计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(06)-1453-10
doi: 10.3778/j.issn.1673-9418.2206052

融合预训练模型和注意力的实体关系抽取方法

李智杰,韩瑞瑞,李昌华+,张　颉,石昊琦
西安建筑科技大学 信息与控制工程学院,西安 710055
+ 通信作者 E-mail: lch304502@126.com

摘　要:实体关系抽取旨在从无结构的文档中检测出实体和实体对的关系,是构建领域知识图谱的重要步骤。针对现有抽取模型语义表达能力差、重叠三元组抽取准确率低的情况,研究了融合预训练模型和注意力的实体关系联合抽取问题,将实体关系抽取任务分解为两个标记模块。头实体标记模块采用预训练模型对句子进行编码,为了进一步学习句子的内在特征,利用双向长短时记忆网络(BiLSTM)和自注意力机制组成特征加强层。采用二进制分类器作为模型的解码器,标记出头实体在句子中的起止位置。为了加深两个标记模块之间的联系,在尾实体标记任务前设置特征融合层,将头实体特征与句子向量通过卷积神经网络(CNN)和注意力机制进行特征融合,通过多个相同且独立的二进制分类器判定实体间关系并标记尾实体,构建出融合预训练模型和注意力的联合抽取模型(JPEA)。实验结果表明,该方法能显著提升抽取的效果,对比不同预训练模型下抽取任务的性能,进一步说明了模型的优越性。

关键词:领域知识图谱;预训练模型;自注意力机制;特征融合
文献标志码:A　**中图分类号:**TP391

Entity Relation Extraction Method Integrating Pre-trained Model and Attention

LI Zhijie, HAN Ruirui, LI Changhua+, ZHANG Jie, SHI Haoqi
School of Information and Control Engineering, Xi'an University of Architectural Science and Technology, Xi'an 710055, China

Abstract: Entity relationship extraction aims to detect the relationship between entities and entity pairs from unstructured text. It is an important step in constructing domain knowledge map. In view of the poor semantic expression ability of the existing extraction models and the low accuracy of overlapping triples extraction, this paper studies the joint extraction of entity relationships by integrating pre-trained model and attention, and divides the entity relationship extraction task into two tag modules. The head entity tagging module uses a pre-trained model to encode sentences. In order to further learn the internal characteristics of sentences, bi-directional long-short term memory and self-attention mechanism are used to form a feature enhancement layer. The binary classifier is used as the decoder of the model to mark the start and end positions of the head entity in the sentence. In order to deepen the relationship between the two marking modules, a feature fusion layer is set up before the tail entity marking task. The head entity features and sentence vectors are fused through convolutional neural networks (CNN) and attention mechanism. The relationship between entities is determined and the tail entity is marked through multiple identical

基金项目:"十三五"国家重点研发计划课题(2019YFD1100904);国家自然科学基金(51878536);陕西省住房城乡建设科技计划(2020-K09)。
This work was supported by the National Key Research and Development Projects in the 13th Five-Year of China (2019YFD1100904), the National Natural Science Foundation of China (51878536), and the Housing and Urban Rural Construction Science and Technology Project of Shaanxi Province (2020-K09).
收稿日期:2022-06-14　**修回日期:**2022-09-07

and independent binary classifiers. A joint model based on pre-trained encoder and attention mechanism (JPEA) is constructed. Experimental results show that this method can significantly improve the extraction effect, and the performance of extraction tasks under different pre-trained models is compared, which further illustrates the superiority of the model.

Key words: domain knowledge graph; pre-trained model; self-attention mechanism; feature fusion

知识图谱[1]一词在2012年被Google公司首次提出,它是一个结构化的语义知识库,可以组织海量信息,用于描述客观世界中的概念以及它们之间存在的关系。知识图谱不仅在语义搜索、深度问答等通用领域发挥着重要作用,在金融、医疗、城市规划等垂直领域中也有着广阔的应用前景[2]。知识图谱的构成单位是关系三元组,因此从非结构化文本中抽取出关系三元组对构建知识图谱十分重要。

实体关系抽取任务的提出就是为了解决关系三元组抽取问题。例如,在句子“姚明出生于上海”中,可以提取出关系三元组(姚明,出生于,上海),其中“姚明”和“上海”分别称为头实体和尾实体,“出生于”称为这两个实体之间的关系。实体关系抽取任务最初采用基于规则和模板的方法[3],由于人工和时间等因素的限制,逐渐发展为基于统计机器学习的方法[4]。近年来,由于深度学习发展迅速,基于深度学习构建实体关系抽取模型已成为信息抽取领域新的研究方向[5],这一方向认为抽取任务应被划分为流水线方法和联合学习方法,这两种方法都基于卷积神经网络(convolutional neural networks,CNN)、循环神经网络(recurrent neural network,RNN)、长短时记忆网络(long short-term memory,LSTM)这三种网络架构进行组织和扩展[6]。流水线方法是在已经抽取出实体的基础上,对每个实体对之间的关系进行分类的方法。虽然采用流水线的方式会使每个子任务更加集中,更容易解决问题,但两个子任务之间的完全分离可能会遇到错误传播问题,同时也忽略了它们之间的相互依赖性[7]。联合学习方法是指在执行两个子任务时,使用同一个编码层,联合检测实体及其关系。联合学习方法能够学习到实体和关系之间潜在的联系,从而在两个子任务中获得更好的性能。

目前多数联合抽取模型使用BERT(bidirectional encoder representations from transformers)[8]进行预训练,它能为联合模型提供一种通用的融合上下文信息的词向量表示。随着自然语言处理任务的发展,研究人员对BERT进行了一系列改进,RoBERTa(robustly optimized BERT pretraining approach)[9]是其中一种变体,相比原生BERT,RoBERTa预训练模型使用了更多的数据集并且训练得更加充分。尽管之前的实体关系抽取工作已经取得了很大的成功,但在以往的多数模型中,关系都被看作需要分配给实体对的离散标签。事实上,在提取出的所有实体对之间,大多都没有形成有效的关系,这导致了很多负样本的产生。另外,如果没有足够的训练样本,分类器很难判断实体参与的关系,导致重叠关系三元组提取不完整。因此,本文提出一种JPEA模型(joint model based on pre-trained encoder and attention mechanism)。本文的工作主要包括以下几点:

