研究与开发

融合混合嵌入与关系标签嵌入的三元组联合抽取方法

戴剑锋，陈星妤，董黎刚，蒋献
（浙江工商大学，浙江 杭州 310018）

摘　要：三元组抽取的目的是从非结构化的文本中获取实体与实体间的关系，并应用于下游任务。嵌入机制对三元组抽取模型的性能有很大影响，嵌入向量应包含与关系抽取任务密切相关的丰富语义信息。在中文数据集中，字词之间包含的信息有很大区别，为了改进由分词错误产生的语义信息丢失问题，设计了融合混合嵌入与关系标签嵌入的三元组联合抽取方法（HEPA），提出了采用字嵌入与词嵌入结合的混合嵌入方法，降低由分词错误产生的误差；在实体抽取层中添加关系标签嵌入机制，融合文本与关系标签，利用注意力机制来区分句子中实体与不同关系标签的相关性，由此提高匹配精度；采用指针标注的方法匹配实体，提高了对关系重叠三元组的抽取效果。在公开的 DuIE 数据集上进行了对比实验，相较于表现最好的基线模型（CasRel），HEPA 的 *F*1 值提升了 2.8%。

关键词：三元组抽取；关系嵌入；BERT；注意力机制；指针标注
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2023021

A triple joint extraction method combining hybrid embedding and relational label embedding

DAI Jianfeng, CHEN Xingyu, DONG Ligang, JIANG Xian
Zhejiang Gongshang University, Hangzhou 310018, China

Abstract: The purpose of triple extraction is to obtain relationships between entities from unstructured text and apply them to downstream tasks. The embedding mechanism has a great impact on the performance of the triple extraction model, and the embedding vector should contain rich semantic information that is closely related to the relationship extraction task. In Chinese datasets, the information contained between words is very different, and in order to avoid the loss of semantic information problems generated by word separation errors, a triple joint extraction method combining hybrid embedding and relational label embedding (HEPA) was designed, and a hybrid embedding means that combines letter embedding and word embedding was proposed to reduce the errors generated by word separation errors. A relational embedding mechanism that fuses text and relational labels was added, and an attention mechanism

收稿日期：2022−07−12；**修回日期**：2023−01−20
通信作者：董黎刚，donglg@zjgsu.edu.cn
基金项目：国家社会科学基金资助项目（No.17BYY090）；浙江省重点研发计划项目（No.2017C03058）；浙江省“尖兵”“领雁”研发攻关计划项目（No.2023C03202）
Foundation Items: The National Social Science Foundation of China (No.17BYY090), Zhejiang Province Key Research and Development Program (No.2017C03058), Zhejiang Province “Top Soldiers” and “Leading Geese” Project (No.2023C03202)

was used to distinguish the relevance of entities in a sentence with different relational labels, thus improving the matching accuracy. The method of matching entities with pointer annotation was used, which improved the extraction effect on relational overlapping triples. Comparative experiments are conducted on the publicly available DuIE dataset, and the *F*1 value of HEPA is improved by 2.8% compared to the best performing baseline model (CasRel).

Key words: triple extraction, relational embedding, BERT, attention mechanism, pointer annotation

0　引言

三元组的自动抽取是自然语言处理领域的一个热门研究课题，它能够从非结构化文本中提取结构化信息，并应用于各类下游任务，如知识图谱、智能问答等。三元组可表示为：<头实体 *s*，关系 *r*，尾实体 *o*>。现有的三元组抽取方法按照建模类型主要可分为两类：流水线法（pipeline）和联合抽取法（joint）。流水线法将三元组抽取任务分割成两个独立的子任务：命名实体识别（named entities recognition，NER）和关系抽取（relation extraction，RE）。首先进行命名实体识别，提取文本中的实体，然后进行关系抽取，使用分类模型匹配各实体对之间的关系。这种串联模型在建模难度上相对简单，但将命名实体识别和关系抽取视作两个独立的任务处理时，存在实体冗余、误差累计、信息丢失等问题，限制了进一步的研究。为了解决流水线法存在的问题，学者们提出用联合抽取法对三元组进行抽取，同时从输入文本中抽取实体及实体间的对应关系[1]。与流水线方法相比，联合抽取法整合了实体和关系信息，有效减少了误差传播，取得了更好的效果。

目前，大部分三元组抽取研究不能较好地处理重叠三元组问题。在三元组抽取任务中，经常会出现同一句子存在多个三元组共享相同的头实体、关系或尾实体的情况。例如“邓超既是《银河补习班》这部电影的导演又是主演。”这句话包含<《银河补习班》，导演，邓超>、<《银河补习班》，主演，邓超>两个三元组，且“《银河补习班》”和“邓超”两个实体间存在多个关系。学者们将这一类共享实体关系的三元组命名为重叠三元组。

重叠三元组按照实体重叠程度可以分为 3 种情况，如图 1 所示，分别为无重叠（normal）、实体对重叠（entity pair overlap，EPO）、单实体重叠（single entity overlap，SEO）。normal 表示

关系重叠类型	文本	实体关系
normal	北京位于中国	北京 —国家→ 中国
EPO	邓超既是《银河补习班》的导演又是主演	邓超 —导演→ 《银河补习班》；邓超 —主演→ 《银河补习班》
SEO	周杰伦，1979年1月18日出生于中国台湾省新北市，是我国知名歌手	周杰伦 —出生地→ 中国台湾省新北市；周杰伦 —国籍→ 中国；周杰伦 —出生日期→ 1979年1月18日

图 1　重叠三元组类型

文本中的实体之间只存在一种关系，不存在关系重叠三元组；EPO 表示相同的两个实体之间存在多个实体关系；SEO 表示一个头实体与多个不同的尾实体存在实体关系。

在联合抽取模型中，对文本中实体进行识别往往选用序列标注的方法。每个字词都会被标注成特有的序列，例如头实体开始、头实体结束、无关词、关系、尾实体开始、尾实体结束。这种标注方法无法将一个词语同时标注成头实体和尾实体，对重叠三元组问题处理效果较差。流水线抽取模型虽然选用遍历所有提取的实体对的方法来解决重叠三元组的提取问题，但太过依赖命名实体识别的准确性，一旦实体识别出错，误差就会累积到下一个任务中，将引入大量错误、冗余的实体对，导致提取性能显著下降。

