基于 RoBERTa-Effg-Adv 的实体关系联合抽取方法

姚飞杨,刘晓静*
(青海大学 计算机技术与应用系,青海 西宁 810016)

摘　要:实体关系抽取是构建知识图谱的关键步骤,其目的是抽取文本中的关系三元组。针对现有中文实体关系联合抽取模型无法有效抽取重叠关系三元组及提取性能不足的问题,该文提出了 RoBERTa-Effg-Adv 的实体关系联合抽取模型,其编码端采用 RoBERTa-wwm-ext 预训练模型对输入数据进行编码,并采用 Efficient GlobalPointer 模型来处理嵌套和非嵌套命名实体识别,将实体关系三元组拆分成五元组进行实体关系联合抽取。再结合对抗训练,提升模型的鲁棒性。为了获得机器可读的语料库,对相关文本书籍进行扫描,并进行光学字符识别,再通过人工标注数据的方式,形成该研究所需要的关系抽取数据集 REDQTTM,该数据集包含 18 种实体类型和 11 种关系类型。实验结果验证了该方法在瞿昙寺壁画领域的中文实体关系联合抽取任务的有效性,在 REDQTTM 测试集上的精确率达到了 94.0%,召回率达到了 90.7%,F1 值达到了 92.3%,相比 GPLinker 模型,在精确率、召回率和 F1 值上分别提高了 2.4 百分点、0.9 百分点、1.6 百分点。

关键词:RoBERTa-wwm-ext;对抗训练;关系抽取;Efficient GlobalPointer;中文实体
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2024)03-0147-08
doi:10.3969/j.issn.1673-629X.2024.03.022

Entity and Relation Joint Extraction Method Based on RoBERTa-Effg-Adv

YAO Fei-yang,LIU Xiao-jing*
(Department of Computer Technology and Application,Qinghai University,Xining 810016,China)

Abstract:Entity and relation extraction is a key step in constructing knowledge graph,its purpose is to extract the relation triples in the text. Aiming at the problem that the current Chinese entity relation joint extraction model cannot effectively extract overlapping relation triples and the extraction performance is insufficient,we propose a entity and relation joint extraction model based on RoBERTa-Effg-Adv. At the encoder,the RoBERTa-wwm-ext pre-training model is used to encode the input data,and the Efficient GlobalPointer model is used to process nested and non-nested named entity recognition. The entity and relation triple is split into five tuples for entity and relation joint extraction. Combined with adversarial training,the robustness of the model is improved. In order to obtain machine-readable corpus,the relevant books are scanned,and optical character recognition is performed,and then the relation extraction dataset REDQTTM required by this study is formed by manually labeling the data. The dataset contains 18 entity types and 11 relationship types. The experimental results verify the effectiveness of the proposed method in the task of entity and relation joint extraction in the field of Qu Tan temple murals. The precision on the test set of REDQTTM reaches 94.0%,the recall reaches 90.7%,and the F1 value reaches 92.3%. Compared with the GPLinker model,the precision,recall and F1 value are improved by 2.4%,0.9% and 1.6% respectively.
Key words:RoBERTa-wwm-ext;adversarial training;relation extraction;Efficient GlobalPointer;Chinese entity

0　引　言

实体关系抽取是自然语言处理领域中一项重要的基础任务,其目的是从结构化、半结构化和非结构化数据中抽取形如<主体,关系,客体>的实体关系三元组。实体关系抽取任务是知识图谱构建、智能推荐、问答系统等众多自然语言处理任务的重要基础工具[1]。因此,实体关系抽取任务准确度的高低决定了自然语言处理领域下游任务效果的好坏。

收稿日期:2023-04-10　　**修回日期**:2023-08-10
基金项目:青海省 2021 年应用基础研究计划项目(2021-ZJ-717)
作者简介:姚飞杨(1996-),男,硕士研究生,研究方向为知识图谱的构建及应用研究;通讯作者:刘晓静(1978-),女,硕士,教授,CCF 会员(33794M),研究生导师,研究方向为信息可视化与媒体计算。

实体关系抽取以流水线方法和联合抽取方法这两类方法为主[2]。流水线方法将实体关系抽取分为命名实体识别和关系抽取这两个独立的任务,先对实体进行识别,再对实体之间的关系进行抽取[3]。流水线方法中每个独立的子任务都依赖前一个任务的结果作为当前任务的输入,这种方法存在着曝光偏差和误差传播等问题[4]。与流水线方法相比,联合抽取方法把三元组抽取看成一个整体任务,可以进一步利用两个任务之间存在的潜在信息,从而获得更好的抽取效果[5]。因此,联合抽取方法成为了当前实体关系抽取领域研究的主流方法。

虽然上述方法在中文实体关系抽取领域取得了较好的效果,但由于中文语言本身的特点,存在嵌套实体的问题,给实体之间的关系抽取带来了挑战。为了更好地获取文本的上下文语义信息,同时更好地提取嵌套实体之间的关系信息,该文提出了 RoBERTa-Effg-Adv 的实体关系联合抽取模型。与传统关系三元组抽取方式不同,该模型采用实体关系五元组抽取思想,将关系抽取任务分为主客体识别,头关系抽取和尾关系抽取,模型使用多头识别嵌套实体的方式,可有效抽取中文文本中重叠三元组。模型结合 PGD(Projected Gradient Descent)[6]对抗训练算法,有效提升了模型的抗扰动能力。

