doi:10.3969/j.issn.1673-9833.2022.05.007

基于位置辅助标记的实体关系抽取方法

欧阳康[1,2]，朱艳辉[1,2]，张 旭[1,2]，孔令巍[1,2]，黄雅淋[1]，金书川[1,2]，沈加锐[2]

（1. 湖南工业大学 计算机学院，湖南 株洲 412007；2. 湖南省智能信息感知及处理技术重点实验室，湖南 株洲 412007）

摘 要：为了解决在抽取过程中出现的关系三元组重叠问题，提出了一种基于位置辅助标记的实体关系联合抽取模型，使用 BERT 作为预训练语言模型，并且通过位置辅助矩阵方法，将关系三元组抽取转换成实体和关系的匹配问题，实现实体和关系的联合抽取，在中文数据集 DuIE 上进行了相关实验。实验结果表明，该模型抽取效果较好，提出的基于位置的辅助标记方法有效解决了关系重叠问题。

关键词：实体关系抽取；重叠关系三元组；位置辅助标记

中图分类号：TP391.1 **文献标志码**：A **文章编号**：1673-9833(2022)05-0050-06

引文格式：欧阳康，朱艳辉，张 旭，等 . 基于位置辅助标记的实体关系抽取方法 [J]. 湖南工业大学学报，2022，36(5)：50-55.

An Entity Relationship Extraction Method Based on Location-Aided Marking

OUYANG Kang[1,2]，ZHU Yanhui[1,2]，ZHANG Xu[1,2]，KONG Lingwei[1,2]，
HUANG Yalin[1]，JIN Shuchuan[1,2]，SHEN Jiarui[2]

（1. College of Computer Science，Hunan University of Technology，Zhuzhou Hunan 412007，China；2. Key Laboratory of Intelligent Information Perception and Processing Technology of Hunan Province，Zhuzhou Hunan 412007，China）

Abstract：In view of a solution of relation triplet overlapping in the extraction process, a joint extraction model of entity relation based on location-aided marking has thus been proposed. Taking BERT as the pre-trained language model, by means of position-aided matrix, the relation triplet extraction is transformed into the matching of entity and relation, thus realizing the joint extraction of entity and relation, with relevant experiments to be carried out on the Chinese data-set DuIE. The experimental results show that the model exhibits an improved performance in extraction, showing that the proposed location-aided marking method can effectively solve the problem of overlapping.

Keywords：entity relationship extraction；overlapping relationship triplet extraction；location-aided marking method

1 研究背景

近年来，各大网络社交及传媒平台带来的信息数不胜数，如何快速、精准地获取其中的关键信息是一个巨大的挑战。信息抽取随之诞生，关系抽取（relation extraction，RE）作为信息抽取的子任务，

收稿日期：2021-10-29

基金项目：湖南省自然科学基金资助项目（2020JJ6089）；湖南省教育厅科研基金资助重点项目（19A133）

作者简介：欧阳康（1998-），男，湖北孝感人，湖南工业大学硕士生，主要研究方向为自然语言处理和知识工程，
E-mail：ouyang982020@163.com

通信作者：朱艳辉（1968-），女，湖南湘潭人，湖南工业大学教授，主要研究方向为自然语言处理与知识工程，
E-mail：swayhzhu@163.com

通过二分类方法判定语句中是否存在关系，并通过多分类的形式判定该语句存在何种类型的关系。在关系分类任务中，往往需要事先定义关系类型，然后判断文本中实体间是否可能存在某种关系类型。该形式可描述为三元组 $<E_1, R, E_2>$，其中 E_1 和 E_2 为实体类型，R 指关系类型。以“利马是秘鲁的首都，同时是最大的港口”为例，先对句子进行预处理，经过命名实体识别出“利马”和“秘鲁”，然后“首都”作为关系触发词，会在这两种实体之间建立联系，说明这两个实体之间可能存在某种特定的关系，最后经过关系抽取模型，输出“首都”作为这两个实体之间的关系类别，从而得到三元组 < 秘鲁 , 首都 , 利马 >。关系抽取可将文本中信息以三元组 $<E_1, R, E_2>$ 存储在数据库中。通过三元组方式存储可以提供进一步的分析和查询，并被广泛应用于知识图谱、信息检索、问答系统和情感分析中。近年来，无论是工业界还是学术界，该技术都引起了广泛关注。

