基于双向门控循环神经网络的事件论元抽取方法

葛唯益[1]，程思伟[2]，王 羽[1]，徐 建[2*]

(1. 中国电子科技集团公司第二十八研究所信息系统工程重点实验室 南京 210007；
2. 南京理工大学计算机科学与工程学院 南京 210094)

【摘要】事件抽取是构建知识图谱的关键前置任务之一，而事件论元抽取是事件抽取的子任务，对事件抽取质量有显著影响。针对现有的流水线式事件抽取方法在论元抽取时忽略了触发词和论元间、论元和论元间相互关系导致抽取质量低的问题，该文提出了一种基于双向门控循环神经网络 (Bi-GRU) 的事件论元抽取方法。该方法融合 Bert 词向量、词性特征、词位置特征和触发词类型特征作为输入，采用 Bi-GRU 网络对文本中的词进行编码，进而应用改进的多注意力机制为句子不同部分分配权重提取句子级别特征，最后通过全连接层实现论元识别和角色分类。在基准数据集上进行了实验验证，结果表明论元识别和角色分类任务的 F1-score 值分别达到了 69.2% 和 61.6%，优于现有方法。

关 键 词 注意力机制； 事件抽取； 循环神经网络； 词嵌入
中图分类号 TP183 **文献标志码** A

Bi-GRU-Based Event Argument Extraction Approach

GE Weiyi[1], CHENG Siwei[2], WANG Yu[1], and XU Jian[2*]

(1. Science and Technology on Information Systems Engineering Laboratory, the 28th Research Institute of China Electronics Technology Group Corporation Nanjing 210007;
2. School of Computer Science & Engineering, Nanjing University of Science & Technology Nanjing 210094)

Abstract Event extraction is one of the important precedent tasks for knowledge graphs, while as a sub-task, event argument extraction has a significant impact on the quality of event extraction. The existing pipelined event extraction approaches usually ignore the relationships between triggers and arguments, or among arguments, which leads to low quality of event extraction. To solve this issue, this paper proposes a bidirectional gated recurrent neural network (Bi-GRU)-based event argument extraction approach. The proposed approach considers Bert-based word vector, word part-of-speech, word position, and trigger types as features, applies Bi-GRU to encode these features of each word in a sentence to get a word vector, leverages the improved multi-attention mechanism to assign weights to different parts of the sentence, and finally identify arguments and their roles in the sentence by a full-connection layer. Experiments are conducted on a benchmark dataset, and experimental results show that argument recognition and role classification tasks achieve 69.2% and 61.6% in F1-score respectively, and are better than compared state-of-the-art approaches.

Key words attention mechanism; event extraction; recurrent neural network; word embedding

文本事件抽取旨在从句子或文档中识别发生的事件，以结构化的方式描述事件的触发词、事件类型、事件论元及其角色，通常是信息检索中的重要前置任务之一，在诸多领域有着广泛应用。如在政府公共事务管理领域，及时捕获社会事件的爆发和掌握演变动态将有助于快速应急响应和事件处置，维护社会安定。因此，面向以自然语言形式存在的文本数据，研究满足应用场景需要的事件抽取方法成为当前热点研究课题之一。

尽管已经开展了相关的研究工作，事件抽取仍然是一项颇具挑战性的任务，主要原因有以下几个方面。首先，自然语言形式表达的文本信息通常具

收稿日期：2021－06－07；修回日期：2021－09－03
基金项目：国家自然科学基金 (61872186)
作者简介：葛唯益 (1985－)，男，博士，高级工程师，主要从事知识图谱、自然语言处理等方面的研究.
*通信作者：徐建，dolphin.xu@njust.edu.cn

有语义歧义和多样化的话语风格，增加了处理难度。其次，事件抽取还依赖于自然语言处理 (natural language processing, NLP) 中若干子任务的性能，如命名实体识别、词性标记和语法解析等。为了应对上述挑战，文献 [1-5] 提出了基于模式匹配的事件抽取方法。该方法先构造一些特定的事件模板，然后执行模板匹配从文本中提取带有参数的事件。代表性的工作有 AutoSlog[1]、GenPAM[3]、BEECON[4] 和 PALKA[5]。虽然由具有专业知识的专家手动构建事件模式质量非常高，且针对特定领域通常可以实现较高的抽取精度，但是手动构建耗时费力，且无法迁移应用到其他领域中。随着机器学习方法在事件抽取方面的广泛应用，研究人员又提出了基于机器学习的事件抽取方法克服人工构建模板的局限性。该方法的基本思路是从训练数据中学习分类器，并将分类器应用于从新文本中提取事件。由于事件抽取可以进一步分为触发词抽取和论元抽取两个子任务，根据两个子任务的完成时间顺序，可以划分为基于流水线式的事件抽取模型[6-10] 和联合抽取模型[11-15]。前者将触发词抽取和论元抽取任务以串行的方式进行，且针对任务特点采用不同的分类器，更注重结构性，针对性模型能够收获更好的效果；而后者同时完成触发词抽取和论元抽取任务，考虑两个任务之间的信息交互，注重任务的整体性。最近，神经网络在 NLP 任务中不断取得突破，基于深度神经网络的事件抽取方法[16-23] 研究得到了很多关注，寻找抽取效果更佳的深度学习模型成为主要难点问题。代表性的工作有：基于卷积神经网络 (convolutional neural networks, CNN) 的事件抽取方法 DMCNN[17]，基于递归神经网络 (recurrent neural networks, RNN) 的事件抽取方法 JRNN[19] 和 dbRNN[22]，基于图卷积神经网络的事件抽取方法 JMEE[23]。基于 CNN 的事件抽取方法的缺点是无法很好地捕捉到距离较远的单词之间的相互关系，因为 CNN 是将单词嵌入级联作为输入的。RNN 刻画可以利用直接或者间接连接的两个任意的词之间的潜在依赖关系，但也存在长距离遗忘的问题。此外，现有的基于深度神经网络的事件抽取方法大多忽略触发词与触发词之间的关联，在多事件句上的效果不佳。

