意力机制 (sentence-level attention), 使用了 PCNN 模型对句包内的句子进行卷积核最大池化操作. 然后对于每个句子与关系向量计算相似度并使用 *Softmax* 函数进行归一化, 最后对包内句子进行加权求和获得句包表征. 本文中的句子编码器部分将采取该模型作为基线. 由于现阶段知识库和训练语料的缺失, 使得远程监督所构建的数据集中部分关系类别示例数目过少, 导致关系抽取模型训练不够充分. 辅助信息增强的方法能很好地缓解这种训练不充分的问题, 基本思路是引入额外知识信息来提升模型的关系判别能力, 如实体关系信息、条件约束、知识表示等. Vashishth 等人[9] 利用知识库中包含的辅助信息 (side information) 和实体类型信息 (entity type) 辅助增强关系抽取效果. Li 等人[10] 使用自注意力机制 (self-attention) 结合实体信息来实现对语义信息的增强, 利用了自注意力机制可以有效地帮助模型关注更为重要的语义信息. Xu 等人[11] 结合了知识表示和文本句子表征, 提出了基于异构表征方法来增强远程监督关系抽取. Liu 等人[3] 结合实体类型对于关系预测的约束作用, 提出了一种多粒度的实体类型约束方法, 并集成到了远程监督关系抽取模型中. 上述工作为辅助信息增强提供了不同思路, 我们可以发现实体类型为 DSRE 模型理解实体提供了有意义的信息. 并在关系抽取过程中发挥了重要作用.

图卷积神经网络 (graph convolutional networks, GCN) 已经应用了到 NLP 的许多经典任务中, 如语义角色标注[12]、依存关系句法分析[13] 和机器翻译等. 本文使用了 GCN 神经网络构建外部知识引入模块, 发现所构建的关系与实体的图结构中边的约束信息也十分重要, 为 DSRE 模型识别噪声示例提供了直接而有效的先验信息.

2 基于外部知识增强的关系抽取模型

2.1 任务定义

给定一个句包 B 和对应实体对 (e_1, e_2), 远程监督关系抽取任务的目的是预测句包所对应的实体对 (e_1, e_2) 的关系 r_i. 构建图结构三元组 $G_{(\text{graph})}=\{T, R, E\}$, 其中 T, R, E 分别表示实体类型、关系和边的集合. 本文的目标是结合外部知识中的实体类型和关系类型, 以

及实体类型与关系类型之间边的约束信息, 来增强远程监督关系抽取模型知识信息的获取, 从而提升模型整体的性能和效果. 在此基础上进一步研究传统句子编码器和预训练模型编码器对 DSRE 任务的影响.

2.2 模型架构

本文模型主要分成两个模块. 1) 句子编码器模块, 该部分主要目标是完成对句包中句子文本的编码, 将句子由文本信息转化为向量表征. 具体流程见第 2.3 节和第 2.4 节. 本文采取传统模型和预训练模型两种句子编码器进行实验探究. 2) 外部知识增强模块, 与一般的 DSRE 模型的管道 (pipeline) 训练形式不同, 本文使用了双通道的训练形式, 利用句子编码器和外部知识增强模块的图卷积编码器分别抽取句包文本特征和知识库类型信息特征, 再进行特征拼接和融合. 外部知识增强模块的主要任务是将知识库中获得的实体类型和关系类型信息通过 GCN 神经网络进行编码, 提取相对应的向量表征, 该模块具体流程见第 2.5 节. 双通道关系抽取模型的整体架构如图 1.

