第 59 卷　第 6 期
2023 年 11 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 59, No. 6
Nov., 2023

DOI:10.13232/j.cnki.jnju.2023.06.011

基于BoBGSAL-Net的文档级实体关系抽取方法

冯超文[1,2],吴瑞刚[1,2],温绍杰[1,2],刘英莉[1,2*]

(1.昆明理工大学信息工程与自动化学院,昆明,650500;
2.云南省计算机技术应用重点实验室,昆明理工大学,昆明,650500)

摘　要:文档级实体关系抽取的主要任务是提取文档中实体之间的关系.相较于句内实体关系提取,文档级实体关系抽取需要对文档中多个句子进行推理.为了解决文档中不同实体之间的复杂信息交互问题,提出一个混合提及级图MMLG(Mixed Mention-Level Graph)策略,用于拟合文档中不同实体之间的复杂信息交互,提高模型对于文档级实体关系的感知能力.此外,为了应对实体关系中存在的关系重叠问题,构建了实体关系图ERG(Entity Relation Graph)模块,该模块融合了路径推理机制,主要针对实体间的多个关系路径进行推理学习,更准确地识别提及级节点实体及关系.通过将MMLG策略与ERG模块聚合到实体关系抽取模型中,构建BoBGSAL-Net(Based on Bipartite Graph Structure Aggregate Logic Network)模型,并在公开数据集DocRED和作者实验室构建的数据集AlSiaRED上开展实验,结果证明BoBGSAL-Net在文档级实体关系抽取任务中性能得到提升,其中BoBGSAL-Net+BERT模型在AlSiaRED数据集上的关系抽取任务中$F1$指标达到66.04%,和其他模型相比,整体性能提升了4.4%,泛化能力突出,综合效果最优.

关键词:文档级实体关系抽取,混合提及级图,实体关系图,BoBGSAL-Net模型

中图分类号:TP183　　**文献标志码**:A

Document-level entity relation extraction method based on BoBGSAL-NET

Feng Chaowen[1,2], Wu Ruigang[1,2], Wen Shaojie[1,2], Liu Yingli[1,2*]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China; 2. Yunnan Key Laboratory of Computer Technology Application, Kunming University of Science and Technology, Kunming, 650500, China)

Abstract: The primary task of document-level entity relation extraction is to extract relationships among entities in a document. Compared to intra-sentence entity relation extraction, document-level entity relation extraction requires reasoning across multiple sentences in the document. To address the challenge of complex information interaction among different entities in the document, this paper proposes a Mixed Mention-Level Graph (MMLG) strategy for modeling intricate information interaction among different entities in the document, thereby enhancing the model's perception of document-level entity relations. Additionally, to handle the issue of relationship overlap within entity relations in documents, an Entity Relation Graph (ERG) module is constructed, incorporating a path reasoning mechanism that focuses on inferring and learning from multiple relationship paths among entities. This module enhances the accurate identification of entity and relation nodes at the mention level.By integrating the MMLG strategy and ERG module into the entity relation extraction model, this paper develops the BoBGSAL-Net (Based on Bipartite Graph Structure Aggregate Logic Network) model. Experimental

基金项目:国家自然科学基金(52061020,61971208),云南计算机技术应用重点实验室开放基金(2020103),云南省重大科技专项计划项目(202302AG050009)

收稿日期:2023-08-20

*通讯联系人,E-mail:lyl@kust.edu.cn

evaluations are conducted on the publicly available DocRED dataset and the AlSiaRED dataset created by the authors' laboratory. The experimental results demonstrate the performance improvement of BoBGSAL-Net in document-level entity relation extraction tasks. Notably, the BoBGSAL-Net+BERT model achieves an $F1$ score of 66.04% in relation extraction tasks on the AlSiaRED dataset, showcasing a 4.4% overall performance improvement compared to other models. The model exhibits exceptional generalization capability, culminating in an optimal comprehensive performance.

Key words: document-level entity relation extraction, mixed mention-level graph, entity relation graph, BoBGSAL-Net model

近年来,随着深度学习算法快速发展,基于神经网络的文档级实体关系抽取[1]方法已经成为研究热点.文档级实体关系抽取是指从整个文档中抽取出实体之间的关系,相较于句子级实体关系抽取[2],文档级实体关系抽取需要处理更大量、更复杂的实体关系信息.因此,需要将多个句子中的实体关系整合起来,以更准确地识别实体之间的关系.目前实体关系抽取的方法主要分为基于传统机器学习和深度学习的方法.基于传统机器学习的文档级实体关系抽取依赖于特征工程,无法处理复杂关系,并且模型的泛化能力有限.相比之下,基于深度学习的实体关系抽取方法可以很好地解决这些问题,对长文本处理更加高效,也具有更强的鲁棒性.