针对上述问题，本文提出一种基于双向门控循环神经网络 (bidirectional gated recurrent neural network, Bi-GRU) 和多注意力机制的事件论元抽取模型，该模型在输入层结合深度上下文词向量和基础特征编码句子，经过Bi-GRU 层特征提取后，输入改进的多注意力机制层，从 3 个方向计算注意力权重，编码语义结构之间的相似度，最后进行分类，完成事件论元抽取任务。

1　事件论元抽取方法

1.1　框架

为了提高论元抽取精度，本文提出了基于 Bi-GRU 和多注意力机制的事件论元抽取模型，命名为 Bi-GRU-MATT，其框架如图 1 所示。该模型由特征编码层、Bi-GRU 层、多注意力机制层和全连接层组成。每一层的输入输出和作用如下。

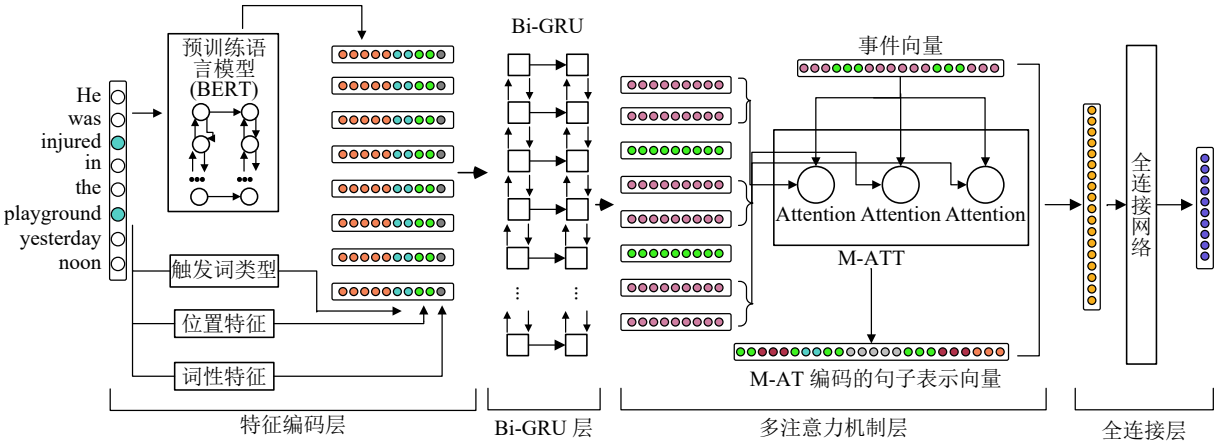

图 1　基于双向门控循环单元和多注意力机制的事件论元抽取模型

1) 特征编码层结合经过预训练的深度语言模型编码的单词复杂特征、触发词类型特征、位置特征和词性特征，将每个单词 token 编码为定长的具有原始句子语义和上下文信息的向量；

2) 将编码层得到的实值向量输入 Bi-GRU 进行进一步编码，GRU 相较于长短期记忆网络 (long

short-term memory, LSTM) 计算效率较高，且模型简单，适合于构建较大的模型。通过 Bi-GRU 进一步捕捉长距离依赖后，输出完整的融合深层语义信息的句子表示。

3) 考虑到同一个单词在不同触发词表示的事件下可能扮演不同的事件论元角色，将 Bi-GRU 编码得到的深层语义信息向量通过改进的注意力神经网络明确编码为句子的表示向量，提取深层的语义信息，输出最终编码向量。

4) 将之前编码得到的深层语义向量和事件向量结合输入全连接网络，结合 Softmax 对句子中的单词 token 进行分类，分类器的输出计为每个论元角色计算置信度得分。

1.2 特征编码层

为了编码深层的上下文信息，在特征编码层考虑 4 个方面的特征对词进行编码，分别是词向量、词性标注 (part-of-speech tagging, POS)、词位置特征和触发词类型特征。具体地，选用当前先进的预训练语言模型 (bidirectional encoder representation from transformers, BERT) 来代替传统的预训练词向量。句子通过 BERT 编码得到的动态词向量表示为$\boldsymbol{e}_{B,i}$，其中i表示句子在第i个位置的单词。考虑到确定论元角色的词性是非常重要的一部分，如“Attack”触发的事件承受者通常是名词，因此将词性特征加入编码，用 one-hot 向量表示为$w_{p,i}$，词性特征的标签共有 46 类 (含标点符号和“<unk>”、“<pad>”)。触发词的事件类型是影响事件论元检测的最重要特征，用 one-hot 向量表示为$w_{t,i}$。同时，将触发词位置信息包括在编码中代表触发词与候选单词的空间关系，这需要给定一个输入的位置序列$\{w_{r,1},w_{r,2},\cdots,w_{r,n}\}$，$w_{r,i}$在 0 和 1 上取值，1 表示触发词，0 表示不是触发词，第 i 个单词表示为$w_{r,i}$。3 个基本特征中的i均表示句子的第 i 个位置的单词，p,t,r用来区分不同的特征。$w_{p,i},w_{t,i},w_{r,i}$可映射成向量，分别为$\boldsymbol{e}_{p,i},\boldsymbol{e}_{t,i},\boldsymbol{e}_{r,i}$：

