文章编号：2096-1618(2025)03-0257-07

融合实体特征和潜在关系的中文关系抽取模型

王 靖[1,2]，余 艳[1,2]，熊 熙[1,2]

(1.成都信息工程大学网络空间安全学院，四川 成都 610225；2.先进密码技术与系统安全四川省重点实验室，四川 成都 610225)

摘要：从非结构化文本中抽取关系三元组信息对于构建知识图谱尤为重要，现有的研究方法通常以先识别实体再抽取关系为主。尽管这些方法取得了良好的性能，但忽略了实体与关系之间的内在联系，且无法有效解决同一文本中实体重叠问题。针对以上问题，提出一种融合实体特征和潜在关系的中文关系抽取模型，以关系作为条件通过主实体映射客实体。首先将实体信息以二维矩阵方式进行标记，进行主实体识别；然后预测出文本可能存在的关系；最后融合实体特征和潜在关系信息进行客实体识别。整个过程采用双向关系三元组抽取框架，即从两个方向上进行关系三元组的抽取，使其双向抽取结果相互补充。该模型有效保留了实体与关系之间的内在联系，增强了对重叠实体的关系识别。实验结果表明，在 DuIE 和 CMeIE 中文数据集上，提出的模型在精确率、召回率和 F1 评测指标上均有一定的提升，证明该模型的有效性。

关 键 词：关系抽取；实体重叠；潜在关系；双向三元组抽取；实体特征

中图分类号：TP391.1 **文献标志码**：A

doi：10.16836/j.cnki.jcuit.2025.03.001

0 引言

关系抽取是信息抽取[1]中的一项子任务，旨在从非结构化的自然语言文本[2]中抽取出关系三元组。这些关系三元组以[主体、关系、客体]的形式存储着事实性知识，其中主体和客体都代表着实体，在语义上通过关系连接起来。例如，三元组[西游记，作者，吴承恩]表示“西游记的作者是吴承恩”。针对关系抽取任务，在早期阶段，提出用流水线方法来解决，它主要包括两个步骤：命名实体识别[3]和关系预测[4]，即首先识别出输入文本中的所有实体，然后预测实体对之间所有存在的关系。然而，这样的方法存在明显的误差传递问题。因为实体识别和关系预测相互独立，忽略了实体识别和关系预测之间的内在联系，如果在实体识别阶段就出现错误和遗漏，那么这种误差在关系预测阶段是无法被纠正和改变的，会导致错误信息传递到关系预测阶段，直接影响到关系抽取的效果。所以，之后的研究工作开始探索一种端到端的联合实体关系抽取方法，旨在同时解决实体识别和关系预测这两个任务。它不同于流水线方法，联合实体关系抽取方法主要将实体抽取和关系预测通过一定的方式进行整合处理，联合学习两个任务，构建端到端的关系抽取模型，以此减少实体识别和关系预测的误差传递问题，提高模型的性能。例如，基于跨度的联合实体关系抽取方法[5]等，取得了较好的成果，并证明了联合实体关系抽取的有效性。尽管这些方法能够提高实体识别和关系预测的内在联系，但会因为训练样本的不足造成参与共享同一实体的多种关系很难被正确识别出来，从而导致在某些复杂的实体重叠问题上仍然达不到理想的效果。实体重叠主要包含：单实体重叠(SEO)表现为同一个实体可能存在于多种关系三元组上；实体对重叠(EPO)表现为同一对实体可能包含多种关系；主客实体重叠(SOO)表现为主实体和客实体相互包含。3种实体重叠的具体区别如表1所示。为解决实体重叠问题，本文提出一种融合实体特征与潜在关系的中文关系抽取模型(IEFALR)。首先预测出文本的主实体信息，然后预测文本可能存在的潜在关系，最后根据主实体和潜在关系信息预测出客实体以组成关系三元组。

表1 实体重叠

类别	文本	三元组
正常	老舍是《骆驼祥子》的作者	(骆驼祥子，作者，老舍)
SEO	【孙悟空】和【猪八戒】是【西游记】中的角色	(西游记，角色，孙悟空) (西游记，角色，猪八戒)
EPO	【北京】是【中国】的首都	(中国，首都，北京)(中国，包含，北京)
SOO	【【北京】大学】是一所综合性大学	(北京大学，地点，北京)

在主实体抽取阶段，引入二维矩阵的实体标注方式，更细致地标记出文本中的实体信息，以加强重叠实

收稿日期：2023-10-31

基金项目：国家自然科学基金资助项目(81901389)；四川省科技计划资助项目(23ZDYF2088)；教育部人文社会科学研究基金资助项目(22YJAZH120)