1 相关工作

基于深度学习的文档级实体关系抽取的方法主要包括基于序列[3]、基于图和基于预训练语言模型[4].Geng et al[5]提出一种基于双向树结构长短期记忆的端到端方法,提取基于句子依赖树的结构特征.Luo et al[6]提出一种基于神经网络的方法,即带有条件随机场层的注意力双向长短时记忆方法,用于文档级别的化学命名实体识别.Tang et al[7]提出一种分层推理网络,充分利用来自实体级、句子级和文档级的丰富信息,将平移约束和双线性变换应用于多个子空间中的目标实体对,以获得实体级的推理信息.Najibi et al[8]提出一种基于卷积神经网络[9-12]的目标检测技术,可以从多尺度网格的固定边界框开始,训练一个回归器,迭代地将网格元素移动和缩放到紧密围绕物体的框中.Huang et al[13]提出一种针对不断变化的大型图而设计的动态图划分算法,该算法与分区算法紧密集成,进一步减少了分区算法切割的边数.尽管以上研究方法已在文档级实体关系抽取任务中取得了一些较好的成果,但仍然存在一定的局限性,具体表现在识别一些不需要一致性的实体类型时可能存在缺点.例如,在文章中有时会使用相同的缩写来指代不同的实体,而且在处理需要捕获更复杂的长距离依赖信息的文章时,这些方法表现不佳.

针对文档级实体关系抽取的研究,主要难点有:(1)文档中不同实体之间的复杂信息交互问题,需要对文档中多个句子进行推理,对于深度学习模型的训练和推理会带来更高的计算复杂度;(2)文档中实体关系中存在的关系重叠问题,一个实体可能有多种不同的含义及解释,一个实体对应多种关系的复杂性.为了解决上述问题,本文提出一种基于双图结构的聚合逻辑网络(Based on Bipartite Graph Structure Aggregate Logic Network,BoBGSAL-Net)的文档级实体关系抽取方法,该方法首先构建一个混合提及级图(Mixed Mention-Level Graph,MMLG)来模拟整个文档中不同提及节点之间的信息交互,然后构建了实体关系图(Entity Relation Graph),针对文档的句内实体进行关系提取.基于MMLG和ERG,本文融合聚合逻辑推理路径来推断实体之间的关系并进行分类预测.最后,在公开的数据集DocRED以及作者实验室构建的数据集AlSia-RED[14]上进行实验,结果证明BoBGSAL-Net在文档级实体关系抽取任务中性能有一定提升.

2 本文方法

BoBGSAL-Net是一种基于双图特征的图聚合和推理网络[15-17],利用异构提及级图来建模文档中不同提及级节点之间的交互,并捕获文档感知功能,从而更好地处理文档级的实体关系提取任务.BoBGSAL-Net采用实体级图,并融合路径推理机制来更明确地推断关系.该模型由四个部

件组成,包括文本编码嵌入机制、混合提及级图策略、实体关系图模块和分类预测模块.其中,文本编码嵌入机制采用了BiLSTM[6],GloVe[18]和BERT[19]三种文本编码嵌入方式.BiLSTM在捕捉局部上下文信息方面效果较好;GloVe生成的词向量利用了全局语料库中的共现信息,对于单个词的语义表示有一定优势;BERT通过双向文本建模捕捉丰富的上下文信息,对于理解复杂实体关系至关重要.BERT的预训练模型能有效编码整个文档上下文,包括长文本中的实体语义关系.相较于BiLSTM和GloVe,采用BERT作为词嵌入模型具有显著优势.混合提及级图策略主要用于不同提及级节点之间的信息交互模拟计算,实体关系图模块对整个文档中的实体关系进行交叉计算.最后,通过实体关系分类预测模块,该模型可以从文档中抽取实体和关系.BoBG-SAL-Net的完整结构如图1所示.

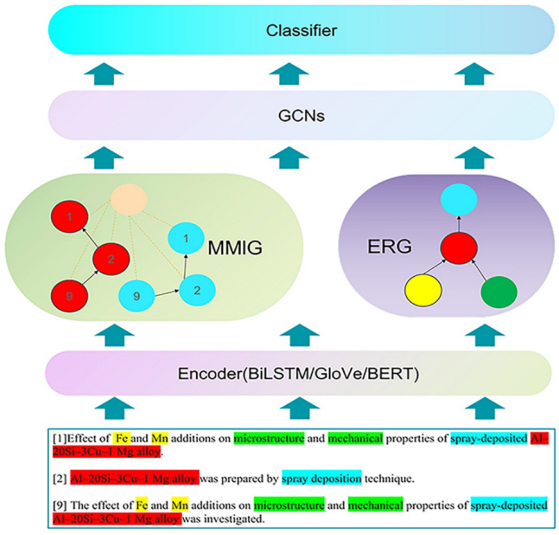

图1　BoBGSAL-Net结构图

Fig. 1　The structure of BoBGSAL-Net

2.1　文本编码嵌入机制　在文本编码嵌入机制中,定义一个文档$D=\{w_i\}_{i=1}^{n}$,其中n为文档中包含的词数量,将D映射为一个向量序列$\{g_i\}_{i=1}^{n}$.对于D中每个词w_i,首先将词嵌入、实体类型嵌入与核心关系嵌入进行拼接,作为一个文本编码向量嵌入,表示方法如式(1)所示:

$$x_i=[E_w(w_i);E_t(t_i);E_c(c_i)] \tag{1}$$

其中,$E_w(\cdot)$,$E_t(\cdot)$和$E_c(\cdot)$分别代表词嵌入矩阵、实体类型嵌入矩阵和核心关系嵌入矩阵,t_i表示命名实体类型,c_i表示实体id.无论是DocRED数据集还是文档级实体关系抽取数据集中都有大量词不属于任何实体,因此本文定义一个None实体类型和id作为这些实体的实体类型嵌入与核心关系嵌入.

