

融合知识图谱的领域事件抽取方法

方 睿, 崔良中

(海军工程大学, 湖北 武汉 430033)

摘 要: 特定领域中专业术语的密集使用增加了理解和抽取事件要素的难度。为此, 提出一种融合知识图谱与预训练模型的事件抽取方法 KG-PMEE。在装备这一特定领域上, 将知识图谱中的实体信息嵌入到模型训练过程中, 以增强预训练模型对专业术语和上下文信息的理解能力, 可提升模型在目标领域的适应性。实验表明, 所提方法在装备领域数据集上的 F1 值相较于基线模型 ERNIE 提升 2.97%, 在公开数据集 ACE2005 上的性能也得到了提升, 验证了引入知识图谱信息对提升领域事件抽取效果的有效性。

关键词: 事件抽取; 知识图谱; 预训练模型; 特定领域; 深度学习

DOI: 10.11907/rjdk.241655

中图分类号: TP301

文献标识码: A

开放科学(资源服务)标识码(OSID):

文章编号: 1672-7800(2025)009-0062-08



An Event Extraction Method in Specific Domains Combining Knowledge Graph

FANG Rui, CUI Liangzhong

(Naval University of Engineering, Wuhan 430033, China)

Abstract: Intensive use of specialized terminology in specific fields increases the difficulty of understanding and extracting event elements. Therefore, a event extraction method KG-PMEE that integrates knowledge graph and pre trained model is proposed. In the specific field of equipment, embedding entity information from the knowledge graph into the model training process enhances the pre trained model's understanding of professional terminology and contextual information, and improves the model's adaptability in the target domain. The experiment shows that the proposed method has an F1 value improvement of 2.97% compared to the baseline model ERNIE on the equipment domain dataset, and its performance has also been improved on the publicly available dataset ACE2005, verifying the effectiveness of introducing knowledge graph information in improving domain event extraction performance.

Key Words: event extraction; knowledge graph; pre-trained model; specific domain; deep learning

0 引言

事件抽取作为信息抽取的一个重要分支,旨在从文本中识别出特定类型的事件,并抽取与事件相关的要素,例如事件触发词、参与者、时间、地点等,对理解和分析特定领域的文本信息具有重要意义,在金融、法律、生物等众多领域具有广泛的应用前景^[1-3]。

然而,特定领域的事件抽取也面临着诸多挑战。在金融领域,万齐智等^[4]针对事件抽取成分缺失、语义出入和放大等问题,利用财经领域的专业词汇、新闻文本表达结构等特点,设计了基于句法语义依存图金融事件抽取模

型;在医疗领域,针对医药不良事件数据量有限的问题时,El-allaly等^[5]使用了迁移学习的方法;在法律领域,Shen等^[6]为了解决论元分类中存在的远距离语义依赖和共指消解问题,提出一种新颖的踏板注意力机制,通过相关的邻近词汇提取两个词之间的语义关系。由此可见,专业术语的复杂性、事件要素的歧义性、语料稀缺、语境依赖等问题,为领域事件抽取带来了一定的限制。

传统的基于特征和机器学习的事件抽取方法,虽然在一定程度上能处理各种文本输入和特征,但受到数据稀疏性、特征工程复杂性及模型泛化能力的制约。随着深度学习技术的快速发展,BERT、GPT等预训练模型展现了强大的文本表示能力和泛化性能,为特定领域的事件抽取提供

收稿日期:2024-08-05

扫描二维码阅读全文:

基金项目:装备预先研究项目(30209040702)

作者简介:方睿(2000-),男,CCF学生会会员,海军工程大学硕士研究生,研究方向为自然语言处理、知识图谱等;崔良中(1979-),男,博士,海军工程大学副教授、硕士生导师,研究方向为知识图谱、数据挖掘与分析等。



了新思路。然而,基于通用域语料训练的大模型虽包含丰富的常识知识,却缺乏垂直领域的专业知识,因此在特定领域单纯依赖预训练模型进行事件抽取时仍存在提升空间,存在缺乏领域知识的指导、对专业术语和上下文信息的理解不足等问题,在一定程度上限制了其在特定领域事件抽取任务中的性能。

为此,本文提出一种结合知识图谱信息与预训练模型的领域事件抽取方法,在装备领域数据上进行实验,主要贡献为:①构建装备领域知识图谱,将知识图谱中的丰富知识整合到预训练模型中,以增强模型对特定领域知识的理解和表示能力;②基于融合后的知识增强表示,设计了一个知识引导的事件抽取框架,利用知识图谱中的信息来指导事件抽取过程,提升模型准确性和鲁棒性。

1 相关工作

传统事件抽取方法包括模式匹配、机器学习方法。基于深度学习的事件抽取方法通过神经网络自动学习文本表示和特征,从而提升事件抽取的性能,主要包括卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)、图神经网络(Graph Convolutional Network, GNN)等^[7-14]。

近年来,BERT、RoBERTa、GPT等预训练语言模型的出现,为事件抽取提供了新方法。预训练模型在大规模无标注文本上进行预训练,以学习丰富的语言知识,包括词汇上下文表示、句法结构和语义关系等,可作为事件抽取任务中特征提取的有力工具。例如,张洪宽等^[15]选择预训练语言模型BERT对文档编码,提出了关于金融领域事件

抽取的端到端联合学习框架,通过BERT的强大表征能力简化了事件抽取任务的复杂度,显著提升了抽取效果。虽然,基于通用领域语料训练的大模型包含丰富的常识,但缺少垂直领域的专业知识,因此在处理特定领域的任务时具有一定的局限性。