$$\boldsymbol{e}_{p,i}=\boldsymbol{M}_p w_{p,i} \tag{1}$$

$$\boldsymbol{e}_{t,i}=\boldsymbol{M}_t w_{t,i} \tag{2}$$

$$\boldsymbol{e}_{r,i}=\boldsymbol{M}_r w_{r,i} \tag{3}$$

式中，$\boldsymbol{M}_p,\boldsymbol{M}_t$和$\boldsymbol{M}_r$表示映射矩阵。映射得到$\boldsymbol{e}_{p,i},\boldsymbol{e}_{t,i},\boldsymbol{e}_{r,i}$后，特征编码层将$\boldsymbol{e}_{B,i}$和映射得到的 3 个特征级联，用矩阵$\boldsymbol{M}_f$映射成维度为$d$的单词嵌入$\boldsymbol{e}_i$：

$$\boldsymbol{e}_i=\boldsymbol{M}_f*\left[\boldsymbol{e}_{B,i};\boldsymbol{e}_{p,i};\boldsymbol{e}_{t,i};\boldsymbol{e}_{r,i}\right] \tag{4}$$

式中，$\boldsymbol{e}_i$为第i个句子的向量表示；$\boldsymbol{M}_f$为映射矩阵。

得到句子中的每个单词x_i编码为实值向量$\boldsymbol{e}_i$后，输入的句子$\boldsymbol{W}$被转换为向量序列E，可表示为$E=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n)$。设词嵌入的维度为$d_w$，触发词嵌入的维度为$d_t$，位置嵌入的维度为$d_s$，词性嵌入的维度为$d_p$，级联之后，$\boldsymbol{e}_i$的维度$d_i$可表示为：

$$d_i=d_w+d_t+d_s+d_p \tag{5}$$

级联起来的包含丰富语义的特征向量作为 Bi-GRU 层的输入，为$n\times d_i$维的矩阵，n为句子中的单词个数。将编码好的特征向量输入后面 Bi-GRU-MATT 的其他层进行进一步的分类任务。

1.3 双向门控循环单元层

得到特征编码层输出的句子表示向量序列$\boldsymbol{W}$后，将向量序列输入一个 Bi-GRU，通过 RNN 编码来进一步捕获长距离的依赖关系和上下文信息。选用 Bi-GRU 作为 RNN 编码层的原因在于与具有相同功效的 LSTM 相比，GRU 计算更容易，具有更高的模型训练效率，能捕获原始输入中包含的长距离依赖信息。

在模型 Bi-GRU-MATT 中，模型中采用的更新门状态和重置门状态分别为：

$$\boldsymbol{z}_i=\sigma(\boldsymbol{W}_z\boldsymbol{e}_i+\boldsymbol{U}_z\boldsymbol{h}_{i-1}+\boldsymbol{b}_z) \tag{6}$$

$$\boldsymbol{r}_i=\sigma(\boldsymbol{W}_r\boldsymbol{e}_i+\boldsymbol{U}_r\boldsymbol{h}_{i-1}+\boldsymbol{b}_r) \tag{7}$$

式中，σ是 sigmoid 函数，负责转换门控信号；$\boldsymbol{W}_z$、$\boldsymbol{W}_r$、$\boldsymbol{U}_z$、$\boldsymbol{U}_r$、$\boldsymbol{b}_z$、$\boldsymbol{b}_r$都是模型自主学习的参数；$\boldsymbol{h}_{i-1}$是第$i-1$步的输出向量；$\boldsymbol{z}_i$是更新门得到的向量；$\boldsymbol{r}_i$是重置门得到的向量。门控信号计算出来后，先用重置门来重置$\boldsymbol{h}_{t-1}$，重置后的$\boldsymbol{h}_{t-1}$记为$\boldsymbol{h}'_{t-1}$，再将其与输入x_t拼接后通过tanh激活函数缩放数据到$[-1,1]$内，如式 (8) 和式 (9) 所示：

$$\boldsymbol{h}'_{t-1}=\boldsymbol{h}_{t-1}\odot\boldsymbol{r}_t \tag{8}$$

$$\widetilde{\boldsymbol{h}}_t=\tanh\left(\boldsymbol{W}\cdot\left[\boldsymbol{h}'_{t-1},\boldsymbol{x}_t\right]\right) \tag{9}$$

模型中该步骤可以表示为：

$$\widetilde{\boldsymbol{h}}_i=\tanh(\boldsymbol{W}\cdot\boldsymbol{e}_i+\boldsymbol{r}_i\boldsymbol{U}\boldsymbol{h}_{i-1}+\boldsymbol{b}) \tag{10}$$

