**文章编号**：1003-0077(2021)10-0073-08

# 基于图注意力卷积神经网络的文档级关系抽取

吴 婷，孔 芳

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

**摘 要**：关系抽取作为信息抽取的子任务，旨在从非结构化文本中抽取出便于处理的结构化知识，对于自动问答、知识图谱构建等下游任务至关重要。该文在文档级的关系抽取语料上开展工作，包括但不局限于传统的句子级关系抽取。为了解决文档级关系抽取中长距离依赖问题，并且对特征贡献度加以区分，该文将图卷积模型和多头注意力机制相融合构建了图注意力卷积模型。该模型通过多头注意力机制为同指、句法等信息构建的拓扑图构建动态拓扑图，然后使用图卷积模型和动态图捕获实体间的全局和局部依赖信息。该文分别在 DocRED 语料和自主扩展的 ACE 2005 语料上进行实验，与基准模型相比，基准模型上融入图注意力卷积的模型在两个数据集上的 $F_1$ 值分别提升了 2.03%和 3.93%，实验结果表明了该方法的有效性。

**关键词**：文档级关系抽取；图卷积网络；图注意力

**中图分类号**：TP391　　**文献标识码**：A

# Document-Level Relation Extraction Based on Graph Attention Convolutional Neural Network

WU Ting, KONG Fang

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

**Abstract**: As a subtask of information extraction, relation extraction aims to extract the structured knowledge from unstructured text, which is very important for the downstream tasks such as automatic question answering and knowledge graph construction. Focused on document-level relation extraction, this paper proposed a graph attention convolution model to deal with long-distance dependence issue. The model uses a multi-head attention mechanism to construct a dynamic topological graph for coreference, syntax and other information. Then it uses the graph convolution model and dynamic graph to capture global and local dependency information between entities. Experiments on the DocRED corpus and the self-expanding ACE 2005 corpus comfirm improvements on $F_1$ values by 2.03 and 3.93, respectively.

**Keywords**: document level relation extraction; graph convolution network; graph attention

## 0 引言

伴随信息时代的高速发展，互联网给人们带来便利的同时也产生数以万计的数据，并呈现指数增长的趋势，给数据的存储和处理造成困难。为了应对信息爆炸带来的挑战，迫切需要一种自动内容抽取的工具帮助人们快速从海量数据中挖掘出感兴趣的信息。在这种背景下，信息抽取(Information Extraction, IE)应运而生[1]。信息抽取研究将非结构化文本转化为便于机器和程序理解的结构化和半结构化信息，并以数据库的形式进行存储，以提高用户的查询效率，也可以为其他自然语言处理任务提供服务。

信息抽取研究从自然语言文本中抽取特定类型的事件和事实信息，通常把特定的事实信息称为实体(Entity)，如组织机构(ORG)、人物(PER)等。实体关系抽取的目标是根据给定的包含实体 $e_1$ 和 $e_2$ 的自然语言文本，识别出 $e_1$ 和 $e_2$ 之间的关系类型 $r$。实体关系抽取作为信息抽取的一项重要子任务，可应用于自动问答[2]、机器翻译[3]、知识图谱[4]等领域，受到了国内外专家学者的广泛关注。

**收稿日期**：2020-02-20　　**定稿日期**：2020-04-03

目前实体关系抽取的相关研究多集中在句子级别[5-9]，即只关注句内两个实体之间的关系，对跨句子的情况关注相对较少。而根据自然语言的表达习惯，实体对分别位于不同句子的情况也十分常见。早在2010年，Swampillai[10-11]等人统计了MUC和ACE 2003语料中跨句子关系的分布情况，分别对应28.5%和9.4%，并于2011年基于SVM模型完成了初步尝试。近几年得益于深度学习的发展，Peng[12]等人于2017年提出了基于graph LSTM的跨句子关系抽取框架，并在生物领域的数据集上验证了该方法的有效性。此后，在文献[12]基础上的一些改进工作相继展开，跨句子实体关系抽取的问题再次进入研究者视野。