为了解决该问题,研究者们将外部知识与预训练模型相结合,以增强模型在特定领域的适应性。裴炳森等^[16]通过微调技术和少量领域数据,构建了适应公安专业的知识抽取方法。Zhuang等^[17]融合事件本体的语义信息来增强事件表示,通过注意力机制增强关键知识的注入,有效缓解了知识缺失问题。PMC-LLaMA、Fin-BERT等模型融入领域知识,增强了在生物医学和金融领域的处理能力^[1,18]。此外,ERNIE、K-BERT、KEPLER、CoLAKE、JACKET、KG-Adapter等模型,都在预训练模型中引入外部知识,提升了模型的适应性和编码能力^[19-24]。本文提出的KG-PMEE(Knowledge Graph and Pre-Trained Model Based Event Extraction)方法将知识图谱中的实体信息嵌入模型训练过程,以增强预训练模型对特定领域知识的理解和表示能力,提升了模型在目标领域的适应性。

2 本文方法

KG-PMEE方法的整体结构如图1所示。

(1)知识融合层。将知识图谱中的信息与输入文本相结合,先识别文本中的实体再将其与知识图谱中的实体关联,构建出一个包含领域背景知识的句子树。

(2)嵌入表示层。该层包括嵌入表示和可见矩阵两个部分,嵌入表示是指将句子树转换为嵌入向量序列,并应

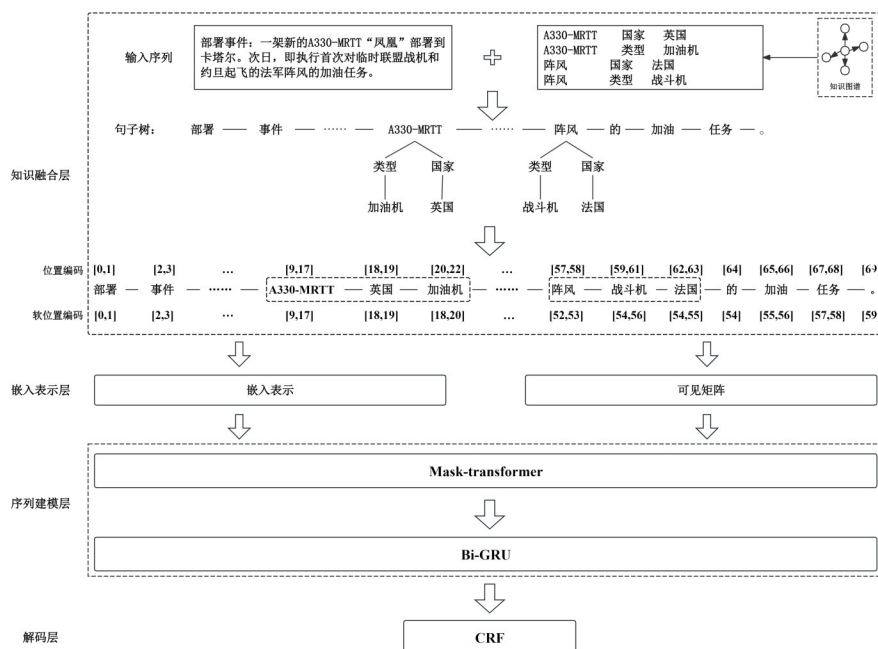


Fig. 1 Specific structure of KG-PMEE method

图1 KG-PMEE方法具体结构

用软位置编码来保持句子的原始结构信息,确保模型能处理非线性结构的数据;可见矩阵用来控制词与词之间的注意力连接,确保模型只关注同一枝干的上下文信息,从而减少知识噪声并提升抽取的准确性。

(3)序列建模层。Mask-transformer中的掩码机制可让模型处理不同长度的序列,防止模型在自注意力计算时考虑填充的部分,确保其只关注实际的文本数据。随后,通过Bi-GRU层进一步捕获序列数据中的深层次时间和上下文依赖关系。

(4)解码层。CRF层将序列建模层的输出转换为最终的标签序列,通过定义标签间的转移概率来优化整个序列的标签分配,从而提升标注准确性。

2.1 知识融合层

知识融合层的主要作用是将知识图谱中的信息与输入文本相结合,包括两个步骤:

(1)对于一段输入文本 s ,通过分词确定每个词的边界得到句子序列 $s = \{c_0, c_1, \dots, c_n\}$,知识图谱 K 中的三元组表示为 $\varepsilon = (e_i, r_i, e_k)$ 。其中, e_i, e_k 代表实体, r_i 代表实体间的关系。在知识图谱 K 中找到句子序列 s 中所有实体相关的实体信息,得到对应的三元组集合 E :

$$E = \{(c_i, r_{i0}, e_{i0}), (c_i, r_{i1}, e_{i1}), \dots, (c_i, r_{ik}, e_{ik})\} \quad (1)$$

(2)将查询到的实体信息 E 注入到句子序列 s 中,构建

一个包含原始文本信息和相关背景知识的句子树 sen_tree ,以丰富句子含义。本文通过收集数据构建了一个包含5 800个装备实体的知识图谱,以“实体—关系—实体”三元组的形式存储,如图2所示。

