计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500133

BGCN:基于 BERT 和图卷积网络的触发词检测

程思伟[1]　葛唯益[2]　王　羽[2]　徐　建[1]
1 南京理工大学计算机科学与工程学院　南京 210094
2 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
(735665705@qq.com)

摘　要　触发词检测是事件抽取的一项基本任务,该任务涉及对触发词进行识别和分类。目前,已有工作主要存在两方面的问题:1)用于触发词检测的神经网络模型只考虑了句子的顺序表示,且通过顺序建模的方法在捕捉长距离依赖关系时效率较低;2)基于表示的方法虽然解决了手动提取特征的问题,但用作初始训练特征的词向量对句子的表示程度有所欠缺,难以捕捉深层的双向表征。因此,文中提出了一种基于 BERT 模型和 GCN 网络的触发词检测模型 BGCN,该模型通过引入 BERT 词向量来强化特征表示,并引入句法结构来捕捉长距离依赖,对事件触发词进行检测。实验结果表明,所提方法在 ACE2005 数据集上的表现优于其他现有的神经网络模型。

关键词:BERT;双向 LSTM;图卷积网络;序列标注;事件触发词
中图法分类号　TP183

BGCN:Trigger Detection Based on BERT and Graph Convolution Network

CHENG Si-wei[1],GE Wei-yi[2],WANG Yu[2] and XU Jian[1]
1 School of Computer Science and Engineering,Nanjing University of Science and Technology,Nanjing 210094,China
2 Key Laboratory of Information System Engineering,28th Research Institute of China Electronic Science and Technology Group Corporation 210007,China

Abstract　Trigger word detection is a basic task of event extraction,which involves the recognition and classification of trigger words. There are two main problems in the previous work:(1)the neural network model for trigger word detection only considers the sequential representation of sentences,and the sequential modeling method is inefficient in capturing long-distance dependencies;(2)although the representation-based method overcomes the problem of manual feature extraction,the word vector used as the initial training feature lacks the degree of representation of the sentence,so it is difficult to capture the deep two-way representation. Therefore,we propose a trigger word detection model BGCN,based on BERT model and GCN network. This model strengthens the feature representation by introducing BERT word vector,and introduces syntactic structure to capture long-distance dependencies and detect event trigger words. Experimental results show that our method outperforms other existing neural network models on ACE2005 datasets.

Keywords　BERT,Bi-LSTM,Graph convolution network,Sequence annotation,Event trigger

1　引言

事件抽取是信息提取(Information Extraction,IE)领域中一项具有挑战性的任务。事件抽取主要研究从描述事件的文本中识别、抽取出事件并以结构化的形式呈现。在给定文本时,事件抽取需要识别具有特定类型的事件触发词以及与触发词相关的论元角色。表 1 是 ACE2005 中定义的事件抽取任务的相关术语。从技术上讲,根据自动内容抽取评测会议 ACE 给出的权威数据集 ACE2005 中的定义,事件抽取可以分为两个子任务:1)事件检测,即对事件触发词进行识别和分类;2)参数提取,即标识事件触发词的参数并标记它的角色。

表 1　事件抽取术语表[1]

Table 1　Event extraction glossary[1]

实体	语义类别中的一个对象或一组对象
实体提及	对实体的引用,通常是名词短语(NP)
事件触发词	最清楚地表示事件发生的主词
事件论元	事件提及中事件所涉及到的要素
论元角色	论元与其参与事件的关系(ACE 定义了 35 种论元角色)
事件提及	描述事件的短语或句子,包括触发词和参数

到稿日期:2020-05-26　返修日期:2020-08-10　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61872186);信息系统工程重点实验室开放基金(05201901)
This work was supported by the National Natural Science Foundation of China(61872186) and Science and Technology on Information System Engineering Laboratory(05201901).
通信作者:徐建(dolphin.xu@njust.edu.cn)

在基于神经网络的 NLP 任务中，词向量是一种非常重要的建模工具，被广泛用于各种 NLP 分类模型。在 NLP 应用中，作为深度学习模型的初始特征输入，很多模型的最终效果很大程度上取决于词向量的效果。word2vec[2] 和 GloVe[3] 是最常见的语言模型。word2vec 明显的缺点是上下文无关，因此采用 word2vec 的模型绝大多数会在下游具体的 NLP 任务中做编码操作；ELMo[4] 是于 2018 年提出的动态词向量模型，具体的方法是使用双向语言模型（Bi-LSTM）来实现，但由于模型的前项和后项两个 LSTM 模型是分别训练的，因此并不是完全的双向，并且 ELMo 要预测的下一个词在给定的序列中已经出现，在词的表示上也存在一定的缺陷。