该文是在中文领域中进行的实体关系联合抽取研究,聚焦瞿昙寺壁画中涉及到的宗教领域中的命名实体识别与实体关系抽取。面向瞿昙寺壁画领域的实体关系联合抽取研究是瞿昙寺壁画知识图谱的建立和基于瞿昙寺壁画知识图谱的智能问答系统研究的基础任务。

主要贡献如下:

(1)通过对专业书籍扫描和手工标注数据等方式构建了瞿昙寺壁画领域的实体关系联合抽取数据集。

(2)在自制的数据集和公开的数据集上的实验证明,RoBERTa-Effg-Adv 模型通过多头识别嵌套实体,并将关系三元组拆分成五元组抽取,通过对抗训练提升模型鲁棒性,在精确率、召回率和 F1 值指标上表现较佳,验证了模型的有效性。

1 相关工作

近年来,深度学习的发展推动了关系抽取方法的不断进步,基于深度学习的实体识别和关系抽取已成为主流研究手段[7]。早期,实体关系抽取以流水线的方式为主,即在命名实体识别已完成的基础上再进行实体之间关系的抽取任务。

Socher 等人[8]在 2012 年将循环神经网络(RNN)应用到实体关系抽取领域中的关系分类,该方法利用循环神经网络对语句进行句法解析,经过不断迭代,从而得到句子的向量表示。这种方法有效地考虑了句子的句法结构。除 RNN 关系分类的方法外,Zeng 等人[9]在 2014 年将卷积神经网络(CNN)应用到关系抽取领域,利用卷积深度神经网络(CDNN)来提取文本语义特征。由于传统的 RNN 无法处理长期依赖,以及存在梯度消失、梯度爆炸等问题,Yan 等人[10]在 2015 年提出了基于长短时记忆网络(LSTM)的句法依存分析树的最短路径方法进行关系抽取研究。

流水线式的实体关系抽取方法中每个独立的任务的输入依赖于前一个任务的输出,因此存在任务之间丢失信息,忽视了两个子任务之间存在的关系信息[11],也可能会产生冗余信息等这些由误差传播引起的问题。实体关系联合抽取方式可以有效利用两个任务之间的潜在信息,同时也避免误差传递等问题。Wei 等人[12]在 2019 年提出一种基于联合解码的实体关系抽取模型 CasRel。CasRel 是层叠指针网络结构,由编码端和解码端组成。编码端使用 BERT[13]模型对输入数据进行编码,所获取的字向量能够利用词与词之间的相互关系有效提取文本中的特征;解码端主要包括头实体识别层、关系与尾实体联合识别层。该模型会先对头实体进行识别,然后在给定关系种类的条件下对尾实体进行识别。CasRel 模型存在曝光偏差问题。Wang 等人[14]在 2020 年提出一种单阶段联合提取模型 TPLinker,该模型解决了曝光偏差和嵌套命名实体识别问题。与 CasRel 模型不同,TPLinker 模型用同一个解码器对实体和关系进行解码,同时对实体和关系进行抽取,保证了训练和预测的一致性。苏剑林在 2022 年提出基于 GlobalPointer[15]的实体关系联合抽取模型 GPLinker。GPLinker 模型将实体关系三元组抽取转变为实体关系五元组(S_h,S_t,P,O_h,O_t)抽取,其中,S_h,S_t 表示主实体的头和尾,P 表示关系,O_h,O_t 表示尾实体的头和尾。与 TPLinker 模型相比,GPLinker 模型计算速度更快,而且显存占用更少。饶东宁等人[16]在 2023 年提出一种基于 Schema 增强的中文实体关系抽取方法。该方法采用字词混合嵌入的方式融合字与词的语义信息来避免中文分词时边界切分出错所造成的歧义问题,并利用指针标注的方式解决关系重叠问题。该方法通过提取出每个数据集的 Schema 进行合并作为先验特征传入模型中,以解决实体冗余及关系种类迁移问题[16]。

2 数据集的制作

本研究制作了瞿昙寺壁画领域的实体关系联合抽取数据集 REDQTTM(Relation Extraction Dataset of Qu Tan Temple Murals)。REDQTTM 原始数据文本来自

研究瞿昙寺壁画的相关专业书籍,对这些书籍进行扫描,并进行光学字符识别(Optical Character Recognition,OCR),从而获得机器可读的语料库。之后,按照预定义的实体和关系种类,使用标注工具对这些文本进行人工标注。标注工具选择 BRAT(Brat Rapid Annotation Tool)[17],BRAT 是基于 Linux 的一款应用于 WebServer 端的文本标注工具。通过对文本进行手工标注,最终得到后缀名为 ann 的标注文件。

