基于主体掩码的实体关系抽取方法

郑慎鹏[1]，陈晓军[1]，向阳[1]，沈汝超[2]
1. 同济大学电子与信息工程学院，上海 201804；
2. 上海国际港务（集团）股份有限公司，上海 200080

摘要
实体关系抽取技术能够自动化地从海量无结构文本中抽取信息，构建大规模知识图谱，丰富现有知识图谱的内容，为知识图谱推理和应用提供支持。目前级联式的实体关系抽取技术已经取得了不错的成绩，但其在主体的向量表示和指针网络解码上存在不足。针对主体向量表示问题，提出利用注意力机制和掩码机制生成主体向量的方法。另外，针对指针网络中因遗漏标注而解码出过长实体的问题，提出引入实体序列标记任务，辅助指针网络解码实体。在大规模实体关系数据集DuIE2.0上进行实验验证得出，相较于先前模型，所提方法取得了0.88%的提升。

关键词
RoBERTa；实体关系抽取；主体掩码

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2021022

An entity relation extraction method based on subject mask

ZHENG Shenpeng[1], CHEN Xiaojun[1], XIANG Yang[1], SHEN Ruchao[2]
1. College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China
2. Shanghai International Port (Group) Co., Ltd., Shanghai 200080, China

Abstract
Entity relationship extraction technology can automatically extract information from massive unstructured texts to construct large-scale knowledge graph, enrich the content of existing knowledge graph, and provide support for reasoning and application of knowledge graph. Although the cascading entity relation extraction technology has achieved good results, it has some shortcomings in the vector representation of the subject and the decoding of pointer network. In order to solve the representation problem of subject vectors, attention mechanism and mask mechanism were used to generate subject vectors. In addition, to solve the problem that long entities have been decoded in pointer network due to missing label, an entity sequence marker task was introduced to assist pointer network decoding entities. There is a 0.88% improvement over the previous model on the large-scale entity relationship dataset DuIE 2.0.

Key words
RoBERTa, entity relation extraction, subject mask

1 引言

目前，网络上保存着海量的非结构化文本，且其规模仍呈指数级上升。同时，知识图谱被广泛应用在政府治理[1]、智能问答[2]、搜索引擎等领域，而知识图谱的内容丰富程度和及时性直接影响其应用效果。因此，作为自动化地从非结构化文本中构建知识图谱的关键技术之一，实体关系抽取技术受到了研究人员的广泛关注。实体关系抽取旨在识别出文本中实体和实体之间的语义关系，并以三元组的形式<主体，关系，客体>表示。比如，“《琴键右角》是张德兰演唱的一首单曲”中包含实体“琴键右角”和“张德兰”，且实体间存在关系“歌手”，用三元组表示为<琴键右角，歌手，张德兰>。

早期的实体关系抽取方法[3-5]和基于传统机器学习的实体关系抽取方法[6-7]需要专家构造大量的规则或者人工特征，难以应对大规模的实体关系提取。随着深度学习的兴起，神经网络模型可以自动提取文本的特征，减少人工提取特征的工作，也能更有力地应对大规模的实体关系提取工作，成为当前实体关系提取的主流方法。目前，基于神经网络的实体关系抽取方法可以分为流水线方法和联合抽取方法两类。

流水线方法将实体关系抽取分解为实体识别和关系分类两个步骤，并用两个独立的模型实现[8-10]。此类方法先用实体识别模型识别出文本中的所有实体，然后用关系分类模型判断所有可能实体对的语义关系。流水线方法能够灵活地选择实体识别模型和关系分类模型，但是其缺点也是显而易见的。首先，流水线方法存在错误传播的问题，实体识别阶段和关系分类阶段的错误会叠加，导致最终的性能下降。再者，实体识别模型和关系分类模型是完全独立的，忽略了实体识别任务和关系抽取任务的内在联系。

联合抽取方法旨在利用一个模型实现实体识别和关系抽取，有效避免流水线方法中存在的两点弊端。联合抽取方法依据解码方式一般可分为独立解码、级联解码和一次解码3类。在独立解码的方法中，实体识别和关系抽取共享文本编码层，在解码时仍然是两个独立的部分。为了使两个任务间建立更加密切的联系，级联解码的方法通常会先抽取主体，再根据主体抽取相关的关系-客体。而一次解码方法则将实体识别和关系抽取统一为一个任务，一次抽取出实体对及其对应关系。目前级联解码的方法和一次解码的方法在实体关系抽取中都取得了不错的成绩。在后两类方法中，实体嵌套问题[11-12]和关系重叠问题相互交织，使情况变得比较复杂，见**表1**。Wei Z P等人[13]提出了一种新颖的级联式标记框架，很好地解决了联合抽取中实体嵌套和关系重叠同时存在的问题。该方法将实体关系抽取看作抽取主体和根据主体抽取关系-客体两个部分，并且采用指针网络的结构标注主体与客体。但是，此方法在表示主体向量时，只是简单地将主体所含的所有字向量做平均，这会导致一些显著特征在平均后会丢失，尤其是在中文中。此外，使用指针网络标注时，模型漏标会导致出现过长且有明显错误的实体。针对该方法中存在的两点问题，本文提出了以下改进：