目前的跨句子关系抽取模型多在文献[12]的基础上进行改进，主要存在两个问题：①跨句子的语料本身序列较长(DocRED中平均198个词)，LSTM在处理长距离依赖上存在局限性，尤其是在长序列中进行信息传递时容易造成信息丢失；②全部采用生物领域的数据集，由于生物领域的特殊性和不同领域之间的差异性，生物领域的研究虽对其他领域具有借鉴意义，但仍然缺乏通用领域的相关尝试。

针对问题①，本文采用一个融入了上下文信息的上下文图卷积(Context Graph Convolutional Network，C-GCN)模型解决长距离依赖不足以及信息丢失的问题；同时，为了对不同依赖特征加以区分，提出多头图注意力卷积模型(Multi-head Attention Graph Convolutional Network，Multi-GCN)进行动态剪枝优化。针对问题②，我们分别在新闻领域DocACE(作者借助同指信息在新闻领域的ACE 2005数据集中构建了跨句子关系数据集)和通用领域的DocRED数据集[13]上进行实验，结果表明了本文方法的有效性。

本文的主要工作包括：首先针对目前文档级关系抽取任务存在的长距离依赖不足、不能较好地利用同指、句法信息等问题，通过构建图注意力卷积模型提高了关系抽取的性能；然后针对目前跨句子关系抽取任务集中在生物领域的应用现状，利用同指信息与相应的筛选策略，对ACE 2005数据集中的跨句子关系进行补充，从一定程度上填补了新闻领域语料不足的空白。

## 1 相关工作

跨句子关系抽取的工作可以追溯到2010年，Swampillai[10]对MUC和ACE 2003两个数据集进行统计，其中跨句子关系的分布对应分别为28.5%和9.4%，如果不进行跨句子关系抽取的工作，在MUC数据集上最高仅能做到71.5%。在2011年，Swampillai[11]尝试用SVM模型进行跨句子关系抽取，并提出了其相对句内关系抽取所面临的挑战，如数据稀疏、句法分析树不能直接利用等问题。随着远程监督在实体关系抽取任务中的有效应用，Quirk[14]等人于2017年借助远程监督生成了生物领域的跨句子关系抽取数据集，为后续的一系列研究奠定了基础。同年，Peng[12]等人提出graph LSTM模型在上述生物语料上进行跨句子关系抽取，核心是借助依存句法分析将文档表示成文档图(document graph)，为了简化和避免形成环，他们把一篇文档表示成前向和后向的两个图。考虑到把依存树进行拆分会造成信息损失，Song[15]等人于2018年在Peng[12]基础上编码拆分前的图结构，实验证明，直接拆分会对性能造成不利影响。Song[15]等人在完整的图结构基础上进一步改进，提出Graph State LSTM模型，将模型从二维空间扩展到三维空间，实现了词与同一时刻的邻居的、不同时刻的自身的信息交换。

Gupta[16]等人于2019年提出iDepNN模型，分别基于最短依存路径(SDP)和子树的增广依存路径(ADP)两类特征进行建模。与前面的工作不同，该工作采用新闻领域的MUC6数据集和生物领域的BioNLP ST 2016数据集进行实验，并对MUC6数据集中的跨句子关系进行了标注。

Verge[17]等人的工作是目前唯一不用图模型的工作，通过引入改进的Transformer模型来解决长序列的问题，并采用多示例学习对数据进行降噪。在文献[17]工作的基础上，Sahu[18]等人通过用GCN替换Transformer模型，进一步解决了依赖捕获不足的问题，两份工作均在生物领域的CDR、CHR数据集上进行。