实体在 ann 文件的格式由 5 列组成,第一列表示实体的编号,第二列表示实体的预定义类别,第三列表示实体在文本的开始下标,第四列表示实体在文本的结束下标,最后一列表示该实体所对应的文本。关系在 ann 文件的格式由 4 列组成,第一列表示关系的编号,第二列表示关系的预定义类别,第三列表示 Subject 实体的实体编号,最后一列表示 Object 实体的实体编号。

REDQTTM 总共包含了 18 种实体类型。瞿昙寺壁画中的神像体系主要有以下类别,分别是佛像、菩萨像、祖师像(或称上师、尊者)、本尊像、护法神像和佛母像[18]。这些神像体系都包含在 REDQTTM 的实体类别中。瞿昙寺壁画对神像的刻画十分详细,包括对神像的法器、服饰、坐骑、台座等细节展示,这些在 REDQTTM 中都有对应的实体种类。表 1 给出了 REDQTTM 中部分预定义的实体种类。

REDQTTM 中包含 11 种关系类型。表 2 给出了 REDQTTM 中预定义的关系种类。

表 1　部分实体类型和举例

实体类型	举例
animal	壁画中的动物,例如四爪金龙、麒麟等
benzun	本尊,例如大轮手金刚、时轮金刚等
bodhisattva	菩萨,例如文殊菩萨、十一面千手千眼观音等
buddha	佛,例如大日如来佛、宝生佛等
deity	佛母,例如积光佛母、妙音天母等
instrument	佛像的法器,例如金刚钺刀、十字金刚杵等
pedestal	佛像的台座,例如莲花座、须弥座等

表 2　关系类型和举例

实体类型	举例
alias	别名关系,例如欲界自在天女的别名是吉祥天女
creative_time	创建时间关系,例如《善财童子第五十二参拜弥勒菩萨》的创建时间是 15 世纪初
decorate	佛与饰品的关系,例如阿閦佛的饰品是宝冠
feature	壁画与面积的关系,例如《释迦净土变》的尺寸是 290 cm * 200 cm
hold	佛与法器的关系,例如大日如来佛的法器是法轮
include	包含关系,例如五方佛包括毗卢遮那佛
is	称号关系,例如额尔德尼罗桑却吉坚参是四世班禅
locate	位置关系,例如《宗喀巴》这幅壁画位于宝光殿西配殿正壁
ride	佛与动物的关系,例如山神的坐骑是战马
shift	佛与手印的关系,例如不空成就佛结期克印
sit	佛与台座的关系,例如释迦牟尼佛坐于莲花座上

通过对 ann 文件进行解析,最终得到本研究所需的数据集 REDQTTM。REDQTTM 分为训练集和测试集,三元组的比例为 8 ∶ 2 左右。如表 3 所示,REDQTTM 同样采用 json 格式,text 字段表示输入文本,predicate 字段表示关系类型,object_type 字段表示 object 实体类型,subject_type 字段表示 subject 实体类型,object 字段表示 object 实体,subject 字段表示 subject 实体。

表 3　A sample data in REDQTTM dataset

Sentence	Spo_list
"text":欲界自在天女,是吉祥天女的一种化身	{["predicate":"alias","object_type":"deity","subject_type":"deity","object":"吉祥天女","subject":"欲界自在天女"]}

3 模 型

3.1 模型整体结构

该文提出的 RoBERTa-Effg-Adv 模型包括 4 个部分:RoBERTa-wwm-ext[19] 编码层,Efficient GlobalPointer[15] 命名实体识别模块,关系抽取模块和对抗训练。模型整体结构如图 1 所示,RoBERTa-wwm-ext 编码层负责将输入的文本转化为词向量,作为模型后续部分的输入。在实体识别方面,使用 Efficient GlobalPointer 对主体和客体进行抽取。在关系抽取方面,将关系实体三元组拆分成五元组来处理,利用 Efficient GlobalPointer 处理 $S(s_h, o_h \mid p)$,其中 s_h 表示主实体的头,o_h 表示尾实体的头,p 表示关系。对于嵌套命名实体识别,需要同时指定起点和结束位置。同理利用 Efficient GlobalPointer 处理 $S(s_t, o_t \mid p)$,其中 s_t 表示主实体的尾,o_t 表示尾实体的尾。模型引入对抗训练来提升模型性能,对抗训练算法使用 PGD 对抗训练策略,该对抗训练算法采用“小步走,走多次”思想找到最优策略。