- 针对主体向量的表示问题，提出基于主体掩码的主体向量生成方法，利用注意力机制和掩码机制，生成主体向量；
- 针对多层指针网络的漏标问题，提出实体序列标注子任务，在解码实体时提供辅助信息。

表1 实体关系抽取复杂情况分析

难点	类型	样例	样例分析
实体嵌套	共享前缀	他的自传《李敖自传》	“李敖”与“李敖自传”
	共享后缀	抗日剧《我的兄弟叫顺溜》	“我的兄弟叫顺溜”与“顺溜”
	包含	《百变梅艳芳之烈焰红唇》	“梅艳芳”与“百变梅艳芳之烈焰红唇”
关系重叠	实体在多个关系中	陈奕迅与妻子徐濠萦共同演唱了《Style》	<Style,歌手,陈奕迅> <陈奕迅,妻子,徐濠萦>
	实体间有多个关系	《无路用的人》是张震岳作词作曲的	<无路用的人,作词人,张震岳> <无路用的人,作曲人,张震岳>

2 相关工作

在知识图谱的构建过程中，实体关系抽取技术起着非常重要的作用。早期基于规则[3]、词典[4]或本体[5]的实体关系抽取方法存在跨领域的可移植性较差、人工标注成本较高以及召回率较低等问题。后来，相比于早期的方法，以统计语言模型为基础的传统机器学习关系抽取方法明显地提高了召回率，具有更强的领域适应性，获得了不错的效果。自从Hinton G E等人[14]首次正式地提出深度学习的概念，深度学习在多个领域取得了突破性进展，也渐渐被研究人员应用在实体关系抽取方面。此外，Transformer结构[15]、BERT（bidirectional encoder representations from transformers）[16]、RoBERTa（robustly optimized BERT pretraining approach）[17]等大规模预训练语言模型也极大地推动了实体关系抽取的进步。现阶段，实体关系抽取主要分为流水线方法和联合抽取方法。

2.1 流水线方法

流水线方法是指在用实体识别模型抽取出所有实体的基础上，利用关系分类模型抽取所有实体对的关系。较早期的基于神经网络的流水线方法主要使用卷积神经网络（convolutional neural network，CNN）或循环神经网络（recurrent neural network，RNN）。Zeng D J等人[18]首次采用CNN提取词级和句子级特征，通过softmax层进行关系分类，提高了关系抽取模型的准确性。

Socher R等人[19]首先使用RNN的方法进行实体关系抽取。该方法利用RNN对标注文本中的句子进行句法解析，经过不断迭代得到句子的向量表示，有效地考虑了句子的句法结构。为了解决RNN的梯度消失和梯度爆炸问题，长短期记忆（long short-term memory，LSTM）神经网络被提出。Yan X等人[8]提出基于LSTM和句法依存分析树的最短路径的方法进行关系抽取。随着图神经网络在近几年取得一些进展，图神经网络也被应用在关系抽取中，Guo Z J等人[9]提出了一种将全依赖树作为输入的注意力引导图卷积网络模型。该模型充分利用了依赖树中的信息，以便更好地提取出相关关系。虽然流水线方法能够取得不错的效果，但也存在以下3个缺点。

- 错误累积：实体识别模块的错误会影响接下来的关系分类性能。
- 忽视了两个子任务之间存在的联系：丢失信息，影响抽取效果。
- 产生冗余实体对：由于要对所有抽取出的实体两两配对，然后再进行关系分