P-3猎户座	国家	美国
P-3猎户座	类型	反潜机
苏-27战斗机	国家	苏俄
苏-27战斗机	类型	战斗机
LCA	国家	印度
LCA	类型	战斗机
直-9	国家	中国
直-9	类型	直升机
SH-60海鹰直升机	国家	美国
SH-60海鹰直升机	类型	直升机

Fig. 2 Equipment knowledge graph triplet

图2 装备知识图谱三元组

$$sen_tree = \{c_0, c_1, \dots, c_i\{(r_{i0}, e_{i0}), \dots, (r_{ik}, e_{ik})\}, \dots, c_n\} \quad (2)$$

对于输入文本“部署事件:一架新的A330-MRTT凤凰部署到卡塔尔。次日,执行首次对临时联盟战机和约旦起飞的法军阵风的加油任务”。

首先,经过分词得到一个句子序列,序列中包含有“A330-MRTT”“阵风”等实体;其次,在装备知识图谱中进行查询,获取与该实体相关的信息,例如“A330-MRTT——类型——加油机”;最后,将查询到的实体信息与文本中的实体进行关联,整合成一个包含领域背景知识的句子树,如图3所示。

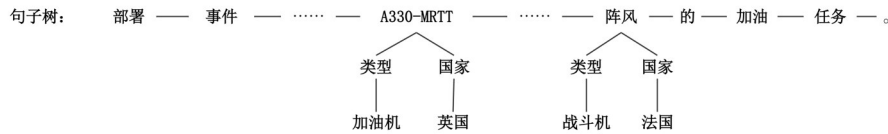


Fig. 3 Sentence tree

图3 句子树

2.2 嵌入表示层

2.2.1 嵌入表示

嵌入表示将句子树转换成向量序列,同时保持句子原本的结构信息,包括词嵌入(Token Embeddings)、段嵌入(Segment Embeddings)和软位置嵌入(Soft-position Embeddings),如图4所示。其中,词嵌入指将文本中的每个字符或词元转换成固定大小的向量;段嵌入用于区分文本中的不同部分,通常为每个部分添加一个不同的嵌入向量来实现,特别是在处理包含多个句子或区分上下文文本时非常

有用。

图4(彩图扫OSID码可见,下同)中只有一个句子,因此将Segment Embedding标记为A。软位置嵌入是指将句子树转换为模型可处理的嵌入向量序列,但利用传统的序列化方法将树状结构的句子树平铺成序列,可能会丢失结构信息,因此使用软位置编码,允许模型在序列化的过程中保留句子树的结构信息。图4中红色数字标记部分为软位置编码,黑色数字标记部分为原本的位置编码。

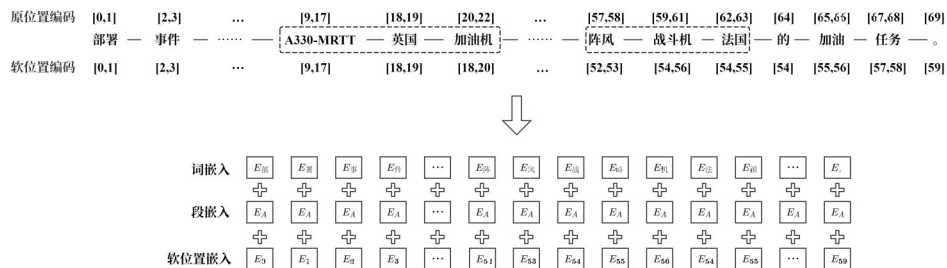


Fig. 4 Embedding representation

图4 嵌入表示

2.2.2 可见矩阵

可见矩阵用于控制模型在处理句子时词与词之间的注意力, 确保模型只关注与当前词相关联的实体, 而非整个句子中的所有词, 如式(3)所示。

$$M_{ij} = \begin{cases} 0 & w_i, w_j \text{ 在同一分支上} \\ \infty & w_i, w_j \text{ 不在同一分支上} \end{cases} \quad (3)$$

对于一个句子而言, 其可见区域包括: 主干上的词之间可见、交叉节点上的词与其分支上的词和主干上的词可见、分支上的词只与交叉节点和对应分支上的词可见。图5中, 深色区域代表词与词之间可见; 浅色区域代表词与词之间不可见。由此可见, 知识图谱中“加油机”“英国”这两个实体只对句子中“A330-MRRT”这个实体产生影响, 对其他词不产生作用, 从而限制了注意力机制的范围, 减少了知识噪声。

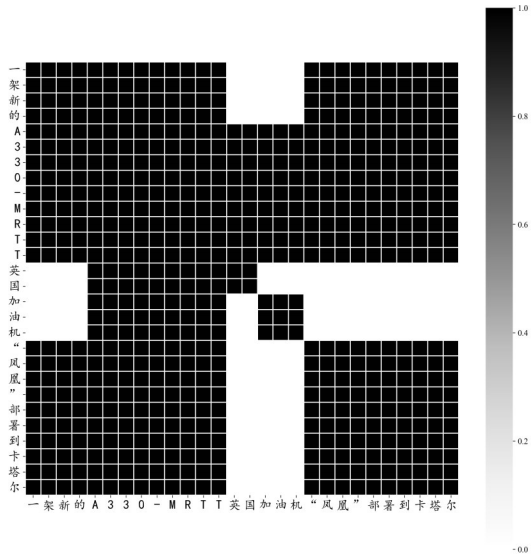


Fig. 5 Visible matrix

图5 可见矩阵

2.3 序列建模层

2.3.1 Mask-Transformer

掩码转换层(Mask-Transformer)的主要功能是将知识图谱中的结构信息整合到语言表示中, 防止因知识注入引起的语义变化。其中, 最关键的结构是掩码—自注意力模块, 结合可见矩阵 M 与自注意力机制可限制自注意力区域。具体实现如下:

(1) 线性变换。对输入的查询(Query)、键(Key)和值(Value)进行线性变换。

$$Q = W^Q X, K = W^K X, V = W^V X \quad (4)$$

式中: X 为输入序列; W^Q, W^K, W^V 为可学习的权重矩阵。

(2) 头分割。将变换后的 Q, K, V 分割成多个头。

$$Q_i = \text{Split}(Q, i), K_i = \text{Split}(K, i), V_i = \text{Split}(V, i) \quad (5)$$

式中: i 为头的索引。

(3) 计算每个头的注意力。

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_{\text{head}}}} + M\right) V_i \quad (6)$$

式中: d_{head} 为头的大小; M 代表上一层得到的可见矩阵, 用来限制注意力范围。

(4) 拼接和输出。

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \quad (7)$$

式中: h 为头数量; W^O 为输出权重矩阵。

掩码转换层由多个掩码—自注意力模块组成, 每个模块都使用可见矩阵 M 来控制自注意区域, 使模型在处理知识注入时能更精确地控制符号间的相互作用。通过限制自注意区域, 掩码转换层能有效防止因知识注入引起的语义变化, 保证模型在利用知识图谱中的信息时不会引入错误的语义关联, 从而保持语言表示的准确性和一致性。

2.3.2 GRU 层

GRU 是长短期记忆递归神经网络(Long Short Term Memory, LSTM)的一种变体, 只保留了重置门和更新门, 如图6所示。

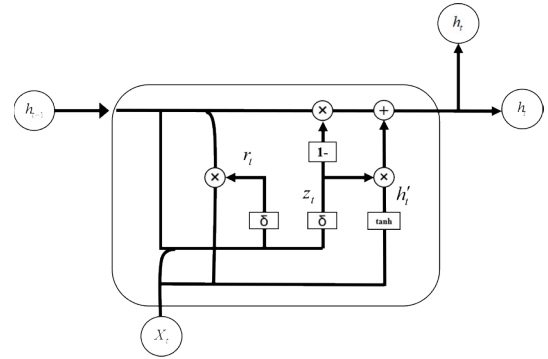


Fig. 6 GRU structure

图6 GRU结构

更新门在功能上相当于LSTM中输入门与遗忘门的结合, 决定了在多大程度上将旧的状态信息与新信息结合, 以更新隐藏状态。GRU内部结构的计算公式如式(8)所示。

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\ \tilde{h}_t &= \tanh(W_h \cdot [r_t * h_{t-1}, x_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \quad (8)$$

式中: x_t 表示当前时刻的输入信息; h_{t-1} 表示上一时刻的隐藏状态, 隐藏状态包含之前节点的数据信息; h_t 表示传递到下一时刻的隐藏状态; $\tilde{h}_t$ 表示候选隐藏状态; r_t 表示重置门; z_t 表示更新门; σ 代表 sigmoid 函数; $\tanh$ 函数将数据归一化到 $[-1, 1]$ 范围。

2.4 解码层

CRF解码层的作用是对Bi-GRU层输出的特征序列进行全局标签推断, 以便在整个序列上进行一致性建模, 还考虑了标签之间的依赖关系, 能更好地捕捉序列标注任务中的局部和全局信息。

首先, 经过Bi-GRU输出的序列为 $X = \{h_1, h_2, \dots, h_t\}$,

每个输入值在各位置的得分函数如式(9)所示;其次,使用 Softmax 函数进行归一化,计算标签序列 y 的概率值,如式(10)所示。

$$S(h, y) = \sum_{i=1}^N P_{i, y_i} + \sum_{i=1}^N R_{y_i, y_{i+1}} \quad (9)$$

$$P(y|h) = \frac{\exp[S(h, y)]}{\sum_{y' \in Y(h)} \exp[S(h, y')]} \quad (10)$$

式中: N 为序列总长度; P_{i, y_i} 表示序列中第 i 个序列输出标签为 y_i 的概率值; $R_{y_i, y_{i+1}}$ 表示字符标签 y_i 到字符标签 y_{i+1} 的转移概率; $Y(h)$ 表示所有可能的标签序列集合。

3 实验结果与分析

3.1 数据集

本文基于 CCKS 2022 开源的多模态军事装备抽取任务中的数据,以构建数据集训练、验证模型,并在公开数据集 ACE2005 上进行测试,具体信息如表 1 所示。装备数据集主要来自于军事领域的新闻文本数据,包含发起者、承受者、使用器械、时间、地点 5 类事件要素。其中,ACE2005 是语言数据联盟(LDC)发布的由实体、关系和事件注释组成的各种类型数据,包括英语、阿拉伯语和中文数据。本文主要使用中文数据,包含生活、移动、事务、商业、冲突、交流、个人职业、司法相关的 8 类事件及 33 种子类别。

Table 1 Datasets information

表 1 数据集信息

数据集	类型	训练集/k	验证集/k	测试集/k
装备数据集	句子	1.60	0.20	0.20
	字符	173.60	16.50	16.00
ACE2005	句子	6.00	0.72	0.73
	字符	260.00	30.00	33.00