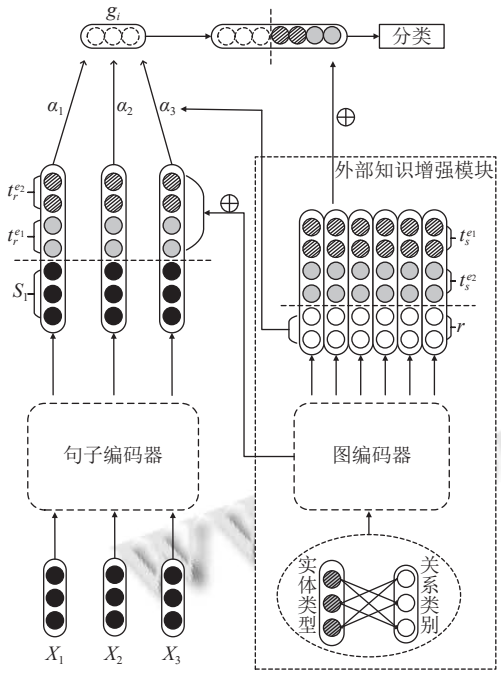

图 1 模型流程图

2.3 传统模型句子编码器

对于句子编码器的输入层, 输入一个句子 $X=\{x_1, \cdots, x_n\}$, n 为句子长度, x_i 表示句子中的每个单词. 通过一个预训练好的嵌入矩阵将每个单词 x_i 转换为一个 d_w 维度的向量 v_i, 然后用词向量 v^{e1} 和 v^{e2} 来表示目标实体 e_1 (头实体) 和 e_2 (尾实体). 位置嵌入使用了两个 d_p 维度的嵌入向量 p^{e1} 和 p^{e2} 来表示单词 x_i 和目标实体对 e_1、e_2 之间的相对距离. 最后通过向量的拼接, 可以得到以下两种类型的词嵌入:

$$x_i^p = [v_i; p_i^{e1}; p_i^{e2}] \in R^{d_w+2d_p} \tag{1}$$

$$x_i^e = [v_i; v_i^{e1}; v_i^{e2}] \in R^{3d_w} \tag{2}$$

其中, x_i^p 表示单词 x_i 的位置嵌入, x_i^e 表示单词 x_i 的词嵌入. d_w 和 d_p 是预定义好的超参数, 表示词嵌入和位置嵌入的维度大小. 最后通过文献 [10] 来嵌入表示每个单词 x_i, 公式如下:

$$M^e = Sigmoid(\lambda \cdot (W_e X^e + b_e)) \tag{3}$$

$$\tilde{X}^p = \tanh(W_p X^p + b_p) \tag{4}$$

$$X = M^e \otimes X^e + (1 - M^e) \otimes \tilde{X}^p \tag{5}$$

其中, $X^e = \{x_1^e, \cdots, x_n^e\}$, $X^p = \{x_1^p, \cdots, x_n^p\}$, n 为句子长度, b_e 和 b_p 是偏置向量. $\otimes$ 表示矩阵每个向量元素相乘, λ 表示平滑系数.

对于输入序列 $X=\{x_1, \cdots, x_n\}$, 一维卷积是权重矩阵 W 与序列 X 的输入矩阵之间的运算. W 为卷积的滤波器, x_i 是与句子中的第 i 个单词相关联的输入向量. 一般来说, 让 $x_{i:j}$ 表示 x_i 到 x_j 的级联, ω 表示滤波器的大小. 定义向量 q_i 作为第 i 个窗口内的 ω 个词向量的级联, 得到新的序列 q_i:

$$q_i = W^{\mathrm{T}} x_{i-\omega+1:j}, 1 \leqslant i \leqslant m \tag{6}$$

q_i 的序列长度是 $n_l-\omega+1$. 因为滑动窗口存在滑出边界的情况, 所以超出长度的部分会填充零元素, 使得向量 q_i 的数量就等于句子 n_l 的长度. 卷积结果是一个特征图矩阵 $Q=\{q_1, \cdots, q_{n_l}\}$. 特征图 Q_i 的数量为 n_f, 其中 n_f 也为滤波器的数量.

分段最大池化的目的是获取句子序列的结构化信息. 经过卷积层后, 每个特征图 Q_i 按两个实体的位置分为 3 个部分 $\{Q_{i_1}, Q_{i_2}, Q_{i_3}\}$. 然后, 对这 3 部分分别进行最大池化操作. 句子表征 h 是所有池化后向量的连接:

$$h = \{q_{i_1}, q_{i_2}, q_{i_3}\} \tag{7}$$

其中, $q_{i_k}=\max(Q_{i_k})$, $k \in \{1, 2, 3\}$, 然后级联所有池化向量 h 得到 $h_{i:m}$, 最后采用非线性激活函数 $ReLU$, 池化层最终输出句子表征 s 为:

$$s = ReLU(h_{i:m}) \in R^{3m} \tag{8}$$

通过上述句子编码器, 我们可以得到一个句包中每个句子的向量表征 s. 传统编码器结构如图 2.