在事件检测任务中，一个句子中存在多个事件是一种普遍现象，而检测出一个句子中的多个事件触发词要比检测单一的事件触发词更困难。很多研究者正致力于解决这个问题，如利用各种特征组合[5-6]、引入记忆向量和矩阵[7] 以及在句子级别的序列模型中保留更多的上下文信息[8]（如 RNN 和 CRF 模型）。然而，句子级别的顺序建模方法在捕捉远距离的依赖上的效率依然受到了很大影响。而且，传统的基于特征的方法需要广泛的人工工程，这也在很大程度上影响了模型的性能。引入句法弧信息可以有效地处理这个问题。与顺序相比，通过句法弧进行建模可以缩短一个句子中从一个触发词到另一个触发词的距离。而对于句法弧建模，则可以使用图卷积网络来完成。

因此，本文提出了一个新的触发词检测模型，通过 BERT[9] 可以直接获得整个输入句子的唯一向量表示。在 BERT 编码特征的基础上，通过引入句法分析树来增强信息流，并用图卷积网络（Graph Convolution Network，GCN）[10-12] 来对句子的句法关系进行建模，捕捉句子中的长距离依赖关系[13]，之后通过一个前馈神经网络来对触发词进行分类和识别。实验结果表明，加入图卷积网络模块后，比单一使用 BERT 词向量编码来进行分类任务的效果更好。在广泛使用的 ACE2005 数据集上评估 BGCN 模型的性能以及在实验中的优势，实验结果表明了 BGCN 模型在触发词检测任务上的有效性。总的来说，本文的贡献如下：

（1）提出了一个新型的基于 BERT 和句法结构的特征抽取框架，它可以增强信息流，无需借助复杂的 NLP 预处理来自动产生句子级别的特征。

（2）设计了一个神经网络触发词检测模型 BGCN，可以对句子中的事件触发词进行识别和分类。

（3）在广泛使用的 ACE2005 数据集上进行了实验，实验结果表明，本文提出的触发词抽取模型的性能优于其他模型。

2 相关工作

传统的基于特征的方法利用词法和全局特征来检测事件[14]，有的方法利用相似度聚类[15] 和模板过滤[16] 的方法来进行事件检测。随着深度学习在 NLP 中的流行，近年来触发词检测在深度学习领域的研究集中在模型改进和关系建模等方面。

大多数方法将触发词抽取看作一个有监督的多分类任务或者序列标注任务来建模。非常典型的基于神经网络的模型如下：Chen 等提出了 DMCNN 模型[8]，利用改进的卷积神经网络（Convolutional Neural Networks，CNN）来实现两阶段的事件提取模型，并改进了 CNN 的池化层，增加了动态多池化机制；Nguyen 等随后提出了一种基于 RNN 的联合模型[7]，用单词词向量、实体类型嵌入向量和二进制向量的级联编码了句子级特征，并将结果送入联合的事件抽取模型中进行训练。部分研究者考虑了对模型的改进，例如：Liu 等提出的 TD-DMN 网络将用于阅读理解的动态记忆网络应用到事件抽取任务上，通过多次处理上下文来更好地利用上下文信息[17]；TBNNAM 模型[18] 提出了一种不需要识别触发词的方法，与以往的 Pipeline 模式不同，直接对整个句子进行事件多分类；Zhang 等[19] 则考虑用基于转移的神经方法来改进建模，构建一个联合抽取模型。

模型创新主要是引入新的建模思路，而关系建模是在序列信息的基础上补充了关系信息，并借助新的关系模型来建模。这些方法又可以分为 3 类。1）RNN 类结构。Orr 等借助 DAG-GRU 建模句法信息[20]，通过注意机制将句法信息与时间结构相结合；dbRNN 是 Sha 等改进的 LSTM 结构[21]，在其基础上引入了远距离关系，不同于 DAG-GRU 借助 Attention 机制，Lei 等是在 LSTM 单元的结构上做了一定修改，引入了 dt 门来控制远距离信息。2）GCN 类结构。Liu 等于 2018 年提出了一种结合 GCN 的联合模型，考虑到了长距离依赖的问题，并率先将事件要素信息作用于事件触发词提取任务[22]，借助 Attention 机制[23] 将事件论元信息输入到事件检测任务中，效果有明显的提升；Yan 等于 2019 年提出了改进的网络 MOGAND[24]，考虑到高阶句法关系，提出了利用一阶句法图和高阶句法图来对候选触发词的多阶表示进行建模并合并；Cui 等于 2020 年提出了关系感知图卷积网络 RA-GCN[25]，在以往的基于 GCN 的事件检测模型上对词与词之间的关系进行建模，弥补了以往方法忽略了句法关系标签的缺陷。3）丰富关系信息。RNN 类和 GCN 类只是建模不一样，不足之处在于仅使用了句法信息，还有很多其他关系信息没有融合。Peng 等构建了一个细粒度的事件分类模型 PP-GCN[26]，引入了丰富的社会关系，如话题、语义等信息。