通信作者：余艳. E-mail：yuyan@cuit.edu.cn

体的识别。同时,预先预测出文本存在的潜在关系,并融合实体特征和潜在关系信息进行客实体识别,增强实体与关系之间的内在联系,也避免因为关系类别过多对文本中未存在的关系进行实体预测而产生的冗余操作。最后采用双向关系三元组抽取框架,即从两个方向上抽取关系三元组,这样可以使双向的抽取结果相互补充,以得到更加准确的关系三元组信息。实验表明,该方法具有一定的有效性,并提高了对重叠实体关系抽取的性能。

本文主要工作如下:

(1) 预测文本可能存在的关系,并融合实体特征与潜在关系的信息,以此增强关系抽取过程中实体与关系之间的内在联系。(2) 构建二维矩阵标记实体方案和双向三元组抽取框架,解决实体重叠问题以提高关系三元组抽取的性能。(3) 在 DuIE 和 CMeIE 中文数据集上验证 IEFALR 模型的有效性。

1 相关工作

早期的关系抽取方法主要采用流水线的方法识别文本中存在的关系三元组,该方法主要是将实体识别和关系预测作为两个相互独立的任务。首先,对文本进行实体识别,然后将识别出的实体再进行关系分类,以此得到关系三元组。例如,关系抽取核方法[6]以及利用句法语义结构进行关系抽取[7]都是采用流水线的方式来实现关系三元组抽取。尽管这些流水线的抽取方法取得了较好的效果,但存在较为严重的误差传递问题,即实体识别阶段出现错误,会直接对关系预测阶段造成影响,并且它还忽略了实体识别与关系预测任务之间的内在联系,直接影响到关系抽取的性能。为解决这个问题,研究工作大多都从联合关系抽取角度出发。例如,用表格表示法建模关系实体及关系抽取[8]以及类型化实体和关系与知识库的联合抽取[9],该方法主要采用基于特征的关系抽取模型,但抽取步骤都过于烦琐,不利于进行高效的关系抽取。此后,出现一种新标注方案的联合实体关系抽取[10],将联合抽取任务转换为序列标记任务。虽然这些方法一定程度上解决了误差传递问题,但无法有效地解决实体重叠问题。

近几年,研究者围绕实体重叠问题,做了许多工作,探索出一种用作关系抽取的二进制标记框架[11]以及基于一种新的分解策略的实体与关系联合抽取方法[12]。首先抽取所有主实体,然后基于抽取的主实体同时抽取客实体和关系。结果表明,此方法取得了有效的成果,提高了包含实体重叠的三元组或多种关系三元组的句子的抽取能力。实际上,这些方法将重叠关系的抽取分解成几个内部依赖步骤,因为解码器需要递归解码过程,并且级联标记必须提前识别主实体。虽然将抽取过程经过分解后使任务易于进行,但在主实体预测阶段发生错误,必然影响后续任务的进行,导致误差累计影响关系抽取。之后的研究中还有一些其他的关系抽取方案,例如,提出一个统一的框架来联合抽取显式和隐式关系三元组[13],采用一个二元指针网络,以顺序方式提取与每个单词相关的重叠关系三元组,并在外部内存中保留先前提取的三元组信息,以此加强对文本关系三元组的识别。还有一种从立体的角度进行关系抽取[14],将关系三元组映射到一个三维(3D)空间,并利用三个解码器来提取任务,旨在同时处理信息丢失、错误传播等问题。这些方法都为关系抽取提供了许多方向以及解决方案。

2 本文模型

融合实体特征和潜在关系的中文关系抽取模型架构如图 1 所示。

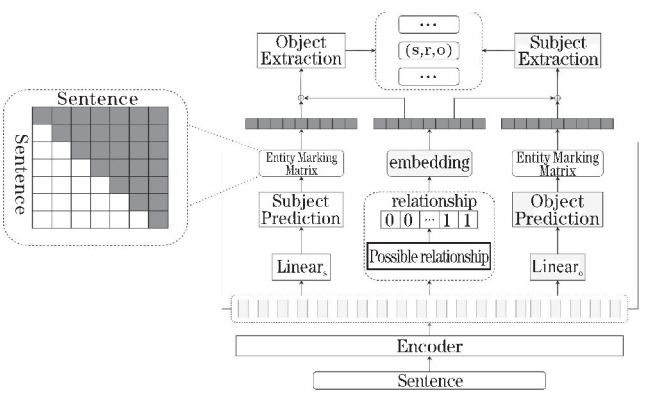

图 1 融合实体特征与潜在关系的中文关系抽取模型(IEFALR)