针对上述问题，本文在联合抽取法的基础上提出一种融合混合嵌入与关系标签嵌入的三元组联合抽取方法（HEPA），融合词句间的文本信息，提高对重叠三元组的抽取精度。本文的主要工作如下：首先针对嵌入方法中忽视字词之间潜在语义关系而导致分词歧义的问题，设计了一种混合嵌入方法，结合字词以及位置信息将输入文本转化为向量，降低由分词错误产生的误差。其次，由于头实体和尾实体间存在位置联系，设计了实体位置注意力机制，赋予实体位置信息权重，从多维度获取文本信息，提高三元组抽取的精度。最后，在 DuIE 数据集上进行了测试，HEPA 相较于其他基线模型在 *F*1 值上有所提升。

1 相关工作

本节主要介绍了三元组抽取中的两种主流方法：流水线法和联合抽取法。

1.1 流水线法

流水线法将三元组抽取的过程分为命名实体识别和关系抽取两个子任务，彼此相互独立。首先通过命名实体识别提取出文本中的实体，再通过关系抽取对每个候选实体进行关系预测，最后以三元组的形式输出预测结果。

Zeng 等[2]首次提出使用具有最大池化（max pooling）的卷积深度神经网络（deep neural network，DNN）算法提取词语和句子级别的特征，将得到的词向量作为模型的原始输入，通过隐藏层和 softmax 层进行关系分类。该模型提出了位置特征来编码当前词与目标词对的相对距离，同时说明位置特征是比较有效的特征。该方法在 SemEval-2010 数据集上达到了最佳效果。Xu 等[3]在 Zeng 等[2]的研究基础上进行改进，使用最短依存路径长短期记忆（the shortest dependency path long short-term memory，SDP-LSTM）网络进行实体关系抽取，把路径节点表示成向量，将词本身、词性信息、句法依存关系、WordNet 上位词等 4 种词信息看作 4 个通道，输入长短期记忆（long short-term memory，LSTM）网络进行前向传播，每一个通道都有一个输出，将所有输出堆叠处理并进行池化操作，最后对 4 个通道输出的隐向量进行拼凑，通过 softmax 层产生最终输出。在训练过程中发现实体间的距离对关系抽取的效果有较大的影响，于是添加了负实体采样策略消除由依存路径分析引入的噪声影响。Socher 等[4]针对单个词向量模型无法捕获长句子合成性信息的问题，设计了一种基于矩阵向量循环神经网络（recurrent neural network，RNN）的抽取模型，提高了模型对任意长度的短语和句子词向量共同表征的学习能力。但 RNN 模型存在长期依赖问题，容易丢失上下文信息。

为了解决这一问题，改善对长难句的建模效果，Li 等[5]提出了一种基于低成本序列特征的 Bi-LSTM-RNN 模型，通过实体周围的分段信息获取更多的语义信息，不需要额外特征帮助。LSTM 模型虽然有效解决了长期依赖问题，但对关键信息的注意不足，难以处理复杂的关系抽取问题。

Su 等[6]在 CNN 模型的池化层加入注意力机制，过滤文本中无关的噪声数据，从而使得模型专注于目标实体特征。Vashishth 等[7]在多实例设置中使用了图卷积神经网络（graph convolutional neural network，GCN）。他们在整个句子依赖树上使用双向门控循环单元（bidirectional gate recurrent unit，Bi-GRU）层和 GCN 层对句子进行编码。将词袋中的句子表示进行聚合并传递给分类器来寻找它们之间的关系。杨帅等[8]提出了一种基于多通道的边学习 GCN，提高了图学习多维边特征学习的能力，拓展了 GCN 在关系抽取领域的应用。

在中文领域中，为了解决流水线方法存在的误差累计问题，李昊等[9]提出一种基于实体边界组合的关系抽取方法，跳过命名实体识别，直接对实体边界信息两两组合来进行关系抽取。由于边界信息性能高于实体性能，所以误差累计的问题得到了一定程度缓解，在 ACE 2005 中文数据集上进行了实验，其 $F1$ 值提高了 13.95%。Zhong 等[10]提出了一种双编码器抽取模型，独立学习两个编码器进行实体识别和关系提取，简单地在两个实体的前后各插入了开始和结束标签，获得了非常好的效果，在多个数据集上都有较大提升，为流水线法提供了新的思路。

虽然流水线方法在建模难度上相对较低，但是存在 3 个主要问题。首先，这种模型容易出现错误传播的情况，命名实体识别环节产生的错误无法及时进行检验纠正，而且会累积到关系抽取环节中，从而影响后续实体关系抽取的效果。其次，不相关的实体对在匹配过程中会产生大量干扰信息，这些干扰信息也会影响模型的性能。最后，分割命名实体识别与实体关系抽取这两个子任务会造成文本信息丢失的问题，影响模型效果。

1.2　联合抽取法

为了解决流水线法存在的问题，越来越多的学者倾向于设计联合模型对三元组进行整体抽取。

不同于流水线法，联合抽取法将命名实体识别和关系抽取两个步骤进行联合建模，在抽取实体的同时分类实体关系。联合抽取法的优点是能够减少误差累计，增强子任务之间的联系。Miwa 等[11]将神经网络应用于联合抽取模型，选用双向序列 LSTM-RNN 对句子的词语顺序信息和依存句法树结构信息进行建模，并将两个模型组合起来，使得关系抽取的过程中可以利用与实体相关的信息。Katiyar 等[12]对 Miwa 等[11]设计的模型进行了改进，引入注意力机制和指针网络，将注意力机制与实体指针、关系指针结合，能够更精准地抽取实体间关系，同时扩展了标签关系类型。Zheng 等[13]选用 LSTM 模型将联合关系抽取任务转化为序列标注任务，选用就近原则进行实体关联。但该模型忽略了句子中存在多个实体关系重叠的问题。

Zeng 等[14]注意到实体关系抽取过程中的关系重叠问题，并提出利用 Seq2Seq 模型进行实体关系联合抽取，在模型中添加了复制机制来解决重叠问题，可以从句子中联合提取关系事实。但该模型太过依赖解码的准确率，可能会导致实体识别不全。Fu 等[15]用依存句法将句子转化为依存树，再通过加权图卷积神经网络（GCN）改进的方法，计算实体对关系的权重，从而解决实体关系重叠的问题，效果比 Zeng 等[14]的模型有所提高。Duan 等[16]提出了一种结合多头注意力机制的图卷积神经网络（MA-DCGCN）模型。在该模型中，多头自注意力机制专门用于将权重分配给实体之间的多个关系类型，以确保多个关系的概率空间不相互排斥，并自适应地提取重叠实体间的多种关系。Wei 等[17]提出一种基于二进制指针序列标注的模型。首先使用两个二进制分类器识别出句子中的所有实体，然后遍历所有实体关系标签，根据语义相似度进行尾实体标注。该模型为重叠三元组抽取提供了新的思路。Wang 等[18]设计了握手标记策略，通过对句子中的主语和谓语的