接着将向量化的单词表征嵌入编码器来获得每个词的上下文敏感表征,嵌入公式如下所示:

$$[g_1,g_2,\cdots,g_n]=Encoder([x_1,x_2,\cdots,x_n]) \tag{2}$$

2.2　混合提及级图策略　本文提出混合提及级图策略对文档级的提及级节点和实体之间的相互作用进行建模.该策略包含两种不同类型的节点,即提及级节点和文档级节点.每个提及级节点表示一个实体的提及表征,文档级节点则用于建模整个文档,类似于一个支点与不同的提及级节点进行交互,以解决长距离节点交互的问题.提及级节点之间的交互采用有向无环图的形式,该表示方式同时代表了节点在文档中的上下文关系.

MMLG模块共包含三种类型的边,包括共指边、实体间边和文档级边.其中,共指边指同一实体类型形成的边,例如实验名——实验名.通过共指边,可以实现文档中同一实体在不同提及方式之间的信息交互和建模.实体间边指两个不同的实体在一个句子中共同出现形成的边,例如合金——元素.通过实体间边,可以对实体之间的信息交互进行建模.共指边和实体间边都属于提及级边,而所有提及级内容都通过文档级边连接到文档节点.

通过以上连接结构,文档级节点可以关注到所有提及级节点,并实现文档和提及之间的互动.同时,使用文档级节点作为支点,两个提及级节点之间的距离最多为两条边,通过这种结构可以很好地避免文档长文本的长距离依赖问题.

接着,在MMLG模块上使用GCN来聚合邻接特征.给定第l层的节点u,图卷积操作的定义如下式所示:

$$h_u^{(l+1)}=\delta\left(\sum_{k\in\kappa}\sum_{v\in N_k(u)}W_k^{(l)}h_v^{(l)}+b_k^{(l)}\right) \tag{3}$$

其中,k代表不同类型的边,$W_k^{(l)}\in\mathbb{R}^{d\times d}$和$b_k^{(l)}\in\mathbb{R}^d$都是可训练参数,$N_k(u)$表示连接在第$k$类边上的节点$u$的邻接,$\delta$表示激活函数.

GCN的不同层表达了不同抽象层次的特征,为了涵盖所有层次的特征,将各隐藏层状态串联起来,形成节点u的最终表示,如式(4)所示:

$$m_u=\left[h_u^{(0)};h_u^{(1)};\cdots;h_u^{(n)}\right] \tag{4}$$

其中,$h_u^{(0)}$是节点u的初始化表征.文档中从第s个词到第t个词的提及如式(5)所示:

$$h_u^{(0)}=\frac{1}{t-s+1}\sum_{j=s}^{t}g_j \tag{5}$$

对于文档级节点,则被初始化为编码模块输出的文档表征.

2.3 实体关系图模块 边连接的实体合并到实体节点,得到ERG中的节点,该模块对文档级节点透明,被提及N次的第i个实体节点利用平均数来表示,如式(6)所示:

$$e_i=\frac{1}{n}\sum_{n}m_n \tag{6}$$

将所有连接两个相同实体提法的实体间边合并,得到ERG中的边.从实体i到实体j的有向边的表示方法如式(7)所示:

$$e_{ij}=\delta\left(W_q\left[e_i;e_j\right]+b_q\right) \tag{7}$$

其中,W_q和b_q为可训练的参数,δ为激活函数.基于向量化的边表示,头实体e_h和尾实体e_t之间经过实体e_o的第i条路径采用如式(8)所示:

$$P_{h,t}^i=\left[e_{ho};e_{ot};e_{to};e_{oh}\right] \tag{8}$$

以上只考虑两次跳转情况的路径,上述公式很容易扩展到多次跳转路径的情况.同时,引入注意力机制[20],使用实体对(e_h,e_t)作为query来融合e_h和e_t之间的不同路径信息.融合公式的表述如式(9)~(11)所示:

$$s_i=\delta\left(\left[e_h;e_t\right]\cdot W_t\cdot P_{h,t}^i\right) \tag{9}$$

$$\alpha_i=\frac{e^{s_i}}{\sum_j e^{s_i}} \tag{10}$$

$$P_{h,t}=\sum_i\alpha_i P_{h,t}^i \tag{11}$$

其中,α_i表示第i条路径的归一化注意力权重,这样会使模型更关注有用的路径.然后,在ERG模块上融合GCN来获取实体关系信息,最大程度上提高模型对实体关系的预测准确率.

通过ERG模块将实体的提及信息进行融合,通常这些信息分布在多个句子中,通过实体之间的不同路径来模拟潜在的推理线索.然后采用自注意力机制结合这些信息,能够更好地利用潜在的逻辑推理链来预测实体之间的关系.