2.1 文本预处理

中文句子中如果通过分词手段来分割句子将无法有效解决实体重叠问题,例如,“北京大学是一所综合性大学”,通过分词器无法确定是分为“北京大学”一个词,还是应该分为“北京”和“大学”两个词,那么在中文句子的预处理上需要逐个字进行分割,使得在后续数据标记阶段可以更好地区分一些重叠实体数据。对于分割好的序列,以 $S=\{c_1,c_2,\cdots,c_n\}$ 表示,其中 n 为以字为单位的文本长度。

2.2 编码层

首先使用预训练模型 BERT[15]生成文本中各个字符的特征表示,然后接入双向 LSTM[16]以加强对文本的上下文理解以及语义表示,最后得到的输出即为文本中各个字符的特征向量:

$$\boldsymbol{Y}_{\text{enc}}=\text{BERT}_{\text{bilstm}}(S) \tag{1}$$

其中,$\boldsymbol{Y}_{\text{enc}}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n \mid \boldsymbol{h}_i\in R^{d\times 1}\}$,$\boldsymbol{h}_i$ 为字符的特征表示向量,d 为特征维度大小。

2.3　主实体预测

该模块是正向关系抽取中的主实体识别模块。在三元组关系中,主实体通常被认为是关系的起点或主要参与者,而客实体则是与主实体相关联的第二个实体。主实体通常在关系中扮演更重要的角色,它是关系的焦点或核心。所以,为更好区分两者,需要先对 Encoder 输出的特征向量经过线性转换,使它们采用不同的特征,如下所示:

$$\boldsymbol{h}_i^S=\boldsymbol{W}_s\boldsymbol{h}_i+\boldsymbol{b}_s \tag{2}$$

其中,$\boldsymbol{W}_s\in R^{d_h\times d_h}$是一个可训练矩阵,$\boldsymbol{b}_s\in R^{d_h}$是一个偏置向量。得到主实体的特征向量后,经过 Subject Prediction 模块,具体内容为将 $\boldsymbol{h}_i^s$ 作为头和尾同时输入到 Biaffine[17] 中。由于在关系抽取任务中,只需要识别实体,不需要识别实体类型,因此在 Biaffine 后只需要得到一个实体二维矩阵输出即可,如下所示:

$$\boldsymbol{M}_s=\sigma(\text{Biaffine}_s(\boldsymbol{h}_1^s,\boldsymbol{h}_2^s,\cdots,\boldsymbol{h}_n^s)) \tag{3}$$

式中,$\boldsymbol{M}_s\in R^{n\times n}$表示二维矩阵,$n$ 为句子长度,σ 为 sigmoid 激活函数。最后,利用 Entity Marking Matrix 方法来获取对应可能的实体。

同理,Object Prediction 为反方向的抽取,相对应的特征向量以及客实体矩阵分别由 $\boldsymbol{h}_i^o$ 和 $\boldsymbol{M}_o$ 表示。该模块主要抽取客实体来进行后续操作,与 Subject Prediction 结构相似,这里不再做过多阐述。

2.4　实体标记矩阵

在对各个实体进行标注时,主要采用矩阵的标注方式。例如,“北京是中国的首都”,包含实体“北京”和“中国”,以实体头部为矩阵的行,实体尾部为矩阵的列,那么可以将“中”作为行,“国”作为列的位置标注为 1,具体方式如图 2 所示。为更好解释实体标注问题,这里对主客实体都进行了标注,在实际情况中会针对主实体与客实体单独标注。

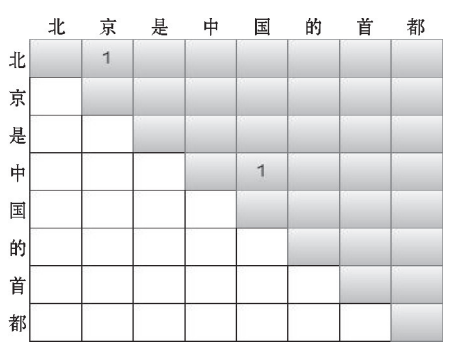

图 2　实体标注矩阵

在处理上一步输出的 $\boldsymbol{M}_s$ 时,为其设定一个阈值,对该矩阵中每一个参数大于该阈值的标记为 1,否则标记为 0,以此来得到实体标记。然而,这样的方式会导致矩阵过于稀疏。因为矩阵的上三角部分才是真正标注实体的位置,而下三角部分不会用于实体的标注。为解决稀疏性问题,将矩阵下半部分舍去,只计算矩阵对角线以及上三角部分。同时,为增加训练过程的计算效率,将矩阵上三角部分按照每行从左到右的顺序进行平铺得到一维特征向量,以 $\boldsymbol{M}_{\text{fla}}^s\in R^{\frac{(1+n)n}{2}\times 1}$ 表示,客实体平铺矩阵表示同理。