首字符建立 3 种标注标签，在给定 scheme 下进行分类训练，通过穷举存在判断的解码实现对重叠关系三元组的抽取。

在中文领域中，联合抽取法也有着广泛的应用。田佳来等[19]采用一种新的标记方案，将关系抽取问题转化成序列标注问题，同时针对三元组重叠问题，采用分层的序列标注方式来解决，在某中文数据集上 $F1$ 值达到 80.84%。苗琳等[20]设计了一种基于图神经网络的实体关系联合抽取模型，将重心放在实体与关系间的相互作用，将实体抽取的范围扩大到每个实体的局部特征，结合图卷积网络对每个实体对进行关系预测，在数据集上进行实验，对比基线模型有 5.2%的提升。针对关系抽取中存在多跳关系的情况，王红等[21]提出了一种基于多跳注意力的实体关系联合抽取方法，先标记头实体，输出其多关系尾实体，然后将尾实体作为下一跳的头实体进行输入，迭代执行关系抽取直到输出最终的实体关系。这一方法充分利用了实体间潜在的隐性关系，对复杂的多跳关系抽取效果极佳，实验表明该方法在民航突发事件数据集中有出色表现。

综上所述，已有较多联合抽取模型在不同领域的研究中取得了不错的成果。但联合抽取模型仍然存在语义信息缺失、精度要求高等问题，而且大多数模型不能较好地处理三元组重叠的情况。针对这些问题，本文提出了一种基于混合关系嵌入的三元组抽取方法。

2 融合混合嵌入与关系标签的三元组联合抽取方法

三元组抽取的目的是从文本中提取出结构化信息，以 (s,r,o) 的形式输出，其中，s 为头实体，o 为尾实体，r 为 s 与 o 的关系。本文提出的三元组抽取方法参考 Seq2Seq 模型的思路，先提取头实体，然后根据实体关系确定尾实体。三元组抽取过程可用式（1）表示：

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

从输入的句子中提取三元组时，将尾实体 o 和关系 r 的分类看成一个整体，根据已提取的头实体 s 进行尾实体 o 和关系 r 的联合预测，如式（2）所示：

$$P(s,r,o\mid x)=P(s\mid x)P(r,o\mid s,x) \tag{2}$$

其中，x 为输入的句子。首先抽取出头实体 s，然后将头实体 s 与句子 x 作为已知条件，联合解码获得关系 r 和尾实体 o。

对式（2）进行分析可知，对一个输入文本进行关系抽取，输出结果只能获得一个三元组，忽略了三元组重叠的情况。因此，选用指针标注的方法对 Seq2Seq 模型的三元组抽取过程进行改进。先对句子中的所有头实体 s 进行抽取，每个实体再遍历所有关系标签 r，匹配出合适的尾实体 o。

HEPA 模型结构如图 2 所示，模型主要可以分为编码层、头实体标注层和关系匹配层 3 个部分。

编码层分别编码输入文本以及关系标签。文本输入到向量混合嵌入层，结合字嵌入的灵活性与词嵌入的语义关系，融合位置信息与关系标签内容进行混合嵌入。得到混合向量后将其输入到采用 BERT 预训练模型的编码层中进行编码，经过多头注意力机制丰富语义特征。

头实体标注层解码由 BERT 编码器产生的编码向量来识别输入语句中的所有可能头实体，经过标签注意力机制标记实体与关系之间的关联程度，其中色块颜色越深，代表标签与实体间的关系越紧密。最后用二进制标注器标注出头实体的开始位置与结束位置。

在实体关系匹配层中，对标注出来的头实体遍历预先设定好的关系标签，为每个关系标签匹配最接近的尾实体，并用标注器标注在文本中的位置。对每个头实体都要进行一次实体关系匹配，最终为所有头实体匹配实体关系与尾实体，并转换为三元组输出。