(1)针对此前大多数实体关系抽取方法在语义特征表示和重叠关系提取方面的缺陷,提出了融合预训练模型和注意力的联合抽取模型JPEA。

(2)为了更准确地抽取出头实体,将预训练模型编码的结果输入BiLSTM网络和自注意力机制进行深层次特征提取,获得更细粒度的语义。

(3)为了增强两个标记模块的依赖性,将提取的头实体特征作为条件信息,利用CNN和注意力模块融合到句子特征向量中,为关系及尾实体的标记增强实体表达能力。

(4)分别在公开的数据集纽约时报(the New York Times,NYT)和WebNLG(Web natural language generation)上针对不同预训练模型进行测试,实验表明,不同预训练模型下JPEA模型的各项评价指标均有较好的表现,$F1$值最高分别可达到92.4%和92.9%。

1 相关工作

根据实体关系抽取的发展来看,该任务主要可以分为基于规则和模板的方法、基于统计机器学习的方法和基于深度学习的方法。

1.1 基于规则和模板的方法

实体关系抽取任务最初通常采用基于规则和模板的方法[10],在该传统方法中,语法和语义规则往往需要通过人工构造的方式获取。这种方式有两个明显的缺点:(1)一般只有对特定领域有深入认知的人员才有资格手动编写模板和规则,这会造成大量的

人力和资源消耗；(2)手动制定的规则具有较差的可移植性，一般难以拓展到其他领域。

1.2 基于统计机器学习的方法

通过早期的相关研究结果可以发现，以特征工程为核心的有监督抽取方法是实施基于统计机器学习的抽取方法的主流[11]。这类方法虽然理论基础已经趋于完善，但是仍然离不开人力的参与，适合用于模型训练的特征集仍需要通过大量的特征工程人工筛选获得。因此，近些年来学术界的研究重点转向了半监督和无监督的抽取方法。Shinyama等人[12]提出了抢占式信息提取的概念，其关键在于找到文本中多个实体之间的并行对应关系，并使用聚类的方式抽取信息。Carlson等人[13]提出的耦合分类和关系实例抽取器的半监督学习方式，能够预防与引导学习方法相关的语义漂移问题，提高抽取精度。Zhang等人[14]提出了基于MBL(memory-based learning)的统一框架，采用无监督学习方法，实现了对实体间多元关系的精准识别。

1.3 基于深度学习的方法

近年来，人们提出了大量的深度神经网络模型来完成实体关系抽取任务，基于深度学习的抽取方法主要采用CNN[15]、RNN[16]和LSTM[17]的变体或组合结构。Socher等人[18]首次在分类任务中引入RNN模型来全面处理词向量空间中的组合性，句法树中的每个节点会被分配到一个向量和一个矩阵，分别用来学习该处的词向量和相邻单词或短语的含义，但是该模型没有考虑实体对的位置信息。Zeng等人[19]首次把CNN应用到关系分类任务中，词汇和句子层面的特征均通过卷积深度神经网络获取，并提出位置特征来指定期望分配关系标签的实体对。Cai等人[20]将CNN和双通道递归神经网络与LSTM单元相结合，同时沿SDP(the shortest dependency path)前向和后向学习具有方向信息的关系表。

上述研究所采用的模型都应用于流水线抽取任务中，近些年，研究人员开始致力于联合抽取模型的研究。Miwa等人[21]首次将LSTM-RNNs神经网络结构应用于联合抽取任务。Zheng等人[22]提出了一种新颖的标记方式，把联合抽取问题建模为端到端的序列标注模型，但是该方法在重叠关系的识别上仍存在不足。Yu等人[23]将联合抽取任务分解为相互关联的两个子任务，采用合理的分解策略，充分捕获不同步骤之间的语义相关性。Wei等人[24]打破了传统的思路，从新的角度理解信息抽取任务，并提出了一个新的级联二进制标记框架，着重处理重叠问题。但该框架抽取头实体时对语义信息获取不够充分，且仅将抽取出的头实体向量与各词向量进行了简单的拼接，忽略了头实体和其他单词之间的细粒度语义关系，存在特征丢失问题。因此，本文在此基础上提出了一种融合预训练模型和注意力的联合抽取模型。

2 融合预训练模型和注意力的联合抽取模型

由于联合抽取方法经常遇到关系重叠和实体嵌套的问题，无法对非结构化文本进行合理的抽取，甚至出现漏抽取的问题。另外，由于静态编码模型无法准确捕获序列时序位置信息，常造成三元组抽取准确率偏低。针对此问题，采用预训练模型动态编码能够有效表述句子序列内在特征的特点，并利用注意力机制捕获头实体信息，提出了一种融合预训练模型和注意力的网络模型用于联合抽取实体关系三元组。

在之前的Seq2seq方法中，关系三元组抽取任务通常建模为式(1)，即先抽取头实体s，然后结合主实体抽取出对应的尾实体o，最后根据抽取出的实体对预测关系r。

$$P(s,r,o)=P(s)P(o|s)P(r|s,o) \tag{1}$$

本文提出的JPEA模型将三元组抽取过程整体建模为式(2)：

$$p((s,r,o)|x)=p(s|x)p((r,o)|s,x)\ =p(s|x)\prod_{r\in T}p_r(o|s,x) \tag{2}$$

其中，x是输入的句子，T是所有关系类型的集合。通过式(2)将三元组的抽取问题转变为指针标注问题，这种建模方式允许模型一次提取出多个关系三元组：首先通过头实体标记器检测出句子中所有的实体，这些实体都是潜在的头实体，即都有可能与其他实体构成关系三元组，然后针对每一个潜在的头实体，通过关系及尾实体标记器来查找所有与该头实体有关的关系和对应关系下的尾实体，若句子中不存在与该头实体有相关关系的尾实体，则舍弃该头实体，最终完整地抽取出句子中包含的三元组。