2.5　潜在关系

该模块为潜在关系的预测,即给定一个句子,首先预测句子中可能存在的潜在关系的子集,然后只需要针对这些潜在关系进行实体抽取,防止冗余地对每个关系都进行实体抽取。具体步骤为:对 Encoder 输出的句子特征向量采用平均池化操作,最后经过一些线性层抽取潜在关系特征:

$$\boldsymbol{h}_{\text{avg}}=\text{Avgpool}(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n) \tag{4}$$

$$\boldsymbol{P}_{\text{rel}}=\sigma(\boldsymbol{W}_{\text{r}}\boldsymbol{h}_{avg}+b_{\text{r}}) \tag{5}$$

式中,Avgpool 是平均池化操作[18],$\boldsymbol{W}_{\text{r}}\in R^{d\times 1}$是可训练权重矩阵,$\sigma$ 表示 sigmoid 函数。首先,在输出关系上,将其建模为多标签二进制分类任务,即设定一个关系分类阈值,如果概率超过这个阈值,则对应关系的索引将被分配为标签 1,否则将被分配为标签 0,见图 1 中 relationship。然后,将所有标签 1 的关系经过 Embedding 层进行特征计算。最后,得到各个潜在关系的特征表示,$\text{Emb}_{\text{r}}=\{\text{r}_1^{\text{emb}},\text{r}_2^{\text{emb}},\cdots,\text{r}_m^{\text{emb}} \mid \text{r}_i^{\text{emb}}\in R^{d\times 1}\}$,$m$ 为潜在关系数量。

2.6　基于主实体的客实体预测

该模块根据得到的主实体以及潜在关系进行客实体的预测。首先,逐个将 $\boldsymbol{M}_s$ 中标签为 1 的部分,以行索引值为 start,列索引值为 end,对句子特征向量 $\boldsymbol{h}_i$ 做平均池化得到主实体特征表示,如下所示:

$$\boldsymbol{p}_S^{\text{start,end}}=\text{Avgpool}(\boldsymbol{h}_{\text{start}},\cdots,\boldsymbol{h}_{\text{end}}) \tag{6}$$

其中 $\boldsymbol{p}_S^{\text{start,end}}\in R^{d\times 1}$。然后,融合主实体特征与关系特征到句子特征上,再经过线性层进行特征提取。最后,经过 Biaffine 得到对应客实体二维特征矩阵:

$$\boldsymbol{h}_i^{\text{s\_r\_conact}}=[\boldsymbol{h}_i;\boldsymbol{p}_s^{\text{start,end}};r_j^{\text{emb}}] \tag{7}$$

$$\boldsymbol{h}_{\text{i}}^{\text{s\_r}}=\boldsymbol{W}_{\text{s\_r}}\boldsymbol{h}^{\text{s\_r\_conact}}+\boldsymbol{b}_{s\_r} \tag{8}$$

$$\boldsymbol{M}_o^{\text{s\_r}}=\sigma(\text{Biaffine}_o^{\text{s\_r}}(\boldsymbol{h}_1^{\text{s\_r}},\cdots,\boldsymbol{h}_n^{\text{s\_r}})) \tag{9}$$

式中 $\boldsymbol{W}_{\text{s_r}}\in R^{3d_h\times d_h}$是可训练权重,$\sigma$ 表示 sigmoid 函数,$\boldsymbol{M}_{\text{s_r}o}\in R^{d\times d}$表示二维输出矩阵。该二维矩阵同样采用 Entity Marking Matrix 标注方法,在取得标签 1 的客实体信息,最终组成关系三元组。同时,将其平铺后的向量以 $\boldsymbol{M}_{\text{fla}}^{\text{s\_r\_o}}\in R^{\frac{(1+n)n}{2}\times 1}$ 表示。同理,反向的抽取方法中 Subject Extraction 与之类似。