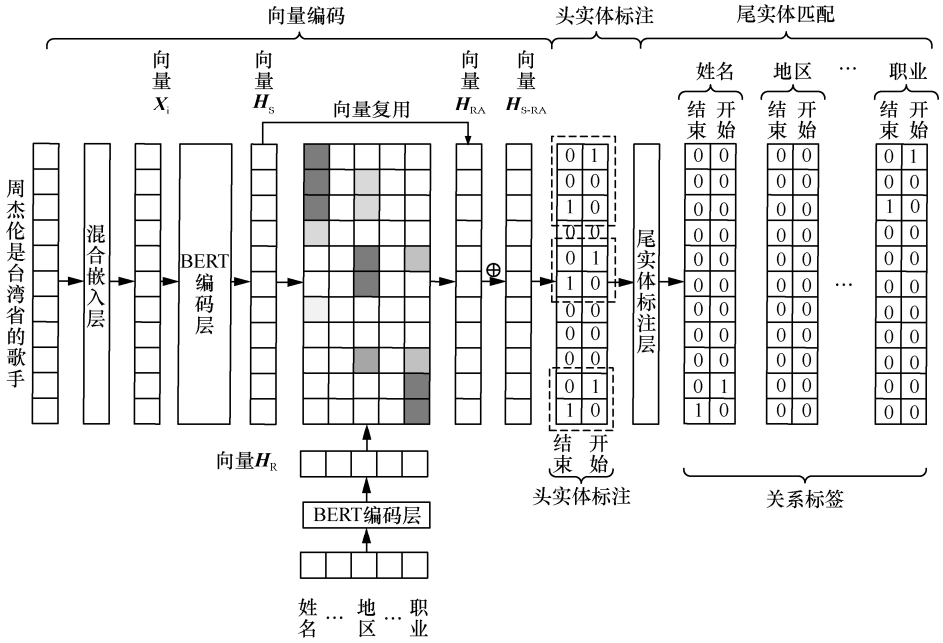

图 2　HEPA 模型结构

2.1　编码层

编码层首先从自然语言语句中提取特征信息，对上下文信息进行编码，将其输入后续的编码模块中。

2.1.1　字词混合嵌入

嵌入的作用是将文本转化为向量的形式，便于后续模型输入。通常文本嵌入选用字嵌入的方法，即将句子中的每个字或单词都单独转换成一个向量，这种方法降低了分词错误引起的误差，是目前主流的嵌入方法。但是字嵌入会导致文本信息缺失，单纯的字嵌入难以存储有效的语义信息。因此选用字嵌入和词嵌入结合的混合嵌入方法，结合文本的位置信息，既保留了字嵌入的灵活的特点，又减少了文本信息的缺失。混合嵌入过程如下：首先输入字 ID 序列到字嵌入层得到字嵌入向量，然后将文本分词，将分词后的结果输入到 Word2Vec 模型中，提取对应的词向量。为了合并字向量与词向量，需要将得到的词向量重复 n 次，n 为词的长度，这一步的目的是得到与字向量长度对齐的词向量序列。为了对齐字向量的维度，将词向量序列输入一个变换矩阵 $\boldsymbol{E}$ 中。变换矩阵 $\boldsymbol{E}$ 是根据字嵌入向量的变化而变化的，规定先保持词嵌入向量不变，通过不断变化字嵌入向量，生成能将词嵌入向量转化为同一维度的变换矩阵，对词嵌入向量进行微调，最终得到与字嵌入向量相同维度的词嵌入向量，再将二者合并，得到混合向量。

$$\boldsymbol{t}_i = (\boldsymbol{w}_i \times \boldsymbol{E}) \oplus \boldsymbol{c}_i \tag{3}$$

其中，$\boldsymbol{t}_i$ 表示第 i 个混合嵌入向量，$\boldsymbol{w}_i$ 表示第 i 个词向量，$\boldsymbol{E}$ 表示变换矩阵，$\boldsymbol{c}_i$ 表示第 i 个字向量。混合嵌入过程如图 3 所示。

在进行嵌入的过程中，常常忽略位置信息的重要性。例如头实体 s 往往出现在句子的开头部分，头实体 s 与尾实体 o 之间的距离不会太远，因此选用位置嵌入的方法加入文本中的位置信息。

首先设定一个文本的最大长度，将文本位置编码输入一个初始化的嵌入层中，输出位置向量 $\boldsymbol{p}$，将位置向量 $\boldsymbol{p}$ 加到混合嵌入向量 $\boldsymbol{t}$ 中，作为完整的嵌入向量输出，该过程可以用式（4）表示：

$$\boldsymbol{x}_i = \boldsymbol{p}_i + \boldsymbol{t}_i \tag{4}$$

其中，$\boldsymbol{x}_i$ 表示第 i 个字符输出的完整混合向量，$\boldsymbol{p}_i$

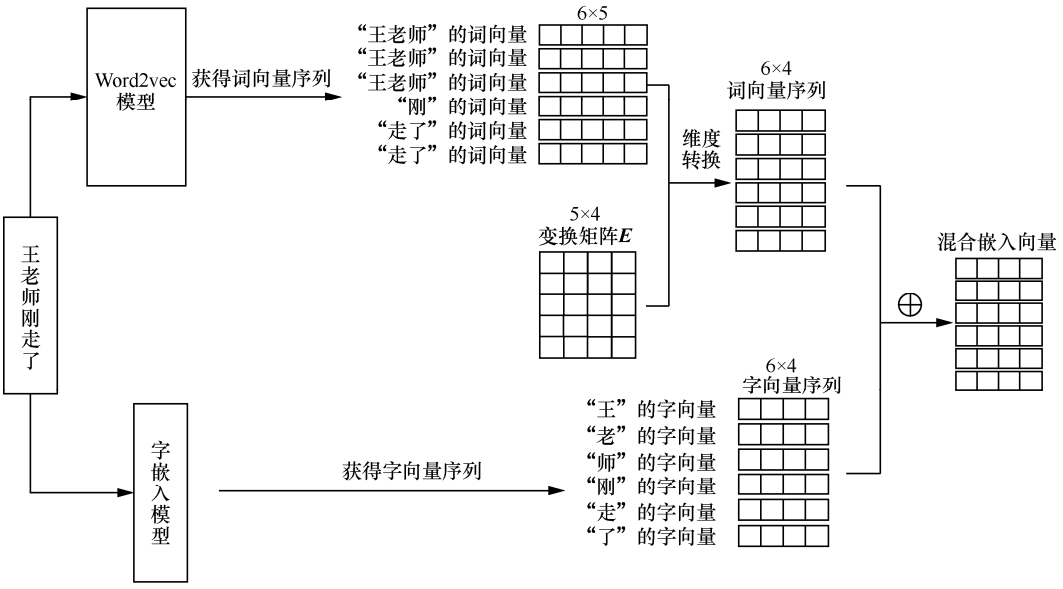

图 3　混合嵌入过程

表示第 i 个字符的位置向量。

2.1.2　BERT 编码

2018 年，Devlin 等[22]提出了经典的 BERT（bidirectional encoder representation from transformers）模型，这是一个预训练的双向编码表征模型。以往模型训练出来的词向量是静态的，也就是与上下文无关，它们没有解决歧义问题。例如“王老师刚刚走了。”中的“走了”可以指代离开的意思，特殊场景下也可以指代去世的意思。BERT 的出现解决了这一问题，BERT 会将每个单词与句子中其他单词计算相关性，以此来获得每个单词的上下信息，根据不同上下文生成对应的词向量，更符合人类的理解逻辑。因此选用 BERT 预训练模型来进行向量编码。

混合嵌入层最终输出的混合向量 $\boldsymbol{x}$ 作为 BERT 的输入，经过 BERT 的编码后输出的特征向量 $\boldsymbol{h}$，该过程用计算式表达为：

$$\boldsymbol{h}_i = \mathrm{BERT}(\boldsymbol{x}_i) \tag{5}$$

$$H_s = \mathrm{sum}(\boldsymbol{h}_i) \tag{6}$$

其中，$\boldsymbol{x}_i$ 表示句子中第 i 个字符的混合向量，$\boldsymbol{h}_i$ 表示 $\boldsymbol{x}_i$ 经过 BERT 模型编码后的特征向量，H_s 是 $\boldsymbol{h}_i$ 的集合。$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} \in X$，$\{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\} \in H$。

2.1.3　关系嵌入

HEPA 模型在对输入语句进行嵌入的同时，还加入了关系嵌入机制。将该机制队的所有关系标签进行编码嵌入，转化为关系向量，利用注意力机制区分不同关系标签与句子中实体的相关程度，从而将关系标签信息整合到字词嵌入中。融合字词嵌入向量与关系嵌入向量，可以利用关系标签信息来丰富给定句子中每个实体间关系，有助于提高对每个三元组抽取的精度。