3 基于 BERT 和图卷积网络的触发词检测方法

3.1 BGCN 触发词检测模型框架

图 1 给出了本文提出的触发词检测模型，整个模型分为 3 层：句子编码层、句法图卷积网络层和触发词识别与分类层。

（1）用 BERT 预训练词向量，结合其他输入特征表示输入句子的句子编码层，该层连接句子中每个单词的词向量，并将词性标记向量作为输入，将句子中每一个单词 token 转换为定长的实值向量。

（2）从句子的句法结构中引入卷积操作的图卷积网络层，该层通过句子 W 的句法分析树来构造正向弧、反向弧和自循环弧，并将其送入多层图卷积网络进行训练，进一步捕捉长距离的依赖。将句子编码层中的实值向量经过双向 LSTM 编码后输入第一层 GCN 网络中，经过图卷积之后输出融合了句法信息的句子表示。

（3）在触发词识别与分类层，将句法图卷积网络层得到的最终句子表示向量输入前馈神经网络，结合 softmax 层对句子中的每一个单词进行分类，捕获句子中的触发词并识别其所属的类别。

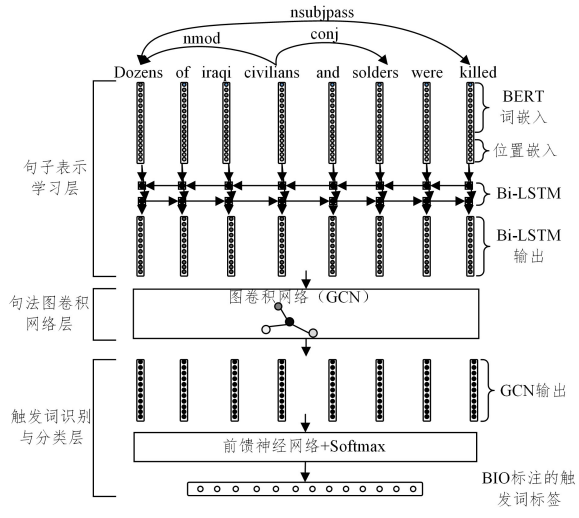

图 1 BGCN 触发词检测模型的架构

Fig. 1 Framework of BGCN trigger detection model

3.2 句子编码层

把一个长度为 n 的句子定义为 $W=w_1,w_2,\cdots,w_n$，w_i 表示第 i 个单词。使用 BIO 标注模式为每个单词 token 分配触发词标签 t_i。在句子编码层中，使用 BERT(Bidirectional Encoder Representations from Transformers)预训练模型将输入的句子转换成定长的向量表示。

句子编码层首先将输入句子 W 中的每一个 w_i 编码成 BERT[9] 需要的 3 个向量：Token Embedding，Segment Embedding 和 Position Embedding。

以句子“Davies is leaving to become chairman of the London School of Economics”为例，需要得到 3 个向量，如图 2 所示。

然后，设定输入句子 W 中需要 mask 的位置，为所有的输入位置都设置 mask。将上述变量输入 BERT 预训练模型就可以得到最终的 BERT 词向量。

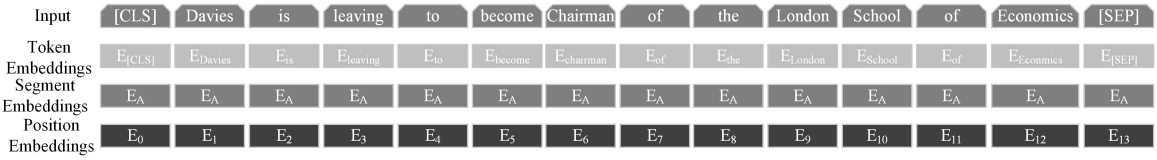

图 2 句子表示学习层编码的 BERT 需要的 3 个向量

Fig. 2 Sentence represents the required vectors of three BERT encoded by learning layer

最后，在 BGCN 模型的句子编码层中，通过连接以下两个向量，将句子中的每一个单词 token 转换为一个实值向量 $\boldsymbol{x}_i$：

(1)w_i 的 BERT 词向量：通过将上述 3 个 embedding 和 mask 输入 BERT 预训练模型得到的 BERT 词向量。

(2)w_i 的词性 POS 标记嵌入向量：这个嵌入向量是通过查找随机初始化的词性 POS 标记标签嵌入矩阵来生成的。

以上从句子的单词 token w_i 到实值向量 $\boldsymbol{x}_i$ 的转换，实质是将输入句子 W 转换为实值向量 $\boldsymbol{X}=(x_1,x_2,x_3,\cdots,x_n)$ 的序列，这个序列将被输入 BGCN 后面的模块中，以获取更多触发词检测的有效表示。