JPEA模型通过预训练模型编码词向量，采用自注意力机制和BiLSTM网络结合来丰富语义特征，捕获更重要的语义信息，再通过归一化标记出所有头实体；其次将多层CNN网络与注意力机制融合，提取头实体特征，计算每个头实体相对于句子中每个词向量的权重，并将加权后的头实体特征与句子向量进行拼接，用于标记与每个头实体之间存在合适关系的全部尾实体及正确的实体间关系。模型总体架构如图1所示，其中预训练模型为RoBERTa，s_start

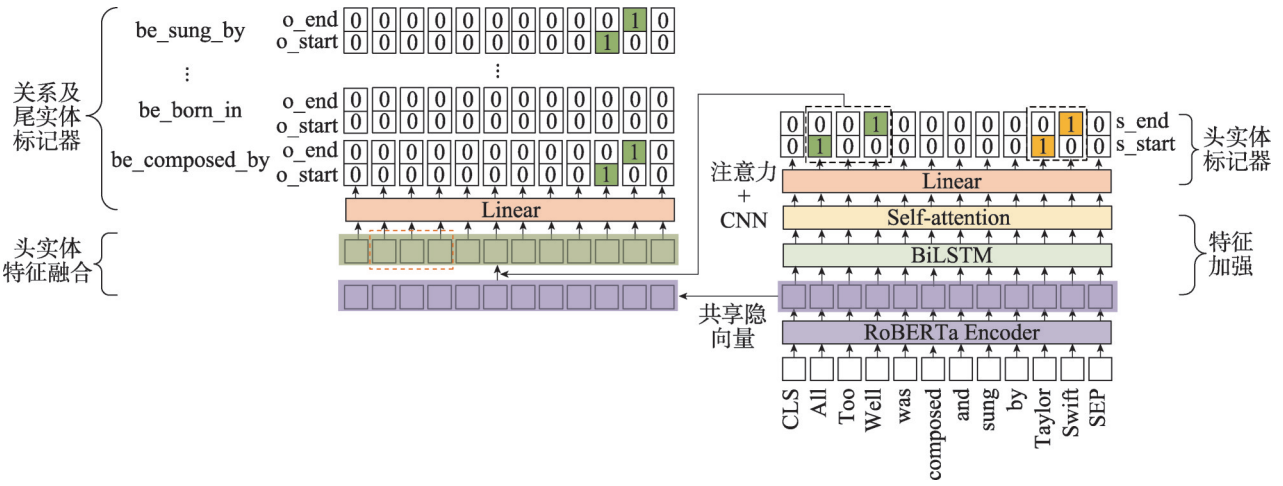


图1 JPEA实体关系联合抽取模型架构

Fig.1 JPEA entity relationship joint extraction model structure

和s_end分别表示头实体的起始和结束，o_start和o_end分别表示尾实体的起始和结束，1/0标记表示该位置是否对应起始或结束位置。以图中抽取出的第一个歌曲实体“All Too Well”为例，在“be_composed_by”和“be_sung_by”关系条件下对应的尾实体均为“Taylor Swift”，在其他的关系条件下没有对应的尾实体，因此头实体“All Too Well”可以抽取出（All Too Well，be_composed_by，Taylor Swift）和（All Too Well，be_sung_by，Taylor Swift）两个三元组。

2.1 RoBERTa编码层

传统的文本编码模型在语义表征能力上有所欠缺，而RoBERTa模型在大量训练数据的基础上，能够准确高效地表达句子的潜在信息。RoBERTa是基于多层双向Transformer的语言表示模型，它通过对每个单词的左右语境进行联合训练来学习深度表征，在许多下游任务中都有着高效的表现，因此所提模型采用RoBERTa预训练模型编码句子向量。

JPEA模型的两个实体标记模块共享同一个编码层，RoBERTa模型从待分析句子序列中提取出语义特征，并将特征传递给两个实体标记模块。首先将输入的文本序列表示成向量形式，对于处理后的文本序列中的第i个字符的向量表示如式（3）所示：

$$\boldsymbol{e}_i=\boldsymbol{W}_{\text{token}}(s_i)+\boldsymbol{W}_{\text{seg}}(seg_i)+\boldsymbol{W}_{\text{pos}}(i) \tag{3}$$

其中，$\boldsymbol{W}_{\text{token}}$、$\boldsymbol{W}_{\text{seg}}$、$\boldsymbol{W}_{\text{pos}}$分别为token嵌入、分句嵌入和位置嵌入。然后通过RoBERTa模型对嵌入结果进行编码，最后一层Transformer输出的值即是文本编码的最终结果，如式（4）所示：

$$\boldsymbol{X}=\text{RoBERTa}(\boldsymbol{E}) \tag{4}$$

其中，$\boldsymbol{E}=\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n\}$为待处理文本序列的向量表示形式，$n$表示文本序列长度，$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$为经过RoBERTa编码得到的具有上下文信息的句子向量。

2.2 特征加强层

传统的循环神经网络在处理时序数据时虽然不受数据长度的限制，但由于无法很好地捕获反向语义，存在严重的信息丢失问题，无法准确描述句子特征。BiLSTM网络是一种特殊的循环神经网络，能够实现从后往前编码，通过利用句子中比较靠后的重要信息，能够很好地捕捉双向的语义依赖。故本模块选取BiLSTM网络对RoBERTa编码层提取的句子向量进行进一步的特征表示。具体操作为：将向量矩阵$\boldsymbol{X}$输入BiLSTM网络进行编码，每一时刻t的输入除了词向量外还有上一时刻的输出向量，每一时刻得到的输出向量$\boldsymbol{h}_t$均为前向编码向量和后向编码向量拼接而成。上述过程如式（5）所示：

$$\boldsymbol{h}_t=\text{BiLSTM}(\boldsymbol{x}_t,\boldsymbol{h}_{t-1}) \tag{5}$$