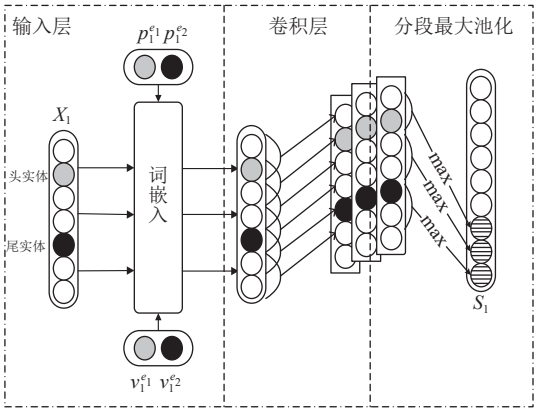

图 2 传统句子编码器

2.4 预训练模型句子编码器

BERT[14] 作为预训练语言模型, 在大量的无监督语料上进行了充分的训练. 并且学习了字符级、单词级、句子级以及句间关系的特征, 很好地缓解了传统的 Word2Vec、FastText、GloVe 等词嵌入方法无法解决的“一词多义”问题. BERT 采用了深层的双向 Transformer[15] 作为组件, 由此学习左右上下文信息, 并生成最终的深层双向语言表征.

本文使用 BERT 预训练语言模型作为句子编码器进行了实验探究, 将输入句子序列转换为特征向量, 并在 DSRE 任务上进行了微调. BERT 采取了 WordPiece 的分词算法, 输入为一个标记 (token) 序列. 在具体的分类任务中, 输入的每一个序列首部插入特定的分类标记 [CLS] 该分类标记对应的最后一个 Transformer 隐藏层输出, 用于表示整个序列的表征信息.

在 BERT 基础上, 本文还采用了 RoBERTa 模型[16] 作为句子编码器进行对比. 相比 BERT 模型, RoBERTa 采取了更为细粒度的 BPE 分词算法, 并且取消了 BERT 的下一个句子预测 (next sentence prediction, NSP) 任务. 在 Transformer 层的堆叠上两个模型相同, 预训练编码器架构如图 3.

2.5 外部知识增强模块

在 FreeBase[17] 知识库中提供了可以利用的可靠类型信息. 在本文的工作中, 使用了从知识库中获得的实体类型和关系类型信息. 通过 NLP 工具对每个句子进行命名实体识别, 预处理句子文本获得实体类型信息. 例如, “广州是广东的省会”, 通过命名实体识别可以得到头实体 e_1 “广州”, 对应实体类型是“城市”, 尾实体 e_2 “广东”, 对应实体类型是“省份”, 而实体类型信息将用于最后句包表征构建. 每个实体之间存在多种关系, 构建如图 4 的图结构, 每个实体之间边的权重就是对应关系. 在预处理过程中会将对应实体转换为知识库中存在的实体类型 (如 ORG、LOC、PERSON 等).

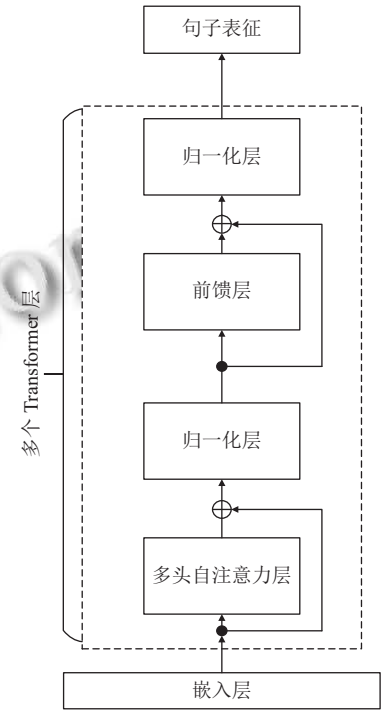

图 3 预训练模型编码器

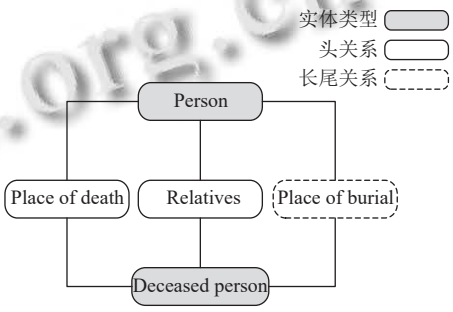

图 4 图结构示意图

同样关系也会转化为知识库中对应的关系类型. 图 4 中头关系表示数据集中样本数目较多的关系类别, 而长尾关系则表示样本数目较少的关系类别. 利用图卷积神经网络能够聚合相邻节点信息的特点, 使得长尾关系能够聚合更多的信息.