3.3 基于句法的图卷积网络层

句子编码层之后是 GCN 层。定义无向图 $G=(V,E)$ 作为句子 W 的句法分析树，其中 $V=v_1,v_2,v_3,\cdots,v_n(|V|=n)$ 是节点的集合，E 是边的集合。在节点集合 V 中，每个 v_i 对应句子 W 中的单词表示 w_i，每条边 $(v_i,v_j)\in E$ 是来自单词 w_i 和单词 w_j 的有向句法弧，这个句法弧 (v_i,v_j) 的标签类型是 $K(w_i,w_j)$。另外，为了实现信息流的反向传输，添加了一个具有类型标签 $K'(w^j,w^j)$ 且和原来的有向句法弧方向相反的反向弧 (v_j,v_i)[10]。同时，还对所有的 $v_i\in V$ 添加了一个自循环，即 (v_i,v_i)。

例如，ACE2005 数据集中有这样一条数据“Police have arrested four people in connection with the killings”。只看“arrested”和“connection”两个单词，在句法分析树中，从“arrested”到“connection”有一条句法弧标记为 nmod，根据上述规则，在单词“arrested”和“connection”之间存在 4 条边：带有类型标签 K(“arrested”“connection”)=nmod 的句法弧、带有附加类型标签 K(“connection”“arrested”)=nmod'的反向句法弧，以及两个单词“arrested”和“connection”的自循环弧(K(“arrested”“arrested”)=loop 和 K(“connection”“connection”)=loop)。因此，在第 k 层的句法图卷积网络中，通过式(1)计算节点 $v\in V$ 的图卷积向量 $\boldsymbol{h}_v^{(k+1)}$：

$$\boldsymbol{h}_v^{(k+1)}=f\left(\sum_{u\in N(v)}\left(\boldsymbol{W}_{K(u,v)}^{(k)}h_u^{(k)}+b_{K(u,v)}^{(k)}\right)\right) \tag{1}$$

其中，$K(u,v)$ 表示边 (u,v) 的类型标签；$\boldsymbol{W}_{K(u,v)}^{(k)}$ 和 $b_{K(u,v)}^{(k)}$ 分别是特定类型标签的 $K(u,v)$ 权重矩阵和偏置；$N(v)$ 是 v 节点的邻近节点集合，因为自循环弧的存在，$N(v)$ 是包括 v 的；f 是激活函数。此外，使用句子编码层 x_i 的输出来初始化第一层 GCN 节点的表示 $h_{v_i}^0$。

使用反向句法弧和自循环弧之后，预定义的定向弧类型标签个数将是原来的两倍，这意味着单层的 GCN 需要提供更多的参数。在实验中，使用 Standford Parser 从句法分析树中生成句法弧。当前的表示包含约 50 种不同的语法关系，这对于单层 GCN 的参数编号而言数量过多，并且与现有的训练数据规模不相符。为了减少参数的数量，将类型标签 $K(w_i,w_j)$ 的定义修改为式(2)：

$$K(w_i,w_j)=\begin{cases}along, & (v_i,v_j)\in\varepsilon\\ rev, & i!=jand(v_i,v_j)\in\varepsilon\\ loop, & i==j\end{cases} \tag{2}$$

这样一来，新的 $K(w_i,w_j)$ 只有 3 种类型的标签。

由于并非所有的弧对于下游任务都具有同等的信息，因此在生成的句法分析结构中也存在噪声。在弧上应用门来衡量每条弧各自的重要程度，为每条弧 (u,v) 计算权重 $g_{u,v}^{(k)}$ 来表示它们在触发词检测任务中所占的权重：

$$g_{u,v}^{(k)}=\sigma(h_u^{(k)}\boldsymbol{V}_{K(u,v)}^{(k)}+d_{K(u,v)}^{(k)}) \tag{3}$$

其中，σ 是 Sigmoid 函数，$\boldsymbol{V}_{K(u,v)}^{(k)}$ 和 $d_{K(u,v)}^{(k)}$ 是门控单元的权重矩阵和偏置。通过添加这种门控机制，最终的句法 GCN 计算可以表述为式(4)：

$$h_v^{(k+1)}=f(\sum_{u\in N(v)}g_{u,v}^{(k)}(\boldsymbol{W}_{K(u,v)}^{(k)}h_u^{(k)}+b_{K(u,v)}^{(k)})) \tag{4}$$