2.4 分类预测模块 BoBGSAL-Net的分类预测模块是该模型的最后一层,用于对文档级实体关系进行分类预测.该模块通过将每个实体对连接起来实现此目的,连接方式如下.

(1)对每个ERG模块中得到的头实体和尾实体表征e_h和e_t,通过对比操作来加强特征,将两个实体表征的绝对值相减,即$|e_h-e_t|$.然后逐元素相乘,即$e_h\odot e_t$.

(2)将每个MMLG模块中的文档级节点表示为m_{doc},利用该节点来聚合跨句间的信息,并提供文档级节点与提及级节点的交互表征信息.

(3)综合以上两步推理路径信息$P_{h,t}$,具体表述如下所示:

$$I_{h,t}=\left[e_h;e_t;|e_h-e_t|;e_h\odot e_t;m_{\text{doc}};P_{h,t}\right] \tag{12}$$

最后,将文档级实体关系抽取任务定位为多标签分类任务,并对实体之间的关系进行预测,公式如下:

$$P(r|e_h,e_t)=\text{sigmoid}\left(W_b\delta\left(W_aI_{h,t}+b_a\right)+b_b\right) \tag{13}$$

其中,W_a,W_b,b_a,b_b为训练参数,δ为激活函数.使用二进制交叉熵作为分类损失来训练该端到端网络,表征连接过程如式(14)所示:

$$\begin{aligned}\ell=-\sum_{D\in S}\sum_{h\neq t}\sum_{r_i\in\mathrm{R}}&\mathbb{I}(r_i=1)\lg P\left(r_i|e_h,e_t\right)+\\&\mathbb{I}(r_i=0)\lg\left(1-P\left(r_i|e_h,e_t\right)\right)\end{aligned} \tag{14}$$

其中,S代表整个语料库,$\mathbb{I}(\cdot)$表示指示函数.

3 实验设置

3.1 实验环境 实验在一台搭载Ubuntu 20.04操作系统的服务器上进行,服务器的相关配置如表1所示.由于BoBGSAL-Net模型是深度学习模型,需要GPU进行模型运算,GPU可以极大地提高模型的运算速度.实验使用的核心依赖工具包如表2所示.

使用NumPy和Matplotlib对数据进行探索性分析,并使用Scikit-learn和Torch构建训练机器学习和深度学习模型.在处理文本数据时,使用

表1　服务器的详细配置

Table 1　Detailed server configuration

操作系统	Ubuntu 20.04 LST
CPU型号	Inter Xeon Gold 5120 (56) CPU @2.2GHZ
CPU存储	256 G
GPU型号	NVIDIA Tesla V100
GPU存储	16 G

表2　核心依赖工具包

Table 2　Core dependency toolkit

安装包	版本
CUDA	10.2
Python	3.7.5
Matplotlib	3.3.5
NumPy	1.19.4
Torch	1.6.0
Transformers	3.1.0
Scikit-learn	0.23.2

Transformers库中预训练的模型来提取特征,然后使用Scikit-learn或Torch进行分类和回归任务.此外,使用CUDA在GPU上加速模型的训练和推断过程,提高计算速度.

3.2　数据集　DocRED是一个大规模数据集,从维基百科和维基数据构建而来.它提供了全面的人工标注,包括实体提及、实体类型、关系事实以及相应的支持证据.共有97个目标关系,每个文档中平均大约有26个实体.数据规模为3053个训练文档,1000个开发集文档和1000个测试文档.此外,DocRED还收集了用于其他研究的远程监督数据.

作者自建数据集AlSiaRED[14]是在铝硅合金研究领域的专家指导下,构建的用于铝硅合金关系抽取的一个数据集,其构建过程包括选择材料科学文献、确定标注内容以及进一步确定数据集的实体类型和关系类型.AlSiaRED数据集共涵盖8226个句子,标注了9362个实体以及6876种关系,可以同时进行实体识别和关系抽取任务.

3.3　实验配置　本文提出的BoBGSAL-Net是一个基于Pytorch和DGL (Deep Graph Library)框架的模型,其中包含两层GCN网络结构,dropout的比率设置为0.6,学习率初始化为0.001.模型优化器采用AdamW,权重衰减为0.0001.

在词嵌入层层面,采用了三种不同的模型,包括BiLSTM,GloVe和BERT.其中,BiLSTM(256 d)和GloVe(100 d)用于词嵌入编码.基于BERT的词嵌入采用官方提供的BERT base和BERT large预训练模型,并将学习率初始化为$1e^{-5}$.

3.4　评估指标　使用$F1$作为评估指标之一.$F1$是精确率和召回率的加权几何平均值,是平衡准确率和召回率的综合指标.精确率、召回率和$F1$如式(15)~(17)所示:

$$precision=\frac{T}{C} \tag{15}$$

$$recall=\frac{T}{A} \tag{16}$$

$$F1=\frac{2\times precision\times recall}{precision+recall} \tag{17}$$

其中,T为一类实体被正确分类的实际个数,C表示被识别为这一类实体的样本总数,A为样本中的实体实例总数.