随着时代发展，信息抽取技术在文本信息化工作中起着越来越不可忽视的作用，也吸引了国内外学者对前沿技术的探索。传统的关系抽取方式大多采用流水线的方式进行，流水线方式抽取包含两个子任务：命名实体识别（named entity recognition，NER）和关系分类 [1−2]。流水线方法指将实体关系抽取以流水线方式进行，即第一步是进行实体识别，在实体识别完成之后，再进行关系抽取。早期工作中，D. Zelenko[3]、Chan Y. S. 等 [4] 以流水线的方式处理关系抽取任务，他们通过两个独立的步骤提取关系三元组：首先，在输入句子上通过命名实体识别的方式识别出文本中所有实体；提取出实体对后，在该基础上进行关系分类。流水线方法通常存在错误传播问题，而且会忽略两个步骤之间的相关性，为了缓解这些问题，现有研究已经提出许多旨在共同学习实体和关系的联合模型。

联合抽取方法的出现，较好地解决了流水线方法中存在的误差传播问题，能够更好地利用子任务之间的关联性。该方法选择对命名实体识别和关系分类采用联合建模，能够较好地利用实体和关系之间的交互信息，同时抽取实体并分类实体对的关系。联合抽取通常是先提取实体对，然后再对关系进行分类或采用统一标注的方式来解决实体关系抽取问题。早期 Yu X. F.[5]、Li Q.[6]、Ren X. 等 [7] 提出的基于特征的联合模型需要复杂的特征工程过程，严重依赖各种 NLP（natural language processing）工具且需手动操作，操作繁琐。为了减少这种手工操作，最近的研究着重于基于神经网络的方法来解决问题。M. Miwa 等 [8] 提出一种端到端的神经网络模型，该方法首次将深度网络和依存树进行结合，通过该方式来进行实体和关系抽取，该论文的成果为后续实体关系联合抽取的研究奠定了基础。Zheng S. C. 等 [9] 通过序列标注的方式解决实体关系联合抽取，该方案可以获得实体信息和它们所持有的关系，摆脱了复杂的特征工程，且该方法可以直接将关系三元组作为一个整体进行建模，但该方法并未解决关系重叠问题。

关系往往指两个实体之间存在的某种联系，关系重叠即某实体和其他实体可能存在一种或多种关系，将这种情况分为单实体多关系（single entity overlap，SEO）和同实体对存在多关系（entity pair overlap，EPO），如图 1，实体“战狼 2”与“吴京”和“卢靖姗”3 个实体之间分别拥有“导演”、“主演”的关系，称之为 SEO 问题；实体“吴京”和实体“战狼 2”分别拥有“导演”和“编剧”两对关系，称之为 EPO 问题。

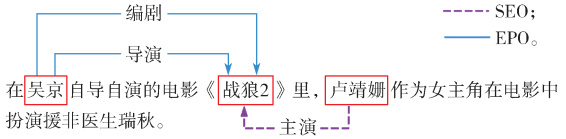


图 1　EPO 及 SEO 举例

Fig. 1　EPO and SEO illustration

针对关系重叠问题，Wei Z. P. 等 [10] 提出了 CASREL 模型，通过一种新的级联二进制框架先预测出文本中所有可能是主体的实体后，下一步通过预测出的主体预测与其相关联的客体。Wang Y. C. 等 [11] 提出了一种 TPLinker 思路，并在英文数据集 NYT 和 WebNLG 上进行实验，通过 3 个握手矩阵将实体与关系进行规则抽取，取得了较好结果。陈仁杰等 [12] 将该两类方法应用在英文数据集 NYT 和 WebNLG 上，都取得了很好的结果，但对关系重叠，还在探索和研究之中。陈仁杰等 [12] 提出了一种融合实体类别信息的实体关系联合抽取方法，该方法在解码阶段融合了头尾实体类别信息的预测，在中文数据集 DuIE 上进行试验，但最终实验结果还有一定的提升空间。为了更有效地抽取出文本中的实体与关系，且能较好地解决实体关系重叠问题，本文做了如下工作：1）针对关系重叠问题提出了位置辅助标记方法；2）运用位置辅助标记方法在中文数据集 DuIE 上进行对比实验证明了该方法的有效性。