3.2 评估指标

本文选择精确度(Precision)、召回率(Recall)和 F1 度量(F1 Score)作为模型评估指标,计算公式如式(11)~(13)所示。精确度是指在所有被模型预测为正类的样本,真正为正类样本所占的比例,以衡量模型预测为正类的可靠性;精确度高的模型意味着其预测为正类的样本中大部分确实为正类;召回率也称为真正率(True Positive Rate, TPR)或灵敏度(Sensitivity),是指在所有实际为正类的样本中,被模型正确预测为正类样本的比例,召回率高意味着模型能找到大部分的正类样本;F1 度量是精确度与召回率的调和平均,以便在精确度与召回率之间进行权衡,反映模型的综合效果,F1 度量高代表模型在精确度与召回率上表现较好。

$$\text{Precision} = \frac{TP}{TP + FP} \quad (11)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (12)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (13)$$

式中:TP(True Positives)表示真正例,即模型预测为正类且实际为正类的样本数;FP(False Positives)表示假正例,即模型预测为正类但实际为负类的样本数;TN(True Negatives)表示真负例,即模型预测为负类且实际为负类的样本数;FN(False Negatives)表示假负例,即模型预测为负类但实际为正类的样本数。

3.3 实验设计与基线方法

本文从两个方面进行实验,验证 KG-PMEE 方法的效果。第一部分设置比较实验,首先,在装备数据集和 ACE2005 数据集上,对 KG-PMEE 与基线方法进行训练和测试;其次,记录、比较不同方法在各数据集上的精确率、召回率和 F1 值;最后,分析所提方法相较于基线方法的优势及原因。第二部分设计消融实验,首先针对本文方法中的知识图谱和可见矩阵设置消融实验,在装备数据集上分别对完整模型和消融模型进行训练和测试;其次,记录、比较不同模型的精确率、召回率和 F1 值;最后,分析知识图谱和可见矩阵对事件抽取性能的影响。

基线方法的选择思路为:融合知识图谱与预训练模型,旨在引入外部知识来增强预训练模型的表示能力,属于知识增强型模型。因此,本文将以下模型作为基线模型:①LSTM-CRF 结合了 LSTM 在处理序列数据方面的能力和 CRF 在标注序列时的优势,是事件抽取任务中常用的模型;②FLAT-Lattice、Soft-Lexicon 是引入词汇信息来增强特征表示的两种模型,可结合预训练模型,其中 FLAT-Lattice 是基于 transformer-encoder 的序列标注模型,通过一种新的位置编码使字符和词汇信息交互更直接,FLAT 扁平结构能支持并行计算,能更好地建模序列的长期依赖关系,Soft-Lexicon 是一种通过词汇分布式表示来增强模型语义理解能力的方法,可在不依赖预定义词汇表的情况下增强模型捕获词汇的灵活性和语义信息^[25,26];③PaddleNLP 是百度飞桨(paddlepaddle)深度学习框架中的一个子模块,用于自然语言处理任务,提供了许多预训练模型和工具,在事件抽取等任务中应用方便;④ERNIE、CK-BERT、LERT 均引入了知识的预训练模型,ERNIE 模型在预训练阶段引入大量知识图谱中的实体、关系信息,CK-BERT 模型融合了语言学信息和知识图谱中的事实知识,LERT 模型在词性标注、命名实体识别和句法依赖关系这 3 种语言特征上进行训练^[27-29]。

上述模型中,LSTM-CRF 可作为基础模型;PaddleNLP 作为 NLP 工具集,既包含先进的算法模型,又注重模型的产业应用,具备较好的灵活性和扩展性,适合作为比较模型来验证模型整体效果和运行效率;其余模型均为知识增强型模型,可作为比较模型验证本文方法在知识融合方面的效果。

3.4 实验环境与参数

本文实验环境、具体参数如表 2、表 3 所示。

Table 2 Experimental environment	
表 2 实验环境	
实验环境	配置
操作系统	Windows 10
CPU	Intel(R) Core(TM) i7-8700K
内存/GB	16
显卡	NVIDIA GeForce RTX 3090
显存/GB	24
开发环境	Python 3.9
开发框架	Pytorch 1.13.0+cu116

Table 3 Hyper-parameter settings	
表 3 超参数设置	
参数	数值
embedding size	768
hidden size	768
learning rate	2e-5
batch_size	10
dropout	0.1
epoch	50

3.5 实验结果

3.5.1 超参数分析

由于实验设备的性能限制,本文选择 batch_size=8、16、32 来分析不同批次大小对模型性能的影响,实验结果如图 7 所示。由此可见,随着 batch_size 增加,模型 F1 值逐渐减小,表明在当前实验条件下,较小的 batch_size 更有利于模型学习,能更好地捕捉数据的细节特征;较大的 batch_size 可能会导致模型发生过拟合,即模型在训练集上表现良好,但在测试集上表现较差。此外,较大的 batch_size 需要更多内存资源,影响了模型训练过程,因此在后续实验中选择 batch_size=10。

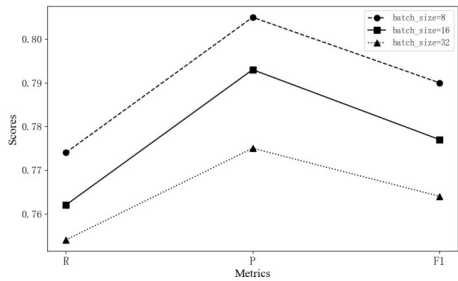


Fig. 7 Performance indicators of different batch_size

图 7 不同 batch_size 的性能指标

3.5.2 比较实验

本文在装备领域数据集、ACE2005 数据集上分别进行比较实验,实验结果如表 4、表 5 所示。其中,在领域数据集上使用装备知识图谱信息,ACE2005 数据集由于事件类型较多,因此使用通用知识图谱 HowNet。表 6 为领域数据集上各事件要素的抽取结果。图 8、图 9 为本文模型在领域数据集上训练得到的指标变化情况。