其中，$\boldsymbol{x}_t$表示t时刻输入的词向量。经过BiLSTM网络编码后得到的向量为$\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$。

为了增强模型的辨别能力，研究人员通常会在神经网络模型中加入自注意力机制，通过为输入信息的每个部分赋予不同的权重，可以抽取出关键信息，使模型做出更加准确的判断。自注意力机制的计算如式（6）所示：

$$\boldsymbol{A}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{6}$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别表示查询矩阵、键矩阵和值矩

阵，$\sqrt{d_K}$ 是键矩阵第一维度的平方根，用以维持梯度的稳定。

自注意力机制不会增加模型的计算开销和结构的复杂度，同时还可以有效弥补BiLSTM网络在解决长距离依赖问题上的缺陷。根据上述原理描述，本模块在BiLSTM层之后拼接一层自注意力网络，根据上下文信息为每个词向量训练相应的权重，以更准确、完整地标记出头实体。操作过程可描述为：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 通过参数矩阵 $\boldsymbol{W}_{\mathrm{Q}}$、$\boldsymbol{W}_{\mathrm{K}}$、$\boldsymbol{W}_{\mathrm{V}}$ 进行线性变换，再通过自注意力运算得到各个位置的注意力大小，最后经过线性变换得到更加丰富的语义信息，这里的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 为上一步的输出向量 $\boldsymbol{H}$。如式(7)所示，$\boldsymbol{M}=\{\boldsymbol{m}_1,\boldsymbol{m}_2,\cdots,\boldsymbol{m}_n\}$ 为经过自注意力层编码后得到的结果。

$$\boldsymbol{M}=\mathrm{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{K}\boldsymbol{W}_{\mathrm{K}},\boldsymbol{V}\boldsymbol{W}_{\mathrm{V}}\right)\boldsymbol{W}^{o} \tag{7}$$

2.3 头实体标记器

头实体标记器用来识别输入句子中所有可能的头实体，它是通过对特征加强层的输出结果 $\boldsymbol{M}$ 进行解码实现的。头实体标记器由两个相同且独立的二进制分类器组成，可以通过为每个位置分配0/1标记来分别检测头实体的开始和结束位置。具体操作如下：

$$p_i^{s\_start}=\sigma(\boldsymbol{W}_{\mathrm{start}}^{s}\boldsymbol{m}_i+\boldsymbol{b}_{\mathrm{start}}^{s}) \tag{8}$$

$$p_i^{s\_end}=\sigma(\boldsymbol{W}_{\mathrm{end}}^{s}\boldsymbol{m}_i+\boldsymbol{b}_{\mathrm{end}}^{s}) \tag{9}$$

其中，$\boldsymbol{m}_i$ 是输入句子中第 i 个单词经特征加强层处理后的向量，$p_i^{s\_start}$ 和 $p_i^{s\_end}$ 均表示第 i 个单词向量经解码层处理后的输出值，两个输出值都是概率值。如果该值大于实验设置的某一限定值，那么该单词所在位置将被分配到标记1，否则将被分配到标记0。可在模型训练的过程中不断调整参数，为头实体标记器确定一个最佳限定值，本文实验在多次调整后将该限定值设为0.5。$\boldsymbol{W}(\cdot)$ 和 $\boldsymbol{b}(\cdot)$ 分别表示训练权重和偏置向量，s 指代头实体，σ 代表sigmoid激活函数。实验设定采用指针对就近匹配原则解决一个句子中存在多个头实体的问题，并且不考虑单词的结束位置在开始位置前面的情况。

2.4 头实体特征融合

为了加强JPEA模型的两个实体标记模块之间的依赖性，在抽取尾实体和实体对间关系之前，还需要对标记出的头实体进行特征处理，但是将头实体表示与句子向量进行简单的拼接不能完整地表达特征信息。本文采用CNN与注意力机制进行头实体与句子向量的融合。首先，获取头实体的起始和结束标记之间所有单词的特征表示 $\boldsymbol{X}_{\mathrm{head}}$，将向量输入到一个多层CNN网络中学习实体级别的特征表示，使用最大池化操作得到最终的特征向量 $\boldsymbol{x}_{\mathrm{head}}$，如式(10)所示：

$$\boldsymbol{x}_{\mathrm{head}}=\mathrm{MaxPooling}(\mathrm{CNN}(\boldsymbol{X}_{\mathrm{head}})) \tag{10}$$

本文认为，头实体的特征对尾实体标记任务的影响主要与当前位置词有关，于是在特征融合的过程中加入了注意力机制，如式(11)所示：

$$\boldsymbol{T}_i=[\boldsymbol{X}_i;(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{x}_{\mathrm{head}})\boldsymbol{x}_{\mathrm{head}}] \tag{11}$$

首先，将编码层输出的句子向量 $\boldsymbol{X}$ 与头实体特征向量 $\boldsymbol{x}_{\mathrm{head}}$ 做点积运算，运算结果即是注意力权重；其次，计算该权重与头实体特征向量相乘的结果；最后，把加权的头实体向量与当前位置的词向量拼接在一起，经过特征融合后得到的向量为 $\boldsymbol{T}=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n\}$。