最后在更新阶段更新记忆，使用之前得到的$\boldsymbol{z}_t$，可同时进行遗忘和选择步骤，得到第i步的输出$\boldsymbol{h}_i$：

$$\boldsymbol{h}_i=(1-\boldsymbol{z}_i)\widetilde{\boldsymbol{h}}_i+\boldsymbol{z}_i\boldsymbol{h}_{i-1} \tag{11}$$

考虑到部分依赖与过去的状态和未来的状态有关，模型在 Bi-GRU 层从正向和反向两个方向使用门控循环单元编码，捕捉丰富的长距离依赖，通过 Bi-GRU 将句子的表示E从两个方向编码为：

$$\overrightarrow{\boldsymbol{p}_t} = \overrightarrow{\mathbf{GRU}}\left(\overrightarrow{\boldsymbol{h}_{t-1}}, \boldsymbol{e}_t\right) \tag{12}$$

$$\overleftarrow{\boldsymbol{p}_t} = \overleftarrow{\mathbf{GRU}}\left(\overleftarrow{\boldsymbol{h}_{t-1}}, \boldsymbol{e}_t\right) \tag{13}$$

经过双向编码之后，第t个单词的编码为$\boldsymbol{e}_t = [\overrightarrow{\boldsymbol{p}_t}, \overleftarrow{\boldsymbol{p}_t}]$，即将双向门控循环单元的两个方向的编码拼接起来得到编码向量序列E=$(\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n)$，这在特征编码层初始特征的基础上融合了更为丰富的长距离依赖信息的句子向量表示。

1.4　多注意力机制层

多注意力机制层为 Bi-GRU-MATT 模型的核心层。事件抽取的难点之一是句子中的某个事件论元可能在两个不同的触发词触发的事件中承担着不同的论元角色。因此，句子的特征与事件触发词、事件候选论元高度相关，在计算句子的特征表示时，这些信息十分重要。所以，在 Bi-GRU-MATT 模型中，使用融合注意力机制的神经网络代替传统的卷积神经网络，进行句子级别特征提取。

注意力机制通常用于将向量序列编码为固定长度的句子表示形式。鉴于同一个句子中可能包含多个事件并且同一个参数可能表示的论元不同，本文采用了一种改进的注意力机制，将变化的触发词明确地编码为句子表示向量，称之为多注意力机制。

句子$\boldsymbol{W} = (w_1, w_2, \cdots, w_n)$经过特征编码层和 Bi-GRU 层编码之后的输出为向量序列$E = (\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n)$，句子中第$i$个单词$w_i$对应的向量编码为$\boldsymbol{e}_1$，通过向量序列$E = (\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n)$可以生成事件向量$\boldsymbol{q}_{\text{event}}$。事件向量代表的是词汇级别的特征表示，考虑到事件向量包含更为丰富的上下文信息，有助于分类准确度的提高，本模型的多注意力机制层的事件向量采用候选触发词和候选事件论元参数的特征编码，以及它们的上一个词和下一个词的特征编码拼接生成的事件向量，如式 (14) 所示：

$$\boldsymbol{q}_{\text{event}} = [\boldsymbol{e}_{i_t-1}; \boldsymbol{e}_{i_t}; \boldsymbol{e}_{i_t+1}; \boldsymbol{e}_{i_c-1}; \boldsymbol{e}_{i_c}; \boldsymbol{e}_{i_c+1}] \tag{14}$$

式中，i_t表示候选触发词的位置；i_c表示候选事件论元的位置。相较于单纯使用候选词 (候选事件触发词和候选事件论元参数)，拼接生成的事件向量包含了候选词的邻近上下文信息，能得到更好的分类效果。

事件向量是词汇级别的特征编码，还需要句子级别的特征向量来完成分类任务。在 Bi-GRU-MATT 模型中采用改进的多注意力机制来得到句子表示$\boldsymbol{s}_{\text{sen}}$。根据候选触发词和候选事件论元，每个句子可以分割为 3 部分，分别与事件向量$\boldsymbol{q}_{\text{event}}$进行注意力运算，得到句子表示$\boldsymbol{s}_{\text{sen}}$。由于候选触发词和候选事件论元的位置$i_t$和$i_c$前后顺序在不同句子中可能有区别，不失一般性，假设$i_t < i_c$，因此，句子表示$\boldsymbol{s}_{\text{sen}}$的计算可表述为：

$$\boldsymbol{s}_{\text{sen}} = \begin{bmatrix} \text{att}(E, 1, i_t); \text{att}(E, i_t+1, i_c); \\ \text{att}(E, i_c+1, n) \end{bmatrix} \tag{15}$$

