文章编号:1671-6833(2026)02-0041-10

基于 RoBERTa 和指针网络的中文实体与关系联合抽取方法

陈 燕[1,2], 韦紫君[2], 廖宇翔[2], 谭志湘[2], 胡小春[3,4], 宋 玲[2]

(1.广西壮族自治区信息中心 广西数字基础设施重点实验室,广西 南宁 530201;2.广西大学 计算机与电子信息学院,广西 南宁 530004;3.广西财经学院 广西财经大数据重点实验室,广西 南宁 530003;4.广西财经学院 大数据与人工智能学院,广西 南宁 530003)

摘 要:为了有效解决非结构化文本中实体与关系联合抽取时的三元组重叠问题,提出了一种基于 RoBERTa 和指针网络的中文实体与关系联合抽取方法。首先,针对实体重叠问题,基于指针网络设计了实体识别模块,将实体识别任务构建为 token-pair 识别问题,通过识别实体的开始和结束位置来提取所有可能的实体;其次,针对三元组重叠问题,设计基于多头注意力机制和 Ptr-Net 的关系抽取模块,将三元组(s, r, o)抽取任务构建为五元组(s_h, s_t, r, o_h, o_t)识别任务;最后,在中文信息抽取数据集 DuIE 上进行大量实验。实验结果表明:所提模型综合性能优于所有基线模型,其精确率、召回率和 $F1$ 值分别为 81.04%、85.82%和 83.36%。

关键词:实体与关系联合抽取; RoBERTa; 指针网络; 自然语言处理; 深度学习

中图分类号: TP391;TP312 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.05.007

在数字化时代背景下,非结构化文本数据的积累达到了前所未有的规模。研究者借助自然语言处理(natural language processing,NLP)技术对文本数据进行处理和分析,并从中抽取具有特定意义和价值的关键信息,为人们提供高质量的智能知识服务,如智能问答、机器翻译和个性化推荐等。面对庞大的文本数据,如何进行数据分类、关键词定位和深层语义关系挖掘是实现智能问答和知识图谱构建技术的关键[1]。

信息抽取旨在通过命名实体识别(named entity recognization,NER)和关系抽取(relation extraction,RE)任务识别非结构化文本中特定的信息,并以结构化的形式来表示。其中命名实体识别旨在从非结构化文本中标注特定的词汇或短语;关系抽取专注于识别实体间的关系。通常关系抽取是基于命名实体识别展开,多任务学习将这两个任务融合,将文本中的实体和关系抽取为主体(subject)、关系(relation)、客体(object)三元组。此类方法能更有效促进命名实体识别和关系抽取任务间信息的传递与理解,从而显著提高实体关系联合抽取的性能。

针对实体关系联合抽取任务没有充分考虑实体识别与关系抽取两个子任务之间的关联信息以及三元组重叠和任务间信息错误传播的问题,本文提出一种基于 RoBERTa(robustly optimized BERT pretraining approach)和指针网络(pointer network,Ptr-Net)的中文实体与关系联合抽取方法[2],主要研究内容如下:

(1)构建基于 RoBERTa 和指针网络的中文实体与关系联合抽取模型(RoBPtr),采用五元组的形式对文本中的实体与关系进行联合抽取,与管道方法相比,该方法能够有效解决实体识别和关系抽取任务间的关联信息未能充分考虑的问题,避免信息错误传播。

(2)借助指针网络将实体识别任务转化为识别 token-pair 的问题,通过确定实体的起始与终止位置,提取所有潜在实体。

(3)利用指针网络与多头注意力机制为每一组

收稿日期:2025-09-06;**修订日期**:2025-11-24

基金项目:国家自然科学基金资助项目 (72461001)

作者简介:陈燕(1975—),女,广西北流人,广西大学教授,博士,主要从事智能算法及应用、数据科学及应用研究,E-mail:cy@gxu.edu.cn。

通信作者:胡小春(1974—),男,广西南宁人,广西财经学院副教授,主要从事大数据与人工智能研究,E-mail:hxch@gxufe.edu.cn。

引用本文:陈燕,韦紫君,廖宇翔,等. 基于 RoBERTa 和指针网络的中文实体与关系联合抽取方法[J]. 郑州大学学报(工学版),2026,47(2):41-50. (CHEN Y, WEI Z J, LIAO Y X, et al. Joint extraction method of Chinese entities and relations based on RoBERTa and pointer network [J]. Journal of Zhengzhou University(Engineering Science),2026,47(2):41-50.)

实体对的多个关系建立专属的概率子空间,实现同一实体对的不同关系被非互斥的识别和提取,解决三元组重叠问题。

1 相关工作

传统的基于神经网络的实体关系抽取方法大多是基于管道方法。管道方法首先使用命名实体识别模型抽取文本中的实体,再使用关系抽取模型预测每个抽取到的实体间的关系[3]。该类方法没有考虑命名实体识别和关系抽取之间存在的紧密联系,同时还会导致信息的错误传播。为解决上述问题,研究者开始在命名实体识别和关系抽取之间设计各种桥梁,以此融合两个任务之间的信息[4-5]。然而,目前大多数方法依然采用参数共享的方式,而非通过联合解码来实现实体和关系的统一学习,仍然需要依赖管道式的处理模式。上述方法将关系抽取视为给实体对分配具体标签的问题,当同一实体存在于多个关系中时,关系分类器通常会产生困扰,无法判断实体参与到哪个关系中,导致抽取到的关系三元组不完整和不准确。