定义图结构为一个三元组 $G_{(\text{graph})}=\{T, R, E\}$, 其中 T 表示实体类型节点, R 表示实体关系节点, E 代表边, 从节点 n_i 到节点 n_j 的一条边被表示为 (i, j, l_r), 其中

l_r 表示关系标签. 节点集合为 $N_{(\text{node})}=T\cup R$, 图结构的边集合为 $E_{(\text{edge})}$, 得到图的邻接矩阵 A_{ij} 定义为:

$$A_{ij}=\begin{cases}1, & \text{if } \{n_i,n_j\}\in E\\ 0, & \text{other}\end{cases} \tag{9}$$

对于每个节点 $n_i\in N$, 随机初始化一个 d_n 维度的嵌入向量$d_i^{(0)}$, 通过上述过程将图结构转化为嵌入矩阵 $N^{(0)}=\{n_1^{(0)},\cdots,n_i^{(0)}\}$与邻接矩阵 A_{ij}.

2.5.1 图编码器

GCN 能够聚合每个节点的相邻节点信息和自身的信息. 为了聚合节点自身的信息, 将自环添加到图结构中, 即 $A_{ii}=1$. 以嵌入节点 $v^{(0)}$ 和邻接矩阵 A 作为输入, 利用 GCN 来提取节点表征. GCN 中第 k 层节点 n_i 的计算可定义为:

$$n_i^{(k)}=ReLU\left(\sum_{i=1}^{n}A_{ij}W^{(k)}n_i^{(k-1)}+b^{(k)}\right) \tag{10}$$

其中, 参数 $W^{(k)}$ 和 $b^{(k)}$ 分别代表第 k 层的权重矩阵和偏置向量. 根据输出表征$N^{(k)}\in R^{n\times d_n}$和节点的类别划分为关系类型表征 R 和实体类型表征 T.

由上述两个编码器我们得到句包表征$B=\{s_1,\cdots,s_{n_s}\}$, 关系类型表征$R=\{r_1,\cdots,r_{n_r}\}$和实体类型表征$T=\{t_1,\cdots,t_{n_t}\}$, 其中 n_s、n_r 和 n_t 分别是句子、关系和实体类型的数量.

对于每个句子 S_i, 本文使用 NLP 工具来识别其目标实体的类型. 对于未识别的实体类型, 将分配特殊类型 Unkown. 假设$t_{r_1}^{e_1}$ (实体 1) 和$t_{r_1}^{e_2}$ (实体 2) 分别是图结构中关系 r_i 的前驱节点和后继节点, 然后通过在类型表征矩阵 T 中查找可以得到关系类型表征$t_r^{e_1}$ (实体 1) 和$t_r^{e_2}$ (实体 2). 最后, 通过将句子表征与关系类型表征进行连接得到最终句子表征, 公式如下:

$$g_i=[s_i;t_{r_1}^{e_1};t_{r_2}^{e_2}]\in R^{3d_n},\quad 1\leqslant i\leqslant n_s \tag{11}$$

同样, 通过在类型表征矩阵 T 中查找可以得到实体类型表征$t_{s_1}^{e_1}$ (实体 1) 和$t_{s_2}^{e_1}$ (实体 2). 然后通过将关系表征及其实体类型表征连接得到图表征, 公式如下:

$$c_i=[r_i;t_{s_1}^{e_1};t_{s_2}^{e_2}]\in R^{3d_n} \tag{12}$$

其中, r_i 和$[t_{s_1}^{e_1};t_{s_2}^{e_2}]$将用于注意力机制层的向量拼接和注意力权重系数计算.

2.5.2 注意力机制层

图卷积编码器是为了得到实体类型$[t_s^{e_1};t_s^{e_2}]$和关系类型$[t_r^{e_1};t_r^{e_2}]$的向量表征, 用于句子表征和句包表征的构建. 定义一个句包示例, 第 i 个示例的对应关系 r 的注意力权重系数α_i计算如下:

$$\alpha_i=\frac{\exp(g_ir_i)}{\sum_{j=1}^{n_s}\exp(g_jr_j)} \tag{13}$$

其中, r 为图表征中的关系标签特征, 经过加权求和后可以得到句包表征, 再与实体类型表征$[t_{s_1}^{e_1};t_{s_2}^{e_2}]$串联后构成最终的句包表征$\tilde{b}$.