之后，按照 Liu 等的方法[13]，利用双向 LSTM 模型来编码句子表示层得到的实值向量序列 $X=(x_1,x_2,x_3,\cdots,x_n)$，将单词表示 X 编码为：

$$\overrightarrow{p_t}=\overrightarrow{LSTM}(\overrightarrow{p_{t-1}},x_t) \tag{5}$$

$$\overleftarrow{p_t}=\overleftarrow{LSTM}(\overleftarrow{p_{t-1}},x_t) \tag{6}$$

经过双向 LSTM 编码之后，第 t 个单词 token 的编码为 $\bar{x}_t=[\overrightarrow{p_t},\overleftarrow{p_t}]$，其中[,]表示连接操作，将两个向量拼接起来。Bi-LSTM 能更好地捕捉每个单词 token 的上下文，从而得到编码向量序列 $\bar{X}=(\bar{x}_1,\bar{x}_2,\cdots\bar{x}_n)$，并将输入到第一个 GCN 层。

3.4 触发词识别与分类层

从 GCN 模块中得到所有单词 token 的表示 D 之后，将得到的向量 $\boldsymbol{D}_i$ 送到一个全连接层中，以识别和预测采用 BIO 标注模式标注的触发词标签：

$$\bar{\boldsymbol{D}}_i=f(W_D\boldsymbol{D}_i+b_D) \tag{7}$$

最后，接上一个 softmax 层来完成最终的触发词分类预测：

$$y_{t_i}=\text{softmax}(W_t\bar{\boldsymbol{D}}_i+b_t) \tag{8}$$

其中，f 是一个非线性的激活函数，y_{t_i} 是第 i 个标记 token 的最终预测输出。softmax 表示 softmax 函数，常用于多分类问题，代表了某个元素被取到的概率，式(9)表示了 softmax 函数：

$$S_i=\frac{e^{V_i}}{\sum_j e^{V_j}} \tag{9}$$

其中，V 表示元素集合，在本模型中即是触发词类别的 BIO 标签集合；i 表示第 i 个元素，也就是当前标签；S_i 表示第 i 个元素的 softmax 值，就是该元素的指数除以所有元素的指数之和，代表当前标签所得的分数。

3.5 损失函数

为了训练 BGCN 网络，需要最小化负对数似然损失函数：

$$J(\theta)=-\sum_{p=1}^{N}(\sum_{i=1}^{n_p}I(y_{t_i})\log(p(y_{t_i}|\theta))) \tag{10}$$

其中，N 是训练语料库中的句子总数；n_p 和 t_p 分别是标记 token 的个数和提取出来的触发词候选个数；$I(y_{t_i})$ 是一个指示函数，如果 y_{t_i} 不是 BIO 标注模式中的 O，则输出一个比 1 大的正浮点数 α，否则输出 1。

4 实验

4.1 实验设置

实验中使用的数据集是 ACE2005 数据集，在其 English 数据集上评估 BGCN 模型。ACE2005 数据集的事件检测子任务 VDR 中标注了 8 个事件类型和 33 个事件子类型。根据 BIO 标注模式和 NONE 标签，将每个标记分类为事件检测中的 67 个类别。而且，为了与之前的研究进行比较，使用与先前工作相同的数据进行分割[3-4,13]，将 40 个新闻类的文本(共包括 881 个句子)作为测试集，30 个其他类型的文本(共包括 1 087 个句子)作为验证集，剩下的 529 个文本(共包括 21 090 个句子)作为训练集。

本模型使用 Stanford CoreNLP 工具包来预处理数据：1)分词，将句子中的句子分成一个个单词标记 token；2)词性标注，对分词分出来的每一个单词 w_i 进行词性标注；3)生成依存句法分析树，用 Stanford CoreNLP 工具包生成依存句法分析树。

本模型的单词表示模块的 BERT 词向量是使用 Google 官方的预训练模型 BERT-Base Cased，包含 12 层 transformer，隐藏层维度为 768 维，参数量为 1.1×10^9。

实验遵循已有研究工作的标准来判断触发词检测预测的正确性。对于所有实验，在单词表示模块中，BERT 词向量的维度为 768 维，单词 w_i 的 pos-embedding 维度为 50 维。在句法 GCN 图卷积模块中，使用一个 3 层的 GCN 网络，GCN 层的输出维度为 400 维，并使用一个单层的 Bi-LSTM 网络，它的隐层维度是 200 维。同时，本文设置 dropout 比率为 0.5。在实验中，batch 的大小为 128，并且本实验中设置了一个句子长度的最大值 $n=50$，比 n 短的句子用 padding 操作补上，比 n 长的句子则进行截断。