2.5 关系及尾实体标记器

关系及尾实体标记器采用多层二进制分类器，在进行尾实体标记时，首先需要预定义若干种关系，关系的数量即为二进制分类器的层数。关系及尾实体标记器的输入是融合了头实体特征的句子向量 $\boldsymbol{T}$，在对向量 $\boldsymbol{T}$ 进行解码时，对于所有可能的关系，标记器将同时为每个检测到的头实体标记出相应的尾实体。详细操作如下：

$$p_i^{o\_start}=\sigma(\boldsymbol{W}_{\mathrm{start}}^{o}\boldsymbol{T}_i+\boldsymbol{b}_{\mathrm{start}}^{o}) \tag{12}$$

$$p_i^{o\_end}=\sigma(\boldsymbol{W}_{\mathrm{end}}^{o}\boldsymbol{T}_i+\boldsymbol{b}_{\mathrm{end}}^{o}) \tag{13}$$

$\boldsymbol{T}_i$ 是第 i 个单词的编码向量经过特征融合后的向量表示，$p_i^{o\_start}$ 和 $p_i^{o\_end}$ 均代表第 i 个融合向量经解码层处理后的输出值，两个输出值都是概率值，$\boldsymbol{W}(\cdot)$ 和 $\boldsymbol{b}(\cdot)$ 分别表示关系条件下训练的权重矩阵和偏置值，o 指代尾实体，σ 代表sigmoid激活函数。

2.6 损失函数

本模型的损失函数可以表示为头实体抽取损失和关系及尾实体抽取损失值的加和，因为两个任务均采用二进制分类器，所以在模型中采用二分类交叉熵损失函数。具体可由式(14)表示：

$$Loss=\sum_{j\in J}-\frac{1}{n}\sum_{i=1}^{n}(y_i^j\cdot\mathrm{lb}\,p_i^j+(1-y_i^j)\cdot\mathrm{lb}(1-p_i^j)) \tag{14}$$

其中，$J=\{s\_start,s\_end,o\_start,o\_end\}$，$n$ 代表句子的长度。y_i^j 表示句子中第 i 个单词是实体的开始或结束位置的样本标签，p_i^j 表示二进制分类器预测开始或结束位置样本标签为正例的概率。

3 实验

3.1 数据集与评价指标

本文选择在NYT和WebNLG两个公开语料库上

进行实验。NYT是摘自New York Times新闻文章的样本，并由远程监督方法进行注释，共包含56 195句用于训练，5 000句用于测试。WebNLG起初被应用于自然语言构建任务，一些学者将其作为关系抽取的数据集进行应用，其包含5 019句用于训练，703句用于测试。为了验证本文提出的JPEA模型在处理重叠关系问题上有更好的表现，将句子类型划分成三部分，分别为：正常（Normal）、实体对重叠（entity pair overlap，EPO）、单一实体重叠（single entity overlap，SEO）。具体划分情况如表1所示。

表1 数据集统计

Table 1 Statistics of datasets

数据集	NYT		WebNLG	
	训练集	测试集	训练集	测试集
Normal	37 013	3 266	1 596	246
EPO	9 782	978	227	26
SEO	14 735	1 297	3 406	457
ALL	56 195	5 000	5 019	703

实验通过准确率（P）、召回率（R）和调和平均值（$F1$）三个指标来评估模型的效果，$F1$为主要指标，各指标的计算公式如式(15)~(17)。

$$P=\frac{\text{正确识别三元组个数}}{\text{被识别的三元组个数}}\times 100\% \tag{15}$$

$$R=\frac{\text{正确识别的三元组个数}}{\text{样本中的三元组个数}}\times 100\% \tag{16}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{17}$$

为了探究JPEA模型各个改进模块的效果，本文针对各个设计做了消融实验，以展示通过BiLSTM与自注意力结合获得语句特征表示、利用CNN与注意力机制融合将头实体特征融入句子向量这两处设计对实验的增益效果。

3.2 实验环境及参数设置

本文的所有实验均在Windows 10操作系统上完成，处理器为Intel® Core i7-10700K@3.80 GHz，显卡为NVIDIA GeForce GTX3090Ti。使用的语言是python3.7，预训练模型均使用Base版本，模型的优化器选择Adam。模型的最优参数设置如表2所示。

3.3 实验结果与分析

3.3.1 模型对比实验分析

为了验证JPEA模型的优越性，本文选取了几个目前在重叠关系抽取方面表现较好的模型进行对比分析，基线模型的实验结果直接摘自原始出版的论文。为了评估引入不同的预训练模型对实体关系抽取任务性能的影响，进一步做了一系列对比实验。$\text{JPEA}_{\text{BERT}}$代表预训练模型改用BERT，$\text{JPEA}_{\text{ALBERT}}$表示编码器采用ALBERT[25]预训练模型，$\text{JPEA}_{\text{ELECTRA}}$表示在ELECTRA[26]预训练模型的基础之上实例化实体关系抽取框架。为了确保对比实验结果的准确性，实验对此类关系抽取模型采用相同的输入，然后比较模型的实验结果。对比情况如表3所示，其中加粗数字表示实验结果的最优值。

表2 模型参数值

Tabel 2 Model parameter values

参数	值
隐向量长度	768
最大输入长度	100
Batch_size	6
学习率	1E−5
实体标记阈值	0.5
迭代次数	60

表3 不同模型在NYT和WebNLG数据集上的实验结果

Table 3 Experimental results of different models on NYT and WebNLG datasets 单位：%

模型	NYT			WebNLG		
	准确率	召回率	$F1$	准确率	召回率	$F1$
CopyRE[27]	61.0	56.6	58.7	37.7	36.4	37.1
GraphRel[28]	63.9	60.0	61.9	44.7	41.1	42.9
CopyRRL[29]	77.9	67.2	72.1	63.3	59.9	61.6
ETL-Span[23]	85.5	71.7	78.0	84.3	82.0	83.1
CasRel[24]	89.7	89.5	89.6	93.4	90.1	91.8
JPEA	91.6	92.5	92.0	**93.8**	91.9	92.8
$\text{JPEA}_{\text{BERT}}$	90.8	91.6	91.1	93.4	91.3	92.3
$\text{JPEA}_{\text{ALBERT}}$	91.9	91.3	91.5	93.2	91.5	92.4
$\text{JPEA}_{\text{ELECTRA}}$	**92.3**	**92.6**	**92.4**	93.6	**92.2**	**92.9**