$$b_r=\sum_{i=1}^{n_s}\alpha_ig_i \tag{14}$$

$$\tilde{b}=[b_r;t_{s_1}^{e_1};t_{s_2}^{e_2}] \tag{15}$$

最后, 将句包表征$\tilde{b}$输入 *Softmax* 分类器, 以计算关系标签上的概率分布, 其中W_{b_r}和 b_r 分别表示权重矩阵和偏置向量. 计算公式如下:

$$P(r|B;G;\theta)=Softmax(W_{b_r}\tilde{b}+b_r) \tag{16}$$

3 实验与分析

3.1 损失函数与优化器

本文实验采取的是交叉熵损失函数, 公式为式 (17), 其中 n 为句包的数量, r_k 为句包 B_k 的标签, θ为模型所有参数. 在训练过程中, 传统句子编码器采取 SGD 梯度下降算法, 预训练编码器采取的 AdamW 梯度下降算法. 在测试过程中, 标签的真值 r 是未知的, 因此, 通过计算相应关系的后验概率, 拥有最高概率的第 k 个关系就是最终的预测结果.

$$J(\theta)=-\frac{1}{n}\sum_{k=1}^{n}\log P(r_k|B_k;G;\theta) \tag{17}$$

$$k^*=\arg\max P(r_k|B_k;G;\theta) \tag{18}$$

3.2 数据集

GDS 数据集: Jat 等人[18] 利用谷歌关系抽取语料集对网页信息进行标注得到的数据集, 其中, 训练集有 13 161 个示例、7 580 个实体对, 测试集包括 5 663 个示例、3 247 个实体对, 一共有 4 种语义关系以及 NA 标签, 具体见表 1.

表 1 GDS 数据集

关系类别	示例数	实体对数
perHasDegree	2 969	1 434
perPlaceOfBirth	3 356	2 159
perPlaceOfDeath	3 469	1 948
NA	4 574	2 667

NYT-10 数据集: Riedel 等人[5] 通过将 FreeBase 和纽约时报语料进行对齐构建了 NYT 数据集. 其中训练

集由 2005 年和 2006 年新闻组成, 测试集由 2007 年新闻组成. NYT 数据集包括了 52 种语义关系和 NA 标签, 训练集共包括 522 611 个示例、281 270 个实体对, 测试集包括 172 448 个示例、96 678 个实体对. GDS 数据集和 NYT 数据集对比见表 2.

表 2 GDS 数据集和 NYT 数据集

数据集	关系类别	Train/Val	示例数	实体对数
GDS	5	Train	11 297	6 498
		Val	1 864	1 082
		Test	5 663	3 247
NYT	53	Train	522 621	281 270
		Test	172 448	76 678

3.3 评估标准

在进行远程监督关系抽取性能评估时, 采用了关系抽取常用评价指标进行评估, 包括精确率 (precision, *P*)、召回率 (recall, *R*), 计算公式如下:

$$P = \frac{TP}{TP + FP} \tag{19}$$

$$R = \frac{TP}{TP + FN} \tag{20}$$

其中. *TP* 为真正例, *FN* 为假反例, *FP* 为假正例. 为了直观呈现本实验效果, 还使用了 *PR* 曲线 (precision-recall curve) 和 AUC 值 (area under curve) 来将本文模型与基线模型进行对比.

本文使用 *P*@*N* 作为模型评估指标, 该指标是指在句包中选取一定数量的示例进行实验时, 将得分最高的前 *N* 个示例进行流出法进而得到精确率. 实验设置为选取一个 (one), 两个 (two), 全部 (all) 示例进行训练, 分别评估 3 种选择下的 *P*@100, *P*@200 和 *P*@300.

3.4 基线模型

(1) Mintz[1]: 一种多分类逻辑回归的远程监督关系抽取模型.

(2) PCNN+ATT[8]: 分段卷积神经网络模型作为句子编码器, 加入句子级别的选择注意力机制的远程监督模型.

(3) RESIDE[9]: 一种借助辅助信息和实体类型信息辅助增强的远程监督关系抽取模型.

(4) PCNN+HATT[19]: 一种基于 PCNN 编码器利用层次化注意力机制的远程监督模型.

(5) PCNN+BATT[20]: 一种基于 PCNN 编码器利用包级别注意力机制的远程监督模型.

(6) DISTRE[21]: 一种基于 GPT-2 编码器在远程监督数据集下进行微调的远程监督模型.