2.7 训练策略

训练期间采用共同训练模型各个模块方式,并且共享 Encoder 参数。由于实体标记以及潜在关系预测都是采用二进制的多标签标记方式,因此损失函数主要采用 Binary CrossEntropy,其中潜在关系预测模块损失函数如下:

$$L_{\mathrm{r}}=-\frac{1}{n_{\mathrm{r}}}\sum_{i=1}^{n_r}\left(y_i\lg P_{\mathrm{rel}}+(1-y_i)\lg(1-P_{\mathrm{rel}})\right) \tag{10}$$

其次,在计算二维实体标记矩阵的损失函数的过程中,由于矩阵稀疏问题,导致数据不平衡,难以训练。为了解决此问题,在原有损失函数基础上加入了 Focal loss[19]。以主实体识别模块为例,具体如下:

$$\mathrm{BCE}_{\mathrm{sub}}=-\frac{1}{n_{\mathrm{fla}}}\sum_{i=1}^{n_{\mathrm{fla}}}\left(y_i\lg M_{\mathrm{fla}}^{s}+(1-y_i)\lg(1-M_{\mathrm{fla}}^{s})\right) \tag{11}$$

$$L_{\mathrm{sub}}=\alpha_t(1-\mathrm{e}^{-\mathrm{BCE}_{\mathrm{sub}}})^{\gamma}\mathrm{BCE}_{\mathrm{sub}} \tag{12}$$

其中 α_t 为控制正负样本的权重,γ 为调节因子。同理,其余 3 个模块损失函数与之类似,分别以 $L_{\mathrm{obj}}^{\mathrm{ex}}$,$L_{\mathrm{obj}}$,$L_{\mathrm{sub}}^{\mathrm{ex}}$表示。最后,总损失函数为所有模块损失总和:

$$L_{\mathrm{total}}=L_{\mathrm{r}}+L_{\mathrm{sub}}+L_{\mathrm{obj}}^{\mathrm{ex}}+L_{\mathrm{obj}}+L_{\mathrm{sub}}^{\mathrm{ex}} \tag{13}$$

3 实验及结果分析

3.1 数据集与评估指标

数据集采用两种公开的中文数据集,包括百度提供关系抽取数据集 DuIE[20] 和中文医学信息抽取数据集 CMeIE[21]。具体信息如表 2 所示。

表 2 数据集信息

类型	DuIE		CMelE	
	训练集	测试集	训练集	测试集
正常	38321	6919	3857	1015
SEO	97504	12253	8842	2160
EPO	12878	1591	259	64
SOO	7228	876	1481	346
总数	155931	21639	14439	3585

DuIE 是一个大规模的高质量中文信息抽取数据集。该数据集包含超过 21 万个句子,43 万个关系三元组实例,以及 50 种预定义的关系类型,这些句子和实例涵盖了现实世界应用程序中的各种领域,例如新闻、娱乐、用户生成的内容。

CMeIE 是一个用于医学关系抽取任务的数据集。该数据集包含从儿科和百余种常见疾病中收集的2.8万个疾病语句和近7.5万个关系三元组数据,共标注了53 个定义的关系类型。

评估指标采用精确率(P)、召回率(R)和 F1 值,分别如下:

$$\mathrm{P}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{14}$$

$$\mathrm{R}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{15}$$

$$\mathrm{F1}=\frac{2\times\mathrm{P}\times\mathrm{R}}{\mathrm{P}+\mathrm{R}} \tag{16}$$

其中,TP(True Positive)表示模型正确预测的正样本数量,FN(False Negative)表示实际为正样本但被错误预测为负样本的数量,FP(False Positive)表示实际为负样本但被错误预测为正样本的数量。

3.2 模型参数设置与实验环境

采用 BERT base Chinese 作为预训练语言模型,句子最大长度设定 128,潜在关系预测以及实体标记分类概率阈值设定为 0.5,Batch Size 设置为 16,学习率为 1E5,设置 epochs 大小为 100。在训练过程中加入 warmup 策略和 early stop 以加快模型收敛速度,提高泛化能力,避免过拟合的风险。实验环境使用 Linux 操作系统,采用 Pytorch 深度学习框架,模型训练在 A40 GPU 上完成。

3.3 实验结果评估与分析

本文使用以下模型和所提出的融合实体特征及潜在关系的中文关系抽取模型进行对比。

CopyMTL[22]:在编码阶段主要采用 BiLSTM 模型来捕获句子的整体信息,以构建句子的特征向量。解码阶段采用 Attention_LSTM 模型来逐步预测和识别关系、头实体和尾实体。