对比表3中的数据可以看出，在三个评价指标上，本文提出的JPEA模型及其变体均取得了较好的实验结果，体现了模型的优越性。CopyRE[27]在实体关系抽取过程中采用动态解码的方式并首次尝试解决重叠关系的抽取问题，但由于RNN展开的固有限制，导致生成的三元组有限。GraphRel[28]在编码的过程加入了GCN（graph convolutional networks），同时获得句子序列和区域依存词的特征，因此在两个数据集上的实验结果都有一定程度的提高。CopyRRL[29]在CopyRE模型的基础上加入了强化学习，考虑了关系三元组的提取顺序对抽取任务的影响，模型整体

有着不错的效果，但其采用的复制机制仍未解决实体复制不完整的问题。ETL-Span[23]模型将抽取任务分解为序列标记问题，在WebNLG上的抽取结果有了很大程度的提升。CasRel[24]更是在此基础上构造了一个全新的指针标注框架，抽取效果达到了领域内最优。在NYT和WebNLG数据集上，$JPEA_{BERT}$模型的$F1$与CasRel模型相比分别提高了1.5个百分点和0.5个百分点，说明对头实体标记器的输入向量做特征加强和对关系及尾实体标记器的输入向量做实体级特征融合对提升模型的抽取效果有很大的贡献。JPEA的抽取性能优于$JPEA_{BERT}$，其原因在于RoBERTa采用了更多的数据进行训练，可以更充分地学习句子的上下文信息。当编码器采用ALBERT预训练模型时，模型的整体表现相较于CasRel有微提升，但与JPEA相比存在差距。原因在于相较于BERT的其他变体来说，ALBERT模型的参数大大减少，导致模型学习得不够充分，不过参数的减少在一定程度上提升了训练的速度。$JPEA_{ELECTRA}$模型在两个数据集上的$F1$值都达到了最佳，因为ELECTRA模型的预训练任务是替换标记检测，这对模型学习能力的提升有一定程度的影响。图2所示为JPEA模型在两个公开数据集上训练时损失函数的变化情况。可以看出随着训练周期的增长，损失值保持下降状态，最终均在第50个训练周期左右损失值降到最低，模型根据早停机制停止训练。

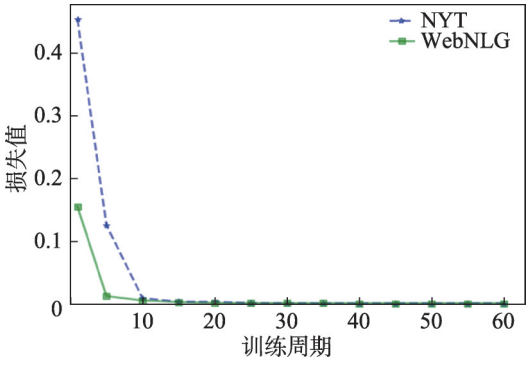


图2 训练损失值

Fig.2 Training loss value

3.3.2 消融实验分析

为了验证BiLSTM网络结合自注意力机制提取特征和CNN与注意力机制特征融合两个模块对JPEA模型性能的增益作用，本文在两个公开语料库上进一步做了消融实验，实验结果如表4所示。其中，JPEA-BAM表示将经过编码层得到的句子表征直接进行头实体标记；JPEA-LAN表示将所有头实体向量的平均值与句子的表征直接拼接。

表4 在两个数据集上的消融实验结果

Table 4 Results of ablation experiments on two datasets 单位：%

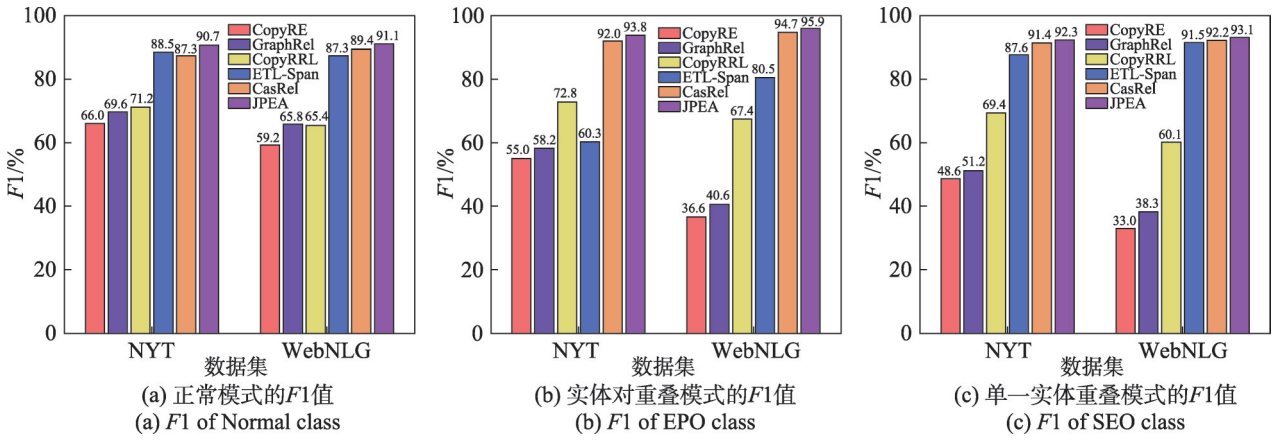
模型	NYT			WebNLG		
	准确率	召回率	$F1$	准确率	召回率	$F1$
JPEA-BAM	91.3	92.4	91.8	93.0	**92.2**	92.6
JPEA-LAN	**92.3**	89.1	90.7	93.2	90.6	91.9
JPEA	91.6	**92.5**	**92.0**	**93.8**	91.9	**92.8**