本文使用 ReLU 作为模型的非线性激活函数，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。

4.2 评价指标

为了评估 BGCN 模型在触发词检测方面的有效性，使用精确率(precision)、召回率(recall)和 F1(F-measure)作为评估指标，并将其分为触发词识别与触发词分类这两大任务来分别进行评估。精确率为模型预测为正类的样本中真正为正类的样本所占的比例，召回率为模型正确预测为正类的样本数量占总的正类样本数量的比值。F-Measure 就是 Precision 和 Recall 的加权调和平均：

$$F\text{-}Measure=\frac{(\alpha^2+1)Precision*Recall}{\alpha^2(Precision+Recall)} \tag{11}$$

当参数 α 为 1 时，也就是通常所使用的 $F1$，$F1$ 的表达式如下：

$$Precision=\frac{TP}{TP+FP} \tag{12}$$

$$Recall=\frac{TP}{TP+FN} \tag{13}$$

$$F1=\frac{2*Precision*Recall}{Precision+Recall} \tag{14}$$

其中，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。

4.3 实验结果

4.3.1 BGCN 模型提取句子级别特征的效果

本实验的目的在于验证 BGCN 在提取句子级别特征时的有效性。为了验证 BGCN 模型提取句子级别特征的效果，将其与 DMCNN，JMEE 以及 Chen 等提出的 DMCNN[8]中提到的两种基线模型 Embedding+T 和 CNN 进行比较。与已有研究工作一样，将测试数据根据句子中的事件数目分成两个部分分别进行评估。在数据中，只包含一

个事件的句子占整个数据集的 72.7%，包含至少两个事件的句子占 27.3%。

表 2 列出了几个模型在测试集上的性能(即 F1 分数)。Embedding+T 使用词汇级的特征和传统的基于人工设计的句子级特征，CNN 与 DMCNN 很相似，它们的区别在于 DMCNN 使用动态多池化层代替了传统的最大池化层。JMEE 是 Liu 等提出的结合图卷积网络和注意力机制的事件抽取模型。从表中可以看出，BGCN 模型基本优于其他方法。在单事件句子(1/1)的触发词抽取上，BGCN 相比两个基线模型 Embedding+T 和 CNN 分别提高了 9.4%和 5.0%；在多事件句子(1/N)的触发词抽取上，相比基线模型其性能提高非常显著。实验结果表明，基于表示的模型的性能优于基于特征的模型，相比 CNN 和 DMCNN，BGCN 模型可以有效捕捉更有价值的线索。在多事件句的实验上，由于 GCN 层结合了句法弧，缩短了句子中多个触发词之间的距离，因此能更有效地捕捉长距离的依赖，使得分类精度更准确。与 Liu 等于 2018 年提出的事件抽取模型 JMEE 相比，BGCN 在单事件句子(1/1)的表现上提升了 3.3%，在多事件句子(1/N)的表现上相差不多，而 JMEE 略优一些。JMEE 在触发词抽取层中加入了自注意力机制，在多事件句抽取上有着更好的效果。总体来说，BGCN 模型的效果最佳。本实验的结果表明，BGCN 模型在提取句子级别特征时效果显著。

表 2 BGCN 模型在单事件句(1/1)和多事件句(1/N)上的表现

Table 2 Performance of BGCN model on single event sentence (1/1) and multiple event sentence (1/N)

	Model	1/1	1/N	all
Event Recognization and Event Classification F1 Score	Embedding+T	68.1	25.5	59.8
	CNN	72.5	43.1	66.3
	DMCNN	74.3	50.9	69.1
	JMEE	75.2	**72.7**	73.7
	BGCN	**77.5**	72.4	**74.2**

4.3.2 不同词向量对 BGCN 模型分类效果的影响

触发词抽取模型都需要在 Embedding 层组合各种特征来编码词汇及句子。如 DMCNN 使用候选触发词及其相邻词的词向量作为词汇级别特征；JRNN 使用单词词向量、实体类型向量和二进制向量 3 个向量进行级联，将单词转换为定长的编码输入等。BGCN 模型的句子编码层使用了 BERT 词向量结合词性 POS 标记嵌入来编码单词的特征输入。本实验比较不同的词向量编码对模型最终效果的影响，分别使用 word2vec，GloVe 和 BERT 词向量对原始输入句子进行编码，并将结果送入后续模块以完成分类任务，使用的 word2vec 词向量和 GloVe 词向量的维度均为 300 维。不同词向量对模型性能的影响如表 3 所列。

表 3 不同词向量对模型性能的影响

Table 3 Influence of different word vectors on the performance of the model

Word Vector	Trigger Recognization/%			Trigger Classification/%		
	P	R	$F1$	P	R	$F1$
word2vec	75.6	69.4	72.3	73.4	68.6	71.0
GloVe	77.2	71.1	74.9	74.5	71.3	72.8
BERT	**81.4**	**74.1**	**77.6**	**75.9**	**72.5**	**74.2**