式中，$\text{att}(E, a, b)$是注意力权重计算函数，表示对句子中所有单词向量做加权的线性组合：

$$\text{attention}(E, a, b) = \sum_{a}^{b} \alpha_i \boldsymbol{e}_i \tag{16}$$

式中，α_i是注意力权重，每个单词的注意力权重为：

$$\alpha_i = \frac{\exp(o_i)}{\sum_j \exp(o_j)} \tag{17}$$

式中，o_i为 Attention 计算的注意力权重：

$$o_i = a(\boldsymbol{e}_i, \boldsymbol{q}_{\text{event}}) \tag{18}$$

式中，$a(x, y)$表示注意力权重函数，是注意力机制的核心，在注意力机制中用于对 Query 和 key 计算注意力权重，在本模型事件论元抽取任务中用于对事件向量$\boldsymbol{q}_{\text{event}}$和句子$W$位置$i$处的单词的匹配程度进行评分，计算注意力权重。注意力权重函数没有固定的形式，只需要对两个输入向量得到一个相似度分数即可。这里使用非线性标度乘积函数，如式 (19) 所示，它考虑了隐藏层的维度和非线性，使得该函数更具有表达性：

$$a(\boldsymbol{s}, \boldsymbol{h}) = \frac{1}{\sqrt{k}} f(\boldsymbol{W}_1 \boldsymbol{s})^{\mathrm{T}} f(\boldsymbol{W}_2 \boldsymbol{h}) \tag{19}$$

式中，$\boldsymbol{W}_1$和$\boldsymbol{W}_2$代表权重矩阵；f表示非线性函数，这里选用ReLU函数。经过多注意力机制层编码后，得到了事件向量$\boldsymbol{q}_{\text{event}}$和 M-ATT 编码的句子表示向量$\boldsymbol{s}_{\text{sen}}$，分别代表了词汇级别的特征和句子级别的特征，共同输入全连接层完成分类任务。

1.5　全连接层

在多注意力层之后，接上一层全连接层完成最

后的分类任务。全连接层的输入k是由事件向量$\boldsymbol{q}_{\text{event}}$和学习到的 M-ATT 编码的句子表示向量$\boldsymbol{s}_{\text{sen}}$级联起来得到的，表示为：

$$k=[\boldsymbol{q}_{\text{event}};\boldsymbol{s}_{\text{sen}}] \tag{20}$$

式中，全连接层输入k的维度是$9d_e$，$k\in R^{9d_e}$，d_e是输入句子$\boldsymbol{W}$中每个单词经过特征编码层和 Bi-GRU 层编码后的输出向量的维度。将k输入全连接层来抽取事件论元的参数标签：

$$\boldsymbol{y}=\text{softmax}(\boldsymbol{W}_t\boldsymbol{k}+\boldsymbol{b}_t) \tag{21}$$

式中，softmax 表示的是softmax函数；$\boldsymbol{y}\in R^m$；$\boldsymbol{W}_t\in R^{m\times 9d_e}$；$\boldsymbol{b}_t\in R^m$；$m$指待抽取的事件论元角色数量，包括非事件论元“NONE”；W_t和b_t是模型待学习的参数；$\boldsymbol{y}$是模型的输出，为每一个事件论元角色提供了置信度得分，并且使用 softmax 归一化。

1.6 损失函数

和事件触发词检测任务相同，使用全连接层输出$\bar{y}$的负对数似然作为整个模型的损失函数：

$$J(\theta)=-\frac{1}{N}\sum_{p=1}^{N}\sum_{i=1}^{n}y_{p_i}\log\left(\bar{y}_{p_i},\theta\right) \tag{22}$$

式中，θ表示整个模型的参数集合；N为输入的句子总数；n为事件论元类型的标签数量，包括 NONE 类型的标签；y_{p_i}是一个二值的指标，当y_i代表真正的事件论元角色时，它的值为 1，其他情况下为 0；$\bar{y}_{p_i}$是模型预测输入实例 p 属于事件论元类别i的概率。

2 实 验

2.1 实验设置

在事件抽取基准数据集 ACE2005 开展实验。该数据集中定义了 35 个事件论元类型，加上 NONE 类型，共 36 个类型。为了与已有研究工作进行比较，使用与它们相同的数据分割方案，即 40 个新闻类的文章 (共有 881 个句子) 作为测试集，30 个其他类型的文本 (共有 1 087 个句子) 作为验证集，剩下的 529 个文本 (共有 21 090 个句子) 用作训练集。

基于 pytorch 框架实现模型，使用 standford CoreNLP 工具包和自然语言处理库 torchtext 来进行数据预处理，将句子分词并获得句子中每个单词w_i的词性标注。使用 Google 官方的预训练模型 BERT-Base 获取特征编码层上下文相关的词向量表示，该预训练模型包含 12 层 transformer，隐藏层维度 768 维，参数量 1.1 亿个。对于编码层的词性 POS 特征、触发词类型特征以及位置特征，维度均为 50，最大句子长度设置为 50，比 50 短的句子用 padding 操作补上，比 50 长的句子则进行截断操作。Bi-GRU 隐藏层维度为 200，dropout 设为 0.5，且 batch 的大小为 64。和大部分模型相同，模型中使用 ReLU 作为非线性激活函数。同时使用 mini-batch 小批量随机梯度下降和 AdaDelta 更新规则，应用反向传播来计算梯度。模型训练 20 个 epoch。Bi-GRU-MATT 模型采用正交矩阵和高斯分布来分别初始化参数矩阵和其他参数。

为了评估 Bi-GRU-MATT 模型在事件论元抽取任务上的性能，使用精确率 (Precision)、召回率 (Recall) 和 F1(F1-score) 作为评价指标：

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{23}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{24}$$