对比实验结果可以发现，两个模块组件都对JPEA模型性能的提高作出了积极贡献。JPEA-LAN模型性能在NYT数据集上下降了1.3个百分点，对模型的影响较大，这说明通过卷积神经网络提取头实体特征并利用注意力机制加权进行特征融合，可以有效地利用头实体信息辅助尾实体及关系标注，忽略其他冗余信息，从而使最终抽取的三元组更加准确。JPEA-BAM模型在两个数据集上的$F1$值与JPEA模型均相差0.2个百分点，可以得出，结合BiLSTM网络与自注意力机制进一步学习句子的内在特征，能够获得细粒度语义信息，更有利于头实体的检测。

3.3.3 重叠问题实验分析

为了进一步验证JPEA模型解决重叠三元组问题的有效性，本文对Normal、EPO和SEO三种模式进行扩展实验，并与基线模型进行对比，在两个数据集上的$F1$值对比情况如图3所示。

由图3可见，在两个数据集上，JPEA模型在三种不同模式下的$F1$值均有很好且较为一致的表现，尤其是在EPO和SEO两种重叠模式下，$F1$值有明显的提高，说明本文所提出的模型在解决重叠关系三元组的提取问题上有优异的表现。其次，可以观察到大多数的基线模型在正常、EPO和SEO三种重叠模式下的抽取性能都呈现依次下降趋势，也就是说，基线模型提取EPO和SEO两种重叠模式的能力有所欠缺。而相比之下，JPEA模型在三种重叠模式下的表现都不一般。这是因为这些基线模型的结构存在一定的缺陷，它们将实体对映射到关系或者选择Sequence-to-Sequence的模型架构。将实体对映射到关系很容易产生冗余实体对问题，导致较高的错误率，难以高效解决关系重叠的问题。并且Sequence-to-Sequence模式设计复杂的解码架构使得局部特征抽取不充分，导致抽取的三元组不够完整。尽管CasRel模型对于重叠关系三元组的提取有



(a) 正常模式的$F1$值 (a) $F1$ of Normal class

(b) 实体对重叠模式的$F1$值 (b) $F1$ of EPO class

(c) 单一实体重叠模式的$F1$值 (c) $F1$ of SEO class

图3 从不同模式的句子中抽取三元组的 $F1$ 值

Fig.3 $F1$-score of extracting relational triples from sentences with different patterns

着良好的表现，但相较于JPEA模型在处理复杂句子方面仍有不足。因为除了采用分层指针标注策略将头实体映射到关系和尾实体上之外，RoBERTa模型改进了优化函数，使用了更大的数据集进行预训练，更完整地提取语句的上下文信息，而BiLSTM网络结合注意力机制更是对句子表征向量进行了深层次的学习。同时在尾实体及关系标记前利用CNN和注意力机制进行头实体特征与句子向量的融合，增强了JPEA模型两个模块之间的依赖。综上所述，JPEA模型具有较强的处理复杂文本的能力，且在解决重叠关系三元组问题上有着良好的高效性。

4 结束语

本文基于预训练模型提出了一种可以解决实体关系抽取过程中三元组重叠问题的JPEA模型。该模型通过预训练模型编码得到包含上下文信息的句子向量，再将句子向量输入BiLSTM网络和自注意力机制得到更精确的句子级特征表示，在关系及尾实体标记任务之前添加特征融合层，利用CNN和注意力机制将头实体特征与句子向量融合，强化头实体标记与关系及尾实体标记模块之间的内在联系。实验结果表明，所提JPEA模型在重叠关系三元组的抽取任务中有着良好表现，当选用不同的预训练模型时，总体模型在两个数据集上执行的抽取工作都能取得不错的结果，其中基于ELECTRA模型的效果最佳。

本文所提模型虽然在抽取的准确率上有一定程度的提升，但是模型的稳定性较差，一旦数据集中带有错误标签的样本过多，JPEA模型的性能就会受到影响并产生波动。因此，在后续的工作中如何进一步提升模型的性能、增强模型的稳定性是亟需解决的主要问题。本文提出的模型目前主要针对公共数据集进行测试，而近年来随着领域知识图谱构建技术的发展，将实体关系抽取技术应用到垂直领域来构建领域知识图谱变得更加有意义。特别地，结合城市规划领域知识图谱对规划活动做决策提供辅助至关重要，因此接下来将会深入城市规划领域对本模型进行改进，为城市规划领域知识图谱的构建及应用做好铺垫工作。

参考文献：

[1] PUJARA J, HUI M, GETOOR L, et al. Knowledge graph identification[C]//LNCS 8218: Proceedings of the 12th International Semantic Web Conference, Sydney, Oct 21-25, 2013. Berlin, Heidelberg: Springer, 2013: 542-557.

[2] 范媛媛, 李忠民. 中文医学知识图谱研究及应用进展[J]. 计算机科学与探索, 2022, 16(10): 2219-2233.
FAN Y Y, LI Z M. Research and application progress of Chinese medical knowledge graph[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(10): 2219-2233.

[3] 薛丽娟, 席梦隆, 王梦婕, 等. 基于规则推理引擎的实体关系抽取研究[J]. 计算机科学与探索, 2016, 10(9): 1310-1319.
XUE L J, XI M L, WANG M J, et al. Entity relation extraction based on rule inference engine[J]. Journal of Frontiers of Computer Science and Technology, 2016, 10(9): 1310-1319.

[4] 张少伟, 王鑫, 陈子睿, 等. 有监督实体关系联合抽取方法研究综述[J]. 计算机科学与探索, 2022, 16(4): 713-733.
ZHANG S W, WANG X, CHEN Z R, et al. Survey of super-

vised joint entity relation extraction methods[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16 (4): 713-733.