基于混淆矩阵衍生出另一个评估指标AUC(Area under Curve),即受试者工作特征曲线下的面积来评估分类模型的性能.评估指标的计算涉及混淆矩阵,主要通过对True Positive (TP), False Positive (FP), True Negative (TN)和False Negative (FN)四个参数进行计算.TP表示模型将样本预测为正例,并且实际标签也为正例,即模型预测正确的标签;FP表示模型将样本预测为正例,但是实际标签为负例,即模型预测错误的标签;TN表示模型将样本预测为负例,并且实际标签也为负例,即模型预测正确的标签;FN表示模型将样本预测为负例,但实际标签为正例,即模型预测错误的标签.

本文采用的评估指标包括$F1$,AUC,Ign $F1$以及Ign AUC.

3.5　基准模型　实验使用的基准模型主要完成实体识别和关系抽取两个任务.对于实体识别任务,选用多种经典模型进行对比,包括LSTM[5],BiLSTM[6],HIN-GloVe[7],CNN[12],Context-Aware[21],CFER-GloVe[27],SSAN-BERT-base[28]和GAIN+SIEF[29].这些模型在文本分类和实体关系抽取任务中表现出色,已被广泛应用于自然语言处理领域.对于关系抽取任务,选择HIN-

BERT-base[7], GCNN[8], LSR-GloVe[22], GAT[23], EOG[24], AGGCN[25], GAIN - GloVe[26], LSR+BERT-base[30]和CGM2IR-RoBERTa[31]作为基准模型. 其中,LSR+BERT-base模型在文档级实体关系抽取任务中具有较高的影响力,已成为该领域的重要研究方向.

总体上,本文实验选用多种经典和代表性模型,对后续研究具有重要的参考价值.

4 实验结果与分析

针对命名实体识别和关系抽取两个任务进行实验,并通过对BoBGSAL-Net模型在DocRED和AlSiaRED数据集上的多方面评估来进行模型性能的分析.

实验1:BoBGSAL-Net在DocRED数据集上的命名实体识别对比实验.

为了评估本文提出的BoBGSAL-Net模型在公开数据集上的命名实体识别性能,在公开数据集DocRED上与基准模型进行对比实验,实验结果如表3所示,表中黑体字表示结果最优. 由表可知,BoBGSAL-Net模型在DocRED数据集上的命名实体识别各项指标均优于基准模型,这可能是因为MMLG策略能够捕捉文档中不同实体间的复杂信息交互,同时ERG模块融合了路径推理机制,能够自动学习实体之间的多个关系路径,导致BoBGSAL-Net模型在DocRED数据集上表现有所提升.

表3 BoBGSAL-Net模型和其他模型在DocRED数据集上的命名实体识别实验结果的对比

Table 3 Experimental results of named entity recognition by BoBGSAL-Net and other models on the DocRED dataset

模型	验证集				测试	
	Ign $F1$	Ign AUC	$F1$	AUC	Ign $F1$	$F1$
CNN[12]	41.58%	36.85%	43.45%	39.39%	40.33%	42.26%
LSTM[5]	48.44%	46.62%	50.68%	49.48%	47.71%	50.07%
BiLSTM[6]	48.87%	47.61%	50.94%	50.26%	48.78%	51.06%
Context-Aware[21]	48.94%	47.22%	51.09%	50.17%	48.40%	50.70%
HIN-GloVe[7]	51.06%	—	52.95%	—	51.15%	53.30%
CFER-GloVe[27]	54.29%	—	55.31%	—	53.70%	54.06%
SSAN-BERT-base[28]	54.03%	—	54.95%	—	53.44%	53.16%
GAIN+SIEF[29]	53.82%	—	54.24%	—	53.87%	54.79%
BoBGSAL-Net	**54.33%**	**53.75%**	**55.84%**	**54.97%**	**54.14%**	**55.08%**

实验2:BoBGSAL-Net模型在AlSiaRED数据集上的命名实体识别对比实验.

对BoBGSAL-Net模型在铝硅合金材料实体识别任务上的性能进行了验证,并在AlSiaRED数据集上进行了命名实体识别实验. 实验结果如表4所示,表中黑体字表示结果最优.

由表可知,BoBGSAL-Net模型在AlSiaRED数据集上的表现优于基准模型,但和其在DocRED数据集上的表现相比,性能有所下降. 这可能是因为作者实验室构建的数据集包含更多的实体类型且文本长度较长,BoBGSAL-Net模型训练和推理的时间开销较大,导致性能指标的下降.

实验3:BoBGSAL-Net模型在DocRED数据集上的关系抽取对比实验.