通过前面的分析可以看出，跨句子关系抽取的研究几乎都集中在生物领域，因此，我们尝试利用同指信息对新闻领域的ACE 2005数据集进行跨句子关系的扩充，构建了DocACE数据集。构造数据集的想法与Yao[13]等人的工作不谋而合，他们于2019年发布了DocRED数据集，该数据集与ACE 2005的标注比较相似，但是在领域覆盖上较前者更丰富，同时提供了标注和远程监督两个版本的数据。无论从领域迁移还是远程监督降噪的角度进行考虑，该

数据集都为相关研究的开展提供了可能。因此，本文的重点工作也将在该数据集上进行。

## 2　模型

### 2.1　基准模型：基于 BiLSTM 的关系抽取模型

为了支持跨句子关系抽取的研究，Yao[13] 等人于 2019 年发布了 DocRED 数据集，同时给出了关系抽取的几个常用模型在该数据集上的表现。根据论文以及线下测试的结果，本文选取 BiLSTM 模型作为基准模型。

关系抽取的通常做法是将实体关系抽取任务看作是分类问题，基准模型仍然沿用这个思路。给定一篇文档 $d$，$[w_1,w_2,\cdots,w_n]$为文档 $d$ 中的第 1，2，…，$n$ 个词，$e_1$和 $e_2$是 $d$ 中的两个实体。跨句子关系抽取的模型以($e_1$，$e_2$，$d$)作为输入，并且返回 $e_1$与 $e_2$之间的关系。

基准模型包括输入层、编码层和分类三部分，下面分别进行介绍。

(1) **输入层**：采用词向量、同指信息、实体类型和实体 $i$ 与 $j$ 之间的距离信息作为基准特征，分别记为$w^w$、$w^c$、$w^t$和$d_{ij}$，将向量进行拼接形成初始词特征 $w$，即$w_i=[w_i^w;w_i^c;w_i^t]$。

(2) **编码层**：采用 BiLSTM 模型对向量化的词序列$w_1,w_2,\cdots,w_n$进行编码，将前向 LSTM 与后向 LSTM 的隐层向量拼接，进而得到融入上下文信息的序列$h_1,h_2,\cdots,h_n$。针对某一时刻 $t$，隐藏层节点 $h_t$的计算与更新如式(1)～式(3)所示。

$$\begin{bmatrix} f_t \\ i_t \\ o_t \\ g_t \end{bmatrix}=\begin{bmatrix} \delta \\ \delta \\ \delta \\ \tanh \end{bmatrix}\left(\begin{bmatrix} W_f \\ W_i \\ W_o \\ W_g \end{bmatrix}\begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}+\begin{bmatrix} b_f \\ b_i \\ b_o \\ b_g \end{bmatrix}\right) \tag{1}$$

$$c_t=f_t * c_{t-1}+i_t * g_t \tag{2}$$

$$h_t=o_t\tanh(c_t) \tag{3}$$

其中，$f_t$、$i_t$、$o_t$分别对应遗忘门、输入门与输出门，$g_t$为单元新值张量，$x_t$为 $t$ 时刻的输入。$c_t$、$h_t$为状态张量、隐层输出。$W_{(\cdot)}$与$b_{(\cdot)}$分别为权重矩阵和偏置项。

(3) **分类**：从编码得到的文本表征中提取出实体表征$e_i$和$e_j$，与距离特征$d_{ij}$、$d_{ji}$拼接后进行双线性变换。最后，经过 sigmoid 函数算出每种类别的概率，从中选出概率最大的作为实体对间的关系，如式(4)～式(7)所示。

$$\ddot{e}_i=[e_i;d_{ij}] \tag{4}$$

$$\ddot{e}_j=[e_j;d_{ji}] \tag{5}$$

$$P(r\mid\ddot{e}_i,\ddot{e}_j)=\text{sigmoid}(\ddot{e}_i W_r \ddot{e}_j+b_r) \tag{6}$$