2　基于位置辅助标记的实体关系抽取方法原理

本文提出一种基于位置辅助标记的方法来抽取

实体关系三元组，整体模型结构如图 2 所示。将实体关系三元组定义为 $<E_1, R, E_2>$，输入语句在编码后会同时经过一个实体抽取通道和一个关系抽取通道，实体抽取通道能将文本中可能存在的实体以头部索引和尾部索引进行存储，记为 E_H 和 E_T；关系抽取通道能够将文本中可能存在的三元组 $<E_1, R, E_2>$ 以关系相关的形式进行存储，分别存放于关系头部矩阵和关系尾部矩阵中，存放形式分别为 $<E_{H1}, E_{H2}>$ 和 $<E_{T1}, E_{T2}>$。存储后，如果 $<E_{H1}, E_{T1}>$ 和 $<E_{H2}, E_{T2}>$ 在实体矩阵中均存在，就可以抽取出对应的三元组 $<E_1, R, E_2>$。

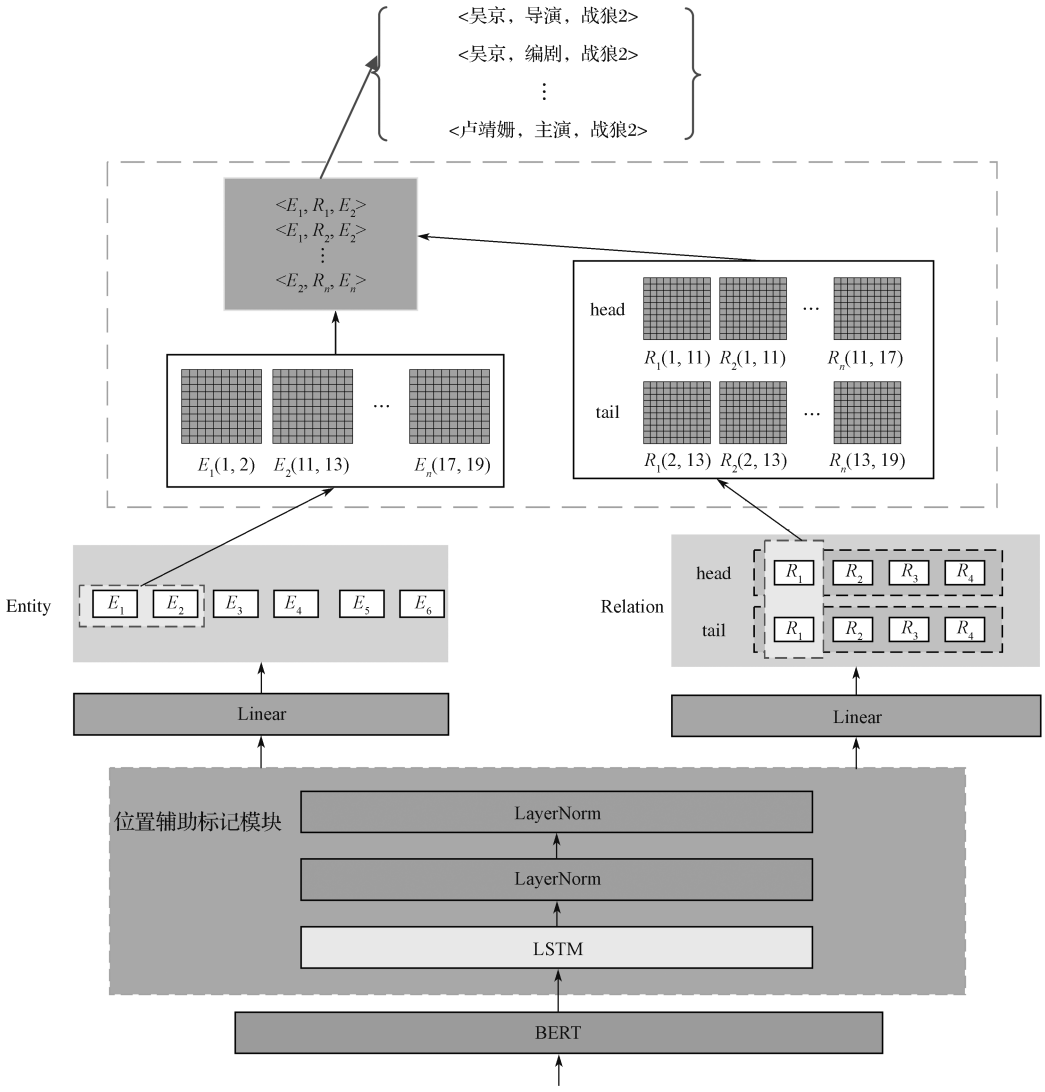


图 2 基于位置辅助标记的实体关系抽取方法模型图

Fig. 2 Model diagram of entity relation extraction method based on location-aided marking

在整个实体关系的抽取过程中，有可能存在 $<E_1, R_1, E_2>$ 和 $<E_1, R_2, E_2>$ 两种重叠关系。在关系抽取通道，根据预设的关系种类设置通道数量，可抽取出该语句中所有的关系头部矩阵和关系尾部矩阵；在实体抽取通道，根据预设的实体类型设置通道，用来抽取出语句中的实体及实体类型。对于三元组 $<E_1, R_1, E_2>$，R_1 在关系头部矩阵中存在 $<E_{H1}, E_{H2}>$，R_1 在关系尾部矩阵中存在 $<E_{T1}, E_{T2}>$。R_2 抽取通道同理，但是由于 R_1 和 R_2 不是在同一通道进行的，故只要 $<E_{H1}, E_{T1}>$，$<E_{H2}, E_{T2}>$ 在实体矩阵中存在，就能输出三元组 $<E_1, R_1, E_2>$ 和 $<E_1, R_2, E_2>$，故该方法能够有效地解决关系重叠问题。