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{25}$$

式中，TP 是混淆矩阵中将正类预测为正类的数目；FP 是混淆矩阵中将负类预测为正类的错误预测数；FN 是混淆矩阵中将正类预测为负类的错误预测数。

2.2 消融实验

为了更清晰地展示 Bi-GRU-MATT 模型每个层对于模型的贡献，进行消融实验，分别从 BERT 预训练语言模型，Bi-GRU 层和多注意力机制层评估了各层的作用。

2.2.1 BERT 对模型性能的影响

本实验旨在揭示 BERT 预训练模型对 Bi-GRU-MATT 论元抽取性能的影响。考虑到训练集标注数据量大和梯度消失的问题，将词嵌入作为可训练的参数去训练模型，会大幅度增加参数量进而引起过拟合问题，因此实验中不直接剔除预训练语言模型来做消融实验，而是替换为其他典型的预训练词向量并比较几种不同词向量编码下模型的抽取效果，结果如表 1。可以看出，采用 BERT 深度预训练上下文语言模型编码句子中各个单词，在事件论元识别和论元角色分类任务中均达到了最佳效果。具体地，在两个具体任务上，采用了上下文相关的词向量编码 BERT 和 ELMo 的结果都显著优于采用传统的上下文无关词向量 word2vec 和 GloVe，这表明包含深层语义和上下文信息的词向量具有更好的

表示能力。进一步地，将 BERT 与 ELMo 相比比较，两个任务的 F1-score 值分别提升了 1.3% 和 1.4%，这得益于 BERT 采用了完全双向信息弥补了 ELMo 的缺陷，且句子级负采样使得 BERT 的编码级别提升至句子级，将句子信息融入编码中使得采用 BERT 的 Bi-GRU-MATT 在实验中取得了最佳效果。

表 1　特征编码层不同的单词编码方式对模型性能的影响

词向量	论元识别/%			论元角色分类/%		
	Precision	Recall	F1	Precision	Recall	F1
word2vec	67.7	60.2	63.7	58.6	53.4	55.8
GloVe	68.3	58.1	62.7	59.4	53.2	56.1
ELMo	70.2	**65.7**	67.9	67.1	54.7	60.2
BERT	**73.3**	65.6	**69.2**	**69.9**	**55.1**	**61.6**

2.2.2　Bi-GRU 层和多注意力机制层对模型性能的影响

本节通过单独移除 Bi-GRU 层和多注意力机制层的方式来评估它们对模型的性能影响，结果如表 2 所示。从表中可以看出，多注意力机制层在事件论元识别和角色分类任务中分别使模型的 F1-score 值提升了 1.7% 和 1.6%，而 Bi-GRU 层在事件论元识别和角色分类任务中分别使模型的 F1-score 值提升了 1.0% 和 0.8%，这表明经过 Bi-GRU 编码后特征向量包含了更加丰富的长距离依赖关系以及句子特征。上述结果验证了多注意力网络和 Bi-GRU 编码的有效性。

表 2　Bi-GRU-MATT 模型相关消融实验

模型	论元识别/%			论元角色分类/%		
	Precision	Recall	F1	Precision	Recall	F1
Bi-GRU-MATT	73.3	65.6	69.2	69.9	55.1	61.6
Multi Attention	70.9	64.5	67.5	68.2	53.6	60.0
Bi-GRU	72.6	64.4	68.2	68.9	54.4	60.8

2.3　多注意力机制层相关分析

多注意力机制层是 Bi-GRU-MATT 模型的核心层。本实验针对多注意力机制层使用的不同注意力权重函数进行对比实验，目的是为了验证选择非线性标度乘积函数作为注意力函数的合理性。

具体地，用$a(\boldsymbol{s},\boldsymbol{h})$代表注意力权重函数，$\boldsymbol{s}$和$\boldsymbol{h}$代表参与注意力计算的两个向量。考虑以下 5 种不同的注意力权重函数来训练模型，其中函数 4 和 5 的非线性激活函数统一使用 ReLU 函数。

1. 乘积函数：$a(\boldsymbol{s},\boldsymbol{h})=\boldsymbol{s}^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{W}_2\boldsymbol{h}$。

2. 加和性函数：$a(\boldsymbol{s},\boldsymbol{h})=v^{\mathrm{T}}\tanh(\boldsymbol{W}_1\boldsymbol{s}+\boldsymbol{W}_2\boldsymbol{h})$。

3. 对称乘积函数：$a(\boldsymbol{s},\boldsymbol{h})=\boldsymbol{s}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{W}\boldsymbol{h}$。

4. 非线性对称乘积函数：$a(\boldsymbol{s},\boldsymbol{h})=f(\boldsymbol{W}s)^{\mathrm{T}}\boldsymbol{D}f(\boldsymbol{W}\boldsymbol{h})$

5. 非线性标度乘积函数：$a(\boldsymbol{s},\boldsymbol{h})=\frac{1}{\sqrt{k}}f(\boldsymbol{W}_1\boldsymbol{s})^{\mathrm{T}}f(\boldsymbol{W}_2\boldsymbol{h})$

在事件论元识别和角色分类两个任务上的实验结果如表 3 所示。可以看出，以 ReLU 为激活函数的非线性标度乘积函数作为注意力权重函数的模型在两个任务上获得了最高的 F1-score 值，表现优于线性的注意力函数，非线性标度乘积函数在两个任务上的 F1-score 值比表现最好的线性注意力函数分别高出 0.4% 和 0.2%。