(7) DSRE-VAE[22]: 一种基于变分自编码器 (VAE) 的 DSRE 模型, 可以通过加入外部知识库的先验知识来进一步改进.

3.5 超参数设置

表 3、表 4 中列出了本文实验所使用的超参数, 其中表 3 为传统模型 PCNN 和 CNN 的参数设置, 表 4 为预训练模型 BERT 和 RoBERTa 的参数设置.

表 3 传统模型参数设置

参数	数值
滑动窗口	3
句子向量维度	230
词向量维度	50
位置向量维度	5
批次大小	160
丢弃率	0.5
学习率	0.5
平滑系数	17

表 4 预训练模型参数设置

参数	数值
隐藏层维度	768
文本最大长度	120
句包大小	4
训练轮数	5
批次大小	32
丢弃率	0.5
学习率	2E–5

3.6 实验结果分析

为了评估本文提出方法的有效性, 将其与第 3.4 节基线模型在 GDS 和 NYT-10 数据集上进行实验探究, 具体如表 5 和表 6. 同时为了验证句子编码器对于本文模型的影响, 也使用了预训练模型 BERT 作为句子编码器并加入外部知识增强模块, 构建成 BERT+GCN 模型, 可以发现预训练模型 BERT 在特征提取能力上相较于传统模型 PCNN 更为优异, 在两个基准数据集上 AUC 值均有一定提升, 其中 GDS 数据集上 *P*@100 和 *P*@200 指标更是达到了 99% 的 SOTA 的效果.

本文模型 PCNN+GCN 与使用了句子级别注意力机制的基线模型 PCNN+ATT, 使用了层次注意力机制的 PCNN+HATT 模型, 以及使用了包级别注意力机制的 PCNN+BATT 模型在两个数据集上 AUC 值均有提升. 说明外部信息中的实体类型和所构建图结构中存

在的正确的关系规则, 对于降噪和长尾关系抽取两个任务上均有显著改进. 同时句包之间的不同粒度的注意力机制会受文本句包的局限性影响, 认为所有句包中至少有一个句子是拥有正确的标签. 然而通过对远程监督数据集进行分析, 其实存在有大量全是噪声示例的句包, 而句子级别和句包级别注意力机制对于都是噪声示例的噪声包是很难甄别的. 外部信息不会被原本知识库通过远程监督构建的错误标签样本所干扰, 能很好地筛选这些噪声句包, 外部信息中的关系类型信息是知识库中提取的正确的关系规则, 对于错误分配标签的句包能通过注意力机制为之分配较低权重, 从而缓解全噪声句包的问题.

表 5 GDS 数据集上模型 AUC 值、*P@N*

模型	AUC	*P*@100 (%)	*P*@200 (%)	*P*@300 (%)
PCNN+ATT	0.799	96.4	93.3	91.5
PCNN+HATT	0.816	94.0	92.9	92.4
PCNN+BATT	0.802	96.3	94.7	93.1
RESIDE	0.868	98.4	96.4	95.2
DSRE-VAE	0.876	96.9	96.7	96.3
PCNN+GCN	0.897	97.0	95.5	94.3
BERT+GCN	**0.933**	**99.0**	**99.0**	**97.3**

注: 加粗模型为本文模型, 加粗指标为最优结果

表 6 NYT-10 数据集上模型 AUC 值、*P@N*

模型	AUC	*P*@100 (%)	*P*@200 (%)	*P*@300 (%)
Mintz	0.107	52.3	48.6	45.0
PCNN+ATT	0.341	81.6	73.0	66.9
PCNN+BATT	0.351	76.9	75.4	72.9
RESIDE	0.415	83.0	**76.7**	71.0
DISTRE	0.422	68.0	65.3	65.0
REDSanT	0.424	78.0	76.5	73.0
DSRE-VAE	0.429	83.0	75.5	73.0
PCNN+GCN	0.387	82.0	71.5	**73.5**
BERT+GCN	**0.438**	**84.0**	75.5	67.7

注: 加粗模型为本文模型, 加粗指标为最优结果

RESIDE 作为外部信息增强的模型基线, 本文模型 PCNN+GCN 在两个基准数据集上与其相比也有一定提升, 主要原因是 PCNN+GCN 具备更高效率和性能, RESIDE 模型使用了 38 种实体类型作为外部知识引入, 而本文模型只采用了 18 种类型却取得了更优的效果. 本文模型相较于 RESIDE 实体类型信息主要针对的长尾实体, 这些长尾实体因为训练示例数目较少从而使得模型难以充分学习到上下文信息. 而且 RESIDE 模型采取的别名信息, 在本文的 BERT-GCN 模型中是可以通过预训练模型在大规模语料的预训练中提前获取的, 因此提供的外部信息会相对冗余. 本文提出的轻量级 PCNN+GCN 因为本身所提供外部实体类型较少, 实际训练过程中效率更高, 且针对性地选取了一些长尾实体类型, 在面对长尾问题上处理更为优秀.