关系嵌入给定一个关系标签类型 r_i 和对应的关系类型集合 R 中的所有关系标签，利用 BERT 预训练模型编码为向量 $\boldsymbol{H}_{r_i}$，具体计算式如下：

$$\boldsymbol{H}_{r_i} = \mathrm{BERT}(r_i) \tag{7}$$

$$h_{r_i} = \mathrm{sum}(\boldsymbol{H}_{r_i}) \tag{8}$$

$$\boldsymbol{H}_{\mathrm{R}} = \mathrm{concat}(h_{r_1}, h_{r_2}, \cdots, h_{r_N}) \tag{9}$$

其中，h_{r_i} 是 $\boldsymbol{H}_{r_i}$ 的集合，R 是所有关系标签的集合，将每个经过编码的 h_{r_i} 组合成关系嵌入向量 $\boldsymbol{H}_{\mathrm{R}}$，$N$ 是嵌入的标签数量。

获得关系嵌入向量 $\boldsymbol{H}_{\mathrm{R}}$ 后，将其输入互注意力机制。互注意力机制是注意力机制的变形应用，

可以通过学习输入语句的内部结构，计算自身与其他字符的关联性。选用互注意力机制来从多维度语句中获得字符的上下文信息，输入特征向量序列，计算字符间的关联度，捕获长距离依赖关系。互注意力机制用于丰富给定句子中每个标记的信息，计算得到关系注意权重 $\text{Attention}_{\text{REL}}$。该权重表示句子中每个实体嵌入与关系嵌入之间的相关性，$\text{Attention}_{\text{REL}}$ 越高，代表实体与关系之间的联系越紧密，反之则越疏远。$\text{Attention}_{\text{REL}}$ 与值矩阵 $\boldsymbol{V}$ 相乘得到 $\boldsymbol{H}_{\text{RA}}$，具体计算式如下：

$$\boldsymbol{Q}=\boldsymbol{H}_{\text{S}}\boldsymbol{W}_{Q} \tag{10}$$

$$\boldsymbol{K}=\boldsymbol{H}_{\text{R}}\boldsymbol{W}_{K} \tag{11}$$

$$\boldsymbol{V}=\boldsymbol{H}_{\text{R}}\boldsymbol{W}_{V} \tag{12}$$

$$\text{Attention}_{\text{REL}}=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d^{k}}}\right) \tag{13}$$

$$\boldsymbol{H}_{\text{RA}}=\boldsymbol{V}*\text{Attention}_{\text{REL}} \tag{14}$$

$$\boldsymbol{H}_{\text{S\_RA}}=\text{concat}(\boldsymbol{H}_{\text{S}},\boldsymbol{H}_{\text{RA}}) \tag{15}$$

其中，$\boldsymbol{H}_{\text{S}}$ 是混合嵌入向量，$\boldsymbol{H}_{\text{R}}$ 是关系嵌入向量，$\boldsymbol{H}_{\text{RA}}$ 是经过注意力机制后输出的关系嵌入向量，Q、K、V 由输入的特征向量分别与权重矩阵 $\boldsymbol{W}_{Q}$、$\boldsymbol{W}_{k}$、$\boldsymbol{W}_{V}$ 相乘后得到。$\boldsymbol{Q}\boldsymbol{K}^{\text{T}}$ 为查询矩阵与键矩阵的乘积，代表当前字符与其他字符间的关联度。$\sqrt{d^{k}}$ 为 $\boldsymbol{Q}$、$\boldsymbol{K}$ 维度键值的平方根，防止内积过大。利用 softmax 函数进行归一化处理后得到注意力权值。最后合并 $\boldsymbol{H}_{\text{S}}$ 和 $\boldsymbol{H}_{\text{RA}}$，得到向量 $\boldsymbol{H}_{\text{S\_RA}}$，作为最终输入头实体标记层中的向量。

2.2　头实体标注层

HEPA 模型选用的标注策略为先标注 BERT 编码序列中的所有头实体，再将头实体作为先验条件输入实体关系匹配层中，遍历所有的实体关系标签，找到一个最匹配的尾实体。选用分层标注的方法对头实体进行标注，设计两个完全一样的二进制标注器，分别对应实体的开始与结束位置，对于语句中的每个字符进行 0/1 标注，确定字符是否为头实体的开始或结束位置。这样做的好处是当语句中存在多个头实体时标注不会重叠，避免出现某个实体的标注结果既为头实体又是尾实体的情况。头实体标注的计算式如下：

$$p_{i}^{\text{start\_s}}=\sigma(\boldsymbol{W}_{\text{start}}\boldsymbol{h}_{i}+b_{\text{start}}) \tag{16}$$

$$p_{i}^{\text{end\_s}}=\sigma(\boldsymbol{W}_{\text{end}}\boldsymbol{h}_{i}+b_{\text{end}}) \tag{17}$$

其中，$p_{i}^{\text{start\_s}}$ 和 $p_{i}^{\text{end\_s}}$ 代表语句中第 i 个字符被标记成头实体开始位置或结束位置的概率，如果概率超过设定的阈值 0.5，则对应位置标记为 1，反之标记为 0。$\boldsymbol{W}_{\text{start}}$ 和 $\boldsymbol{W}_{\text{end}}$ 代表权重参数矩阵，b_{start} 和 b_{end} 代表偏置参数，σ 代表 Sigmoid 函数。

对句子中的主语进行抽取的概率函数如下：

$$p_{\theta}(s\mid\boldsymbol{h})=\prod_{t\in\{\text{start\_s},\text{end\_s}\}}\prod_{i=1}^{L}\left(p_{i}^{t}\right)^{I\{y_{i}^{t}=1\}}\left(1-p_{i}^{t}\right)^{I\{y_{i}^{t}=0\}} \tag{18}$$

其中，L 是句子的长度。如果 z 为真，则 $I\{z\}$=1，反之则 $I\{z\}$=0。y_{i}^{t} 表示主语开始和结束的位置，1 代表开始，0 代表结束。参数 $\theta\in\{\boldsymbol{W}_{\text{start}},b_{\text{start}},\boldsymbol{W}_{\text{end}},b_{\text{end}}\}$。