2.1 BERT 预训练语言模型

BERT（bidirectional encoder representation from transformers），2018 年由 Google 首次提出，是一种预训练语言模型。BERT 采用双向 Transformer 的 Encoder，通过新的遮蔽语言模型（masked language modeling，MLM）来生成深度的双向语言表征，这

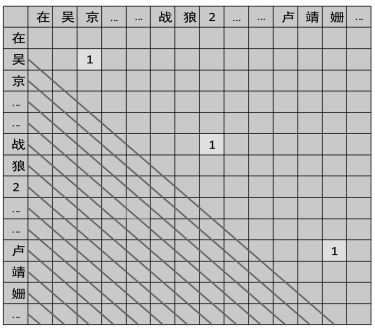
不同于以往采用单向语言模型或将两个单向语言模型进行浅层拼接，且 BERT 在预训练完成之后，只需要一个额外的输出层进行 fine-tune，就能在各种下游任务中取得 state-of-the-art 的表现。基于此，本文采用 BERT 来作为 embedding 层，以较好地获取上下文信息，将 BERT 对应输入文本使用词向量的方式进行表示。

$$\boldsymbol{V}(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n)=embedding(S(w_1,w_2,\cdots,w_n))。\quad(1)$$

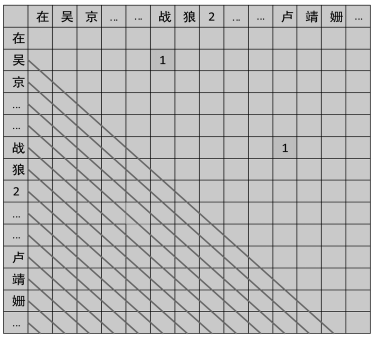
式中：S 为模型文本；$(w_1, w_2, \cdots, w_n)$ 为 S 的每个字；$\boldsymbol{V}$ 为文本 S 经过 BERT 预训练语言模型后的字向量表现形式；$(\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n)$ 为 $(w_1, w_2, \cdots, w_n)$ 经过 BERT 预训练语言模型后的每个字对应的字向量表现形式。