CopyRE+RL[23]:该模型在 CopyRE 模型的基础上,引入了强化学习方法,将三元组的生成过程类比与强化学习过程,并采用 REINFORCE 算法优化模型,使模型预测三元组更加高效准确。

CasRel[24]:该模型采用一种创新的级联二元标注框架,将三元组抽取任务分解为主实体、关系和对象实体三个级别的问题。同时,将关系抽取视为两个实体之间的映射关系,从而有效解决句子中三元组重叠的挑战。这种方法为实体关系联合抽取提供了一种新的思路,并在性能上取得了显著的提升。

Seq2UMTree[25]:该模型在 CasRel 模型基础上做出了改进,引入一种树形结构的解码层,用于实现句子中实体关系的联合抽取。这一改进解决了数据中标签不平衡的问题,提高模型的鲁棒性和性能。通过使用树形结构,该模型能够更好地处理实体关系的复杂性,并为实体关系联合抽取任务带来新的突破。

FETI[26]:该模型首先利用 BiLSTM 模型在编码层捕获输入句子的全局特征,以便更好地理解输入句子的上下文信息。接着,在解码层采用树形结构来进行句子中实体和关系的联合抽取。这一设计策略成功地提升了模型的性能表现。

在 DuIE 数据集上,本文提出的融合实体特征和潜在关系的中文关系抽取模型与对比模型的性能对比见表 3和图 3。可见,融合实体特征和潜在关系的中文关系抽取模型在 F1 得分上相比于最好的模型提升了 1.9,证明该模型在处理关系抽取上的有效性,以及从两个方向上抽取实体关系的优势。

表 3　各个模型在 DuIE 数据集上的结果　　单位:%

模型	P	R	F1
CopyMTL	49.9	39.4	43.9
CopyRE+RL	76.2	79.8	77.9
CasRel	80.1	77.2	78.6
Seq2UMTree	75.6	73.0	74.3
FETI	75.7	76.0	75.8
本文模型	**81.2**	**79.9**	**80.5**

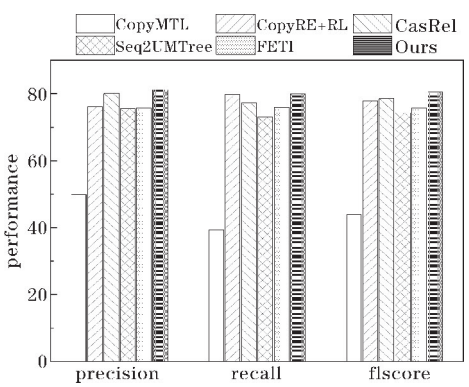

图 3　各个模型在 DuIE 数据集上对比

在 CMeIE 数据集上,本文提出的融合实体特征和潜在关系的中文关系抽取模型与对比模型的性能对比见表 4 和图 4。由于 CMeIE 医疗数据涉及多种医疗的专有名词,再加上 Bert 是在大规模的通用语料库上训练得来的,导致模型无法得到更好的训练,因此 CMeIE 数据集上的效果明显低于 DuIE 数据集效果。不过,本文提出的模型与其他的模型相比有一定的提升,在 F1 得分上比最好的模型提高了1.1。

表 4　各个模型在 CMeIE 数据集上的结果　　单位:%

模型	P	R	F1
CopyMTL	31.7	26.8	29.1
CopyRE+RL	54.0	55.7	54.6
CasRel	56.3	56.7	56.5
Seq2UMTree	46.6	37.4	41.5
FETI	48.8	40.3	44.1
本文模型	**57.7**	**57.6**	**57.6**

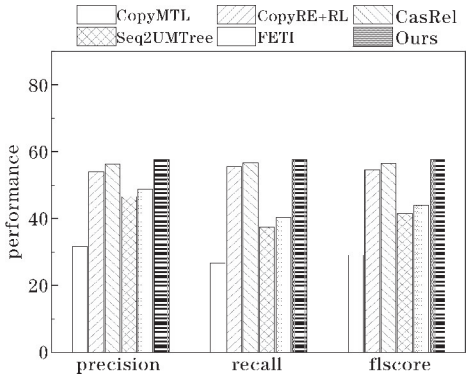

图 4　各个模型在 CMeIE 数据集上对比

为验证模型在实体关系抽取中实体重叠问题上的效果,选取 CopyR+RL 和 CasRel 模型,在 DuIE 数据集上分别进行了单实体重叠、实体对重叠和主客实体重叠 3 种情况的对比实验,以验证其有效性。其中,单实体重叠对、实体对重叠和主客实体重叠对比结果如表 5 ~7所示。