2.3　实体关系匹配层

向量序列经过头实体标注层处理后会产生多个头实体标记，如何为头实体匹配合适的尾实体成为提高模型处理效率亟需解决的问题。通常在一段完整的文本中，匹配的头实体与尾实体在距离上不会相距太远，因此本文在头实体标注层中加入实体位置注意力机制，将文本当前位置信息加入注意力机制中，筛选合适的实体关系进行匹配。实体位置注意力机制如下：

$$h_{i}'=\left[\boldsymbol{h}_{i};\left(\boldsymbol{h}_{i}^{\text{T}}\boldsymbol{h}_{\text{head}}\right)\boldsymbol{h}_{\text{head}}\right] \tag{19}$$

其中，$\boldsymbol{h}_{i}$ 代表当前字符的向量序列，$\boldsymbol{h}_{\text{head}}$ 代表头实体的向量序列，h_{i}' 代表当前字符与头实体的实体位置注意力权值。将得到的结果 $H=\left(h_{1}',h_{2}',\cdots,h_{n}'\right)$ 输入尾实体标注器中，尾实体标注方法与头实体标注方法相同，由两个二进制标注器组成。尾实体标注的计算式如下：

$$p_{i}^{\text{o\_start}}=\sigma(\boldsymbol{W}_{\text{start}}^{r}h_{i}'+b_{\text{start}}^{r}) \tag{20}$$

$$p_i^{\text{o\_end}} = \sigma(\boldsymbol{W}_{\text{end}}^r h_i' + b_{\text{end}}^r) \tag{21}$$

其中，$p_i^{\text{o\_start}}$ 和 $p_i^{\text{o\_end}}$ 代表语句中第 i 个字符被标记成尾实体开始位置或结束位置的概率，阈值与头实体标注相同为 0.5。$\boldsymbol{W}_{\text{start}}^r$ 和 $\boldsymbol{W}_{\text{end}}^r$ 代表权重参数矩阵，b_{start}^r 和 b_{end}^r 代表偏置参数，σ 代表 Sigmoid 函数。

在给定主语和特征向量情况下对句子中宾语进行抽取的概率计算式如下：

$$p_{\varphi_r}(o \mid s, \boldsymbol{h}) = \prod_{t\in\{\text{start\_s,end\_o}\}} \prod_{i=1}^{L} \left(p_i^t\right)^{I\{y_i^t=1\}} \left(1-p_i^t\right)^{I\{y_i^t=0\}} \tag{22}$$

其中，L 是句子的长度。如果 z 为真，则 $I\{z\}$=1，反之则 $I\{z\}$=0。y_i^t 表示主语开始和结束的位置，1 代表开始，0 代表结束。参数 $\varphi_r \in \{\boldsymbol{W}_{\text{start}}, b_{\text{start}}, \boldsymbol{W}_{\text{end}}, b_{\text{end}}\}$。

2.4 损失函数

HEPA 模型主要分为头实体标注与实体关系匹配两个部分，因此总损失函数由这两个部分的损失函数之和构成，选用二分类交叉熵损失函数。计算过程如计算式所示：

$$\text{Loss} = \sum_j \left\{ -\frac{1}{L} \sum_{i=1}^{L} t_i^j \text{lb} p_i^j + (1-t_i^j)\text{lb}(1-p_i^j) \right\} \tag{23}$$

其中，$j \in \{\text{start\_s,start\_o,end\_s,end\_o}\}$，$L$ 代表输入文本的长度，t_i^j 代表第 i 个字符为真实的头实体或尾实体的开始或结束位置。

3 实验分析

3.1 数据集与实验设置

为了验证 HEPA 模型的效果，本文选择在 DuIE 数据集上设计实验进行验证。DuIE 数据集是目前中文关系抽取领域中规模最大的数据集之一，来自2019年百度举办的语言与智能技术竞赛。DuIE 数据集包含 48 个已定义的 schema 约束，其中有43个简单知识schema，5个复杂知识schema，超过 21 万条中文语句和 45 万个三元组实例，并且包含大量的重叠三元组。DuIE 数据集的数据来自各领域，包括但不限于游戏、影视、教育，对模型的泛化性有较高要求。

在模型验证过程中，超参数设置如下：输入句子的最大长度设置为 256 字符；头、尾实体的标注阈值均设置为 0.5；batchsize 设置为 32 条；学习率设置为 1×10^{-5}；epoch 设置为 10 次；使用 Adam 优化器进行自适应学习；BERT 预训练模型使用 BERT-base 版本。

3.2 数据集与实验设置

为了验证 HEPA 模型在三元组抽取任务中的有效性，本文选用精确率（precision）、召回率（recall）和 $F1$ 值（$F1$-score）3 个主要指标来评价模型的效果，计算式如下：

$$\text{precision} = \frac{\text{正确识别的三元组个数}}{\text{被识别的三元组总数}} \times 100\% \tag{24}$$

$$\text{recall} = \frac{\text{正确识别的三元组个数}}{\text{样本中的三元组总数}} \times 100\% \tag{25}$$

$$F1\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{26}$$

3.3 对比实验

在 DuIE 数据集上设计实验，将 HEPA 模型在与其他基线模型进行对比，融合混合嵌入与关系标签的三元组抽取模型与基线模型对比见表 1。

表 1 融合混合嵌入与关系标签的三元组抽取模型与基线模型对比

模型	精确率	召回率	$F1$ 值
CopyMTL[23]	0.496	0.394	0.439
WDec[24]	0.641	0.542	0.587
CoType[25]	0.659	0.609	0.633
MHS[26]	0.772	0.623	0.690
CasRel[27]	0.786	0.689	0.734
HEPA	0.793	0.733	0.762

（1）CopyMTL 是在 CopyRE 的研究基础上提出的基于 copy+Seq2Seq 的三元组抽取模型，针对 CopyRE 无法区分文本中头、尾实体的问题进行了改进，通过多任务学习获取实体特征。

（2）WDec 是一个标准的 Seq2Seq 模型，具有动态掩蔽功能，能对实体标记（token）进行逐个解码，对实体关系重叠的问题有较大优化。

（3）CoType 是基于远程监督和弱监督的三元组联合抽取模型，充分利用数据集中句子级别的局部信息，降低了人工标注的要求，具有较好的泛用性。

（4）MHS 是一个联合抽取模型，使用 CRF 将实体识别任务和关系提取任务共同建模，将关系抽取任务转化为多头选择任务。该模型的优势是不需要依赖外部 NLP 工具进行标注。