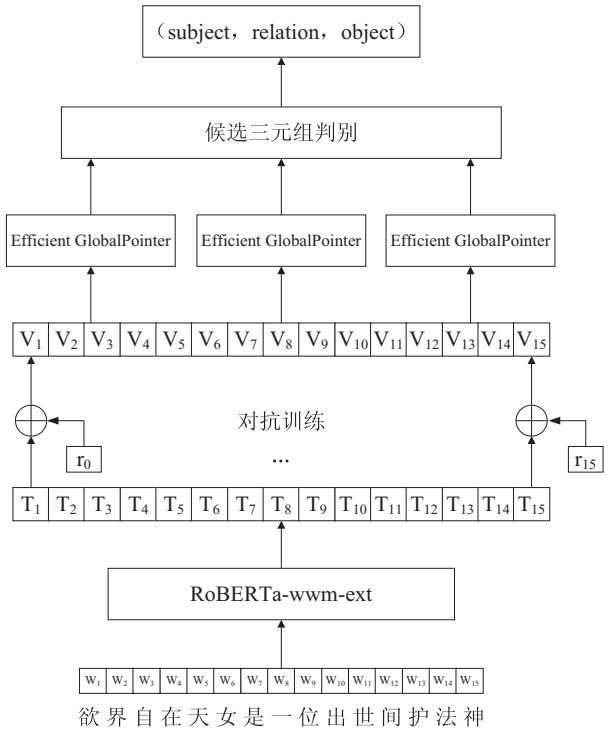

图 1 模型整体结构

3.2 RoBERTa-wwm-ext 编码层

编码端主要负责将输入文本转化为词向量,所获取的词向量能够利用词与词之间的相互信息提取文本中的特征信息。BERT 预训练模型的架构为 Transformer[20] 中的 Encoder,是目前使用最广泛的编码端模型,但原始的 BERT 模型不是最佳选择。文中编码端使用 RoBERTa-wwm-ext 预训练模型,该模型是在 RoBERTa[21] 模型的基础上做了一些优化,相比 BERT 预训练模型,能达到更好的编码效果。

(1)RoBERTa-wwm-ext 预训练模型在预训练阶段采用 wwm(whole word masking)策略进行 mask,而 BERT 模型是随机进行 mask,采用 wwm 策略的效果更好,具体示例如表 4 所示。

表 4 wwm 策略和 BERT 原始策略

	wwm 策略	BERT 原始策略
mask 前	两臂护法,肤色灰白, 黄色头发上扬,头戴宝冠	
mask 后	两臂护法,肤色灰白,黄色[MASK]上扬,头戴[MASK]	两臂护法,肤色灰白,黄色头[MASK]上扬,头戴[MASK]冠

(2)RoBERTa-wwm-ext 预训练模型取消了 NSP(Next Sentence Prediction)任务。取消了 NSP 任务后,模型性能得到提升。

(3)RoBERTa-wwm-ext 预训练模型采用更大的 Batch Size,这样有助于提高性能。ext(extended data)表示增加了训练数据集的大小。

3.3 Efficient GlobalPointer 命名实体识别模块

GlobalPointer 将实体的首尾视为一个整体去识别。如图 2 所示,在“欲界自在天女是一位出世间护法神”这句话中,对于实体类型“佛母”,该类型实体在文本中只有一个,是“欲界自在天女”;对于实体类型“称号”,该类型实体共有两个,分别是“出世间护法神”和“护法神”,从这里可以看出,GlobalPointer 可以识别嵌套类型实体。综上所述,假设待识别文本序列长度为 n,待识别实体个数为 k,那么在该序列中会有 $n(n+1)/2$ 个候选实体。在 GlobalPointer 中,命名实体识别任务可以看成“$n(n+1)/2$ 选 k”的多标签分类问题。如果一共有 m 种实体类型需要识别,那么可以看成 m 个“$n(n+1)/2$ 选 k”的多标签分类问题。GlobalPointer 是一个 token-pair 的识别模型,用一种统一的方式处理嵌套和非嵌套命名实体识别。

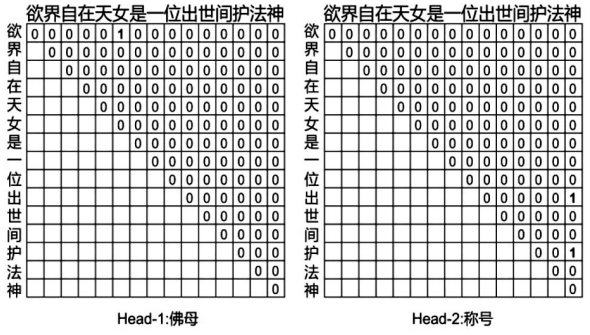

图 2 GlobalPointer 多头识别嵌套实体示意图

定义:

$$s_\alpha(i,j) = q_{i,\alpha}^{\mathrm{T}} k_{j,\alpha} \tag{1}$$

式 1 作为从 i 到 j 的连续片段是类型为 α 的实体的打分函数。其中,$q_{i,\alpha} = w_{q,\alpha} h_i + b_{q,\alpha}$ 和 $k_{i,\alpha} = w_{k,\alpha} h_i +$

$b_{k,\alpha}$ 是长度为 n 的输入 t 经过编码后得到的向量序列 $[h_1,h_2,\cdots,h_n]$ 变换而来。得到用于识别第 α 种类型实体所用的序列 $[q_{1,\alpha},q_{2,\alpha},\cdots,q_{n,\alpha}]$ 和 $[k_{1,\alpha},k_{2,\alpha},\cdots,k_{n,\alpha}]$。

Efficient GlobalPointer 主要针对 GlobalPointer 参数利用率不高的问题进行改进,优化了打分函数,达到了降低 GlobalPointer 的参数量的效果。