$$\hat{y_{ij}}=\text{argmax}(P(r\mid\ddot{e}_i,\ddot{e}_j)) \tag{7}$$

### 2.2　基于依赖图的 BiLSTM-GCN(C-GCN)模型

与句内关系抽取相比，跨句子关系抽取面临更多的挑战：首先是序列长度较前者有明显的变化，而 BiLSTM 在捕获长距离依赖方面具有局限性；其次，Peng[12] 等人在进行跨句子关系抽取时利用了 30 多种特征，可见该任务需要考虑更全面的信息，即充分利用句内依赖与句间依赖；最后，文献[19]证明了句法信息在句子级关系抽取任务中的有效性，基准模型缺乏对句法信息的考虑。针对上面提到的问题，本文在 BiLSTM 获取上下文信息的基础上，引入图卷积(Graph Convolutional Network，GCN)模型来加入句法、同指等特征，便于捕获局部和全局依赖信息，如图 1 所示。

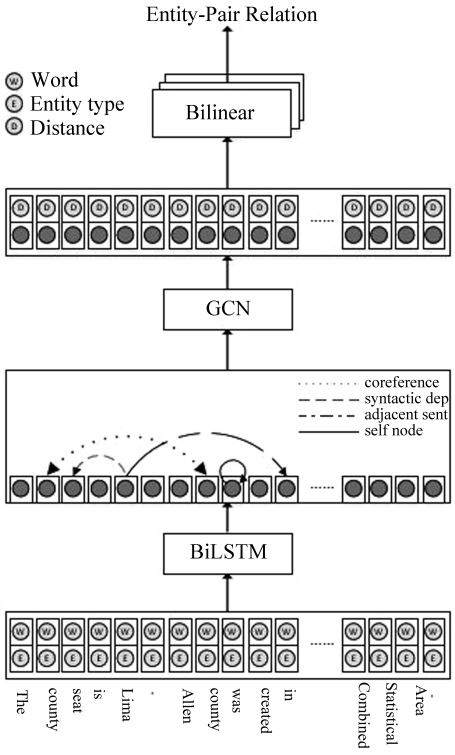


图 1　基于依赖图的 BiLSTM-GCN 模型

与基准模型相比，本文并非在输入层简单地编码同指特征，而是利用图卷积对具有同指关系的单词编码进行迭代更新。首先把一篇文档表示为图 $G(V,E)$，其中，$V$ 表示顶点集合，$E$ 表示边集合。本文中顶点对应文档中的单词，边对应不同词之间的关系依赖，如依存关系、同指信息、相邻边等。为

了避免引入过多特征产生过拟合的问题，本文借鉴Vashishth[20]等人的工作，采用一种简化的GCN模型，下面对该部分进行介绍。

2.2.1 建图

为了将文档转化为图表示，本文以词为顶点选取如下4种依赖特征，对应图1中GCN的输入部分的不同类型的边。

（1）**依存关系边**：作为语法特征，依存关系在关系抽取任务中得到了广泛应用[19]。为了丰富句内信息，通过依存关系获取句内局部依赖。借鉴Vashishth[20]的工作，为了简化模型，我们不区分依存类型，只区分方向。

（2）**同指依赖边**：作为篇章级任务，同指可以有效捕获局部依赖和全局依赖。为了缩短词之间的距离，减少信息远距离传输中的损失，在图中引入同指依赖边。

（3）**相邻边**：跨句子关系抽取要考虑不同句子的实体对之间的关系，为了对同指覆盖不到的全局依赖进行补充，本文借助虚根对相邻句的依存句法树的根（root）节点进行桥接，缩短相邻句子间实体的距离。

（4）**自反边**：在GCN模型中，每个节点可以学到其邻居的信息，为了防止丢失节点自身携带的信息，为每个节点添加一个指向自身的自反边。

2.2.2 GCN层

GCN是运行在图结构上的卷积神经网络，通过接收邻域信息扩大感受野，下面对多层GCN的工作流程进行介绍。给定由$n$个节点构成的图，可以得到邻接张量$A \in R^{n\times n\times k}$，其中$k$为依赖边的数量。根据2.2.1中的特征，当节点$i$与$j$之间存在由$i$到$j$的第$k$种依赖边时，$A_{ij}^{k}=1$，否则$A_{ij}^{k}=0$，$A_{ji}^{k}$同理。经过GCN后，节点$i$携带的信息形式化如式(8)所示。

$$h_i^{(l)}=f\left(\sum_{j=1}^{n}A_{ij}\,W^{(l)}\,h_j^{(l-1)}+b^{(l)}\right) \tag{8}$$