（5）CasRel 是一个二进制级联抽取模型，它提出了一种将实体与关系建模为映射函数的三元组抽取方法。

分析表 1 的结果可知，HEPA 模型在精确率、召回率和 *F*1 值共 3 项评估指标中结果都优于最佳基线模型（CasRel），分别有 0.7%、4.4%、2.8%的提升，在召回率上有较大提升，说明在处理关系重叠三元组时有较好效果。HEPA 模型能取得优秀的效果依赖于混合嵌入带来的更多语义信息，模型能够充分利用上下文信息；加入标签嵌入机制能够增强实体之间的关联度。

为了验证标签嵌入机制对模型效果的帮助效果，设计了对三元组中不同元素抽取的对比实验，各模型提取不同元素的 *F*1 值对比见表 2。

表 2　各模型提取不同元素的 *F*1 值对比

模型	(*s*,*r*)	(*r*,*o*)	(*s*,*o*)	(*s*,*r*,*o*)
CopyMTL	0.482	0.489	0.502	0.439
WDec	0.623	0.633	0.652	0.587
CoType	0.673	0.669	0.675	0.633
MHS	0.725	0.737	0.742	0.690
CasRel	0.761	0.782	0.789	0.734
HEPA（无关系嵌入）	0.731	0.722	0.735	0.709
HEPA	0.792	0.803	0.809	0.762

分析表 2 结果可知，添加关系嵌入机制后能够加强头实体 *s*、关系 *r* 和尾实体 *o* 之间成对甚至三元组之间的联系。首先，在 4 组实验(*s*, *r*)、(*r*, *o*)、(*s*, *o*)和(*s*, *r*, *o*)中，HEPA 模型在 DuIE 数据集上的表现优于所有的对比模型。其次，当 HEPA 模型去除关系嵌入机制后，每组实验的效果都大幅降低，说明关系嵌入机制加强了实体与关系间的联系。最后，虽然关系嵌入机制同时编码大量关系标签存在引入噪声的问题，但从实验结果上看，该机制的加入对模型效果改进整体上还是利大于弊。

3.4　消融实验

为了进一步验证本文创新部分对模型效果的影响，在 DuIE 数据集上设计了消融实验进行对比，基于混合关系嵌入的三元组抽取模型消融实验结果见表 3。

表 3　基于混合关系嵌入的三元组抽取模型消融实验结果

模型	精确率	召回率	*F*1 值
去除词嵌入	0.742	0.682	0.711
去除字嵌入	0.763	0.692	0.725
去除实体位置注意力机制	0.775	0.659	0.704
去除关系嵌入机制	0.732	0.694	0.709
HEPA	0.793	0.733	0.762

HEPA 模型在去除字、词嵌入、实体位置注意力机制和关系嵌入机制后，在精确率、召回率和 *F*1 值评价指标上都有不同程度的下降，证明了本文添加的机制对改进模型效果有一定帮助。其中，只使用字嵌入或词嵌入时，模型精确率下降较多，说明字词混合嵌入对于模型准确抽取三元组帮助较大；去除实体位置注意力机制后召回率大幅下降，说明实体位置注意力机制能够有效匹配头实体与相应尾实体之间的关系，减少实体关系重叠情况对模型的影响；去除标签嵌入机制后，精确率下降 6.1%，*F*1 值下降 5.3%。

3.5　重叠三元组实验

为了验证 HEPA 模型在不同重叠三元组中的抽取效果，在 DuIE 数据集上分别对不同三元组重叠情况 normal、EPO 和 SEO 设计并进行了实验。不同重叠情况的模型表现如图 4 所示。

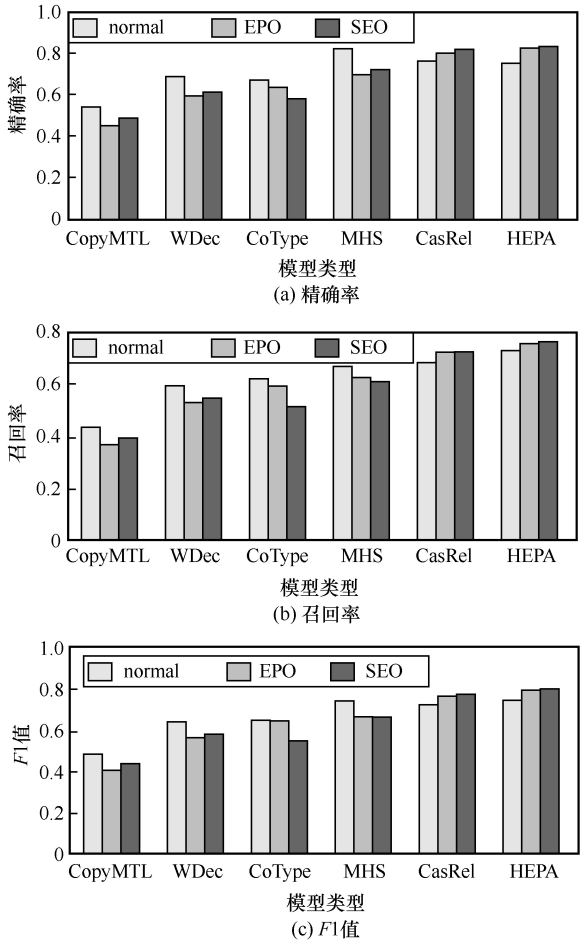

图 4　不同重叠情况的模型表现

图 4 显示了在 DuIE 数据集上各模型在不同重叠类型上的精确率、召回率和 *F*1 值。分析图 4 结果可知，HEPA 在 normal、EPO 和 SEO 3 种重叠情况下各项评价指标都取得了最好的效果。与 CasRel 对比，HEPA 在 EPO、SEO 情况下有着 2.9%和 3.6%的提升，但在 normal 情况下效果不如 CasRel。另外，大多数基线模型在不同的三元组重叠情况下的性能有不同程度的下降，原因是这些基线模型对实体关系的建模是离散的，无法较好地识别参与多个关系的实体。而 HEPA 对 EPO 和 SEO 的抽取效果呈现上升的趋势，原因是指针标注将关系抽取转化为实体与关系之间一对一的映射，无论文本有多复杂，都能为头实体匹配最相近的实体关系与尾实体。此外，注意力机制能够从不同的维度提取句子中的关键信息，帮助模型理解复杂文本。与基线模型对比 HEPA 更加适合复杂文本下的三元组抽取，稳定性更佳。

4　结束语

本文设计了一种融合混合嵌入与关系嵌入的三元组联合抽取方法 HEPA，能够降低嵌入过程中由分词错误引起的语义信息缺失问题，在复杂的文本环境中取得较好的效果，同时对抽取重叠三元组的效果有显著提升。该模型通过字嵌入结合词嵌入的混合嵌入方法融入更多的语义信息，减少由于分词错误造成的误差，在将标签信息加入文本输入中，提高了关系匹配精度，在实体匹配层中添加了注意力机制，多维度地捕获文本语义特征，在实体关系匹配过程中加入实体位置注意力机制，为头实体匹配最合适的尾实体。将 HEPA 与其他模型进行对比实验后，发现 HEPA 能够较好地解决重叠三元组问题，相比于其他基线模型在性能上有较大提升。

参考文献：

[1] 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.
LI D M, ZHANG Y, LI D Y, et al. Overview of entity relationship extraction methods[J]. Computer Research and Development, 2020, 57(7): 1424-1448.