本文模型相较于主流最优模型 DSRE-VAE 在两个基准数据集上 AUC 值均有提升. 本文提出的外部知识增强模块与 DSRE-VAE 引入的外部 KB 知识不同. 外部知识增强模块主要利用了图结构中的关系类型和实体类型信息, 而 DSRE-VAE 利用了 KB 知识库的先验知识, 通过图嵌入的方式来进行外部知识引入. 但 DSRE-VAE 构建的先验知识规模要远远大于本文所提出的外部知识增强模块所构建的图结构信息. 因此在 *P*@200 和 *P*@300 指标上都要优于本文模型. 但本文模型在 AUC 值上的提升意味着模型整体鲁棒性更强, 主要原因是因为外部信息规模太大导致的, 大量的外部信息可能存在冗余, 使得模型训练效率不高, 且没有针对性的选取长尾实体类型, 使得缓解长尾问题任务上不占优势. 但 DSRE-VAE 模型所提出的概率模型, 对于句子表征空间的描述类似于知识库, 可解释性更强, 未来工作也会考虑将该先验知识空间集成到本模型进行优化.

3.7 消融实验

本文对于句子编码器部分进行了对比试验, 如图 5, 主要采取了传统模型 CNN、PCNN 以及预训练模型 BERT、RoBERTa 作为句子编码器进行对比. 句子编码器与外部知识增强模块结合方式如图 2, 编码器会将句包文本转为向量表征, 与外部知识增强模块的类型信息表征进行融合, 再继续进行关系抽取的多分类任务. 实验发现传统模型 CNN 与 PCNN, 在 NYT-10 数据集上 AUC 值有明显提升. 主要原因是 NYT-10 数据集的噪声影响严重, 而 PCNN 相较于 CNN 引入了位置编码和分段最大池化技术, 对句包文本特征提取和降噪的效果更优, 鲁棒性更强. 预训练模型 BERT、RoBERTa 相比于 PCNN 模型 AUC 值均有一定提升, 说明预训练模型相比于传统模型捕捉上下文信息的能力更强, 因为在大量语料上进行过预训练, 从而泛化能力更优.

本文在句子编码器部分相对于传统模型 PCNN 添加了 (实体-感知) 词嵌入[10] 的方式, 同时还增加了基于图卷积 GCN 的知识增强模块, 针对这两部分改进进行了相关消融实验来验证有效性. 根据表 7 和表 8 可知, 去除了知识增强模块 (w/o GCN) 后, PCNN+GCN 和

BERT+GCN 在 GDS 和 NYT-10 数据集上 AUC 值均有明显下降. 说明 GCN 可以学习到关系标签之间更明确的相关性, 即正确的关系规则, 提高了 DSRE 任务的抽取性能. 去除了 (实体-感知) 词嵌入 (w/o ENT) 后, 在 GDS 和 NYT-10 数据集上 PCNN+GCN 和 BERT+GCN 模型 AUC 值都有一定下降, 说明实体信息的嵌入对于编码器在特征抽取过程中句包文本信息进行了很好的补充.

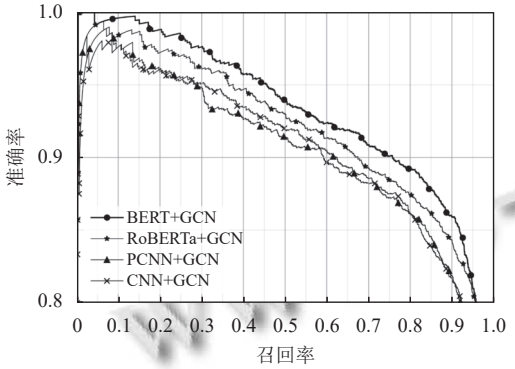

图 5 GDS 数据集上模型对比 *PR* 曲线图

表 7 GDS 数据集上模型 AUC 值 (%)