其中，$h_j^{(l-1)}$是第$j$个词经过$l-1$层GCN之后的表示，$W^{(l)}$是权重矩阵，$A_{ij}$为邻接张量，$b^{(l)}$是偏移量，$f$为激活函数，如ReLU。$h_j^{(0)}$初始化为BiLSTM的输出。

## 2.3 基于图注意力卷积模型的文档级关系抽取

在2.2节的BiLSTM-GCN模型中，采用5种特征（依存关系具有方向性）进行建图，充分利用了局部和非局部的依赖。但是，这种建图方式对不同类型的依赖（如同指、相邻句）给予同等的关注，而根据直观感受，同指相比相邻句会更重要一点。由于不同依赖对关系抽取的贡献程度不同，本文提出一种基于图注意力卷积模型（BiLSTM-Multi-GCN）的动态筛选策略进行特征优选。

注意力机制（Attention）可以满足对不同类型区别对待的需求，而多头自注意力机制（Multi-head Attention）可以将模型划分为多个子空间，帮助模型关注不同方面的信息。因此，为了对类型特征加以区分，同时考虑多层次信息，本文将GCN与Multi-head Attention进行结合称为图注意力卷积模型（Multi-GCN），用以替换图1模型中的GCN部分，下面只对修改部分进行展开，其他模块同2.1节，此处不再赘述。

2.3.1 Multi-GCN层

出于对不同类型特征的关注不同，本文提出一种基于图注意力卷积模型的筛选策略对特征进行优选，下面给出Multi-GCN的细节图，如图2所示。

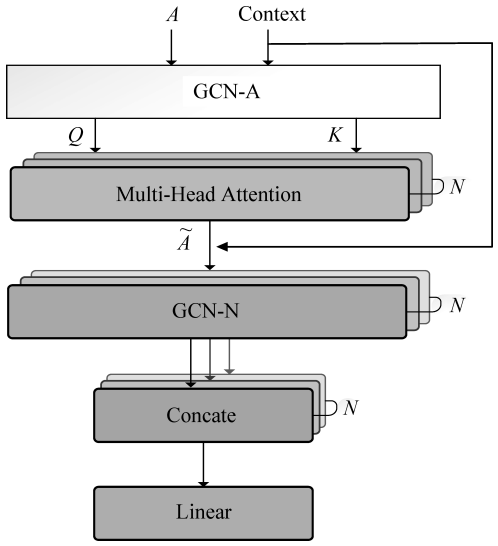


图2 Multi-GCN细节图

Multi-GCN包括GCN-A、Multi-Head Attention和GCN-N三部分。首先，将BiLSTM的输出Context$\in \boldsymbol{R}^{n\times \text{hidden}}$和初始的邻接张量$A \in R^{n\times n\times k}$作为输入，通过GCN-A获取到融入上下文的初始图信息$O_{\text{GCN-A}} \in \boldsymbol{R}^{n\times k\times \text{hidden}}$。其中，GCN-A的计算公式与式(8)相同，与GCN的不同点在于：GCN中的$W \in \boldsymbol{R}^{(k\times \text{hidden})\times \text{hidden}}$，GCN-A中的$W \in \boldsymbol{R}^{\text{hidden}\times \text{hidden}}$，所以GCN的输出为$O_{\text{GCN}} \in \boldsymbol{R}^{n\times \text{hidden}}$，而GCN-A的输出为$O_{\text{GCN-A}} \in \boldsymbol{R}^{n\times k\times \text{hidden}}$。然后，将$O_{\text{GCN-A}}$分别经过两个线性函数得到$Q$和$K$，最后通过Multi-Head Attention得到$N$（$N$为head的数目）个不同的动态优