Miwa 等[6]提出一种基于历史信息的结构化实体与关系联合抽取学习方法,引入实体与关系表,用于表示句子中整个实体和关系的结构,将实体和关系联合抽取问题视为表格填充问题。Wei 等[7]提出一种用于关系三元组抽取的级联二进制标注框架,将关系建立为句子中主语映射到宾语的函数,并结合预训练语言模型,解决了传统方法中因实体共享导致的三元组重叠问题。Wang 等[8]提出一种名为 TPLinker 的单阶段模型,通过创新的令牌对链接机制,将实体识别与关系分类统一,建立 3 个矩阵的链接操作模型,有效解决传统方法中的暴露偏差和重叠关系难题。Yan 等[9]提出一种分区过滤网络(partition filter network,PFN)方法,用于正确建立 NER 与 RE 两个任务之间的双向交互关系,缓解了联合抽取中 NER 与 RE 任务之间特征交互不平衡的问题。Zheng 等[10]基于潜在关系和全局对应的关系三元组抽取框架 PRGC,首先筛选文本中的候选关系,然后针对每个潜在关系独立进行主客实体序列标注,最后通过构建字符级相关矩阵来快速匹配并组合主客实体对,从而组成三元组。Li 等[11]提出了一种用于实体和关系联合抽取的高效翻译解码模式 TDEER,通过将关系建立为主体到宾语的翻译操作,自然地处理重叠三元组问题,并通过引入负样本来增强模型的鲁棒性。Sui 等[12]提出了集合预测网络(set prediction networks,SPNs),以 BERT(bidirectional encoder representation from transformers)[13]作为编码器,利用基于 Transformer 的非自回归解码器作为集合生成器,可以一次性预测所有的三元组,避免了三元组排序的问题,并提出了受运筹学中分配问题启发的二分匹配损失函数。Gao 等[14]提出了一种新颖的轻量级联合抽取模型,通过基于仿射变换的全局实体匹配策略、候选关系注意力机制及负采样策略,在很大程度上简化了模型结构,并在一定程度上解决了三元组重叠的问题。Li 等[15]提出了一种基于分解策略的联合抽取框架,通过引入指针机制提高边界特征提取效率,增强边界感知与分类能力。

上述方法通过参数共享和联合解码等机制,为实体识别和实体关系抽取两个子任务建立了协同关联,从而实现两个子任务之间的交互信息建模,在一定程度上缓解了传统管道方法无法考虑实体识别与实体关系抽取的关联信息以及任务间信息错误传播的问题。然而,对于包含大量三元组嵌套的句子,这些方法的抽取性能仍有较大的提升空间。

2 RoBPtr 模型

2.1 模型结构

本文提出一种基于 RoBERTa 和指针网络的中文实体与关系联合抽取方法,将实体与关系联合抽取任务构建为 token-pair 识别问题,即把三元组(s, r, o)抽取任务重构为五元组(s_h, s_t, r, o_h, o_t)识别任务。该方法主要分为通用领域中文实体与关系抽取数据集构建模块、预训练模块和下游实体与关系联合抽取模型训练 3 个部分,本文方法流程如图 1 所示。

RoBPtr 模型实现了实体与关系联合抽取。模型框架主要分为 3 部分:编码层、实体识别模块和关系抽取模块。首先,使用 RoBERTa 获得字符的深度双向表示;其次,利用实体识别模块提取文本中所有潜在的主体和客体,形成候选实体集合;最后,使用关系抽取模块解析所有可能的实体关系三元组。模型结构如图 2 所示,图 2 中黄色标注表示实体开始和结束位置的 token-pair,蓝色标注表示 subject 和 object 开始位置的 token-pair,粉色标注表示 subject 和 object 结束位置的 token-pair。

2.2 矩阵平坦化

使用连接标签方式对 token-pair 进行标注,然后借助连接矩阵对不同的标注结果进行解码,得到所有实体及其对应的嵌套关系。传统连接矩阵标注方法难以用一个矩阵表示同一实体对的多个关系,即不能有效解决实体对重叠(entity pair overlap,