2.2　位置辅助标记

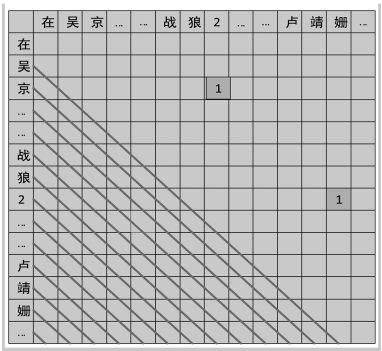
通过位置辅助标记矩阵，可以通过索引标记将实体对之间建立关系，如图 3 所示。



a）实体矩阵



b）关系头部矩阵



c）关系尾部矩阵

图 3　位置辅助矩阵

Fig. 3　Location-aided matrix

输入语句 S：“在吴京自导自演的电影《战狼 2》里，卢靖姗作为女主角在电影中扮演援非医生瑞秋。”在该例句中，以三元组 < 吴京 , 导演 , 战狼 2> 为例，实体矩阵中，“吴京”实体索引位置为 <1, 2>，“战狼 2”实体索引位置为 <11, 13>；关系矩阵中分为关系头部矩阵和关系尾部矩阵，关系头部矩阵中，“吴京”实体中的第一个字“吴”与“战狼 2”实体中第一个字“战”索引位置为 <1, 11>，关系尾部矩阵中，“吴京”实体的最后一个字“京”与“战狼 2”实体最后一个字“2”索引位置为 <2, 13>。解码时，实体矩阵中存在索引 <1, 2> 与 <11, 13>，通过关系矩阵中存在 <1, 11> 与 <2, 13> 可以得到索引 <1, 2> 与 <11, 13>，故实体“吴京”和“战狼 2”存在关系，能够构成三元组 < 吴京，导演，战狼 2>。此外，对于实体矩阵而言，因为实体尾部索引必然大于实体头部索引，矩阵存储的方式采用上三角矩阵的方式进行存储，即对于句子 S，设句长为 n，则位置辅助矩阵的长度为 $n(n+1)/2$，通过该方式，可以更好地利用空间。该模块的操作流程通过以下公式表示：

$$H_{i,j}=\tanh\left(\boldsymbol{W}_h\cdot\left[h_i,h_j\right]+\boldsymbol{b}_h\right),\ j\geqslant i。\quad(2)$$

式中：$\boldsymbol{W}_h$ 为参数矩阵；$[h_i, h_j]$ 为在位置辅助矩阵中的索引位置；$\boldsymbol{b}_h$ 为可在训练中进行更新的偏差向量。

2.3　实体抽取器

本文针对每一种实体类型设置一个实体抽取器，对于 N 个预定义的实体类型，共有 N 个实体抽取器。实体抽取器内使用一个全连接层加 softmax 的方式，标签采用 1/0 的方式，若在文本中某一位置上被判定为某一类型实体的可能性大于 0，那么就将该值标记为 1，否则标记为 0。得到标记结果的情况下，会根据位置辅助上三角矩阵来获取每个实体抽取器中的所有实体。操作流程如公式（3）（4）所示。

$$p(y_{i,j})=\text{soft}\max\left(\boldsymbol{W}_e\cdot\boldsymbol{h}_{i,j}+\boldsymbol{b}_e\right),\quad(3)$$

$$Link(w_i,w_j)=\arg\max P(y_{i,j}=m)。\quad(4)$$

式（3）（4）中：$\boldsymbol{W}_e$ 为参数矩阵；$\boldsymbol{h}_{i,j}$ 为位置向量；$\boldsymbol{b}_e$ 为可以在训练中进行更新的偏差向量；$P(y_{i,j}=m)$ 为在 $[i, j]$ 位置链接实体类型为 m 的概率；$Link(w_i, w_j)$ 为 $[i, j]$ 位置链接概率最大的实体类型。

2.4　关系抽取器

与实体抽取器相似，根据预定义关系种类设立关系抽取通道，但本文针对每个关系设置 4 个关系抽取器。以导演为例，将会设置导演 SH2OH（主实体头部与客实体头部），导演 ST2OT（主实体尾部与客实体尾部），导演 OH2SH（客实体头部与主实体