[2] ZENG D J, LIU K, LAI S W, et al. Relation classification via convolutional deep neural network[C]//Proceedings of International Conference on Computational Linguistics. [S.l.:s.n.], 2014.

[3] XU K, FENG Y, HUANG S, et al. Semantic relation classification via convolutional neural networks with simple negative sampling[J]. Computer Science, 2015, 71(7): 941-9.

[4] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]//Proceedings of Joint Conference on Empirical Methods in Natural Language Processing & Computational Natural Language Learning. Hongkong: EMNLP Press, 2012.

[5] LI F, ZHANG M, FU G, et al. A Bi-LSTM-RNN model for relation classification using low-cost sequence features:

10.48550/arXiv.1608.07720[P]. 2016.

[6] SU Z, JIANG J. Hierarchical gated recu0rrent unit with semantic attention for event prediction[J]. Future Internet, 2020, 12(2): 39.

[7] VASHISHTH S, JOSHI R, PRAYAGA S S, et al. RESIDE: improving distantly-supervised neural relation extraction using side information[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. [S.l.:s.n.], 2018.

[8] 杨帅, 王瑞琴, 马辉. 基于多通道的边学习图卷积网络[J]. 电信科学, 2022, 38(9): 95-104.
YANG S, WANG R Q, MA H. Multi-channel based edge-learning graph convolutional network[J]. Telecommunications Science, 2022, 38(9): 95-104.

[9] 李昊, 陈艳平, 唐瑞雪, 等. 基于实体边界组合的关系抽取方法[J]. 计算机应用, 2022, 42(6): 6.
LI H, CHEN Y P, TANG R X, et al. Relationship extraction method based on entity boundary combination [J]. Computer Applications, 2022, 42 (6): 6.

[10] ZHONG Z, CHEN D. A frustratingly easy approach for entity and relation extraction[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics. [S.l.:s.n.], 2021.

[11] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures[J]. arXiv preprint, arXiv: 1601.00770, 2016.

[12] KATIYAR A, CARDIE C. Going out on a limb: joint extraction of entity mentions and relations without dependency trees[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.:s.n.], 2017: 917-928.

[13] ZHENG S, F WANG, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[J]. arXiv preprint, arXiv:1706.05075, 2017.

[14] ZENG X, ZENG D, HE S, 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 (Volume 1: Long Papers). [S.l.:s.n.], 2018: 506-514.

[15] FU T J, MA W Y. GraphRel: modeling text as relational graphs for joint entity and relation extraction[C]//Meeting of the Association for Computational Linguistics. [S.l.:s.n.], 2019: 1409-1418.

[16] DUAN G, MIAO J, HUANG T, et al. A relational adaptive neural model for joint entity and relation extraction[J]. Frontiers in Neurorobotics, 2021(15): 635492.

[17] 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. [S.l.:s.n.], 2020: 1476-1488.

[18] WANG Y, YU B, ZHANG Y, et al. TPLinker: single-stage joint extraction of entities and relations through token pair linking[J]. arXiv preprint, arXiv:2010.13415, 2020.

[19] 田佳来, 吕学强, 游新冬, 等. 基于分层序列标注的实体关系联合抽取方法[J]. 北京大学学报：自然科学版, 2021, 57(1): 53-60.
TIAN J L, LYU X Q, YOU X D, et al. A joint extraction method of entity relations based on hierarchical sequence annotation[J]. Journal of Peking University (Natural Science Edition), 2021, 57(1): 53-60

[20] 苗琳, 张英俊, 谢斌红, 等. 基于图神经网络的联合实体关系抽取[J]. 计算机应用研究, 2022, 39(2): 424-431.
MIAO L, ZHANG Y J, XIE B H, et al. Joint entity relationship extraction based on graph neural network[J]. Proceedings of the Computer Application Research, 2022, 39 (2): 424-431

[21] 王红, 吴燕婷. 基于多跳注意力的实体关系联合抽取方法及应用研究[J]. 太原理工大学学报, 2022, 53(1): 63-70.
WANG H, WU Y T. Joint extraction of entity relationships based on multi-hop attention and its application [J]. Proceedings of the Journal of Taiyuan University of Technology, 2022, 53(1): 63-70.

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

[23] 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.

[24] NAYAK T, NG H T. Effective modeling of encoder-decoder architecture for joint entity and relation extraction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8528-8535.

[25] REN X, WU Z, HE W, et al. CoType: joint extraction of typed entities and relations with knowledge bases[J]. Proceedings of the 26th International Conference on World Wide Web. New York: ACM Press, 2017: 1015-1024.

[26] GIANNIS B, JOHANNES D, THOMAS D, et al. Joint entity recognition and relation extraction as a multi-head selection problem[J]. Expert Systems with Application, 2018, 114(11): 34-45.

[27] WEI Z, SU J, WANG Y, et al. A novel hierarchical binary tagging framework for relational triple extraction[J]. arXiv preprint, arXiv:1909.03227v4, 2020.

[作者简介]

戴剑锋（1997– ），男，浙江工商大学信息与电子工程学院（萨塞克斯人工智能学院）硕士生，主要研究方向为智慧教育、自然语言处理。

董黎刚（1973– ），男，博士，浙江工商大学信息与电子工程学院（萨塞克斯人工智能学院）党委书记、教授、博士生导师，浙江省计算机学会理事，主要研究方向为新一代网络和分布式系统。

陈星妤（1999– ），女，浙江工商大学信息与电子工程学院（萨塞克斯人工智能学院）硕士生，主要研究方向为智慧教育、自然语言处理。

蒋献（1988– ），男，浙江工商大学信息与电子工程学院（萨塞克斯人工智能学院）讲师、实验员，主要研究方向为智慧教育和智慧网络。