从整体性能上看,KG-PMEE 方法在指标上均优于其他基线方法,说明引入知识图谱可在一定程度上增强事件论元的识别效果。在特征表示层面,LSTM-CRF 主要依赖

Table 4 Experimental results on domain datasets			
表 4 在领域数据集上的实验结果 (%)			
模型	精确率(P)	召回率(R)	F1 值
LSTM-CRF	67.39	65.96	66.67
FLAT-Lattice	71.01	59.96	65.02
Soft-Lexicon	79.66	68.71	73.78
PaddleNLP	75.58	74.57	75.07
ERNIE	81.92	73.78	77.64
CK-BERT	80.85	68.71	74.29
LERT	80.97	73.78	77.21
KG-PMEE	82.04	79.53	80.61

Table 5 Experimental results on the ACE2005 dataset			
表 5 在 ACE2005 数据集上的实验结果 (%)			
模型	精确率(P)	召回率(R)	F1 值
LSTM-CRF	42.14	33.23	37.16
FLAT-Lattice	45.61	33.28	38.48
Soft-Lexicon	60.22	50.08	54.68
PaddleNLP	44.79	55.47	49.56
ERNIE	52.28	52.68	52.48
CK-BERT	56.64	51.61	54.01
LERT	51.24	53.60	52.40
KG-PMEE	60.60	68.90	64.50

Table 6 Experimental results of event elements in domain datasets			
表 6 领域数据集事件要素的实验结果 (%)			
事件要素	精确率(P)	召回率(R)	F1 值
时间	91.30	85.71	88.42
地点	83.91	73.00	78.07
装备	70.83	80.95	75.56
发起者	89.58	86.43	87.98
承受者	74.58	71.54	73.03

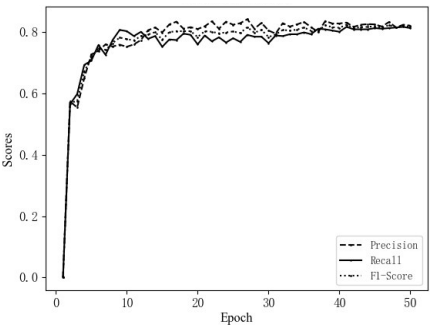


Fig. 8 Performance metrics on domain dataset

图 8 领域数据集上性能指标

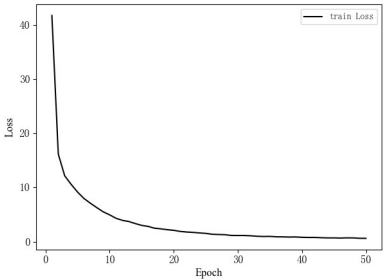


Fig. 9 Average loss

图 9 平均损失值

模型学习到的语义特征;FLAT-Lattice、Soft-Lexicon 通过词汇信息来增强特征表示;PaddleNLP 使用预训练模型提供的特征表示;ERNIE、CK-BERT、LERT 则通过预训练阶

段的知识融入,生成更丰富、精确的文本表示。在知识利用层面,LSTM-CRF没有显式地利用外部知识;FLAT-Lattice、Soft-Lexicon利用了词汇信息;PaddleNLP使用了预训练模型中蕴含的背景知识;ERNIE、CK-BERT利用了外部的知识图谱信息;LERT未直接引入外部知识图谱,但通过训练过程中关注的语言特征,间接利用了文本中的知识。

分析发现,LSTM-CRF主要关注文本的序列信息,难以捕捉事件元素之间的复杂关系;FLAT-Lattice、Soft-Lexicon虽通过词汇信息增强特征表示,但对知识的利用程度有限,主要关注词汇的语义信息,对文本的结构化信息利用不足;PaddleNLP使用预训练模型提供的特征表示,ERNIE、CK-BERT、LERT是知识增强型预训练模型,主要学习的是通用语言知识,对特定领域的专业术语和领域知识的理解程度有限,效果略差于KG-PMEE模型。

KG-PMEE在利用预训练模型提供丰富的语言知识和文本表示能力的基础上,结合知识图谱信息与文本信息,相较于其他知识增强型的预训练模型能根据文本中的实体,有针对性地引入相关知识图谱信息,从而构建包含领域背景知识的句子树,并使用软位置编码和可见矩阵等技术利用结构化信息,提升模型对文本的理解能力,因此在事件抽取任务中的效果优于基线模型。

由表4、表5可知,各模型在特定领域数据集上的表现优于ACE2005数据集,反映了事件类型集中性和领域知识依赖性对事件抽取效果具有积极作用。在特定领域数据集上,模型可更容易学习到特定领域的事件模式和特征,性能更好;在ACE2005数据集上,由于事件类型多样且涉及多个领域知识和概念,模型难以适应、捕捉所有可能的事件模式和特征。KG-PMEE模型结合了知识图谱与预训练模型的优势,能有效利用领域知识来增强模型的语义理解能力,从而在面对不同事件类型时具有一定的鲁棒性,因此在特定领域上的效果更优。