头部），导演 OT2ST（客实体尾部与客实体头部），因为在位置辅助标记中，统一采用了上三角矩阵进行存储。这样，对于某些输入语句而言，可能会存在客实体在主实体之前，那么，上三角矩阵就无法标记出所有的实体对，无法确认主实体与客实体，故设置了 OH2SH 和 OT2ST 通道，将其设置为输出种类，操作流程与实体抽取流程一样。

$$P\left(y_{i,j}\right)=\mathrm{softmax}\left(\boldsymbol{W}_r\cdot\boldsymbol{h}_{i,j}+\boldsymbol{b}_r\right), \tag{5}$$

$$Link\left(w_i,w_j\right)=\arg\max P\left(y_{i,j}=n\right)。 \tag{6}$$

式（5）（6）中：$\boldsymbol{W}_r$ 为参数矩阵；$\boldsymbol{b}_r$ 为训练中进行更新的偏差向量；$P(y_{i,j}=n)$ 为在 $[i,j]$ 位置链接关系 n 的概率；$Link(w_i, w_j)$ 为 $[i, j]$ 位置链接概率最大的关系类型。

2.5 损失函数

由于该任务用于多分类任务，所以选择了多分类交叉熵损失函数作为损失函数，公式定义如下：

$$L_{\text{link}}=-\frac{1}{N}\sum_{i=1,\ j\geqslant i}^{N}\ \sum_{*\in\{m,n\}}\log P\left(y^*_{i,j}=\hat{l}^*\right)。 \tag{7}$$

式中：N 为实体类型数加上 4 倍的关系类型数；m 为实体类型集合；n 为关系类型集合。

3 实验结果

3.1 数据集介绍

实验数据集来自百度提供的数据集 DuIE[13]。DuIE 是一个大规模的高质量数据集，该数据集采用从粗到精的过程，结合远程监控和众包标注。本文使用的是 DuIE 中的训练集和验证集，由于实验设备受限，且所提出模型复杂程度与句长呈指数级关系，故将文本长度大于 130 的句子全部舍弃，最终实验所用数据情况如表 1 所示。

表 1 实验数据集

Table 1 Experimental dataset

属 性	值	属性	值
数据集	DuIE	验证集大小	21 050
训练集大小	168 580	关系类别	49
实体类别	28		

3.2 评价指标

基于位置辅助标记方法的实体关系抽取模型评价指标通过精准率（P，precision），召回率（R，recall）和 F_1 值来进行衡量，具体计算方式如下：

$$P=\frac{TP}{TP+FP}\times 100\%, \tag{8}$$

$$R=\frac{TP}{TP+FN}\times 100\%, \tag{9}$$

$$F_1=\frac{2PR}{P+R}\times 100\%。 \tag{10}$$

式（8）（9）中：TP 为预测出的三元组与原文本中存在三元组一致的个数；FP 为预测出来的三元组与原文本中存在三元组不一致的个数；FN 为原文本中存在的三元组但预测结果没有该三元组的个数。

3.3 实验环境

实体关系抽取模型实验环境如表 2 所示。

表 2 实验环境

Table 2 Experimental environment

参数	值	参数	值
数据集	DuIE	内存	16 GB
操作系统	Ubuntu 16.04 LTS	硬盘	1 TB
CPU	i7-10750H @ 2.60 GHz	Python 版本	3.8
GPU	NVIDIA GeForce RTX2080Ti	Pytorch 版本	1.9.0

3.4 实验参数设计

本文进行了一系列实验，各实验参数设置如表 3 所示。

表 3 各模型参数设置

Table 3 Parameter settings of each model

模 型	参 数 设 置
BERT+sigmoid	采用 BERT-roberta 版预训练语言模型，学习率设置为 0.00 003，最大句长为 128，批次大小（batch_size）为 16，优化算法使用 Adam
FETI（融合实体类别信息）	词向量维度、LSTM 隐藏层维度、关系嵌入维度、头实体嵌入维度、尾实体嵌入维度均为 200，dropout 为 0.5，优化算法使用 Adam
BERT+CNN+CRF	采用 BERT-base-chinese 版预训练语言模型，采用序列标注方法进行抽取，学习率设置为 0.000 03，批次大小（batch size）为 20，dropout 为 0.1，优化算法使用 Adam
BERT+ 位置辅助标记 +sigmoid	采用 BERT-roberta 版预训练语言模型，学习率设置为 0.000 03，批次大小（batch size）为 10，epoch 为 20，最大句长设置为 128，优化算法使用 Adam