类，那些不存在关系的实体对就会带来多余信息，提升错误率。

2.2 联合抽取方法

联合抽取方法[20]能够在一个模型中实现实体关系抽取，此类方法能够利用实体和关系间的联系，减少或者避免流水线方法带来的问题。

早期联合抽取方法通常是基于人工构造特征的结构化学习方法。Miwa M等人[21]首次将神经网络的方法用于联合抽取实体和关系，该方法将实体关系抽取分解为实体识别子任务和关系分类子任务。在模型中使用双向序列LSTM对文本进行编码，将实体识别子任务当作序列标注任务，输出具有依赖关系的实体标签。同时，在关系分类子任务中捕获词性标签等依赖特征和实体识别子任务中输出的实体序列，形成依存树，并根据依存树中目标实体间的最短路径对文本进行关系抽取。在该模型中，关系分类子任务和实体识别子任务的解码过程仍然是独立的，它们仅仅共享了编码层的双向序列LSTM表示，并不能完全地避免流水线方法的问题。Zheng S C等人[22]认为之前的联合抽取方法虽然将两个任务整合到一个模型中并共享了一部分参数，但是实体识别与关系抽取任务仍是两个相对独立的过程。于是Zheng S C等人[22]提出了一种基于新的标注策略的实体关系抽取方法，把原来涉及实体识别和关系分类两个子任务的联合学习模型完全变成了一个序列标注问题。在该方法中，实体的位置标签和关系标签被统一为一个标签，通过一个端到端的神经网络模型一次解码就可得到实体以及实体间的关系，解决了独立解码的实体关系联合抽取方法的交互不充分和实体冗余问题。但是，该方法没有能力应对普遍存在的实体嵌套和关系重叠的情况，这使得该方法在实际应用中难以取得好的效果。

为了应对实体嵌套和关系重叠的问题，Li X Y等人[23]提出将实体关系联合抽取的任务当作一个多轮问答类问题来处理，该方法需要构造不同的问题模板，通过一问一答的形式依次提取出主体、关系、客体。这种多轮问答的方法能够很好地解决实体嵌套和关系重叠的问题，但是其需要为每一种主体类型、每一种关系都设计一个问答模板，并进行多次问答，这会产生很多计算冗余，非常消耗计算资源。Wei Z P等人[13]提出了一种级联式解码实体关系抽取方法，并用多层二元指针网络标记实体。不同于独立解码模型，该方法将任务分解为主体识别子任务和依据主体抽取关系-客体子任务，而且将两个子任务都统一为序列标注问题。该方法解决了实体嵌套和关系重叠的问题，同时没有引入太多的冗余计算。但是，此方法简单地将主体所含的所有字向量取平均后作为主体向量，导致一些显著特征在平均后会丢失。此外，因为指针网络仅标记实体的首尾位置，当出现漏标时，会导致模型解码出比较长的错误实体。为此，本文提出基于主体掩码的主体向量生成方法，并利用实体序列标注辅助指针网络解码实体。

3 基于主体掩码的实体关系抽取模型

实体关系抽取旨在抽取出文本中所有的<主体,关系,客体>三元组，而这些三元组间可能会存在实体嵌套和关系重叠的情况，为了应对这类情况，Wei Z P等人[13]提出了一种基于新型的级联式二元标注结构的实体关系抽取模型。不同于以往的模型，该模型将任务分解为主体识别阶段和关系-客体识别阶段。在该模型的基础上，

本文提出基于主体掩码的主体向量生成方法，利用注意力机制和掩码机制，生成主体向量。此外，为了排除因模型漏标产生的长度过长的实体，增加实体序列标注任务，以辅助实体解码。

模型结构如**图1**所示。首先，该模型将文本输入编码层，得到文本向量序列$\boldsymbol{h}_N$。然后，将文本向量序列$\boldsymbol{h}_N$输入主体指针网络，得到主体k。接着，根据主体得到主体掩码序列$\boldsymbol{m}^k_{\text{sub}}$，将文本向量序列$\boldsymbol{h}_N$和主体掩码序列$\boldsymbol{m}^k_{\text{sub}}$输入Transformer层，在该层中计算出与每个字相关的主体向量序列$\boldsymbol{v}^k_{\text{sub}}$，与文本向量序列$\boldsymbol{h}_N$相加，得到关系-客体向量序列$\boldsymbol{h}^k_{\text{obj}}$。最后，通过客体识别网络预测得到主体$k$的所有相关客体，而主体和客体之间的关系由客体所在指针网络的层数决定。此外，模型还会取出编码层的中间某一层作为实体序列标注的输入，该子任务的目标是标记出所有属于主体或客体的字。**图1**中主体掩码序列和关系-客体标记结构中使用“1”和“2”区分不同主体和与不同主体相关的客体。

3.1 编码层

本文使用RoBERTa作为编码层，将文本编码为文本向量序列$\boldsymbol{h}_N$。RoBERTa是基于BERT提出的一种效果更好的预训练模型。RoBERTa由多层Transformer堆叠而成，本文将取RoBERTa的最后一层输出作为文本向量序列$\boldsymbol{h}_N$，中间某一层输出作为实体序列标注的输入。

3.2 实体识别

实体标记有多种不同的方案，从简单地使用0/1标记到使用OBIE（O表示非实体，B表示开始，I表示内部，E表示实体尾部）标记。但是这些方案不能解决实体嵌套问题，这在相当程度上降低了实体关