3.5.3 训练时间

图10为不同模型在领域数据集上训练一轮的所需时间。由此可见,耗时较短的是FLAT-Lattice、PaddleNLP模型;其次是KG-PMEE模型;耗时最长的是LSTM-CRF模型。原因为:FLAT-Lattice的扁平化结构使模型可并行计算,训练时间显著降低;PaddleNLP作为一个自然语言处理工具,其模型架构优化较好,训练更高效;Soft-Lexicon结合了词汇特征,增加了额外的计算量,时间相对较长;ERNIE、CK-BERT、LERT基于BERT,参数量较大、训练时间也显著增加;LSTM-CRF虽然参数相对较少,但针对长序列时的计算复杂度较高,因此在所有模型中训练耗时最长;KG-PMEE结合了知识图谱,额外增加了知识查询和注入步骤,但总体时间相对可接受。KG-PMEE并非简单地将整个知识图谱嵌入模型,而根据文本中的实体提及,有针对性地查询、引入相关知识图谱信息,避免了处理整个知识图谱的庞大计算量,从而保持了一定的计算效率。

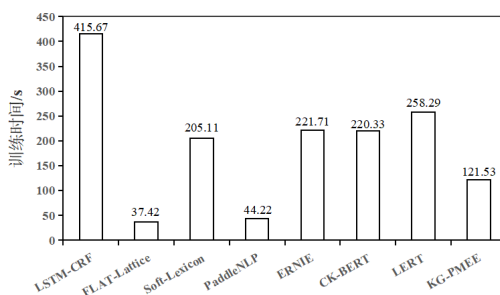


Fig. 10 Training time

图10 训练时间

3.6 消融实验

3.6.1 知识图谱

为了验证知识图谱信息对事件抽取性能的影响,消融实验结果如表7所示。其中,KG-PMEE(none)作为基线模型,不引入任何知识图谱信息;KG-PMEE(HowNet)代表引入HowNet知识图谱信息的模型;KG-PMEE(CnDbpedia)代表引入CnDbpedia知识图谱信息的模型;KG-PMEE(equipment)代表引入装备领域知识图谱信息的模型。

Table 7 Ablation experiment results of knowledge graph

表7 知识图谱消融实验结果 (%)

模型	精确率(P)	召回率(R)	F1值
KG-PMEE(none)	78.30	75.60	76.90
KG-PMEE(HowNet)	82.40	78.30	80.30
KG-PMEE(CnDbpedia)	82.40	78.00	80.20
KG-PMEE(equipment)	83.10	81.70	82.40

由表7可见,引入外部知识图谱能在一定程度上提升KG-PMEE模型性能;HowNet、CnDbpedia作为通用知识图谱,对提升模型性能有一定的帮助,但效果存在差异;特定领域的知识图谱对模型性能提升最显著,表明领域特定的知识图谱能为模型提供更准确的信息,从而在事件抽取任务中取得更好的效果。

3.6.2 可见矩阵

为了验证可见矩阵对模型性能的影响,消融实验结果如表8所示。其中,no_vm表示不使用可见矩阵引导的注意力机制;use_vm表示使用可见矩阵引导的注意力机制。由此可见,使用可见矩阵的模型在准确率、召回率和F1值方面更优秀,可见矩阵可有效控制模型的注意力机制,使其聚焦于与当前词相关的实体信息,减少知识噪声,提升事件元素的识别准确率。

Table 8 Visible matrix ablation experimental results

表8 可见矩阵消融实验结果 (%)

模型	精确率(P)	召回率(R)	F1值
no_vm	82.20	76.00	79.00
use_vm	83.10	81.70	82.40

4 结语

本文提出一种融合知识图谱与预训练模型的特定领

域中文事件抽取方法,结合知识图谱丰富的语义信息与预训练模型强大的文本表示能力,克服了单纯使用预训练模型在领域专业性和专业术语理解等方面的局限性。

实验表明,在引入知识图谱特别是领域知识图谱信息后,所提方法相较于基线方法的F1值提升5.5%,证明了知识图谱信息能提升特定领域事件抽取任务的性能。然而,所提方法性能高度依赖知识图谱的质量和覆盖范围,如果知识图谱中存在信息缺失、错误或过时的情况,可能会对模型性能产生负面影响;结合知识图谱与预训练模型会增加计算资源,在一些资源受限的环境中应用受限。

后续,将考虑设计更有效的知识图谱注入与知识蒸馏机制,减少信息干扰、优化模型架构和算法来降低计算资源,以进一步提升模型的事件抽取性能。

参考文献:

- [1] LI Y, GENG C Y, YANG D. A Fin-BERT-based event extraction method for chinese financial domain[J]. Computer Engineering and Applications, 2024, 60(14): 123-132.
李熠, 耿朝阳, 杨丹. 基于Fin-BERT的中文金融领域事件抽取方法[J]. 计算机工程与应用, 2024, 60(14): 123-132.
- [2] QU X Y, LI B, WEN L Q. A research on event extraction for administrative law enforcement case texts[J]. Computer Engineering, 2024, 50(9): 1-11.
屈满雅, 李兵, 温立强. 面向行政执法案件文本的事件抽取研究[J]. 计算机工程, 2024, 50(9): 63-71.
- [3] SU F F, LI F, JI D H. Generative biomedical event extraction based on controllable decoding[J]. Journal of Chinese Information Processing, 2023, 37(11): 68-80.
苏方方, 李霏, 姬东鸿. 基于可控解码策略的生成式生物学事件抽取[J]. 中文信息学报, 2023, 37(11): 68-80.
- [4] WAN Q Z, WAN C X, HU R, et al. Chinese financial event extraction based on syntactic and semantic dependency parsing[J]. Chinese Journal of Computer, 2021, 44(3): 508-530.
万齐智, 万常选, 胡蓉, 等. 基于句法语义依存分析的中文金融事件抽取[J]. 计算机学报, 2021, 44(3): 508-530.
- [5] EL-ALLALY E, SARROUTI M, EN-NAHNAHI N, et al. MTLADE: a multi-task transfer learning-based method for adverse drug events extraction[J]. Information Processing and Management, 2021, 58(3): 102473.
- [6] SHEN S R, QI G L, LI Z, et al. Hierarchical Chinese legal event extraction via pedal attention mechanism[C]// Proceedings of the 28th International Conference on Computational Linguistics, 2020: 100-113.
- [7] CAO K, LI X, FAN M, et al. Improving event detection with active learning[C]// Proceedings of the International Conference Recent Advances in Natural Language Processing, 2015: 72-77.
- [8] BETHARD S, MARTIN J H. Identification of event mentions and their semantic class[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2006: 146-154.
- [9] YU X, WANG X, LUO X, et al. Multi-scale event causality extraction via simultaneous knowledge-attention and convolutional neural network[J]. Expert Systems, 2022, 39(5): 12952.
- [10] WANG L, CAO H, YUAN L, et al. Child-sum EATree-LSTMs: enhanced attentive child-sum tree-LSTMs for biomedical event extraction[J]. BMC Bioinformatics, 2023, 24(1): 253.
- [11] YU C M, DENG B, TAN L Y, et al. Syntax-enhanced event extraction model based on XLNET and GAT[J]. Data Analysis and Knowledge Discovery, 2024, 8(4): 26-38.
余传明, 邓斌, 谈腊云, 等. 基于XLNET和GAT的句法信息增强事件抽取模型[J]. 数据分析与知识发现, 2024, 8(4): 26-38.
- [12] SUN H, ZHOU J, KONG L, et al. Seq2EG: a novel and effective event graph parsing approach for event extraction[J]. Knowledge and Information Systems, 2023, 65(10): 4273-4294.
- [13] WANG X, GUI L, HE Y. Document-level multi-event extraction with event proxy nodes and hausdorff distance minimization[C]// 61st Annual Meeting of the Association for Computational Linguistics, 2023: 10118-10133.
- [14] NGUYEN C, MAN H, NGUYEN T. Contextualized soft prompts for extraction of event arguments[C]// Findings of the Association for Computational Linguistics, 2023: 4352-4361.
- [15] ZHANG H K, SONG H, XU B, et al. A BERT-based end-to-end model for Chinese document-level event extraction[J]. Journal of Chinese Information Processing, 2022, 36(10): 97-106.
张洪宽, 宋晖, 徐波, 等. 基于BERT的端到端中文篇章事件抽取[J]. 中文信息学报, 2022, 36(10): 97-106.
- [16] PEI B S, LI X, JIANG Z T, et al. Research on public security professional small sample knowledge extraction method based on large language model[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(10): 2630-2642.
裴炳森, 李欣, 蒋章涛, 等. 基于大语言模型的公安专业小样本知识抽取方法研究[J]. 计算机科学与探索, 2024, 18(10): 2630-2642.
- [17] ZHUANG L, FEI H, HU P. Knowledge-enhanced event relation extraction via event ontology prompt[J]. Information Fusion, 2023, 100: 101919.
- [18] WU C, ZHANG X, ZHANG Y. PMC-LLaMA: further finetuning LLaMA on medical papers[DB/OL]. <https://arxiv.org/abs/2304.14454v1>.
- [19] ZHANG Z, HAN X, LIU Z, et al. ERNIE: enhanced language representation with informative entities[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1441-1451.
- [20] LIU W, ZHOU P, ZHAO Z, et al. K-Bert: enabling language representation with knowledge graph[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 2901-2908.
- [21] WANG X, GAO T, ZHU Z, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 176-194.
- [22] SUN T, SHAO Y, QIU X, et al. CoLAKE: contextualized language and knowledge embedding[C]// Proceedings of the 28th International Conference on Computational Linguistics, 2020: 3660-3670.
- [23] YU D, ZHU C, YANG Y, et al. Jacket: joint pre-training of knowledge graph and language understanding[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 11630-11638.
- [24] TIAN S, LUO Y, XU T, et al. KG-Adapter: enabling knowledge graph integration in large language models through parameter-efficient fine-tuning[C]// Findings of the Association for Computational Linguistics, 2024: 3813-3828.
- [25] LI X, YAN H, QIU X, et al. FLAT: Chinese NER using flat-lattice transformer[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6836-6824.
- [26] MA R, PENG M, ZHANG Q, et al. Simplify the usage of lexicon in chinese NER[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 5951-5960.
- [27] SUN Y, WANG S, FENG S, et al. Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation[DB/OL]. <https://arxiv.org/abs/2107.02137>.
- [28] ZHANG T L, DONG J W, WANG J N, et al. Revisiting and advancing chinese natural language understanding with accelerated heterogeneous knowledge pre-training[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 560-570.
- [29] CUI Y, CHE W, WANG S, et al. Lert: a linguistically-motivated pre-trained language model[DB/OL]. <https://arxiv.org/abs/2211.05344>.

(责任编辑:刘嘉文)