3.5 实验结果分析

为了验证本文提出模型的有效性，利用上述模型按照上述参数进行了对比实验，结果如表 4 所示。

如表 4 所示，本文提出的方法与基于 BERT 的模型相比，F_1 值提高了 0.107；与 FETI（融合实体类型信息）的实体关系联合抽取方法相比，本文提出的

方法 F_1 值提高了约 0.079；相较于序列标注的方法，本方法 F_1 值提高了约 0.027。采用 BERT+sigmoid 先识别实体再关系分类，使用该方式很难获取到实体和关系之间的相互依赖性。相比之下，本文提出的模型，实体和关系共享编码层参数，可以较好地利用实体和关系之间的交互信息。FETI 融合了实体类别信息进行关系抽取，将实体类别信息作为辅助信息做实体关系联合抽取，使用该方式，会给联合抽取任务提供大量的背景知识，同时也会引入更多噪声。本文提出的方法不会引入与抽取出的三元组无关的噪声，可以有效提高抽取的准确率。BERT+CNN+CRF 采用了序列标注的方法进行学习，通过该类方法模型需要额外学习复杂的 BIEOS（B-beginner, I-inside, E-end, O-outsde, S-single）对性能有一定影响。而采用多个二分类器的方法进行实体关系联合抽取过程较为简单，在解码阶段引入相应位置辅助信息，而且能解决 SEO 和 EPO 问题。

表 4　各模型实验结果

Table 4　Experimental results of each model

模　型	P	R	F_1
BERT+sigmoid	0.734	0.726	0.730
FETI (BERT+ 融合类别信息)	0.757	0.760	0.758
BERT+CNN+CRF	0.825	0.795	0.810
BERT+ 位置辅助标记 +sigmoid	0.829	0.845	0.837

综上所述，本研究提出的 BERT 结合位置辅助标记和 sigmoid 模型的总体性能最好，优于其他模型。

4　结语

为了解决实体关系抽取任务中的关系重叠问题，本文采用了基于位置辅助标记 [14−16] 的方法进行实体关系抽取，可以更准确地抽取出实体关系三元组。该方法使用了 BERT 来获取文本中的上下文语义，位置辅助信息可以更好地定义规则，为后续学习建立一个更明确的目标，解决了关系重叠问题。经过实验论证，本文在中文数据集 DuIE 上表现良好，取得了不错的效果。在后续研究中，由于位置辅助矩阵引入参数过多，模型花费时间会较长，因而将对模型进行改进，以减少参数，提高效率。

参考文献：

[1] ZHANG L，ZHAO H. Named Entity Recognition for Chinese Microblog with Convolutional Neural Network[C]//2017 13th International Conference on Natural Computation，Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). Guilin：IEEE，2017：87−92.

[2] 陈　宇，郑德权，赵铁军 . 基于 Deep Belief Nets 的中文名实体关系抽取 [J]. 软件学报，2012，23(10)：2572−2585.
CHEN Yu，ZHENG Dequan，ZHAO Tiejun. Chinese Relation Extraction Based on Deep Belief Nets[J]. Journal of Software，2012，23(10)：2572−2585.

[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 - EMNLP’02. Morristown：Association for Computational Linguistics，2002：71−78.

[4] CHAN Y S，DAN R . ACL’11 Exploiting Syntactico-Semantic Structures for Relation Extraction[EB/OL]. [2021−09−02]. https://aclanthology.org/P11−1056.pdf.

[5] YU X F，LAM W. Jointly Identifying Entities and Extracting Relations in Encyclopedia Text via a Graphical Model Approach[EB/OL]. [2021−09−02]. https://aclanthology.org/C10-2160.pdf.

[6] 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：Association for Computational Linguistics，2014：402−412.

[7] REN X，WU Z Q，HE W Q，et al. CoType：Joint Extraction of Typed Entities and Relations with Knowledge Bases[C]//Proceedings of the 26th International Conference on World Wide Web. Perth：International World Wide Web Conferences Steering Committee，2017：1015−1024.

[8] MIWA M，BANSAL M. End-to-End Relation Extraction Using LSTMS on Sequences and Tree Structures[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1：Long Papers). Berlin：Association for Computational Linguistics，2016：1105−1116.

[9] ZHENG S C，WANG F，BAO H Y，et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1：Long Papers). Vancouver：Association for Computational Linguistics，2017：1227−1236.

[10] WEI Z P，SU J L，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. Stroudsburg：Association for Computational Linguistics，2020：1476−1488.

（下转第 62 页）