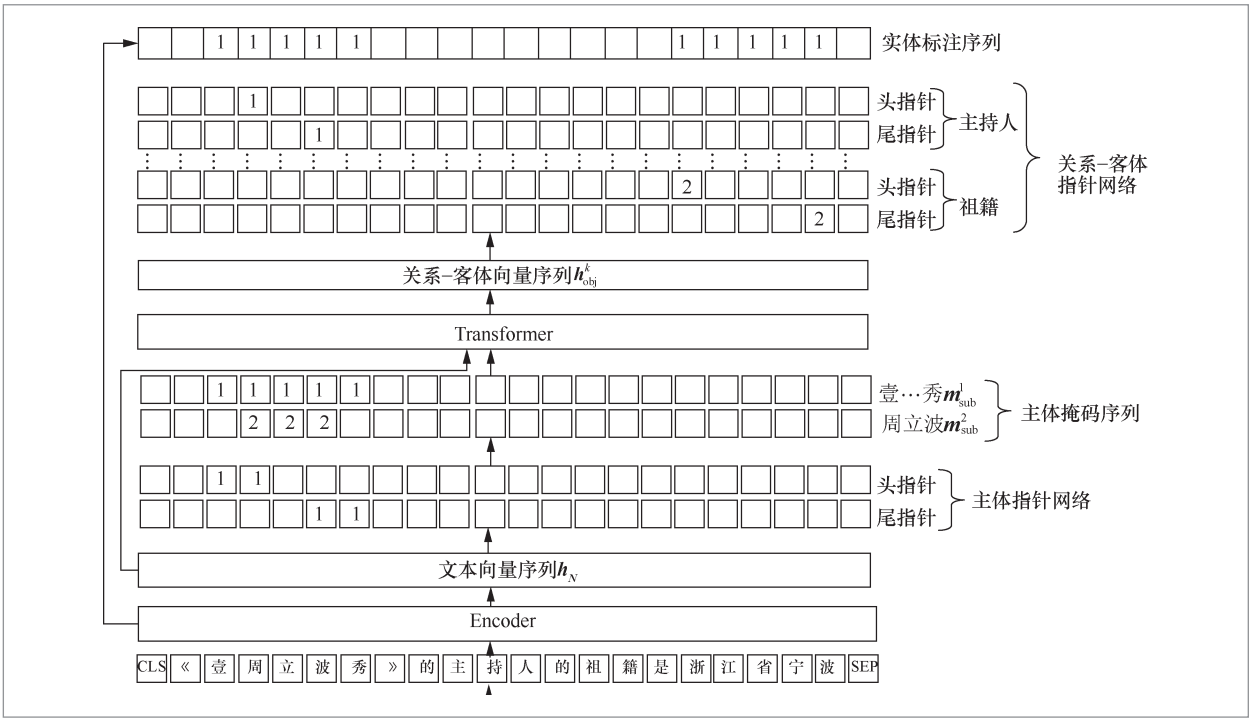

图1 基于主体掩码的实体关系抽取模型

系抽取的准确率。为了解决实体嵌套的问题，本文采用Wei Z P等人[13]使用的指针网络标记方案。指针网络在解码时一般采用就近原则，头指针与后文的第一个尾指针配对，尾指针与前文中最近的头指针配对。若出现头指针、头指针、尾指针、尾指针的序列模式，则认定为包含模式，将处于最前和最后的头尾指针配对，中间的头尾指针配对，如**图1**中例子所示。

在主体识别的指针网络中仅使用一对头尾指针进行标注，计算式如下：

$$p_i^{\text{sub}_{\text{start}}} = \sigma\left(W_{\text{start}}\boldsymbol{x}_i + b_{\text{start}}\right) \quad (1)$$

$$p_i^{\text{sub}_{\text{end}}} = \sigma\left(W_{\text{end}}\boldsymbol{x}_i + b_{\text{end}}\right) \quad (2)$$

其中，$p_i^{\text{sub}_{\text{start}}}$和$p_i^{\text{sub}_{\text{end}}}$分别表示输入文本的第$i$个字为主体开始位置和结束位置的概率，若概率大于阈值，对应位置将会被标记为1，否则标记为0。$\boldsymbol{x}_i = \boldsymbol{h}_N[i]$为文本向量序列对应位置的向量，$W_{\text{start}}$、$b_{\text{start}}$、$W_{\text{end}}$和$b_{\text{end}}$是可训练参数，$\sigma$为sigmoid函数。

关系-客体识别的指针网络由N对头尾指针标注序列构成，计算式如下：

$$p_i^{r\text{-}\text{obj}_{\text{start}}} = \sigma\left(W_{\text{start}}^r\boldsymbol{z}_i^k + b_{\text{start}}^r\right) \quad (3)$$

$$p_i^{r\text{-}\text{obj}_{\text{end}}} = \sigma\left(W_{\text{end}}^r\boldsymbol{z}_i^k + b_{\text{end}}^r\right) \quad (4)$$