表 5　DuIE 数据集上 SEO 对比结果　　单位:%

模型	P	R	F1
CopyR+RL	71.3	69.4	70.3
CasRel	82.6	80.5	81.5
Ours	**83.1**	**81.7**	**82.3**

表 6　DuIE 数据集上 EPO 对比结果　　单位:%

模型	P	R	F1
CopyRRL	76.9	77.4	77.1
CasRel	**80.5**	74.3	77.3
Ours	79.7	**78.3**	**78.9**

表 7　DuIE 数据集上 SOO 对比结果　　单位:%

模型	P	R	F1
CopyRRL	60.3	62.8	61.5
CasRel	62.9	64.3	63.6
Ours	**63.2**	**64.5**	**63.8**

根据各个实体重叠问题对比实验,可发现本文模型在 EPO 上的精确率略低于 CasRel 模型,这是因为潜在关系预测与主实体预测独立进行,可能会存在预测出的关系类别与主实体不相关的情况,进而影响结果。但从整体上来看本文模型相比于其他模型都有一定的提升,表明了该模型在解决实体重叠问题上有一定的可行性。这是由于采用二维矩阵的实体标记方式能够针对所有重叠实体进行标注,并在潜在关系的作用下,将标注得到的实体逐个映射到与之对应的关系三元组上,从而解决实体重叠问题。

3.4　消融实验

为验证本文提出的方法中潜在关系和双向抽取对

性能的影响,提出了两个变体:变体 1,将潜在关系预测替换为对所有关系进行预测;变体 2,去掉反向三元组抽取,只根据正向抽取预测三元组。

将两个变体在 DuIE 数据集上做消融研究,实验结果如表 8 所示。在变体 1 上可以看到各个指标分数都有一定的下降,这是因为在关系抽取中进行了许多冗余的预测,导致模型预测出了一些不存在的关系,降低了模型的识别效果。在变体 2 上将反向预测去除后,由于正向预测可能会预测不够充分,使得模型预测准确性有所降低。

表 8 DuIE 数据集上的消融实验 单位:%

模型	P	R	F1
本文模型	**81.2**	**79.9**	**80.5**
变体 1	79.8	78.4	79.1
变体 2	79.7	78.1	78.9

经过消融实验,印证了融合潜在关系预测以及加入反向预测模型都对模型识别效果有一定的提升,从而证明了该两个模块的有效性。

4 结束语

针对关系三元组抽取中的实体重叠问题,提出一种融合实体特征和潜在关系的中文关系抽取模型框架。以关系作为条件通过主实体映射客实体。在实体识别阶段,采用二维矩阵实体标记方式有效地解决关系抽取中存在的重叠实体识别不充分问题。同时,融合潜在关系与实体特征,加强实体与关系之间的内在联系,也避免对文本中未存在的关系进行实体预测而产生的冗余操作。最后,在中文关系数据上验证了该方法的有效性。下一步将在各个特定领域,如金融、医疗领域下测试该模型的性能以及泛化能力。

参考文献:

[1] 郭喜跃,何婷婷. 信息抽取研究综述[J]. 计算机科学,2015,42(2):14-17+38.

[2] 王江鹏. 基于深度学习的自然语言处理技术发展分析[J]. 中国安防,2022,201(12):40-43.

[3] Ratinov L,Roth D. Design challenges and misconceptions in named entity recognition[C]. CoNLL ´09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. 2009.

[4] Wang Z,Wen R,Chen X,et al. Finding Influential Instances for Distantly Supervised Relation Extraction[C]. International Conference on Computational Linguistics. ICCL,2022:2639-2650.

[5] Eberts M,Ulges A . Span-based Joint Entity and Relation Extraction with Transformer Pre-training [J]. arXiv,2019.

[6] Zelenko D,Aone C,Richardella A. Kernel Methods for Relation Extraction[C]. Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2002.

[7] Chan Y S,Roth D . Exploiting Syntactico-Semantic Structures for Relation Extraction[C]. The 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies, Proceedings of the Conference,19-24 June,2011, Portland, Oregon, USA. Association for Computational Linguistics,2011.

[8] Miwa M,Sasaki Y . Modeling Joint Entity and Relation Extraction with Table Representation[C]. Conference on Empirical Methods in Natural Language Processing. 2014.