命名实体识别可以分为抽取和分类两部分,对于抽取部分,用一个打分矩阵表示,即用 $(w_q h_i)^{\mathrm{T}}(w_k h_j)$ 表示抽取出为实体的片段;对于分类部分,用 $w_\alpha^{\mathrm{T}}[h_i;h_j]$ 表示每个实体的类型。于是可以将两项组合起来,作为新的打分函数:

$$s_\alpha(i,j)=(w_q h_i)^{\mathrm{T}}(w_k h_j)+w_\alpha^{\mathrm{T}}[h_i;h_j] \tag{2}$$

对于抽取部分,所有实体类型共享这部分参数,所以在公式 2 的基础上,记 $q_i=w_q h_i$,$k_i=w_k h_i$,用 $[q_i;k_i]$ 来代替 h_i 以此进一步地减少参数量,此时

$$s_\alpha(i,j)=q_i^{\mathrm{T}}k_j+w_\alpha^{\mathrm{T}}[q_i;k_i;q_j;k_j] \tag{3}$$

得到的公式 3 作为 Efficient GlobalPointer 最终的打分函数,相比于公式 1 来说,参数利用率得到提升,参数量也降低了。

3.4　关系抽取模块

GPLinker 模型将实体关系三元组抽取转变为实体关系五元组 $(S_{\mathrm{h}},S_{\mathrm{t}},P,O_{\mathrm{h}},O_{\mathrm{t}})$ 抽取,其中,$S_{\mathrm{h}},S_{\mathrm{t}}$ 表示主实体的头和尾,P 表示关系,$O_{\mathrm{h}},O_{\mathrm{t}}$ 表示尾实体的头和尾。关系抽取流程如图 3 所示。

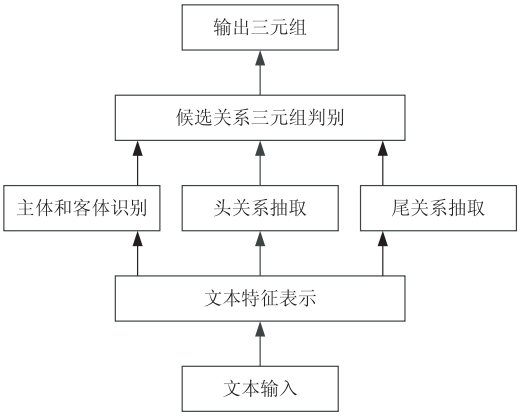

图 3　关系抽取流程

$$S(s_{\mathrm{h}},s_{\mathrm{t}},p,o_{\mathrm{h}},o_{\mathrm{t}})=S(s_{\mathrm{h}},s_{\mathrm{t}})+S(o_{\mathrm{h}},o_{\mathrm{t}})+S(s_{\mathrm{h}},o_{\mathrm{h}}\mid p)+S(s_{\mathrm{t}},o_{\mathrm{t}}\mid p) \tag{4}$$

模型训练时,对于标注的五元组让公式 4 中 $S(s_{\mathrm{h}},s_t)$,$S(o_{\mathrm{h}},o_{\mathrm{t}})$,$S(s_{\mathrm{h}},o_{\mathrm{h}}\mid p)$ 和 $S(s_{\mathrm{t}},o_{\mathrm{t}}\mid p)$ 皆大于 0,其他五元组这四项皆小于 0。模型预测时,枚举所有可能的五元组,找出 $S(s_{\mathrm{h}},s_{\mathrm{t}})>0$,$S(o_{\mathrm{h}},o_{\mathrm{t}})>0$,$S(s_{\mathrm{h}},o_{\mathrm{h}}\mid p)>0$ 和 $S(s_{\mathrm{t}},o_{\mathrm{t}}\mid p)>0$ 的部分,取它们的交集部分。

$S(s_{\mathrm{h}},s_t)$、$S(o_{\mathrm{h}},o_{\mathrm{t}})$ 分别是 subject 实体、object 实体的首尾打分函数,通过 $S(s_{\mathrm{h}},s_t)>0$,$S(o_{\mathrm{h}},o_{\mathrm{t}})>0$ 来得到所有的 subject 实体和 object 实体。至于函数 $S(s_{\mathrm{h}},o_{\mathrm{h}}\mid p)$ 和 $S(s_{\mathrm{t}},o_{\mathrm{t}}\mid p)$,则是 predicate 关系的匹配,$S(s_{\mathrm{h}},o_{\mathrm{h}}\mid p)$ 表示以 subject 和 object 的首特征作为它们自身的表征来进行一次匹配,考虑到存在嵌套实体,需要对实体的尾再进行一次匹配,即 $S(s_{\mathrm{t}},o_{\mathrm{t}}\mid p)$ 这一项。由于 $S(s_{\mathrm{h}},s_t)$,$S(o_{\mathrm{h}},o_{\mathrm{t}})$ 是用来识别 subject,object 对应的实体的,用一个 Efficient GlobalPointer 来完成;至于 $S(s_{\mathrm{h}},o_{\mathrm{h}}\mid p)$,它是用来识别关系为 p 的 $(S_{\mathrm{h}},O_{\mathrm{h}})$ 对,也可以用 Efficient GlobalPointer 来完成,最后对于 $S(s_{\mathrm{t}},o_{\mathrm{t}}\mid p)$ 这一项,处理和 $S(s_{\mathrm{h}},o_{\mathrm{h}}\mid p)$ 原理相同。