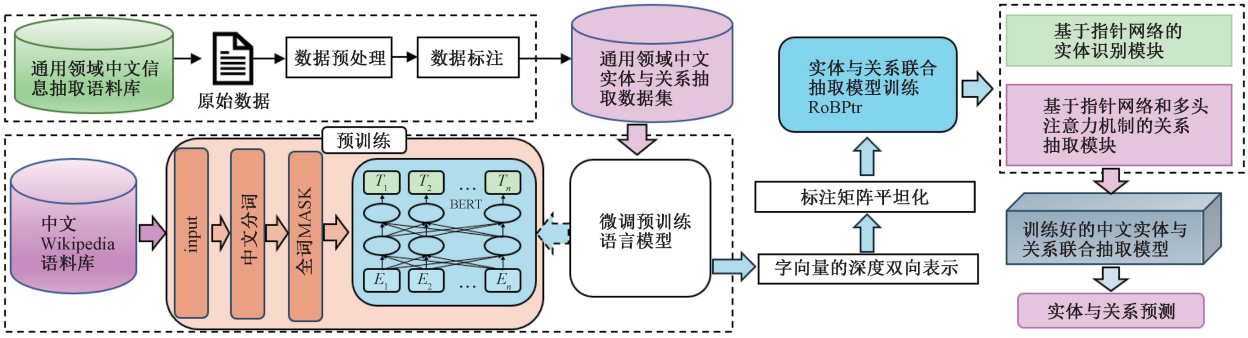

图 1　本文方法流程图

Figure 1　Flowchart of proposed method

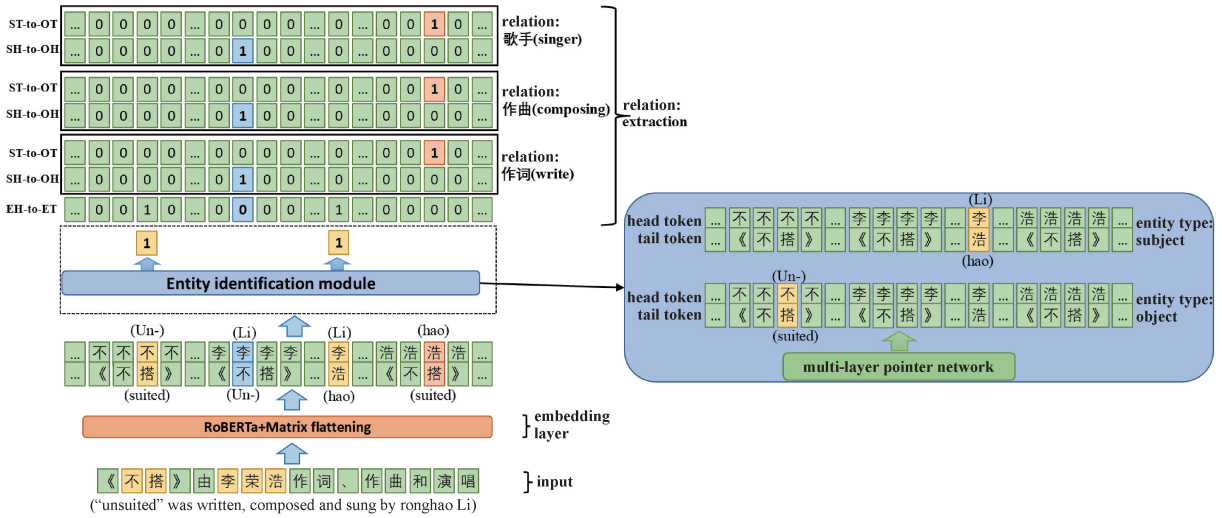

图 2　RoBPtr 模型结构

Figure 2　Model architecture of RoBPtr

EPO)问题。为此,本文依据关系对实体边界进行对齐,以序列的方式将每种关系的连接矩阵平坦化,并设计了 3 种不同的连接标签方式来标注 token-pair,实现使用同一个矩阵将每种关系都进行标注[16]。3 种连接标签方式如下。

(1) EH-to-ET(entity head to entity tail):token-pair 表示单个实体在文中的起始和结束位置。

(2) SH-to-OH(subject head to object head):token-pair 表示 subject 和 object 在文中的起始位置。

(3) ST-to-OT(subject tail to object tail):token-pair 表示 subject 和 object 在文中的结束位置。

token-pair 标注示例如表 1 所示。

表 1　token-pair 标注示例

Table 1　Tagging examples of token-pair

标注类型	示例
原句子	《不搭》由李荣浩作词、作曲和演唱
EH-to-ET	(不,搭)、(李,浩)
SH-to-OH	(李,不)
ST-to-OT	(浩,搭)

在传统的 token-pair 标注方法中,为了表示同一实体对间的不同关系,需要对应多个独立的连接矩阵,如图 3 所示。图 3 中黄色标注表示实体开始和结束的 token,蓝色标注表示 subject 和 object 的开始 token,粉色标注表示 subject 和 object 的结束 token。

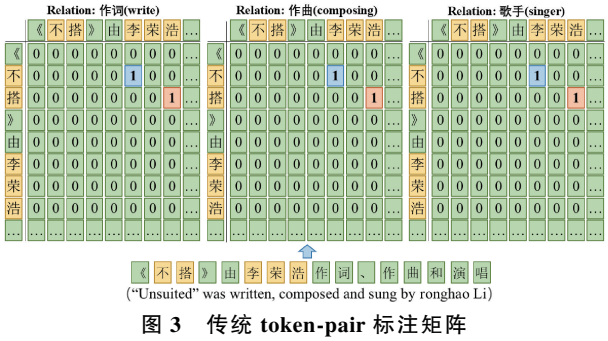

图 3　传统 token-pair 标注矩阵

Figure 3　Traditional token-pair tagging matrix

经过矩阵平坦化处理后,同一实体对 3 种不同关系标记只需一个连接矩阵,具体参见图 4。图 4 中黄色标注表示实体在文中的起始和结束位置的 token-pair,蓝色标注表示 subject 和 object 在文中的起始位置的 token-pair,粉色标注表示 subject 和 ob-

ject 在文中的结束位置的 token-pair。

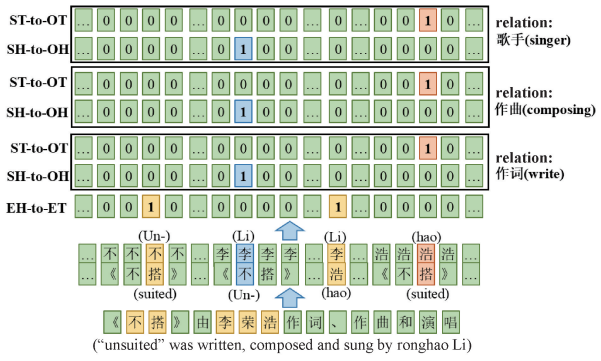

图 4 矩阵平坦化后的 token-pair 标注矩阵

Figure 4 Token-pair tagging matrix after matrix flattening

2.3 RoBERTa 层

本文采用 RoBERTa 作为嵌入层学习字符的深度双向表示,从而充分考虑上下文语义和位置信息的字符嵌入。RoBERTa 使用分词器对输入文本进行分词操作,随后从得到的分词结果中随机选择多个连续词汇实施全词遮蔽(whole word masking, WWM)[17],进而使模型依据上下文来推测被遮蔽的单词,以此增强模型对上下文及词语的理解与表征能力。此外,RoBERTa 还将位置嵌入与字符嵌入相融合,使其最终输出的嵌入能够涵盖上下文语义和位置信息,从而充分表征字符的特征。得到的嵌入向量将在下游实体识别任务与关系抽取任务中共用。

2.4 实体识别模块

传统基于指针网络的实体识别方法常分别识别实体的 head-token 和 tail-token,容易引起训练阶段与预测阶段的不匹配问题。本文将 head-token 与 tail-token 统一考虑来进行实体识别,从而将该任务转换为 token-pair 的连接问题。通过构建多层指针网络,高效地识别实体边界,其中每一层指针网络专注于一种特定类型的实体识别,基于多层指针网络的实体识别方法的结构如图 5 所示。

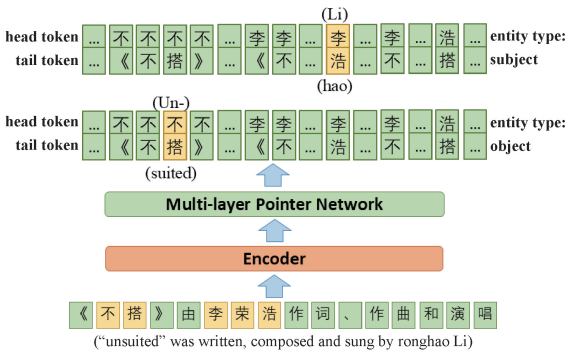

图 5 基于多层指针网络的实体识别方法的结构

Figure 5 Structure of entity recognition method based on multi-layer pointer network