从表 3 中的数据可以看出，BERT 词向量所表现出来的性能优于传统的 word2vec 和 GloVe。同为根据共现信息编码词汇的静态词向量，Word2vec 和 GloVe 词向量在触发词识别和分类任务上的表现差别不大，而 BGCN 模型所使用的 BERT 表现出来的效果在各方面均优于 word2vec 和 GloVe。在触发词识别任务上，本文方法的 $F1$ 分数比 word2vec 高 5.3%，在触发词分类任务上，本文方法的 $F1$ 分数比 word2vec 高 3.2%。相比 GloVe，本文方法在两个任务上的 $F1$ 分数也分别高了 2.7%和 1.4%。这得益于 BERT 大规模的训练和优秀的双向表征能力。BERT 模型进一步增强了词向量模型的泛化能力。本实验表明，采用 BERT 词向量编码句子信息能更好地提升模型的性能。

4.3.3 模型整体表现

将 BGCN 模型与以下最先进的方法进行对比，比较各方法在触发词识别与触发词分类任务上的表现，如表 4 所列。

表 4 BGCN 模型与最先进的方法的整体性能比较

Table 4 Overall performance of the BGCN model compared with the most advanced methods

Model	Trigger Recognization/%			Trigger Classification/%		
	P	R	$F1$	P	R	$F1$
Cross-Event	—	N/A	—	68.7	68.9	68.8
Cross-Entity	—	N/A	—	72.9	64.3	68.3
DMCNN	80.4	67.7	70.4	75.6	63.6	69.1
JRNN	68.5	75.7	71.9	66.0	73.0	69.3
DLRNN	—	N/A	—	77.2	64.9	70.5
ANN-S2	—	N/A	—	78.0	66.3	71.7
GCN-ED	—	N/A	—	77.9	68.8	73.1
GMLATT	80.9	68.1	74.1	78.9	66.9	72.4
dbRNN	—	N/A	—	74.1	69.8	71.9
JMEE	80.2	72.1	75.9	76.3	71.3	73.7
TBNNAM	—	N/A	—	76.2	64.5	69.9
BGCN model	**81.4**	**74.1**	**77.6**	75.9	**72.5**	**74.2**

(1)Cross-Event：是基于特征的方法，利用了同一文档中同一类型事件的一致性信息和同一文档中不同事件类型的共现信息。

(2)Cross-Entity：基于特征的方法，它将实体共现作为预测事件提及的关键特征。

(3)DMCNN：由 Chen 于 2015 年提出，改进了传统的最大池化，用动态多池化网络来捕获更多有价值的信息。

(4)JRNN：由 Nguyen 于 2016 年提出，使用双向 RNN 来进行事件抽取。

(5)DLRNN[27]：是 Duan 等于 2017 年提出的一个文档级事件检测模型，自动学习句子以外的特征。

(6)ANN-S2[22]：由 Liu 等于 2017 年提出，通过有监督的注意机制来显式地利用 ED 的参数信息。

(7)GCN-ED[12]：是 Nguyen 于 2018 年提出的事件检测模型，集成了语法信息，并且提出了一种新的基于实体提及的池化机制。

(8)GMLATT[28]：是 Liu 等提出的一种多语言注意力机制事件检测框架，利用多语言数据中的一致性信息来缓解数据稀疏问题，用跨语言注意力机制来解决单语言歧义问题。

(9)dbRNN：由 Lei 等于 2018 年提出，是在循环神经网络上构建模型并利用依存关系训练每个词的语法相关信息。

(10)JMEE：由 Liu 等于 2018 年提出，通过句法弧来代替句子级别的顺序建模，能够捕捉更长距离的依赖。

(11)TBNNAM[18]：Liu 等 2019 年提出的一种不需要识别触发词的事件抽取方法，用注意机制的类型感知偏向神经网络来对整个句子直接进行事件多分类。

从表 4 的实验对比结果可以看出，在 ACE2005 数据集上，BGCN 模型在对比模型中有着最高的 F1 分数。基于特征的模型 Cross-Event 和 Cross-Entity 的效果弱于基于表示的模型 DLRNN 等。这表明人工设计的特征不足以进行事件检测，而基于神经网络的特征自动提取可以捕捉到更丰富的语义线索。而 BGCN 的性能优于其他基于表示的模型，相比现有的神经网络模型，BGCN 有着更高的性能。相比通过 RNN 利用文档信息的 DLRNN 和利用论元信息的 ANN-S2，BGCN 在触发词分类任务上表现出了优越性，F1 分数分别提高了 3.7%和 1.1%，这得益于其能够充分捕获上下文信息的大规模预训练 BERT 向量。与同样利用了 GCN 网络的 JMEE 和 GCN-ED 模型相比，BGCN 在触发词分类任务上的 F1 分数分别高出了 1.1%和 0.5%，JMEE 的分类效果也优于 GCN-ED，这说明通过句法弧结合 GCN 的方法无论是在捕捉长距离依赖上还是在句子表示上都有很好的效果。BGCN 使用的 BERT 向量的编码能力也优于 JMEE 的 GloVe。这些结果表明了 BERT 和 LSTM 结合图卷积网络和句法分析树来完成触发词检测任务的有效性。