3.5　对抗训练

对抗训练是一种引入噪声的训练方式,可以对参数进行正则化,提升模型的鲁棒性和泛化能力[22]。对嵌入层的字向量添加一些较小的扰动,生成对抗样本,将获得的对抗样本再反馈给模型,从而提升模型的抗扰动能力。本研究使用的是 PGD 对抗训练算法。该算法通过多次迭代,以“小步走,走多次”的策略找到最优策略,并且通过设置扰动半径来防止扰动过大。扰动项 r_{adv} 的计算公式如下:

$$r_{\mathrm{adv}}=\varepsilon\cdot g(x)/\|g(x)\|_2 \tag{5}$$

$$g(x)=\nabla_x L(\theta,x,y) \tag{6}$$

其中,x 表示输入,y 表示标签,θ 表示模型参数,ε 表示扰动半径,$L(\theta,x,y)$ 表示单个样本的 loss。

PGD 算法步骤如下所示:

(1)计算 x 前向 loss,然后反向传播计算梯度并备份;

(2)对于每个步骤 t:根据 embedding 层的梯度,计算其 norm,然后根据公式计算出 r_{adv},再将 r_{adv} 累加到原始 embedding 的样本上,即 $x+r_{\mathrm{adv}}$,得到对抗样本;

(3)如果 t 不是最后一步,将梯度归 0,根据 $x+r_{\mathrm{adv}}$ 计算前后向并得到梯度;

(4)如果 t 是最后一步,恢复步骤 1 时的梯度值,计算最后的 $x+r_{\mathrm{adv}}$ 并将梯度累加到步骤 1 上,跳出循环;

(5)将被修改的 embedding 恢复到步骤 1 时的值;

(6)根据步骤 4 时的梯度对模型参数进行更新。

3.6　损失函数

损失函数选择稀疏版多标签分类的交叉熵损失函数。P,N 分别是正负类的集合,$A=P\cup N$,S 为对应的分数。

$$\log\Big(1+\sum_{i\in N}\mathrm{e}^{S_i}\Big)=\log\Big(1+\sum_{i\in A}\mathrm{e}^{S_i}-\sum_{i\in P}\mathrm{e}^{S_i}\Big)=\log\Big(1+\sum_{i\in A}\mathrm{e}^{S_i}\Big)+$$

$$\log(1-(\sum_{i\in P}e^{S_i})/(1+\sum_{i\in A}e^{S_i})) \tag{7}$$

令 $a=\log(1+\sum_{i\in A}e^{S_i})$，$b=\log(\sum_{i\in P}e^{S_i})$，那么可以写成：

$$\log(1+\sum_{i\in N}e^{S_i})=a+\log(1-e^{b-a}) \tag{8}$$

4 实 验

4.1 实验环境及参数设置

实验在 Linux 集群环境下进行，机器配置为 5 块 NVIDIA A100 80GB PCIe 显卡，代码使用 Python 语言编写。

实验主要参数设置如表 5 所示。

表 5 实验主要参数设置

参数	参数值
学习率	$1e^{-5}$
批处理数量	32
最大文本长度	400
优化器	Adam

4.2 实验数据集

为了验证文中方法的有效性，先后在 REDQTTM 和 DuIE[23] 数据集上进行实验。其中 DuIE 的训练集含有 173 108 条句子，验证集含有 21 639 条语句。

4.3 评价指标

使用精确率(Precision)、召回率(Recall)和 F1 值作为评估模型性能的指标。其中，精确率是模型预测正确的关系三元组数与预测出的三元组总数的比值；召回率则是模型预测正确的关系三元组数与实际三元组数的比值；F1 值是精确率和召回率的调和平均值，可以对模型的整体性能进行综合评价。Precision，Recall 和 F1 值的计算方式如公式 9～11 所示。

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{9}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{10}$$

$$\text{F1}=2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{11}$$

其中，TP 表示正例预测为正例的数量，FP 表示负例预测为正例的数量，FN 表示正例预测为负例的数量。

4.4 REDQTTM 数据集检测结果分析

该文选择多个基线模型在 REDQTTM 数据集上进行对比实验，这些模型包括 CasRel 模型、PRGC 模型、TPLinker 模型和 GPLinker 模型。

(1)CasRel：一种基于联合解码的实体关系抽取模型。该模型首先对头实体进行识别，然后在给定关系种类的条件下对尾实体进行命名实体识别。

(2)PRGC[24]：基于潜在关系和全局对应关系的实体关系抽取模型，将关系抽取分解为关系判断、实体抽取和主客体对齐三个任务。

(3)TPLinker：一种单阶段联合提取模型，该模型解决了曝光偏差和嵌套命名实体识别问题。TPLinker 模型保证了训练和预测的一致性，因其用同一个解码器对实体和关系进行解码，同时对实体和关系进行抽取。

(4)GPLinker：基于 GlobalPointer 的实体关系联合抽取模型。GPLinker 模型将实体关系三元组抽取转变为实体关系五元组(S_h,S_t,P,O_h,O_t)抽取。GPLinker 模型有着计算速度快、显存占用少等优点。