对于一个具有 n 个 token 的序列,每个实体都是该序列中的一个或多个连续片段,其长度不固定,并允许嵌套结构存在。在该序列中共有 $n(n+1)/2$ 个可能的子序列,每一个都可被视为潜在的实体。模型的任务是在这些候选实体中精确地确定实体边界。当需要识别 m 种不同类型的实体时,则将问题构建为从 $m \cdot n(n+1)/2$ 个实体中选择 k 的多标签分类任务。算法如式(1)~式(3)所示:

$$\boldsymbol{q}_{i,a} = \boldsymbol{W}_{q,a}\boldsymbol{h}_i + \boldsymbol{b}_{q,a}; \tag{1}$$

$$\boldsymbol{K}_{i,a} = \boldsymbol{W}_{k,a}\boldsymbol{h}_i + \boldsymbol{b}_{k,a}; \tag{2}$$

$$P_a(i,j) = \boldsymbol{q}_{i,a}^{\mathrm{T}}\boldsymbol{K}_{j,a}。 \tag{3}$$

式中:$\boldsymbol{h}_i$ 表示第 i 个 token 的向量表示;$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别指代模型训练过程中可学习的权重参数和偏置;$P_a(i,j)$ 指代从序列中第 i 个到第 j 个 token 所组成的连续片段属于类型 a 实体的概率。

2.5 关系抽取模块

借助多头注意力机制与指针网络[18]实现关系抽取模块,实体对的各类关系通过多头注意力机制分配独立的概率子空间,实现对同一实体对可能存在的多种关系进行非互斥的识别,进而有效地处理 EPO 问题。关系抽取模块如图 6 所示。

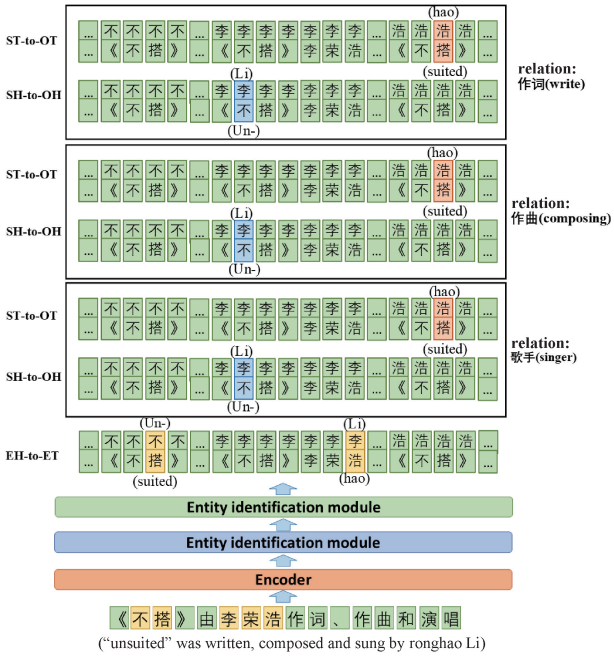

图 6 关系抽取模块

Figure 6 Module of relation extraction

将三元组(s, r, o)识别任务构建为五元组(s_h, s_t, r, o_h, o_t)识别问题。其中 h 表示实体的开头位置,t 表示实体的结尾位置。对于实体识别模块中抽取的所有 subject 和 object,利用预定义的 relation 作为条件对齐 subject 和 object 实体的边界以实现关系抽取,即在 relation 条件下对(s_h, o_h)和(s_t, o_t)

进行打分,其中所有关系共享 EH-to-ET 序列。算法将计算三元组(s, r, o)的得分重构为计算五元组(s_h, s_t, r, o_h, o_t)的得分,由式(1)~式(3)可推导得到 token-pair 的打分函数,算法如式(4)~式(8)所示:

$$S(s_h,s_t)=S_{sub}(i,j)=\boldsymbol{q}_{i,sub}^{T}\boldsymbol{K}_{j,sub};\tag{4}$$

$$S(o_h,o_t)=S_{obj}(i,j)=\boldsymbol{q}_{i,obj}^{T}\boldsymbol{K}_{j,obj};\tag{5}$$

$$S(s_h,o_h|r)=S_r(s_h,o_h)=\boldsymbol{q}_{s_h,r}^{T}\boldsymbol{K}_{o_h,r};\tag{6}$$

$$S(s_t,o_t|r)=S_r(s_t,o_t)=\boldsymbol{q}_{s_t,r}^{T}\boldsymbol{K}_{o_t,r};\tag{7}$$

$$S(s_h,s_t,r,o_h,o_t)=S(s_h,s_t)+S(o_h,o_t)+S(s_h,o_h|r)+S(s_t,o_t|r)。\tag{8}$$

式中:$S(s_h, s_t)$表示以 s_h 和 s_t 作为开始和结束 token 的 subject 的得分;$S(o_h, o_t)$表示以 o_h 和 o_t 作为开始和结束 token 的 object 的得分;$S(s_h, o_h|r)$表示分别以 s_h 和 o_h 作为开头 token 的 subject 和 object 之间的关系为 r 的得分;$S(s_t, o_t|r)$表示分别以 s_t 和 o_t 作为结束 token 的 subject 和 object 之间的关系为 r 的得分;$S(s_h, s_t, r, o_h, o_t)$表示五元组的得分。

模型训练目标为让真实五元组满足条件 $S(s_h, s_t, r, o_h, o_t)>0$、$S(s_h, s_t)>0$、$S(o_h, o_t)>0$、$S(s_h, o_h|r)>0$、$S(s_t, o_t|r)>0$,其余五元组 $S(s_h, s_t, r, o_h, o_t)<0$。预测阶段当且仅当 $S(s_h, s_t, r, o_h, o_t)>0$ & $S(s_h, s_t)>0$ & $S(o_h, o_t)>0$ & $S(s_h, o_h|r)>0$ & $S(s_t, o_t|r)>0$ 为真时,将对应的五元组作为输出。

2.6　损失函数

本文联合训练实体识别模块和关系抽取模块,在训练期间优化组合目标函数。损失函数使用多标签交叉熵[19]。

实体识别模块损失函数为

$$L_{NER}=\log\left(1+\sum_{(i,j)\in P_a}e^{-s_a^{(i,j)}}\right)+\log\left(1+\sum_{(i,j)\in N_a}e^{s_a^{(i,j)}}\right)。\tag{9}$$