表 3　不同注意力权重函数对模型性能的影响

注意力函数	论元识别/%			角色分类/%		
	Precision	Recall	F1	Precision	Recall	F1
Multiplicative	72.7	64.9	68.6	69.2	**55.2**	61.4
Additive	72.8	64.7	68.5	68.8	54.7	60.9
Symmetric multi	73.0	65.2	68.8	69.4	54.9	61.3
Symmetric multi (ReLU)	**73.6**	65.1	69.0	69.7	55.0	61.5
Scaled multi(ReLU)	73.3	**65.6**	**69.2**	**69.9**	55.1	**61.6**

2.4　模型在多论元事件句上的表现

为了进一步验证 Bi-GRU-MATT 模型在事件论元抽取任务上的有效性，特别是对于不止一个论元的句子。根据句子中论元的数量将句子分成两部分，其中仅有一个论元的事件句占整个数据集的 76.8%，包含至少两个论元的事件句占整个数据集的 23.2%。将 Bi-GRU-MATT 与基线模型 Embedding+T、CNN，以及 DMCNN、JRNN 和 JMEE 3 个前沿事件抽取模型进行对比，获得的 F1-score 值如表 4 所示。

表 4　Bi-GRU-MATT 模型在单论元事件句 (1/1) 和多论元事件句 (1/*N*) 上的抽取性能

模型	性能/%		
	1/1	1/*N*	F1-score
Embedding+T	37.4	15.5	32.6
CNN	51.6	36.6	48.9
DMCNN	54.6	48.7	53.5
JMEE	59.3	57.6	60.3
JRNN	50.0	55.2	55.4
Bi-GRU-MATT	**60.1**	**58.7**	**61.6**

从表 4 可以看出，Bi-GRU-MATT 模型无论是在单论元事件句 (1/1) 还是多论元事件句 (1/*N*) 上都有最高的 F1-score 值。在多论元事件句上，Bi-GRU-MATT 比动态多池化网络 DMCNN 的 F1-score 值高出了 7.1%，这验证了 Bi-GRU-MATT 方法的有效性。和同样使用了循环神经网络的模型 JMEE 和 JRNN 相比，F1-score 值分别提高了 1.3% 和 5.2%，这是因为本模型采用包含丰富语义的 BERT 模型编码单词，并且多注意力机制有助于学

习到更多的语义信息，提高模型的精度。

2.5 对比实验

将 Bi-GRU-MATT 与当前先进的事件抽取方法在事件论元识别和论元角色分类任务上进行对比。采用的对比方法分为 3 类，基于特征的抽取模型、基于流水线式的抽取模型和联合抽取模型，其中基于特征的抽取模型包括 Cross-Event、Cross-Entity 和 RBPB，基于流水线式的抽取模型有 DMCNN、JRNN、dbRNN，而联合抽取模型有 JMEE、S-CNNs[24]、Ding's model[25] 和 Joint3EE[26]。

表 5 给出了 Bi-GRU-MATT 模型与这些对比方法在事件论元抽取任务上的性能。可以看出，提出的 Bi-GRU-MATT 模型在事件论元识别和角色分类任务上均取得了最佳的 F1-score 值。Bi-GRU-MATT 模型和代表性的基于特征的抽取模型相比，精确率、召回率和 F1-score 值均显著优于后者，在两大任务上的 F1-score 值比最佳的基于特征的模型 (RBPB) 高 8.0% 和 7.8%，性能提升显著。与联合抽取模型对比，Bi-GRU-MATT 的 F1-score 值也优于它们。在事件论元检测任务上 F1-score 值比表现最好的联合抽取模型 (JMEE) 高 0.8%，且精确率和召回率也有提升。在论元角色分类任务上，单独执行触发词抽取和论元抽取任务的性能优于联合抽取的，主要原因在于事件类型是时间论元抽取任务中的重要特征，同时标记触发词和参数的联合模型容易忽视触发词类型特征相关的信息。

表 5 Bi-GRU-MATT 与其他先进方法的性能比较

模型	论元识别/%			论元角色分类/%		
	Precision	Recall	F1	Precision	Recall	F1
Cross-Event	50.9	49.7	50.3	45.1	44.1	44.6
Cross-Entity	53.4	52.9	53.1	51.6	45.5	48.3
DMCNN	68.8	51.9	59.1	62.2	46.9	53.5
S-CNNs	69.2	50.8	58.6	63.3	45.8	53.1
RBPB	63.2	59.4	61.2	54.1	53.5	53.8
JRNN	61.4	64.2	62.8	54.2	**56.7**	55.4
dbRNN	71.3	64.5	67.7	66.2	52.8	58.7
JMEE	71.4	65.6	68.4	66.8	54.9	60.3
Ding's model	64.7	65.0	64.8	57.4	55.8	56.6
Joint3EE	59.9	59.8	59.9	52.1	52.1	52.1
Bi-GRU-MATT	**73.3**	**65.6**	**69.2**	69.9	55.1	**61.6**