选后的邻接张量 $\widetilde{A}\in\boldsymbol{R}^{N\times n\times n\times k}$，计算如式(9)所示，这里仅使用 $Q$、$K$ 计算更新的邻接张量。

$$\widetilde{A}^{z}=\mathrm{softmax}\left(\frac{QW^{Q}\ (KW^{K})^{\mathrm{T}}}{\sqrt{d}}\right) \tag{9}$$

其中，$\widetilde{A}^{z}$ 为第 $z$ 个 head 对应的邻接张量，$W^{Q}$、$W^{K}$ 为模型参数。

在每个 head 内部，对邻接张量的最后一维(依赖关系对应的维度)进行注意力计算，对不同的关系类型分配不同的权重。假设节点 $i$ 与节点 $j$ 之间的关系矩阵为[1 1 0 0 0]，经过注意力计算后关系矩阵变为[0.5 0.3 0.06 0.1 0.04]。这样会存在一个问题，节点 $i$ 与 $j$ 之间原本没有第 3、4、5 种关系，但是经过注意力计算后，节点间存在了上述 3 种关系。因此，我们使用一个掩码矩阵，将最终的关系矩阵变为[0.5 0.3 0.0 0.0 0.0]，新的邻接图是对 $k$ 种依赖边的贡献度调和的结果。

在模型的 GCN-N 部分以更新后的邻接张量 $\widetilde{A}$ 和 Context 为输入，学习融入权重的邻域信息，最终得到 $N$ 个隐层向量表示。其中，GCN-N 为 $N$ 个 head 对应的图卷积。将隐层向量在最后一维进行拼接，记为 $h_{m\_h}$。经线性变换得到最终的输出，线性变换定义如式(10)所示。

$$h_{m}=W_{m}\,h_{m\_h}+b_{m} \tag{10}$$

其中，$h_{m}$ 为经过 Multi-GCN 层之后的隐层输出，$W_{m}$、$b_{m}$ 为模型参数。

### 2.4　BERT-GCN、BERT-Multi-GCN 模型

众所周知，关系抽取是偏向语义层的任务，即在不理解语义的基础上很难有效地解决问题。预训练的 BERT 模型从大规模的无监督语料中学到了许多先验知识，比如语言本身的逻辑规律等。BERT 编码了丰富的语言学层次信息：底层网络关注浅层特征，中层网络倾向于句法信息，语义特征集中在高层网络[21]。另一方面，随着序列长度的增加，BiLSTM 长距离依赖捕获不足的问题更加严重。因此，我们尝试用 BERT 代替 BiLSTM 进行编码，并在此基础上引入 2.2 节和 2.3 节的 GCN 与 Multi-GCN 模型，设置对比实验。其中，GCN 与 Multi-GCN 的细节见 2.2 节和 2.3 节，此处不再赘述。

## 3　实验配置

### 3.1　语料的划分与预处理

实验选用的语料包括两个：①清华大学发布的 DocRED 语料；②我们借助同指扩展 ACE 2005 语料得到的 DocACE 语料。两个语料都涵盖跨句的情况，为文档级关系抽取任务的开展提供数据支撑。下面对两个语料进行介绍。

(1) 针对 DocRED 语料，本文目前只考虑有监督标注的数据部分，暂时未涉及远程监督部分。DocRED 共计标注 5 053 篇维基百科文档，132 392 个实体和 63 443 个实体关系。本文采用与基准模型相同的实验设置，将数据集划分为训练集 3 053 篇，验证集和测试集各 1 000 篇，如表 1 所示。