式中:P_a 表示所有类型为 a 的实体首尾集合;N_a 表示所有类型为非 a 的实体首尾集合或非实体;$\boldsymbol{S}_a^{(i,j)}$ 表示第 i 到第 j 个 token 组成的片段为类型 a 的实体的得分。

关系抽取模块损失函数为

$$L_{RE}=\log\left(1+\sum_{i\in P}e^{-S_i}\right)+\log\left(1+\sum_{i\in N}e^{S_i}\right)。\tag{10}$$

式中:P 表示真实五元组的集合;N 表示非真实五元组的集合或非五元组;S_i 表示五元组的得分。

模型总体损失函数为

$$L_{TOTAL}=L_{NER}+L_{RE}。\tag{11}$$

3　数据集与评价指标

3.1　数据集

本文使用 DuIE 数据集[20],该数据集是当前中文信息抽取领域内依据 Schema 构建的最大规模的公开可用数据集。数据集中包含 50 种常见关系类型,例如作曲、目和民族等。图 7 展示了前 15 种占比最高的关系类型分布情况。

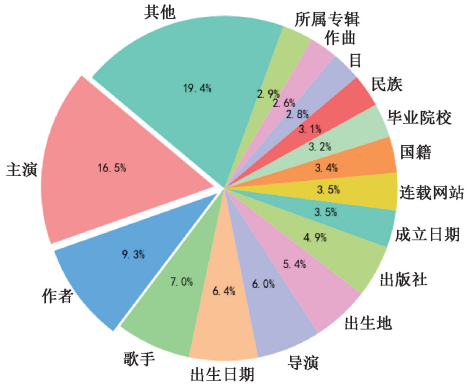

图 7　DuIE 数据集关系类型分布

Figure 7　Relationship type distribution of DuIE dataset

DuIE 数据集包括 214 590 个中文句子和 457 886 个实例,每个实例包含原始句子("text")、主体("subject")、主体类型("subject_type")、客体("object")、客体类型("object_type")、关系列表("spo_list")及谓词("predicate")。统计数据见表 2。

3.2　数据预处理

对 DuIE 数据集进行预处理操作,将原始语料转化为以文本、实体列表和三元组列表形式的数据结构,同时以 token-pair 形式标注实体的开始和结束位置,DuIE 数据示例如表 3 所示。

表 2　DuIE 数据集统计信息

Table 2　Statistics of DuIE datasets

类型	不同重叠模式句子数			不同三元组数量的句子数				
	Normal	SEO	EPO	$N=1$	$N=2$	$N=3$	$N=4$	$N=5$
Train	75 534	109 746	9 160	74 118	56 559	21 288	10 174	10 831
Valid	9 396	13 698	1 128	9 215	7 153	2 594	1 290	1 374

注:Normal 表示无重叠;SEO 表示单实体重叠;EPO 表示实体对重叠;N 表示三元组数量。

表 3 DuIE 数据预处理示例

Table 3 Data preprocessing example of DuIE

原始数据	预处理后数据
{"text":"《逐风行》是百度文学旗下纵横中文网签约作家清水秋风创作的一部东方玄幻小说,小说已于 2014-04-28 正式发布","spo_list":[{"predicate":"连载网站","object_type":"网站","subject_type":"网络小说","object":"纵横中文网","subject":"逐风行"}]}	{"text":"《逐风行》是百度文学旗下纵横中文网签约作家清水秋风创作的一部东方玄幻小说,小说已于 2014-04-28 正式发布","id":"train_4","relation_list":[{"subject":"逐风行","object":"纵横中文网","subj_char_span":[1, 4],"obj_char_span":[12, 17],"predicate":"连载网站","subj_tok_span":[1, 4],"obj_tok_span":[12, 17]}],"subj_tok_span":[1, 4],"obj_tok_span":[21, 25]}],"entity_list":[{"text":"逐风行","type":"网络小说","char_span":[1, 4],"tok_span":[1, 4]},{"text":"纵横中文网","type":"网站","char_span":[21, 25],"tok_span":[21, 25]}]}

3.3 模型评估指标

实验以召回率 R、精确率 P 及 $F1$ 值作为评价标准,用于评估模型的性能,对应计算公式见式(12)~式(14):

$$R = \frac{TruePositive}{ActualPositive} \times 100\%; \tag{12}$$

$$P = \frac{TruePositive}{PredictPositive} \times 100\%; \tag{13}$$

$$F1 = \frac{2PR}{P + R} \times 100\%。 \tag{14}$$

式中:*TruePositive* 为准确识别的五元组数量;*ActualPositive* 为数据集中实际存在的五元组总数;*PredictPositive* 为识别出的五元组总数。

4 实验与结果分析

4.1 实验参数设置

基于数据集 DuIE 进行实验,并将数据集按 8∶2 的比例划分为训练集与验证集。所用的 RoBERTa-large 模型包含 24 层网络,多头注意力机制中自注意力头数量为 16,隐藏层维度为 1 024。学习率为 2e-5,批处理大小为 32,训练的迭代次数为 30,模型的优化器选择 AdamEMA。整个模型的参数量约为 3.55e-8。

4.2 对比实验结果与分析

4.2.1 不同编码器的性能分析

用 BERT[13]、RoBERTa[2]、ALBERT[21]和 ELECTRA[22]4 种预训练语言模型作为编码器时的联合模型性能进行了比较,选择性能最佳的预训练模型作为模型的编码器。实验结果如表 4 所示。

实验结果显示,编码层为 RoBERTa 时,RoBPtr 模型性能表现最佳,$F1$ 值为 83.36%。ALBERT、ELECTRA 和 RoBERTa 均为在 BERT 基础上进行优化的衍生版本,故选用 BERT 作为参照标准来对比

表 4 不同预训练语言模型对比实验

Table 4 Experimental results of the different pre-trained language models

模型	P/%	R/%	$F1$/%	t/h
BERT	81.15	83.71	82.42	1.5
ALBERT	**84.32**	79.33	81.79	**1.2**
ELECTRA	81.52	83.70	82.65	1.8
RoBERTa	81.04	**85.82**	**83.36**	2.2