其中，$p_i^{r\text{-}\text{obj}_{\text{start}}}$和$p_i^{r\text{-}\text{obj}_{\text{end}}}$分别表示第$i$个字与主体存在关系$r$的开始位置和结束位置的概率。$\boldsymbol{z}_i^k = \boldsymbol{h}_{\text{obj}}^k[i]$为实体$k$的关系-客体向量序列$\boldsymbol{h}_{\text{obj}}^k$对应位置的向量，$W_{\text{start}}^r$、$b_{\text{start}}^r$、$W_{\text{end}}^r$和$b_{\text{end}}^r$是与关系类型相关的可训练参数。

3.3 Transformer结构

Transformer结构由Vaswani A等人[15]在2018年提出。Transformer将矩阵$\boldsymbol{H} \in \mathbb{R}^{l\times d}$作为输入，其中$l$为文本长度，$d$为输入维度。接着，$\boldsymbol{H}$分别与$\boldsymbol{W}_q$、$\boldsymbol{W}_k$和$\boldsymbol{W}_v$ 3个矩阵相乘，得到查询矩阵$\boldsymbol{Q}$、关键字矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$。通常，$\boldsymbol{W}_q$等3个矩阵的大小都为$\mathbb{R}^{d\times d_k}$，d_k为一个超参数。Transformer中注意力机制的计算如下所示：

$$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} = \boldsymbol{H}\boldsymbol{W}_q,\boldsymbol{H}\boldsymbol{W}_k,\boldsymbol{H}\boldsymbol{W}_v \quad (5)$$

$$A_{i,j} = \boldsymbol{Q}_i\boldsymbol{K}_j^{\top} - N_{\text{MAX}}(1-M_{i,j}) \quad (6)$$

$$\text{Attn}(\boldsymbol{K},\boldsymbol{Q},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{A}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (7)$$

其中，$\boldsymbol{Q}_i$为第i个字的查询向量，$\boldsymbol{K}_j$为第j个字的关键字向量。$A_{i,j}$为注意力分数矩阵$\boldsymbol{A}$的第i行第j列的元素，表示第i个字在第j个字上的注意力分数。N_{MAX}为一个极大的正数，$M_{i,j}$为掩码矩阵$\boldsymbol{M}$的第i行第j列的元素，表示第i个字对第j个字的查询掩码，使注意力集中在指定的关键字上。softmax函数在最后一个维度上计算出归一化后的注意力分数。在Transformer中，为了提升模型性能，使用多组参数实现注意力机制，再将通过不同组参数得到的向量进行拼接，然后通过$\boldsymbol{W}_o$做线性变换得到多头注意力机制的输出。这类使用多组参数的注意力机制被称为多头注意力机制，如下所示：

$$\boldsymbol{Q}^{(h)},\boldsymbol{K}^{(h)},\boldsymbol{V}^{(h)} = \boldsymbol{H}\boldsymbol{W}_q^{(h)},\boldsymbol{H}\boldsymbol{W}_k^{(h)},\boldsymbol{H}\boldsymbol{W}_v^{(h)} \quad (8)$$

$$\mathbf{head}^{(h)} = \text{Attn}\left(\boldsymbol{Q}^{(h)},\boldsymbol{K}^{(h)},\boldsymbol{V}^{(h)}\right) \quad (9)$$

$$\text{MultiHead}(\boldsymbol{H}) = \left[\mathbf{head}^{(1)},\cdots,\mathbf{head}^{(n)}\right]\boldsymbol{W}_o \quad (10)$$

其中，h表示第几个“头”，n表示“头”的数量，且会使$d_k \times n = d$，$\boldsymbol{W}_o$的大小为$\mathbb{R}^{d\times d}$。

为了应对深度学习中的网络退化问题，Transformer中采用残差连接的方式，并加入归一化层，如下所示：

$$\boldsymbol{Z} = \text{Norm}\left(\text{MultiHead}(\boldsymbol{H}) + \boldsymbol{H}\right) \quad (11)$$