为了评估BoBGSAL-Net模型在DocRED数据集上的关系抽取任务性能,进行了相应的对比实验,结果如表5所示,表中黑体字表示结果最优. 由表可知,BoBGSAL-Net模型在DocRED数据集上的关系抽取任务中,性能比基准模型更好,主要原因是该模型中的MMLG模块和ERG模块都具有针对实体之间关系信息的感知结构. 与GAT和GCNN相比,BoBGSAL-Net具有更好的全局上下文建模能力,能够更好地理解多个句子之间的实体关系. BoBGSAL-Net结合了图结构、实体关系路径推理和注意力机制,能够自动学习

表4　BoBGSAL-Net模型和其他模型在AlSiaRED数据集上的命名实体识别实验结果的对比

Table 4　Experimental results of named entity recognition by BoBGSAL-Net and other models on the AlSiaRED dataset

模型	验证集				测试	
	Ign $F1$	Ign AUC	$F1$	AUC	Ign $F1$	$F1$
CNN[12]	39.53%	31.47%	40.15%	32.44%	38.73%	39.20%
LSTM[5]	41.34%	40.43%	43.03%	41.09%	41.26%	42.97%
BiLSTM[6]	44.08%	43.65%	46.57%	45.13%	43.24%	45.16%
Context-Aware[21]	46.09%	45.36%	48.85%	47.33%	46.13%	48.17%
HIN-GloVe[7]	48.38%	—	50.35%	—	48.24%	50.18%
CFER-GloVe[27]	53.34%	—	54.27%	—	52.45%	53.60%
SSAN-BERT-base[28]	53.45%	—	53.25%	—	52.34%	53.27%
GAIN+SIEF[29]	**53.82%**	—	54.24%	—	**53.87%**	53.29%
BoBGSAL-Net	53.66%	**53.19%**	**55.39%**	**55.23%**	52.55%	**54.83%**

表5　BoBGSAL-Net模型和其他模型在DocRED数据集上的关系抽取实验结果的对比

Table 5　Experimental results of relation extraction by BoBGSAL-Net and other models on the DocRED dataset

模型	验证集				测试	
	Ign $F1$	Ign AUC	$F1$	AUC	Ign $F1$	$F1$
GAT[23]	45.17%	—	51.44%	—	47.36%	49.15%
GCNN[8]	46.22%	—	51.52%	—	49.59%	51.62%
EOG[24]	45.94%	—	52.15%	—	49.48%	51.82%
AGGCN[25]	46.29%	—	52.47%	—	48.89%	51.45%
LSR-GloVe[22]	48.82%	—	55.17%	—	52.15%	54.18%
GAIN-GloVe[26]	53.05%	52.57%	55.29%	55.44%	52.66%	55.08%
HIN-BERT-base[7]	54.29%	—	55.43%	—	53.70%	55.60%
LSR+BERT-base[30]	58.93%	—	60.89%	—	57.71%	59.94%
CGM2IR-RoBERTa[31]	62.03%	—	63.95%	—	61.96%	62.89%
BoBGSAL-Net	54.32%	53.47%	55.20%	54.43%	53.62%	54.57%
BoBGSAL-Net+GloVe	56.15%	54.39%	57.33%	57.63%	54.35%	56.97%
BoBGSAL-Net+BiLSTM	60.62%	58.27%	61.45%	59.72%	58.47%	60.54%
BoBGSAL-Net+BERT	**65.20%**	**64.47%**	**64.38%**	**64.58%**	**62.43%**	**65.32%**

实体之间的多个关系路径.与EOG和AGGCN相比,BoBGSAL-Net在捕捉实体之间的多层语义关系时表现更为突出.由表可知,BoBGSAL-Net模型在DocRED数据集上的性能不如LSR+BERT-base和CGM2IR-RoBERTa,这可能是因为BoBGSAL-Net具有更复杂的模型结构,导致在训练过程中需要更多的计算资源和参数调优,而不当的调优会影响性能.

此外,在引入词嵌入模型后,性能与BoBGSAL-Net相比,有显著提升,尤其在BoBGSAL-Net与BERT相结合的BoBGSAL-Net+BERT模型中,性能表现最为出色.可能因为BoBGSAL-Net+BERT模型将图结构与BERT的预训练语义表示相结合,从而更加充分地整合不同层次的信息.通过ERG模块的路径推理机制,该模型能够更准确地学习实体关系的多个关系路径,增强对复杂关系的抽取能力,使得该模型在关系抽取任务中表现出色.

实验4:BoBGSAL-Net模型在AlSiaRED数据集上的关系抽取对比实验.

为了评估BoBGSAL-Net模型在作者实验室构建的数据集上的关系抽取性能,设置该实验对

模型性能进行测试,实验结果如表6所示,表中黑体字表示结果最优. 由表可知,在AlSiaRED数据集上的关系抽取任务中,BoBGSAL-Net模型的性能和其他模型相比,提升更显著. 此外,BoBGSAL-Net模型结合了MMLG策略和BERT的全局上下文建模,能够更准确地捕捉整个文档的实体关系,在语义和语法更复杂以及长句子更多的AlSiaRED数据集中表现更好.