表 6 实验结果

方法名称	Precision/%	Recall/%	F1/%
CasRel	89.8	79.5	84.3
PRGC	85.0	86.5	85.7
TPLinker_plus	92.5	88.7	90.5
GPLinker	91.6	89.8	90.7
Ours	94.0	90.7	92.3

从表 6 可以看出，在 REDQTTM 数据集上，提出的方法无论是在 Precision，还是在 Recall 和 F1 上都是最优的。相比 GPLinker 模型，在 Precision 上提高了 2.4 百分点，在 Recall 上提高了 0.9 百分点，在 F1 上提高了 1.6 百分点。可见，提出的方法在瞿昙寺壁画实体关系联合抽取任务上取得了较好的效果。

为了验证各个模块的有效性，在 REDQTTM 数据集上进行了消融实验。-RoBERTa-wwm-ext 表示不使用此预训练模型，改为使用 BERT；-Pgd 表示不使用对抗训练；-Efficient GlobalPointer 表示不使用此模块，改用 GlobalPointer。实验结果如表 7 所示，去掉各模块后的性能都有所下降，验证了各模块的有效性。

表 7 消融实验结果

方法名称	Precision/%	Recall/%	F1/%
Ours	94.0	90.7	92.3
-RoBERTa-wwm-ext	93.1	90.2	91.6
-Pgd	93.1	89.9	91.5
-Efficient GlobalPointer	93.9	90.0	91.9

将相关文本输入到模型,抽取文本中的实体关系三元组。表 8 展示了模型对关系三元组的抽取效果。三元组的抽取是建立瞿昙寺壁画领域知识图谱的关键步骤。

表 8　三元组抽取结果部分示例

原始文本	实体关系三元组
例 1:毗那夜迦,即誐那钵底,也称象鼻天、欢喜天。毗那夜迦在印度教中奉为智慧之神,作为护法神。在佛教密宗中,它则变成性格暴戾、为害世界的恶神,人称“大荒神”。观音幻变成其明妃与彼抱合而生欢喜心,皈依佛法,得名“大圣欢喜天”	[“毗那夜迦”,“alias”,“大圣欢喜天”], [“毗那夜迦”,“alias”,“大荒神”], [“毗那夜迦”,“alias”,“象鼻天”], [“毗那夜迦”,“alias”,“誐那钵底”], [“毗那夜迦”,“is”,“智慧之神”], [“毗那夜迦”,“alias”,“欢喜天”]
例 2:金刚杵佛,身绿色,一面二臂,右手扬,手持十字金刚杵,左手当胸施期克印,黄发上竖,髯如火焰,张口龇牙,呈忿怒相。身饰璎珞、钏镯,披蓝绢,腰束虎皮裙	[“金刚杵佛”,“shift”,“期克印”], [“金刚杵佛”,“hold”,“十字金刚杵”], [“金刚杵佛”,“decorate”,“璎珞”], [“金刚杵佛”,“decorate”,“虎皮裙”], [“金刚杵佛”,“decorate”,“钏镯”], [“金刚杵佛”,“decorate”,“蓝绢”]
例 3:《四世班禅额尔德尼罗桑却吉坚参》,时间是 18-19 世纪,位于宝光殿西配殿西壁,大小为 220cm * 350cm。四世班禅额尔德尼罗桑却吉坚参(1570-1662)	[“额尔德尼罗桑却吉坚参”,“is”,“四世班禅”], [“《四世班禅额尔德尼罗桑却吉坚参》”,“feature”,“220cm * 350cm”], [“《四世班禅额尔德尼罗桑却吉坚参》”,“creative_time”,“18-19 世纪”], [“《四世班禅额尔德尼罗桑却吉坚参》”,“include”,“额尔德尼罗桑却吉坚参”], [“《四世班禅额尔德尼罗桑却吉坚参》”,“locate”,“宝光殿西配殿西壁”]

4.5　DuIE 数据集检测结果分析

文中模型在 DuIE 训练集上训练,在验证集上进行评估。MultiR[25]、CoType[26]、指针标注模型[27]、FETI[28]、CasRel、字词混合模型[29]和 BSCRE[30]模型的实验结果来自禹克强等人[30]的实验结果,如表 9 所示。

表 9　DUIE 数据集上的实验结果

方法名称	Precision/%	Recall/%	F1/%
MultiR	57.7	35.6	44.0
CoType	66.1	60.5	63.2
指针标注模型	69.4	63.9	66.5
FETI	75.7	76.0	75.8
CasRel	77.2	76.4	76.8
字词混合模型	81.3	78.1	79.7
BSCRE	81.6	79.5	80.5
Ours	82.3	83.2	82.7

从表 9 中可以看出,该文提出的方法相较于禹克强等人提出的 BSCRE 模型,在 DuIE 训练集上,Precision,Recall 和 F1 值分别提高了 0.7 百分点,3.7 百分点和 2.2 百分点。验证了 RoBERTa-Effg-Adv 模型在其它中文领域的实体关系联合抽取任务的有效性。