[9] Ren X,Wu Z,He W,et al. CoType:Joint Extraction of Typed Entities and Relations with Knowledge Bases[C]. International conference on world wide web. ACM,2017:1015-1024.

[10] Zheng S,Wang F,Bao H,et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme[C]. Association for Computational Linguistics. ACL,2017:1227-1236.

[11] Wei Z,Su J,Wang Y ,et al. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[12] Yu B,Zhang Z,Shu X ,et al. Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy[J]. arXiv,2019.

[13] Chen Y,Zhang Y,Hu C ,et al. Jointly Extracting Explicit and Implicit Relational Triples with Reasoning Pattern Enhanced Binary Pointer Network [C]. North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics,2021.

[14] Xuetao Tian,Liping Jing,Lu He,Feng Liu. StereoRel:Relational Triple Extraction from a Stereoscopic Perspective[C]. International Joint Conference on Natural Language Processing; Annual Meeting of the Association for Computational Linguistics. 2021.

[15] Devlin J,Chang M W,Lee K ,et al. BERT:Pre-

training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv,2018.

[16] 石文浩,孟军,张朋,等. 融合 CNN 和 Bi-LSTM 的 miRNA-lncRNA 互作关系预测模型[J]. 计算机研究与发展,2019,56(8):1652-1660.

[17] Bohnet B,Yu J,Poesio M. Named Entity Recognition as Dependency Parsing[C]. Association for Computational Linguistics. ACL,2020:6470-6476.

[18] Lin M,Chen Q,Yan S. Network In Network[Z]. International Conference on Learning Representations. ICLR,2014.

[19] Lin T Y,Goyal P,Girshick R ,et al. Focal Loss for Dense Object Detection[J]. arXiv e-prints,2017.

[20] 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. Baidu Inc. Beijing 100193,China,2019.

[21] Guan T,Zan H,Zhou X,et al. CMeIE:Construction and Evaluation of Chinese Medical Information Extraction Dataset[C]. Natural Language Processing and Chinese Computing:9th CCF International Conference. NLPCC,2020:272-282.

[22] Zeng D,Zhang H,Liu Q . CopyMTL:Copy Mechanism for Joint Extraction of Entities and Relations with Multi-Task Learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence,2020,34(5):9507-9514.

[23] Zeng X,He S,Zeng D ,et al. Learning the Extraction Order of Multiple Relational Facts in a Sentence with Reinforcement Learning[C]. Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2019.

[24] Wei Z,Su J,Wang Y ,et al. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[25] Zhang R H,Liu Q,Fan A X,et al. Minimize Exposure Bias of Seq2Seq Models in Joint Entity and Relation Extraction[C]. Findings of the Association for Computational Linguistics. EMNLP,2020:236-246.

[26] 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):8.

Integrating Entity Features and Latent Relation Model for Chinese Relation Extraction

WANG Jing[1,2], YU Yan[1,2], XIONG Xi[1,2]

(1. College of Cybersecurity,Chengdu university of Information Technology,Chengdu 610225,China;2. Advanced Cryptography and System Security Key Laboratory of Sichuan Province,Chengdu 610225,China)

Abstract:Extracting relationship triad information from unstructured text is especially important for constructing knowledge graphs,and existing research methods usually focus on identifying entities before extracting relationships. Although these methods have achieved good performance,they ignore the intrinsic connection between entities and relations,and cannot effectively solve the problem of overlapping entities in the same text. To address the above problems,a Chinese relationship extraction model that integrates entity features and potential relationships is proposed,and the main idea is to map the guest entities through the main entities with the relationships as the conditions. The main idea is to map the guest entities by using the relationships as conditions. Firstly,the entity information is labeled in a two-dimensional matrix to recognize the main entities; then the possible relationships in the text are predicted; finally,the entity features,and potential relationship information are fused to recognize the guest entities. The whole process adopts a bidirectional relationship ternary extraction framework,i. e. ,the relationship ternary is extracted from two directions,so that the bidirectional extraction results are complementary to each other. The model effectively preserves the intrinsic connection between entities and relationships and enhances the relationship recognition of overlapping entities. The experimental results show that the model proposed in this paper has some improvement in precision rate,recall rate,and F1 evaluation metrics on DuIE and CMeIE Chinese datasets,which proves the effectiveness of the model.

Keywords:relation extraction;overlap entity;potential relationships;bidirectional triplet extraction;entity feature