注:t 表示模型训练时长,下同。

这 4 个模型的性能表现。ALBERT 通过显著减少参数数量提高训练效率,但性能也随之下降,$F1$ 值较 BERT 降低了 0.63 百分点。ELECTRA 通过引入 RTD(replaced token detection)任务代替 BERT 原有的 MLM(masked language model)机制,从而增强了模型学习到的表示能力,$F1$ 值相比 BERT 提升了 0.23 百分点。RoBERTa 则通过使用更广泛的数据集、增加批处理大小以及更长的输入序列,并结合动态掩码策略来深化对语义信息的理解,从而在上述 3 种模型中脱颖而出。使用不同预训练模型作为 RoBPtr 模型嵌入层的性能对比分别如图 8 与图 9 所示。

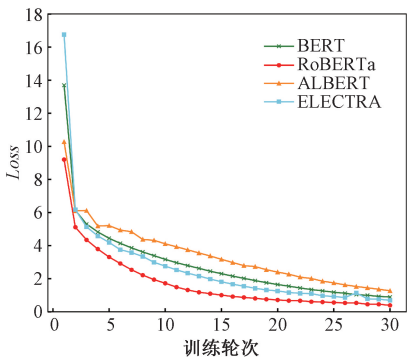

图 8 联合抽取模型使用不同语言模型作为编码器在损失曲线的性能对比

Figure 8 Performance comparison of joint extraction model using different LM as encoders in loss curve

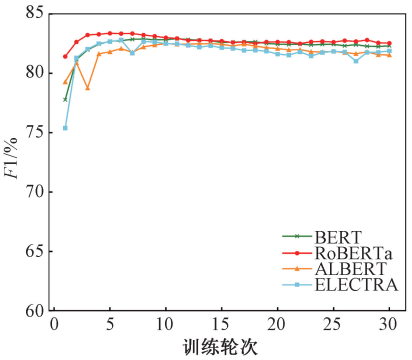

图 9　联合抽取模型使用不同语言模型作为编码器在 *F*1 值曲线的性能对比

Figure 9　Performance comparison of joint extraction model using different LM as encoders in *F*1 score curve

4.2.2　不同模型性能的对比分析

为了评估 RoBPtr 模型在实体关系联合抽取任务上的性能表现,挑选 5 个经典模型(TPLinker[8]、TPLinker_plus[8]、PRGC[10]、TDEER[11]和 CasRel[7])为基线进行比较,对比实验结果如表 5 所示。实验结果显示,与所有基线模型相比,RoBPtr 模型在 DuIE 数据集上的表现较佳,其精确率、召回率及 *F*1 值分别为 81.04%、85.82%和 83.36%。在对比的基准模型中,CasRel 的 *F*1 值达到 82.64%,但其两阶段抽取方法在训练过程中容易受到信息错误传播的影响。TDEER 模型引入负样本以减少错误累积,但仍未能完全消除阶段间的信息暴露偏差和错误传播问题,模型的精确度、召回率和 *F*1 值分别为 80.38%、82.29%和 81.32%。RoBPtr 模型通过联合建模有效解决信息错误传播和暴露偏差等问题,性能表现优于 CasRel 和 TDEER 模型。相较于专门为处理复杂句式和三元组重叠问题而设计的 TPLinker(BERT)、TPLinker_plus、PRGC 及 TDEER 模型,RoBPtr 展现出极具竞争力的性能,相较于上述 4 个基线模型,RoBPtr 模型 *F*1 值分别提升了 3.24 百分点、1.49 百分点、3.25 百分点和 2.04 百分点。图 10 展示了各模型在整个训练过程中 *F*1 变化趋势。

表 5　不同模型对比实验结果

Table 5　Experimental results of different models

模型	*P*/%	*R*/%	*F*1/%	*t*/h
TPLinker(BiLSTM)	**83.38**	72.04	77.30	**0.8**
TPLinker(BERT)	78.05	82.30	80.12	1.4
TPLinker_plus	80.80	82.97	81.87	1.5
PRGC	76.37	82.41	80.11	1.5
TDEER	80.38	82.29	81.32	1.9
CasRel	81.57	83.73	82.64	2.3
RoBPtr(本文)	81.04	**85.82**	**83.36**	2.2

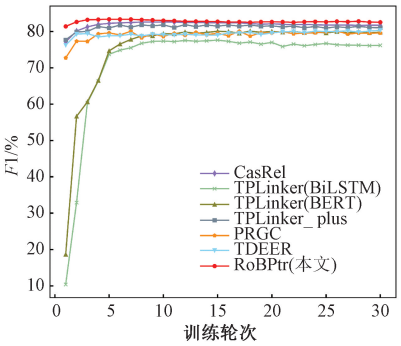

图 10　不同模型的 *F*1 值曲线

Figure 10　*F*1 values curve of different models

4.2.3　不同复杂度句子下模型性能的比较分析

为评估模型在处理复杂句子方面的性能,对包含不同数量三元组的句子进行实验,以分析模型从不同复杂度句子中抽取三元组的能力。通常,句子中三元组数量越多,其复杂度越高。实验结果如表 6 所示。实验结果表明,在处理具有不同数量三元组(N=1,2,3,4 和 $N\geqslant5$)的语句时,RoBPtr 模型的性能超过了所有对比基准模型,*F*1 值分别为 80.41%、86.70%、85.63%、86.43 和 87.12%。针对不同复杂度级别的比较(N=1~5),RoBPtr 模型展现了其优越性:当 N=1 时,与 TPLinker(BiLSTM)相比,RoBPtr 的 *F*1 值提高 0.31 百分点;对于 N=2 和 N=3 的情形,RoBPtr 相对于 CasRel,*F*1 值分别提高了 0.78 百分点和 1.06 百分点;而对于 N=4 情形,RoBPtr 的性能表现略逊于表现最佳的 PRGC 模型;但当 $N\geqslant5$ 时,与 PRGC 模型相比,RoBPtr 的 *F*1 值提升了 0.35 百分点。同时,实验结果表明,在文中存在多个三元组时,RoBPtr 模型性能表现最优,并随着三元组个数的增加呈上升趋势,具体变化趋势参考图 11。由图 11 可知,RoBPtr 模型能有效处理复杂句子,并在复杂场景下的实体与关系联合抽取任务中优于基线模型。