在得到归一化层的输出$\boldsymbol{Z}$后，将$\boldsymbol{Z}$通过一层全连接层和归一化层得到Transformer最后的输出。

3.4 主体表示

在识别出输入文本中的所有主体后，根据每个主体的位置构造出各自的主体掩码$\boldsymbol{m}_{\text{sub}}$。具体来说，构造一个与输入文本等

长的掩码序列，将主体开始到结束对应的位置全部标记为1，其余位置标记为0，在一个句子中，一个实体可能会出现多次，则每个对应位置都需要标记为1。然后，将掩码序列m_{sub}和文本向量序列h_N输入注意力层，m_{sub}使注意力层在计算文本各位置的注意力分数时，只关注掩码序列中标为1的位置，如**图2**所示。在计算“周立波”的主体向量时，每个字仅和“周立波”交互计算注意力分数，由此得到“周立波”在各个字上的主体向量序列。将主体向量序列与文本向量序列h_N相加，就可得到与各个主体相关的关系-客体向量序列h^k_{obj}。Transformer中的注意力机制和残差连接恰好可以实现生成主体向量序列并与文本向量序列h_N相加的操作，因此本文将主体掩码m_{sub}和文本向量序列h_N输入Transformer就可得到关系-客体向量序列h^k_{obj}。

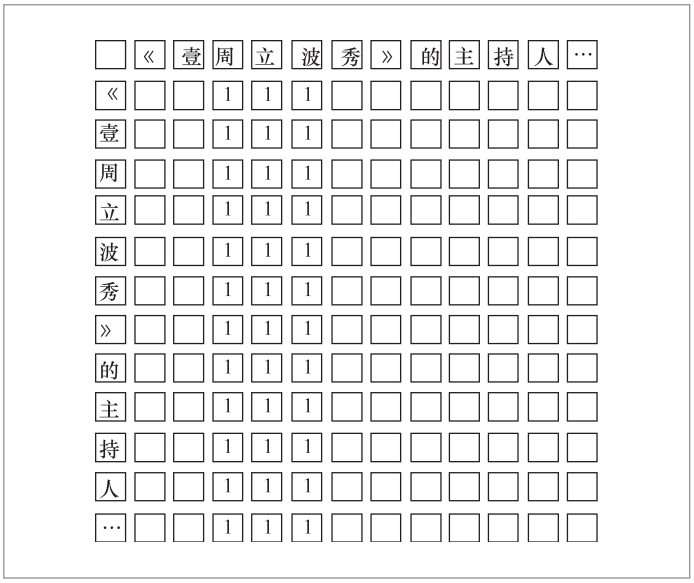

图 2 注意力计算可视矩阵

3.5 实体序列标注

实体序列标注的任务是标记出文本中的所有主体和客体，在这个任务中仅使用一行标注序列，将所有属于主体和客体的字都标记为1，如**图1**所示。在主体识别阶段或者关系-客体识别阶段，对于一些过长的实体，需要进一步判断该实体包含范围内的实体序列标注结果是否全部为1。若有一个位置为0，则判断该实体为错误实体，并忽略该实体。该任务要求标记出实体的所有位置，而不仅是实体的首尾位置，因此为了避免对主任务产生不利的影响，本文选取RoBERTa的中间层的输出，通过简单的全连接层和Sigmoid得到每个位置的实体概率。

3.6 损失函数设计

本文模型为多任务学习模型，包含主体识别、关系-客体识别和实体序列标注3个子任务。每个任务的基础都是做二分类任务，因此将二元交叉熵作为各个任务的损失函数：

$$\text{loss} = \alpha \text{loss}_{sub} + \beta \text{loss}_{obj} + \gamma \text{loss}_{entity} \quad (12)$$

其中，loss_{sub}是主体识别损失，loss_{sbj}是关系-客体识别损失，loss_{entity}是实体序列标注损失，α、β、γ分别为各项损失的权重。

4 实验与分析

4.1 数据集与评测指标

本文使用的数据是由百度提供的DuIE2.0数据集[24]。该数据集是业界规模最大的基于模式的中文信息抽取数据集，其包含超过21万个中文句子及48个已定义好的模式，数据集中包含43个简单知识的模式和5个复杂知识的模式。数据集中的句子来自百度百科、百度贴吧和百度信息流文本。17万个句子被划分为训练集，2万

个句子被划分为验证集，2万个句子被划分为测试集，将micro-F1值作为评测指标。当提取的三元组的主体、关系、客体与标注一致时，视为正确。

4.2 预处理

DuIE2.0数据集中包含43种简单模式和5种复杂模式，复杂模式会有多个客体值。例如，<主体(影视作品),关系1(票房),客体1(数值),关系2(地点),客体2(地点)>，这个模式要求抽取出一个影视作品在某地的票房。本文将所有的复杂模式分解为简单模式，仍以上面这个例子为例，其会被分解为<影视作品,票房,数值>、<影视作品,上映地点,地点>、<数值,票房-地点,地点>，对于其他的复杂模式，也进行类似的处理。另外，删除了简单模式中<人物,丈夫,人物>的模式，因为模式<人物,妻子,人物>也能表达出这个概念。最后得到54种简单模式。