5　结束语

该文自建了瞿昙寺壁画领域的实体关系联合抽取数据集 REDQTTM,其中包含 18 种实体类型和 11 种关系类型。针对瞿昙寺壁画领域的实体关系联合抽取任务,提出了一种实体关系联合抽取模型 RoBERTa-Effg-Adv,其编码端使用 RoBERTa-wwm-ext 预训练模型,并采用 Efficient GlobalPointer 对命名实体进行识别,总体上使用实体关系五元组策略进行实体关系联合抽取。再结合对抗训练,提升模型整体的鲁棒性。由于该数据集包含的实体关系数量较少,后期会增加更多的预定义实体关系类别和数量来扩充数据集,也会在该实体关系联合抽取的基础上,开展建立瞿昙寺壁画领域的知识图谱、基于瞿昙寺壁画知识图谱的智能问答等研究。

参考文献:

[1] ZHANG Q,CHEN M,LIU L. A review on entity relation extraction[C]//2017 second international conference on mechanical, control and computer engineering (ICMCCE). Harbin:IEEE,2017:178-183.

[2] 衡红军,苗　菁. 语义与句法信息加强的二元标记实体关系联合抽取[J]. 计算机工程,2023,49(4):77-84.

[3] GAO C,ZHANG X,LIU H,et al. A joint extraction model

of entities and relations based on relation decomposition[J]. International Journal of Machine Learning and Cybernetics, 2022,13(7):1833-1845.
[4] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述[J]. 软件学报,2019,30(6):1793-1818.
[5] 刘雅璇,钟 勇. 基于头实体注意力的实体关系联合抽取方法[J]. 计算机应用,2021,41(9):2517-2522.
[6] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv:1706.06083,2017.
[7] WANG H,QIN K,ZAKARI R Y,et al. Deep neural network-based relation extraction:an overview[J]. Neural Computing and Applications,2022,34(6):4781-4801.
[8] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces [C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Stroudsburg: ACL, 2012: 1201-1211.
[9] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network [C]//Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. Stroudsburg: ACL, 2014:2335-2344.
[10] YAN X, MOU L, LI G, et al. Classifying relations via long short term memory networks along shortest dependency path [J]. arXiv:1508.03720,2015.
[11] 李冬梅,张 扬,李东远,等. 实体关系抽取方法研究综述[J]. 计算机研究与发展,2020,57(7):1424-1448.
[12] WEI Z, SU J, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[J]. arXiv: 1909.03227,2019.
[13] LEE J D M C K, TOUTANOVA K. Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805,2018.
[14] WANG Y, YU B, ZHANG Y, et al. TPLinker: single-stage joint extraction of entities and relations through token pair linking[J]. arXiv:2010.13415,2020.
[15] SU J, MURTADHA A, PAN S, et al. Global pointer: novel efficient span-based approach for named entity recognition [J]. arXiv:2208.03054,2022.
[16] 饶东宁,李 冉. 基于 Schema 增强的中文实体关系抽取方法[J]. 软件导刊,2023,22(2):47-52.
[17] STENETORP P, PYYSALO S, TOPIĆ G, et al. BRAT: a web-based tool for NLP-assisted text annotation[C]//Proceedings of the demonstrations at the 13th conference of the European chapter of the association for computational linguistics. Avignon: EACL, 2012: 102-107.
[18] 金 萍. 瞿昙寺壁画的艺术考古研究[D]. 西安:西安美术学院,2012.
[19] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.
[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st international conference on neural information processing systems. Red Hook: Curran Associates Inc., 2017: 5998-6008.
[21] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach [J]. arXiv: 1907.11692, 2019.
[22] 朱 红,牛浩然,朱 彤. 基于字词融合与对抗训练的行业人物实体识别[J]. 计算机工程,2023,49(5):56-62.
[23] LI S, HE W, SHI Y, et al. Duie: a large-scale chinese dataset for information extraction [C]//CCF international conference on natural language processing and Chinese computing. Dunhuang: Springer, 2019: 791-800.
[24] ZHENG H, WEN R, CHEN X, et al. PRGC: potential relation and global correspondence based joint relational triple extraction[J]. arXiv:2106.09895,2021.
[25] HOFFMANN R, ZHANG C L, LING X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]//Proceedings of the 49th annual meeting of the association for computational linguistics. [s. l.]: ACL, 2011: 541-550.
[26] REN X, WU Z, HE W, et al. Cotype: joint extraction of typed entities and relations with knowledge bases[C]//Proceedings of the 26th international conference on world wide web. New York: ACM, 2017: 1015-1024.
[27] 王勇超,穆华岭,周灵智,等. 基于指针网络的实体与关系联合抽取方法[J]. 计算机应用研究,2021,38(4):1004-1007.
[28] 陈仁杰,郑小盈,祝永新. 融合实体类别信息的实体关系联合抽取[J]. 计算机工程,2022,48(3):46-53.
[29] 葛君伟,李帅领,方义秋. 基于字词混合的中文实体关系联合抽取方法[J]. 计算机应用研究,2021,38(9):2619-2623.
[30] 禹克强,黄 芳,吴 琪,等. 基于双向语义的中文实体关系联合抽取方法[J]. 计算机工程,2023,49(1):92-99.