表6 BoBGSAL-Net模型和其他模型在AlSiaRED数据集上的关系抽取实验结果的对比

Table 6 Experimental results of relation extraction by BoBGSAL-Net and other models on the AlSiaRED dataset

模型	验证集				测试	
	Ign $F1$	Ign AUC	$F1$	AUC	Ign $F1$	$F1$
GAT[23]	46.33%	—	48.20%	—	45.54%	47.39%
GCNN[8]	48.46%	—	50.36%	—	47.85%	49.83%
EOG[24]	45.57%	—	46.91%	—	45.31%	46.32%
AGGCN[25]	49.19%	—	50.95%	—	48.89%	49.63%
LSR-GloVe[22]	51.35%	—	53.44%	—	51.27%	53.29%
GAIN-GloVe[26]	57.88%	56.47%	59.29%	57.89%	57.57%	59.14%
HIN-BERT-base[7]	53.62%	—	54.44%	—	52.56%	54.72%
LSR+BERT-base[30]	59.23%	—	61.47%	—	59.62%	60.20%
CGM2IR-RoBERTa[31]	63.53%	—	62.74%	—	63.38%	63.26%
BoBGSAL-Net	55.43%	54.64%	56.51%	55.78%	54.84%	55.73%
BoBGSAL-Net+GloVe	60.45%	56.47%	59.29%	57.89%	57.57%	59.14%
BoBGSAL-Net+BiLSTM	61.58%	59.73%	62.50%	60.48%	59.76%	61.48%
BoBGSAL-Net+BERT	**66.14%**	**65.59%**	**65.40%**	**65.32%**	**64.73%**	**66.04%**

实验5:BoBGSAL-Net模型在DocRED数据集上的实体抽取对比实验.

为了评估文档级实体抽取相对文档-句子-语言三级实体抽取在DocRED数据集上的实体抽取性能,本文设置该实验对模型性能进行测试,实验结果如表7所示,表中黑体字表示结果最优. 由表可知,BoBGSAL-Net+BERT模型在DocRED数据集上的实体抽取任务性能优于其他模型. 相较于文档-句子-语言三级实体抽取模型,BoBGSAL-Net+BERT模型不仅是将不同模块简单地串联起来,而且将图结构与语义表示紧密结合,使模型更深入地理解实体关系. 在文档级实体抽取中,BoBGSAL-Net综合考虑整个文档的语境,更好地理解实体的上下文关系,由于直接在文档级别进行抽取,相对于独立处理文档、句子和语言级别的模型,其整体处理速度可能更快. 这种整合性使得BoBGSAL-Net+BERT能够更好地理解文本中的复杂关系,提升了抽取质量.

表7 BoBGSAL-Net模型和其他模型在DocRED数据集上的实体抽取实验结果的对比

Table 7 Experimental results of entity extraction by BoBGSAL-Net and other model on the DocRED dataset

模型	验证集				测试	
	Ign $F1$	Ign AUC	$F1$	AUC	Ign $F1$	$F1$
DocRED-CNN[32]	40.27%	32.75%	43.35%	34.17%	36.44%	42.33%
MRN+BERT[33]	59.47%	—	60.20%	—	59.52%	61.74%
DRN-GloVe[34]	54.61%	—	56.49%	—	54.35%	56.33%
BoBGSAL-Net	55.43%	54.64%	56.51%	55.78%	54.84%	55.73%
BoBGSAL-Net+GloVe	60.45%	56.47%	59.29%	57.89%	57.57%	59.14%
BoBGSAL-Net+BiLSTM	61.58%	59.73%	62.50%	60.48%	59.76%	61.48%
BoBGSAL-Net+BERT	**66.14%**	**65.59%**	**65.40%**	**65.32%**	**64.73%**	**66.04%**

5　结论

本文提出一种文档级实体关系抽取方法,即基于双图结构的聚合逻辑网络BoBGSAL-Net.该方法首先构建一个MMLG模块,模拟整个文档中不同提及之间的复杂信息交互,提高模型对文档级实体关系的感知能力.其次,构建了ERG模块,该模块融合路径推理机制,主要针对实体间的多个关系路径进行推理学习,更准确地识别提及级节点实体及关系.

本文基于MMLG和ERG提出聚合逻辑推理路径以推断实体之间的关系,并进行分类预测.在公开数据集DocRED以及作者实验室构建的数据集AlSiaRED上进行对比实验,结果表明BoBGSAL-Net+BERT在文档级实体关系抽取任务中,性能优于其他所有模型,与CGM2IR-RoBERTa模型相比,$F1$指标提升2.66%,在文档级关系抽取任务中性能得到提升.

未来将探索并优化本文模型,进一步提高实体关系抽取性能.针对多语言文档的场景,通过跨语言模型迁移等技术实现对不同语言的文档级实体关系抽取,提高模型的通用性和可扩展性.

参考文献

[1] Yuan C S, Huang H Y, Feng C, et al. Document-level relation extraction with entity-selection attention. Information Sciences, 2021 (568): 163－174.

[2] Zhang Q Q, Chen M D, Liu L Z. A review on entity relation extraction//Proceedings of the 2nd International Conference on Mechanical, Control and Computer Engineering. Harbin, China: IEEE, 2017: 178－183.

[3] Li Z H, Yang Z H, Xiang Y, et al. Exploiting sequence labeling framework to extract document-level relations from biomedical texts. BMC Bioinformatics, 2020, 21(1): 125.

[4] Han X Y, Wang L. A novel document-level relation extraction method based on BERT and entity information. IEEE Access, 2020(8): 96912－96919.

[5] Geng Z Q, Chen G F, Han Y M, et al. Semantic relation extraction using sequential and tree-structured LSTM with attention. Information Sciences, 2020(509): 183－192.