4.2.4　不同重叠模式下模型性能的比较分析

为评估模型在不同重叠模式下的性能,对包含不同重叠模式的句子展开实验,分析模型在不同模式下的性能表现。实验结果如表 7 所示,不同重叠模式下三元组抽取的 *F*1 如图 12 所示。表 7 的实验结果表明,在 SEO 和 EPO 模式下,RoBPtr 模型的性能均优于其他基线模型,精确率分别为 84.79%和 83.13%,召回率分别为 84.02%和 83.89%,*F*1 值分别为 84.44%和 83.51%。由图 12 可以看出,RoBPtr 模型在处理包含重叠三元组的句子时,*F*1 值相比所有基线模型都有显著提高,证明 RoBPtr 模型在实体关系联合抽取任务中,能够高效地识别并正确处理包含重叠现象的三元组。

表6 不同模型对包含不同三元组数量的句子的抽取性能

Table 6 Extraction performance of different models for sentences containing different number of triple 单位:%

Model	$N=1$			$N=2$			$N=3$			$N=4$			$N\geqslant 5$		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
CasRel	80.82	77.83	79.30	86.29	85.55	85.92	85.10	84.04	84.57	84.76	82.83	83.78	85.36	82.60	83.95
TPLinker(BiLSTM)	**86.36**	74.68	80.10	86.36	73.15	79.21	**83.64**	55.75	66.91	74.99	20.51	32.28	84.37	59.29	69.64
TPLinker(BERT)	82.00	77.03	79.44	**86.44**	81.36	83.83	85.31	71.12	77.62	80.27	23.81	36.73	76.97	72.27	74.54
TPLinker_plus	82.89	76.87	79.77	85.41	84.29	85.34	83.71	82.79	83.25	80.30	78.52	79.40	79.94	84.19	82.01
PRGC	78.04	78.64	78.34	85.01	80.50	82.69	85.81	81.12	83.40	**88.51**	87.64	**88.07**	86.17	**87.38**	86.77
TDEER	80.60	78.94	79.76	83.99	81.82	82.89	83.58	80.57	82.04	82.58	81.78	82.18	85.06	84.11	84.58
RoBPtr(本文)	80.52	**80.31**	**80.41**	85.51	**87.92**	**86.70**	85.03	**86.23**	**85.63**	85.20	**87.68**	86.43	**89.88**	84.52	**87.12**

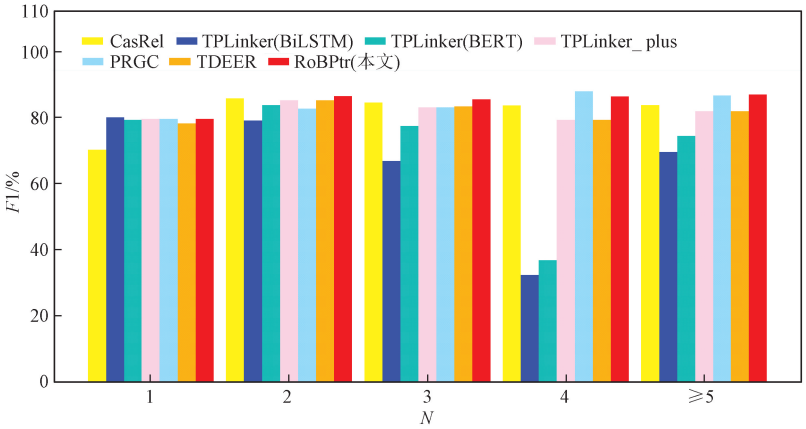

图11 包含不同三元组数的句子的 $F1$ 值对比

Figure 11 Comparison of $F1$ values for sentences containing different number of triple

表7 不同重叠模式的三元组抽取性能

Table 7 Performance of triple extraction with different overlap patterns 单位:%

模型	Normal			SEO			EPO		
	P	R	$F1$	P	R	$F1$	P	R	$F1$
CasRel	79.29	76.59	77.92	83.46	83.12	83.29	81.67	81.21	81.45
TPLinker(BiLSTM)	**87.02**	74.11	**80.12**	**85.71**	75.15	80.08	81.43	54.87	65.56
TPLinker(BERT)	81.50	74.11	79.22	82.17	82.20	82.19	76.00	77.47	76.73
TPLinker_plus	82.40	75.52	78.81	87.85	33.31	48.30	82.68	24.93	38.30
PRGC	76.40	76.08	76.24	84.37	82.21	83.28	81.20	75.75	78.38
TDEER	78.27	75.93	77.08	83.86	80.63	82.21	64.83	78.33	70.94
RoBPtr(本文)	80.40	**78.70**	79.54	84.79	**84.02**	**84.44**	**83.13**	**83.89**	**83.51**

注:Normal 表示无重叠;SEO 表示单实体重叠;EPO 表示实体对重叠。

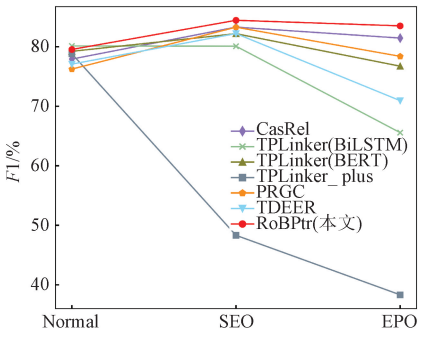

图12 不同重叠模式下三元组抽取的 $F1$ 值

Figure 12 $F1$ values for triple extraction in different overlap patterns

5 结论

本文针对中文实体与关系联合抽取任务中存在的三元组重叠问题,提出一种基于 RoBERTa 和指针网络的中文实体与关系联合抽取方法。首先,利用预训练语言模型 RoBERTa 作为嵌入层,以获取充分考虑上下文语义和位置信息的字符嵌入;其次,使用基于多层指针网络的实体识别模块,将实体识别任务建模为 token-pair 识别问题,通过识别实体的开始和结束位置来解析出所有可能的实体;最后,使用基于指针网络和多头注意力机制的关系抽取模块,把

三元组(s, r, o)抽取任务建模为五元组(s_h, s_t, r, o_h, o_t)识别问题。借助多头注意力机制为每个实体对的不同关系分配独立的概率子空间,实现对同一实体对多个关系的不互斥抽取。在 DuIE 数据集上,与所有基线模型相比,RoBPtr 模型的性能较佳,其精确率、召回率和 $F1$ 值分别达到 81.04%、85.82%和 83.36%。此外,本文还针对重叠模式(SEO、EPO)和不同复杂度的句子展开实验。实验结果显示,与基线模型相比,RoBPtr 模型在处理不同重叠模式和高复杂度的句子时,性能表现出色,进一步证明 RoBPtr 模型能够有效应对复杂场景和重叠模式下的实体与关系联合抽取任务。

参考文献:

[1] 陈宏, 陈新财, 巩晓赟, 等. 基于知识图谱的风电机组诊断系统构建与应用[J]. 郑州大学学报(工学版), 2023, 44(6): 54-60, 98.
CHEN H, CHEN X C, GONG X B, et al. Construction and application of wind turbine diagnosis system based on knowledge graph[J]. Journal of Zhengzhou University (Engineering Science), 2023, 44(6): 54-60, 98.