[5] JI S X, PAN S R, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494-514.

[6] 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.
E H H, ZHANG W J, XIAO S Q, et al. Survey of entity relation extraction based on deep learning[J]. Journal of Software, 2019, 30(6): 1793-1818.

[7] ZHUANG C Z, ZHANG N Y, JIN X L, et al. Joint extraction of triple knowledge based on relation priority[C]//Proceedings of the 2020 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking, Exeter, Dec 17-19, 2020. Piscataway: IEEE, 2020: 562-569.

[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805, 2018.

[9] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv:1907.11692, 2019.

[10] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 3(3): 1083-1106.

[11] HENDRICKX I, KIM S N, KOZAREVA Z, et al. SemEval-2010 Task 8: multi-way classification of semantic relations between pairs of nominals[C]//Proceedings of the 5th International Workshop on Semantic Evaluation, Los Angeles, Jul 15-16, 2010. Stroudsburg: ACL, 2010: 33-38.

[12] SHINYAMA Y, SEKINE S. Preemptive information extraction using unrestricted relation discovery[C]//Proceedings 2006 of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, Jun 4-9, 2006: 304-311.

[13] CARLSON A, BETTERIDGE J, WANG R C, et al. Coupled semi-supervised learning for information extraction[C]//Proceedings of the 3rd ACM International Conference on Web Search and Web Data Mining, New York, Feb 4-6, 2010. New York: ACM, 2010: 101-110.

[14] ZHANG Y, ZHOU J F. A trainable method for extracting Chinese entity names and their relations[C]//Proceedings of the 2nd Workshop on Chinese Language Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, Oct 8, 2000. Stroudsburg: ACL, 2000: 66-72.

[15] KAMATH S, KARIBASAPPA K G, REDDY A, et al. Improving the relation classification using convolutional neural network[J]. IOP Conference Series: Materials Science and Engineering, 2021, 1187(1): 012004.

[16] GUO X Y, ZHANG H, YANG H J, et al. A single attention-based combination of CNN and RNN for relation classification[J]. IEEE Access, 2019, 7: 12467-12475.

[17] RONRAN C, LEE S. Effect of character and word features in bidirectional LSTM-CRF for NER[C]//Proceedings of the 2020 IEEE International Conference on Big Data and Smart Computing, Busan, Feb 19-22, 2020. Piscataway: IEEE, 2020: 613-616.

[18] 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, Jeju Island, Jul 12-14, 2012. Stroudsburg: ACL, 2012: 1201-1211.

[19] ZENG D J, LIU K, LAI S W, et al. Relation classification via convolutional deep neural network[C]//Proceedings of the 25th International Conference on Computational Linguistics, Dublin, Aug 23-29, 2014. Stroudsburg: ACL, 2014: 2335-2344.

[20] CAI R, ZHANG X D, WANG H F. Bidirectional recurrent convolutional neural network for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Aug 7-12, 2016. Stroudsburg: ACL, 2016: 756-765.

[21] 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, Berlin, Aug 7-12, 2016. Stroudsburg: ACL, 2016: 1105-1116.

[22] ZHENG S C, FENG W, 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, Vancouver, Jul 30-Aug 4, 2017. Stroudsburg: ACL, 2017: 1227-1236.

[23] YU B W, ZHANG Z Y, SHU X B, et al. Joint extraction of entities and relations based on a novel decomposition strategy[C]//Proceedings of the 24th European Conference on Artificial Intelligence, Santiago de Chile, Aug 29-Sep 8, 2020. Amsterdam: IOS Press, 2020: 2282-2289.

[24]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, Jul 5-10, 2020. Stroudsburg: ACL, 2020: 1476-1488.

[25] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[J]. arXiv:1909.11942, 2019.

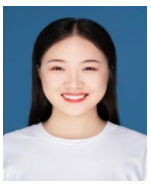
[26] CLARK K, LUONG M T, LE Q V, et al. ELECTRA: pre-training text encoders as discriminators rather than generators[J]. arXiv:2003.10555, 2020.

[27] ZENG X R, ZENG D J, HE S Z, et al. Extracting relational facts by an end-to-end neural model with copy mechanism [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Jul 15-20, 2018. Stroudsburg: ACL, 2018: 506-514.

[28] FU T J, LI P H, MA W Y. GraphRel: modeling text as relational graphs for joint entity and relation extraction[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Jul 28-Aug 2, 2019. Stroudsburg: ACL, 2019: 1409-1418.

[29] ZENG X R, HE S Z, ZENG D J, et al. Learning the extraction order of multiple relational facts in a sentence with reinforcement learning[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language, Hong Kong, China, Nov 3-7, 2019. Stroudsburg: ACL, 2019: 367-377.

李智杰(1980—),男,河南人,博士,副教授,主要研究方向为人工智能及其行业应用、数字建筑。
LI Zhijie, born in 1980, Ph.D., associate professor. His research interests include artificial intelligence and its industrial application, digital architecture.



韩瑞瑞(1998—),女,河南人,硕士研究生,主要研究方向为人工智能及其行业应用。
HAN Ruirui, born in 1998, M.S. candidate. Her research interests include artificial intelligence and its industrial application.

李昌华(1963—),男,宁夏人,博士,教授,主要研究方向为人工智能及其行业应用、数字建筑。
LI Changhua, born in 1963, Ph.D., professor. His research interests include artificial intelligence and its industrial application, digital architecture.

张颉(1989—),男,陕西人,硕士研究生,助教,主要研究方向为数字建筑、模式识别。
ZHANG Jie, born in 1989, M.S. candidate, assistant. His research interests include digital architecture and pattern recognition.

石昊琦(1992—),男,宁夏人,硕士,主要研究方向为模式识别、数字建筑、优化算法。
SHI Haoqi, born in 1992, M.S. His research interests include pattern recognition, digital architecture and optimization algorithm.