4.3 具体实现

模型编码器使用RoBERTa-base版本。文本最大长度限制为205个字。批大小设置为16，训练12轮。RoBERTa部分的学习率设置为3×10^{-5}，除RoBERTa外的网络学习率设置为9×10^{-5}。同时使用学习率线性衰减的策略，学习率计算式如下：

$$\mathrm{lr}_i = \mathrm{lr}_{\mathrm{init}} \times \left(1-(i-1)/e_a\right) \qquad (13)$$

其中，i表示训练轮次，e_a为总训练轮次，lr_i表示第i个训练轮次的学习率，$\mathrm{lr}_{\mathrm{init}}$表示初始学习率。

在训练的初始阶段，预训练模型以外的部分是随机初始化的，会产生很大的梯度。为了防止预训练模型的参数分布受到无意义的破坏，将除RoBERTa预训练模型外的网络先训练500步，再一起训练全部的网络。此时，模型损失相比初始时已经下降了一个数量级。损失函数中α、β、γ分别设置为1、1、0.05。在实体序列标注任务中，取RoBERTa的第9层输出作为输入。

在训练时，若一个句子中有多个不同主体的三元组需要抽取，Wei Z P等人[13]会随机抽取一个主体进行训练。但是，同一个句子内的三元组是有联系的，这种操作会丢失这部分信息，且需要训练更多的轮数。因此，本文将有多个主体的样本复制多份，并各自计算主体向量，实现在一个批中训练全部的三元组。

4.4 实验结果

实验结果见**表2**。Official为百度官方的基线模型，该模型一次解码出主体与客体，并根据就近原则配对；Pipline是一种常规的流水线模型，用实体识别模型抽取出所有实体，在关系判断模型中判断所有实体对间是否存在关系；Mean表示取主体的所有向量做平均，并将其作为主体的向量表示；Max表示取主体所有向量做Max操作，得到主体的向量表示；Transf表示使用主体掩码生成主体；Transf+seq表示在Transf的基础上增加实体序列标注任务辅助实体的识别。从**表2**可以看出，基于主体掩码的主体向量表示方法相比于简单取平均值和取最大值的方法，F1值都有一定的提高。另外，在Transf方法的基础上引入实体序列标注子任务使模型在测试集上的F1值进一步提高了0.003 3。最终，模型在测试集上F1值为0.750 8。

4.5 样例分析

在例子“安徽四创电子股份有限公司，

表 2 不同模型在 DuIE2.0 数据集的 F1 值

模型	Dev-精确率	Dev-召回率	Dev-F1	Test-F1
Official	–	–	–	0.710 6
Pipline	0.715 2	0.748 8	0.731 7	0.720 3
Mean	0.706 0	0.788 6	0.745 0	0.742 0
Max	0.702 8	0.795 0	0.746 0	0.742 7
Transf	0.719 2	0.786 3	0.751 3	0.747 5
Transf + seq	0.742 2	0.766 4	0.754 1	0.750 8

现注册资本1.59亿元，总资产36.62亿元”中有实体“安徽四创电子股份有限公司”和“1.59亿元”，存在关系“注册资金”。使用Mean方法生成主体向量可以理解为将注意力分散到了每个字上。本文所述方法显示了其较为不同的过程。在本例子中，“1.59亿元”会将注意力更加集中在“股份有限”上，而其他字则会集中在“四”“公司”或是分散在各个字上。该例子表明，不同的字会注意到实体的不同位置，尤其是在实体中包含一些有具体含义的词语时，见**表3**。

5 结束语

利用实体关系抽取技术自动化地抽取无结构化文本中的三元组信息，能够及时更新知识图谱的内容，为知识图谱推理和应用提供支持。本文提出利用注意力机制和掩码机制生成主体表示，同时提出利用实体序列标注结合指针网络进行实体抽取。最终的实验结果表明，合适的主体向量生成方法能够在不改变模型整体结构的前提下，有效地提升模型效果。另外，结合实体序列标注，可以解决在指针网络中因漏标而解码出一些较长的实体的问题。最终，模型在DuIE2.0数据集上的F1值为0.750 8，因为该数据还未被太多研究者使用，所以在百度官方排名上暂列第一。另外，在实验过程中发现，基于远程监督和人工修正的DuIE2.0数据集中存在较多的漏标三元组。在后面的工作中，笔者计划研究漏标、错标的噪声问题。

参考文献:

[1] 邹艳珍，王敏，谢冰，等. 基于大数据的软件

表 3 样例的注意力分布

	安	徽	四	创	电	子	股	份	有	限	公	司
安	0.06	0.09	0.14	0.02	0.03	0.02	0.06	0.05	0.06	0.03	0.24	0.18
徽	0.14	0.03	0.05	0.01	0.03	0.01	0.08	0.05	0.09	0.06	0.25	0.19
…	…	…	…	…	…	…	…	…	…	…	…	…
1	0.06	0.07	0.03	0.04	0.04	0.03	0.10	0.13	0.14	0.16	0.13	0.07
.	0.11	0.03	0.05	0.01	0.03	0.02	0.12	0.11	0.11	0.15	0.17	0.11
59	0.10	0.04	0.04	0.02	0.02	0.03	0.13	0.11	0.14	0.12	0.15	0.10
亿	0.03	0.03	0.08	0.06	0.06	0.04	0.12	0.09	0.20	0.16	0.10	0.03
…	…	…	…	…	…	…	…	…	…	…	…	…

项目知识图谱构造及问答方法[J]. 大数据, 2021, 7(1): 22-36.
ZOU Y Z, WANG M, XIE B, et al. Software knowledge graph construction and Q&A technology based on big data[J]. Big Data Research, 2021, 7(1): 22-36.

[2] 陈成, 陈跃国, 刘宸, 等. 意图知识图谱的构建与应用[J]. 大数据, 2020, 6(2): 57-68.
CHEN C, CHEN Y G, LIU C, et al. Constructing and analyzing intention knowledge graphs[J]. Big Data Research, 2020, 6(2): 57-68.

[3] AITKEN J S. Learning information extraction rules: an inductive logic programming approach[C]//Proceedings of ECAI. [S.l.:s.n.], 2002: 355-359.

[4] AONE C, RAMOS-SANTACRUZ M. REES: a large-scale relation and event extraction system[C]//Proceedings of the 6th Conference on Applied Natural Language Processing. [S.l.:s.n.], 2000: 76-83.

[5] IRIA J. T-rex: a flexible relation extraction framework[C]//Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics. [S.l.:s.n.], 2005.

[6] JIANG J, ZHAI C X. A systematic exploration of the feature space for relation extraction[C]//Proceedings of the Main Conference on Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2007: 113-120.

[7] SUN X, DONG L H. Feature-based approach to Chinese term relation extraction[C]//Proceedings of the 2009 International Conference on Signal Processing Systems. Piscataway: IEEE Press, 2009: 410-414.

[8] YAN X, MOU L L, LI G, et al. Classifying relations via long short term memory networks along shortest dependency paths[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. [S.l.:s.n.], 2015: 1785-1794.

[9] GUO Z J, ZHANG Y, LU W. Attention guided graph convolutional networks for relation extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2019: 241-251.

[10] ZHONG Z X, CHEN D Q. A frustratingly easy approach for joint entity and relation extraction[J]. arXiv preprint, 2020, arXiv:2010.12812.

[11] WANG J, SHOU L D, CHEN K, et al. Pyramid: a layered model for nested named entity recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2020: 5918-5928.

[12] ZHENG C M, CAI Y, XU J Y, et al. A boundary-aware neural model for nested named entity recognition[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: ACL Press, 2019: 357-366.

[13] 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: ACL Press, 2020: 1476-1488.

[14] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017.

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

understanding[J]. arXiv preprint, 2018, arXiv:1810.04805.
[17] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized bert pretraining approach[J]. arXiv preprint, 2019, arXiv:1907.11692.
[18] 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. Stroudsburg: ACL Press, 2014: 2335–2344.
[19] 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. Stroudsburg: ACL Press, 2012: 1201–1211.
[20] 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. Stroudsburg: ACL Press, 2019: 1409–1418.
[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. Stroudsburg: ACL Press, 2016: 1105–1116.
[22] 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. Stroudsburg: ACL Press, 2017: 1227–1236.
[23] LI X Y, YIN F, SUN Z J, et al. Entity–relation extraction as multi–turn question answering[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2019: 1340–1350.
[24] LI S J, HE W, SHI Y B, et al. DuIE: a large–scale Chinese dataset for information extraction[C]//Proceedings of the 8th CCF International Conference on Natural Language Processing and Chinese Computing. [S.l.:s.n.], 2019: 791–800.

作者简介

郑慎鹏（1995– ），男，同济大学电子与信息工程学院硕士生，主要研究方向为自然语言处理。

陈晓军（1995– ），男，同济大学电子与信息工程学院博士生，主要研究方向为自然语言处理。

向阳（1962- ），男，同济大学电子与信息工程学院教授，主要研究方向为数据挖掘、自然语言处理、智能决策支持系统。

沈汝超（1989- ），男，上海国际港务（集团）股份有限公司工程师，主要研究方向为港口科技管理。

收稿日期： 2021-01-30

通信作者： 向阳，shxiangyang@tongji.edu.cn

基金项目： 国家自然科学基金资助项目（No.72071145）；国家重点研发计划资助项目（No.2019YFB1704402）

Foundation Items: The National Natural Science Foundation of China(No.72071145), The National Key Research and Development Program of China(No.2019YFB1704402)