[2] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach [EB/OL]. (2019-07-26)[2025-07-08]. https://doi.org/10.48550/arXiv.1907.11692.

[3] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2002: 71-78.

[4] YU X F, LAM W. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg: ACL, 2010: 1399-1407.

[5] LI Q, JI H. Incremental joint extraction of entity mentions and relations[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2014: 402-412.

[6] MIWA M, SASAKI Y. Modeling joint entity and relation extraction with table representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1858-1869.

[7] WEI Z P, SU J L, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[EB/OL]. (2019-09-07)[2025-07-08]. https://doi.org/10.48550/arXiv.1909.03227.

[8] WANG Y C, YU B W, ZHANG Y Y, et al. TPLinker: single-stage joint extraction of entities and relations through token pair linking[EB/OL]. (2020-10-26)[2025-07-08]. https://doi.org/10.48550/arXiv.2010.13415.

[9] YAN Z H, ZHANG C, FU J L, et al. A partition filter network for joint entity and relation extraction[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2021, 185-197.

[10] ZHENG H Y, WEN R, CHEN X, et al. PRGC: potential relation and global correspondence based joint relational triple extraction[EB/OL]. (2021-06-18)[2025-07-08]. https://doi.org/10.48550/arXiv.2106.09895.

[11] LI X M, LUO X T, DONG C H, et al. TDEER: an efficient translating decoding schema for joint extraction of entities and relations[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2021: 8055-8064.

[12] SUI D B, ZENG X R, CHEN Y B, et al. Joint entity and relation extraction with set prediction networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(9): 12784-12795.

[13] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2025-07-08]. https://doi.org/10.48550/arXiv.1810.04805.

[14] GAO C, ZHANG X, LI L Y, et al. ERGM: a multi-stage joint entity and relation extraction with global entity match [J]. Knowledge-Based Systems, 2023, 271: 110550.

[15] LI R, LA K J, LEI J S, et al. Joint extraction model of entity relations based on decomposition strategy[J]. Scientific Reports, 2024, 14(1): 1786.

[16] 宋玲, 韦紫君, 陈燕, 等. 基于 RoBERTa 和指针网络的中文实体与关系联合抽取方法及系统: CN116663539A[P]. 2023-08-29.
SONG L, WEI Z J, CHEN Y, et al. Joint extraction method and system of Chinese entities and relations based on RoBERTa and pointer network: CN116663539A[P]. 2023-08-29.

[17] CUI Y M, CHE W X, 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.

[18] VINYALS O, FORTUNATO M, JAITLY N. Pointer networks[EB/OL]. (2015-06-09)[2025-07-08]. https://doi.org/10.48550/arXiv.1506.03134.

[19] 张强, 曾俊玮, 陈锐. 基于对比学习与梯度惩罚的实体关系联合抽取模型[J]. 吉林大学学报(理学版),

2024, 62(5): 1155-1162.
ZHANG Q, ZENG J W, CHEN R. Entity-relation joint extraction model based on contrastive learning and gradient penalty[J]. Journal of Jilin University (Science Edition), 2024, 62(5): 1155-1162.
[20] LI S J, HE W, SHI Y B, et al. DuIE: a large-scale Chinese dataset for information extraction[C]//Natural Language Processing and Chinese Computing Natural Language Processing and Chinese Computing: 8th CCF International Conference. Cham: Springer, 2019: 791-800.
[21] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[EB/OL]. (2019-09-26)[2025-07-08]. https://doi.org/10.48550/arXiv.1909.11942.
[22] CLARK K, LUONG M T, LE Q V, et al. ELECTRA: pre-training text encoders as discriminators rather than generators[EB/OL]. (2020-03-23)[2025-07-08]. https://doi.org/10.48550/arXiv.2003.10555.

Joint Extraction Method of Chinese Entities and Relations Based on RoBERTa and Pointer Network

CHEN Yan[1,2], WEI Zijun[2], LIAO Yuxiang[2], TAN Zhixiang[2], HU Xiaochun[3,4], SONG Ling[2]

(1. Guangxi Key Laboratory of Digital Infrastructure, Guangxi Zhuang Autonomous Region Information Center, Nanning 530201, China; 2. School of Computer and Electronic Information, Guangxi University, Nanning 530004, China; 3. Guangxi Key Laboratory of Finance and Economics Big Data, Guangxi University of Finance and Economics, Nanning 530003, China; 4. School of Big Data and Artificial Intelligence, Guangxi University of Finance and Economics, Nanning 530003, China)

Abstract: To effectively solve the problem of triple overlap in the joint extraction of entities and relations in unstructured text. A Chinese entity and relation joint extraction method was proposed based on RoBERTa and pointer network. Firstly, for the entity overlap problem, an entity recognition module was based on the pointer network, and the entity recognition task was constructed as a token-pair recognition problem, which extracted designed all possible entities by recognizing the start and end positions of the entities. Secondly, for the triplet overlap problem, a relation extraction module was designed based on the multi-head attention mechanism and Ptr-Net to construct the triple (s, r, o) extraction task as a quintuple (s_h, s_t, r, o_h, o_t) identification problem. Finally, extensive experiments on the Chinese information extraction dataset DuIE showed that the comprehensive performance of the proposed model was better than all baseline models, with the precision, recall and $F1$ values of 81.04%, 85.82% and 83.36% respectively.

Keywords: entity and relation joint extraction; RoBERTa; pointer network; natural language processing; deep learning