结束语 基于 BERT 和 GCN 图卷积网络，提出了一种新的神经网络模型 BGCN，用于事件提取中的触发词检测任务。在 BGCN 模型中，采用最先进的预训练词向量模型 BERT 来为该网络提供词嵌入向量，并引入句法图卷积网络来增强向量的信息表示，之后通过全连接的分类网络来进行触发词的检测。实验结果表明了 BGCN 模型在相应任务上的有效性。

参 考 文 献

[1] GRISHMAN R, WESTBROOK D, MEYERS A. NYU's English ACE 2005 system description [J/OL]. ACE, 2005, 5. http://www.researchgate.net/publication/228638184_NYU's_English_ACE_2005_system_description.

[2] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781, 2013.

[3] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[4] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv: 1802.05365, 2018.

[5] LIU S, LIU K, HE S, et al. A probabilistic soft logic based approach to exploiting latent and global information in event classification[C]// Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[6] LI X, NGUYEN T H, CAO K, et al. Improving event detection with abstract meaning representation[C]// Proceedings of the First Workshop on Computing News Storylines. 2015: 11-15.

[7] 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. 2016: 300-309.

[8] CHEN Y, XU L, LIU K, et al. Event extraction via dynamic multi-pooling 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 (Volume 1: Long Papers). 2015: 167-176.

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

[10] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv: 1609.02907, 2016.

[11] MARCHEGGIANI D, TITOV I. Encoding sentences with graph convolutional networks for semantic role labeling[J]. arXiv: 1703.04826, 2017.

[12] NGUYEN T H, GRISHMAN R. Graph convolutional networks with argument-aware pooling for event detection[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[13] LIU X, LUO Z, HUANG H. Jointly multiple events extraction via attention-based graph information aggregation[J]. arXiv: 1809.09078, 2018.

[14] 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 (Volume 1: Long Papers). 2013: 73-82.

[15] ZHANG X F, GUO Z G, LIU S, et al. Self-similarity Clustering Event Detection Based on Triggers Guid [J]. Computer Science, 2010, 27(3): 212-214.

[16] XU X, LI P F, ZHU Q M. Pattern Filtering and Conversion Methods for Semi-supervised Chinese Event Extraction. [J]. Computer Science, 2015, 42(2): 253-255.

[17] LIU S, CHENG R, YU X, et al. Exploiting contextual information via dynamic memory network for event detection[J]. arXiv: 1810.03449, 2018.

[18] LIU S, LI Y, ZHANG F, et al. Event Detection without Triggers [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 735-744.

[19] ZHANG J, QIN Y, ZHANG Y, et al. Extracting Entities and Events as a Single Task Using a Transition-Based Neural Model [C]// IJCAI. 2019: 5422-5428.

[20] ORR J W, TADEPALLI P, FERN X. Event detection with neural networks: A rigorous empirical evaluation[J]. arXiv: 1808.08504, 2018.

[21] SHA L, QIAN F, CHANG B, et al. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[22] LIU S, CHEN Y, LIU K, et al. Exploiting argument information to improve event detection via supervised attention mechanisms [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1789-1798.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[24] YAN H R, JIN X L, MENG X B, et al. Event Detection with Multi-Order Graph Convolution and Aggregated Attention [C]// The 9th International Joint Conference on Natural Language Processing. 2019: 5770-5774.

[25] CUI S Y, YU B W, LIU T W, et al. Event Detection with Relation-Aware Graph Convolutional Networks [J]. arXiv: 2002.10757, 2020.

[26] PENG H, LI J, GONG Q, et al. Fine-grained event categorization with heterogeneous graph convolutional networks [J]. arXiv: 1906.04580, 2019.

[27] DUAN S, HE R, ZHAO W. Exploiting document level information to improve event detection via recurrent neural networks [C]// Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2017: 352-361.

[28] LIU J, CHEN Y, LIU K, et al. Event detection via gated multilingual attention mechanism [C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

CHENG Si-wei, born in 1997, postgraduate. His main research interests include natural language processing and machine learning.

XU Jian, born in 1979, Ph. D, professor, master director. His main research interests include data mining and knowledge graph.