3 结 束 语

本文提出了一个基于 Bi-GRU 和改进注意力机制的事件论元抽取模型 Bi-GRU-MATT。该模型在特征编码层同样使用了深度的上下文预训练语言模型 BERT，并结合词性特征和位置特征，以及触发词特征来编码单词向量，之后送入 Bi-GRU 网络中编码长距离的依赖关系，再输入多注意力机制层计算注意力权重，生成事件向量和句子表示向量，级联输入全连接层完成最后的分类工作。实验表明该模型可以显著提升事件论元抽取的效果，在事件论元识别和论元角色分类任务上达到了较好的效果，F1-score 值分别为 69.2% 和 61.6%。

本文的研究工作得到了信息系统工程重点实验室开放基金项目 (05201901) 的支持，在此深表感谢！

参 考 文 献

[1] RILOFF E. Automatically constructing a dictionary for information extraction tasks[C]//Proceedings of the Eleventh National Conference on Artificial Intelligenc. Menlo Park, CA: AAAI, 1993, 1(1): 811-816.

[2] RILOFF E, SHOEN J. Automatically acquiring conceptual patterns without an annotated corpus[C]//The 3rd Workshop on Very Large Corpora. [S.l.]: Association for Computational Linguistics, 1995: 148-161.

[3] 姜吉发. 一种事件信息抽取模式获取方法[J]. 计算机工程, 2005, 31(15): 96-98.
JIANG J F. An event IE pattern acquisition method[J]. Computer Engineering, 2005, 31(15): 96-98.

[4] ARENDARENKO E, KAKKONEN T. Ontology-based information and event extraction for business intelligence[C]//International Conference on Artificial Intelligence: Methodology, Systems, and Applications. Heidelberg: Springer, 2012: 89-102.

[5] KIM J T, MOLDOVAN D I. Acquisition of linguistic patterns for knowledge-based information extraction[J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(5): 713-724.

[6] CHIEU H L, NG H T. A maximum entropy approach to information extraction from semi-structured and free text[C]//The 18th National Conference on Artificial Intelligence. [S.l.]: AAAI, 2002: 786-791.

[7] LLORENS H, SAQUETE E, NAVARRO B. TimeML events recognition and classification: Learning CRF models with semantic roles[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2010: 725-733.

[8] 丁效, 宋凡, 秦兵, 等. 音乐领域典型事件抽取方法研究[J]. 中文信息学报, 2011, 25(2): 15-21.
DING X, SONG F, QIN B, et al. Research on typical event extraction method in the field of music[J]. Journal of Chinese Information Processing, 2011, 25(2): 15-21.

[9] LIAO S S, GRISHMAN R. Using prediction from sentential scope to build a pseudo co-testing learner for event extraction[C]//Proceedings of 5th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 714-722.

[10] LIU S L, LIU K, HE S Z, et al. A probabilistic soft logic based approach to exploiting latent and global information in event classification[C]//The 30th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 2993-2999.

[11] LI P F, ZHU Q M, DIAO H J, et al. Joint modeling of trigger identification and event type determination in Chinese event extraction[C]//Proceedings of COLING 2012. Mumbai: The Coling 2012 Organizing Committee, 2012: 1635-1652.

[12] LI Q, JI H, HUANG L. Joint event extraction via structured prediction with global features[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2013: 73-82.

[13] MCCLOSKY D, SURDEANU M, MANNING C D. Event extraction as dependency parsing[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011: 1626-1635.

[14] RIEDEL S, SAETRE R, CHUN H W, et al. Bio-molecular event extraction with Markov logic[J]. Computational Intelligence, 2011, 27(4): 558-582.

[15] VENUGOPAL D, CHEN C, GOGATE V, et al. Relieving the computational bottleneck: Joint inference for event extraction with high-dimensional features[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 831-843.

[16] TONG M H, XU B, WANG S, et al. Improving event detection via open-domain trigger knowledge[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 5887-5897.

[17] CHEN Y B, XU L H, LIU K, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]//Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 167-176.

[18] NGUYEN T H, GRISHMAN R. Event detection and domain adaptation with convolutional neural networks [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 365-371.

[19] NGUYEN T H, CHO K, GRISHMAN R. Joint event extraction via recurrent neural networks[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2016: 300-309.

[20] ZHANG W B, DING X, LIU T. Learning target-dependent sentence representations for Chinese event detection[C]//China Conference on Information Retrieval. Switzerland: Springer, 2018: 251-262.

[21] DUAN S Y, HE R F, ZHAO W L. Exploiting document level information to improve event detection via recurrent neural networks[C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017: 352-361.

[22] LEI S, QIAN F, CHANG B B, et al. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 5916-5923.

[23] LIU X, LUO Z C, HUANG H Y. Jointly multiple events extraction via attention-based graph information aggregation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 1247-1256.

[24] ZHANG Z K, XU W R, CHEN Q Q. Joint event extraction based on skip-window convolutional neural networks[M]//Natural Language Understanding and Intelligent Applications. Switzerland: Springer, 2016: 324-334.

[25] DING R X, LI Z J. Event extraction with deep contextualized word representation and multi-attention layer[C]//International Conference on Advanced Data Mining and Applications. Switzerland: Springer, 2018: 189-201.

[26] NGUYEN T M, NGUYEN T H. One for all: Neural joint modeling of entities and events[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 6851-6858.

编　辑　税　红