[6] Luo L, Yang Z H, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 2018, 34(8): 1381－1388.

[7] Tang H Z, Cao Y N, Zhang Z Y, et al. HIN: Hierarchical inference network for document-level relation extraction//Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2020: 197－209.

[8] Najibi M, Rastegari M, Davis L S. G-CNN: An iterative grid based object detector//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 2369－2377.

[9] Gu J X, Wang Z H, Kuen J, et al. Recent advances in convolutional neural networks. Pattern Recognition, 2018(77): 354－377.

[10] Li Z W, Liu F, Yang W J, et al. A survey of convolutional neural networks: Analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 6999－7019.

[11] O'Shea K, Nash R. An introduction to convolutional neural networks. 2015, arXiv: 1511.08458.

[12] Lavin A, Gray S. Fast algorithms for convolutional neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 4013－4021.

[13] Huang J W, Abadi D J. Leopard: Lightweight edge-oriented partitioning and replication for dynamic graphs. Proceedings of the VLDB Endowment, 2016, 9(7): 540－551.

[14] 刘英莉,吴瑞刚,么长慧,等.铝硅合金实体关系抽取数据集的构建方法.浙江大学学报(工学版), 2022, 56(2): 245－253. (Liu Y L, Wu R G, Yao C H, et al. Construction method of extraction dataset of Al-Si alloy entity relationship. Journal of Zhejiang University (Engineering Science), 2022, 56(2): 245－253.)

[15] Sheng D M, Wang D, Shen Y, et al. Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion

recognition//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: International Committee on Computational Linguistics, 2020: 4153－4163.

[16] Auten A, Tomei M, Kumar R. Hardware acceleration of graph neural networks//Proceedings of 2020 57th ACM/IEEE Design Automation Conference (DAC). San Francisco, CA, USA: IEEE, 2020: 1－6.

[17] Abadal S, Jain A, Guirado R, et al. Computing graph neural networks: A survey from algorithms to accelerators. ACM Computing Surveys, 2022, 54 (9): 191.

[18] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL, 2014: 1532－1543.

[19] Tanvir R, Shawon T R, Mehedi H K, et al. A GAN - BERT based approach for bengali text classification with a few labeled examples// Proceedings of the 19th International Symposium on Distributed Computing and Artificial Intelligence. Springer Berlin Heidelberg, 2022: 20－30.

[20] Niu Z Y, Zhong G Q, Yu H. A review on the attention mechanism of deep learning. Neuro - computing, 2021(452): 48－62.

[21] Harter A, Hopper A, Steggles P, et al. The anatomy of a context-aware application. Wireless Networks, 2002, 8(2－3): 187－197.

[22] Mrityunjay K, Ravindra G. Learning to fingerprint the latent structure in question articulation//2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Orlando, FL, USA: IEEE, 2018: 73－80.

[23] Veličković, Cucurull G, Casanova A, et al. Graph attention networks. 2017, arXiv: 1710.10903.

[24] Chen L, Tian F L. Skew-rank of an oriented graph with edge - disjoint cycles. Linear and Multilinear Algebra, 2016, 64(6): 1197－1206.

[25] Li Z X, Sun Y R, Zhu J W, et al. Improve relation extraction with dual attention - guided graph convolutional networks. Neural Computing and Applications, 2021, 33(6): 1773－1784.

[26] Zeng S, Xu R, Chang B, et al. Double graph based reasoning for document - level relation extraction. 2020, arXiv: 2009.13752.

[27] Dai D M, Ren J, Zeng S, et al. Coarse-to-fine entity representations for document - level relation extraction. 2020, arXiv: 2012.02507.

[28] Xu B F, Wang Q, Lyu Y J, et al. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction// Proceedings of the 35th AAAI Conference on Artificial Intelligence. Online: AAAI Press, 2021: 14149－14157.

[29] Xu W, Chen K H, Mou L L, et al. Document-level relation extraction with sentences importance estimation and focusing//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, WA, United States: ACL, 2022: 2920－2929.

[30] Nan G S, Guo Z J, Sekuli I, et al. Reasoning with latent structure refinement for document - level relation extraction//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020: 1546－1557, DOI: 10.18653/v1/2020.acl-main.141.

[31] Zhao C, Zeng D J, Xu L, et al. Document-level relation extraction with context guided mention integration and inter-pair reasoning.2022, arXiv: 2201.04826.

[32] Yao Y, Ye D M, Li P, et al. DocRED: A large-scale document - level relation extraction dataset// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 764－777, DOI: 10.18653/v1/P19 -1074.

[33] Li J Y, Xu K, Li F, et al. MRN: A locally and globally mention - based reasoning network for document - level relation extraction//Proceedings of the Findings of the Association for Computational Linguistics. Online: ACL, 2021: 1359－1370.

[34] Xu W, Chen K H, Zhao T J. Discriminative reasoning for document - level relation extraction// Proceedings of the Findings of the Association for Computational Linguistics. ACL, 2021: 1653－1663, DOI: 10.18653/v1/2021.findings-acl.144.

（责任编辑 杨可盛 高善露）