模型	AUC
CNN+GCN	89.6
PCNN+GCN	89.7
BERT+GCN	93.3
RoBERTa+GCN	91.8
PCNN+GCN w/o GCN	79.9
BERT+GCN w/o GCN	83.5
PCNN+GCN w/o ENT	88.9
BERT+GCN w/o ENT	92.8

表 8 NYT 数据集上模型 AUC 值 (%)

模型	AUC
CNN+GCN	38.2
PCNN+GCN	41.2
BERT+GCN	43.8
RoBERTa+GCN	41.8
PCNN+GCN w/o GCN	34.5
BERT+GCN w/o GCN	39.8

4 总结

本文提出了一种基于外部知识增强的远程监督关系抽取模型, 能有效地处理远程监督所带来的噪声和长尾问题. 通过句子编码器中词嵌入的优化以及外部知识引入的实体类型和关系类型信息, 提升了模型对于句包文本的特征提取效果, 从而缓解了噪声示例对 DSRE 任务的影响. 同时外部知识增强模块利用了 GCN 神经网络, 能很好地将头节点的信息通过神经网络的边传递到长尾节点, 丰富长尾节点信息的同时, 缓解了长尾示例带来的模型训练不充足的问题. 但是外部知识的引入不能从根本上解决实际场景的远程监督数据集的问题, 因为外部引入的实体类型和关系类型信息, 不能够囊括所有领域. 未来希望在跨领域且低资源的远程监督数据下进行关系抽取研究.

参考文献

1 Mintz M, Bills S, Snow R, *et al*. Distant supervision for relation extraction without labeled data. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec: ACL, 2009. 1003–1011.

2 Ji GL, Liu K, He SZ, *et al*. Distant supervision for relation extraction with sentence-level attention and entity descriptions. Proceedings of the 31th AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017. 3060–3066.

3 Liu Y, Liu K, Xu LH, *et al*. Exploring fine-grained entity type constraints for distantly supervised relation extraction. Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. Dublin: ACL, 2014. 2107–2116.

4 Wang GY, Zhang W, Wang RX, *et al*. Label-free distant supervision for relation extraction via knowledge graph embedding. Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 2246–2255.

5 Riedel S, Yao LM, McCallum A. Modeling relations and their mentions without labeled text. Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases. Barcelona: Springer, 2010. 148–163.

6 Zhou ZH. Multi-instance learning: A survey. Technical Report, Nanjing: Nanjing University, 2004.

7 Zeng DJ, Liu K, Chen YB, *et al*. Distant supervision for relation extraction via piecewise convolutional neural networks. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 1753–1762.

8 Lin YK, Shen SQ, Liu ZY, *et al*. Neural relation extraction with selective attention over instances. Proceedings of the 54th Annual Meeting of the Association for Computational

Linguistics. Berlin: ACL, 2016. 2124–2133.
9 Vashishth S, Joshi R, Prayaga SS, *et al*. RESIDE: Improving distantly-supervised neural relation extraction using side information. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 1257–1266.
10 Li Y, Long GD, Shen T, *et al*. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 8269–8276.
11 Xu P, Barbosa D. Investigations on knowledge base embedding for relation prediction and extraction. arXiv:1802.02114, 2018.
12 Bastings J, Titov I, Aziz W, *et al*. Graph convolutional encoders for syntax-aware neural machine translation. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 1957–1967.
13 Strubell E, McCallum A. Dependency parsing with dilated iterated graph CNNs. Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing. Copenhagen: ACL, 2017. 1–6.
14 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186.
15 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
16 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.
17 Bollacker K, Evans C, Paritosh P, *et al*. Freebase: A collaboratively created graph database for structuring human knowledge. Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver: ACM, 2008. 1247–1250.
18 Jat S, Khandelwal S, Talukdar P. Improving distantly supervised relation extraction using word and entity based attention. arXiv:1804.06987, 2018.
19 Han X, Yu PF, Liu ZY, *et al*. Hierarchical relation extraction with coarse-to-fine grained attention. Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 2236–2245.
20 Ye ZX, Ling ZH. Distant supervision relation extraction with intra-bag and inter-bag attentions. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Minneapolis: ACL, 2019. 2810–2819.
21 Alt C, M Hübner, Hennig L. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 1388–1398.
22 Christopoulou F, Miwa M, Ananiadou S. Distantly supervised relation extraction with sentence reconstruction and knowledge base priors. Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 11–26.

(校对责编: 孙君艳)