**表 1　DocRED 数据集统计**

| 统计 | 训练集 | 开发集 | 测试集 | 总数 |
|---|---|---|---|---|
| 文章数目 | 3 053 | 1 000 | 1 000 | 5 053 |
| 实体数目 | 79 481 | 26 207 | 26 704 | 132 392 |
| 关系数目 | 38 269 | 12 332 | 12 842 | 63 443 |
| 关系类型 | 96 种 | | | |

(2) 在 ACE 2005 中，英文语料共计 599 篇，涵盖 6 种不同形式的新闻题材(广播、新闻、广播对话等)。对于 ACE 语料、普遍沿用 Li[22] 2014 年的数据集划分方法，删除 CTS (Conversational Telephone Speech) 和 UN (Usenet Newsgroups/Discussion Forum)两种形式的语料(篇数太少，共 88 篇)，将剩余的 511 篇语料划分为训练集(351)、验证集(80)、测试集(80)，如表 2 所示。

**表 2　ACE 2005 数据集统计**

| 统计 | 训练集 | 开发集 | 测试集 | 总数 |
|---|---|---|---|---|
| 文章数目 | 351 | 80 | 80 | 511 |
| 实体数目 | 14 757 | 3 512 | 3 075 | 21 344 |
| 关系数目 | 6 114 | 1 466 | 1 439 | 9 019 |
| 关系类型 | 6 种 | | | |

### 3.2　性能评价方法

对于关系抽取问题，本文采用 $F_1$ 值作为最终的评价指标，相关定义如式(11)所示。

$$F_{1}=\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{11}$$

其中，准确率(Precision)和召回率(Recall)的定义为：

$$\mathrm{Precision}=\frac{\text{某类被正确预测的实例数}}{\text{预测的某类实例总数}} \tag{12}$$

$$\mathrm{Recall}=\frac{\text{某类被正确预测的实例数}}{\text{正确的某类实例总数}} \tag{13}$$

### 3.3 模型训练

模型训练以整个文档为单位，通过预训练的词向量将输入转换为低维稠密表示，首先基于BiLSTM或者BERT模型获取包含语境的序列信息，然后将该序列表示送入GCN或者Multi-GCN中融入图结构，完成邻域信息交换。最后，将包含上下文信息与邻域信息的实体表征与距离向量进行拼接，送入双线性变换，经sigmoid函数得到预测概率分布。其中，词向量维度为100，BiLSTM隐层输出维度为128，Multi-GCN中head个数为2。模型采用交叉熵作为损失函数，定义如式(14)所示。

$$L=-\frac{1}{|T|}\sum_{(\vec{e}_i,\vec{e}_j)\in T} y_{ij}\log P(r\mid \vec{e}_i,\vec{e}_j) \quad (14)$$

其中，$T$是实体对集合，$y_{ij}$为实体$\vec{e}_i$与$\vec{e}_j$之间标注的关系，$P(r|\vec{e}_i,\vec{e}_j)$为模型判定的关系。采用Adam优化器优化模型参数，学习率设置为0.000 01并随训练次数的变化而动态调整。

## 4 实验结果与分析

### 4.1 基于依赖图BiLSTM-GCN模型的性能

我们首先分析在所有依赖特征(详见2.2.1节)建图基础上，GCN层数对任务性能的影响。然后，通过消融实验，对不同特征进行分析，找出文档级关系抽取任务的有效特征组合。下面给出表格中相关符号的说明：

(1) **特征**：采用词向量、同指信息、实体类型和实体间的距离信息作为基准特征，记为Ⅰ；建图特征包括依存信息、同指依赖、相邻句和自反边，记为Ⅱ，详见2.2.1节；基准特征去掉同指信息记为Ⅲ；

(2) **模型**：① BiLSTM；② BiLSTM-GCN；表3给出了BiLSTM-GCN与BiLSTM基准模型的结果，通过实验1与实验2的对比，验证了引入图信息的有效性。通过设置2、3、4、5的对比实验，说明GCN层数对模型效果是有影响的，在DocRED语料中3层效果最好，而DocACE中1层效果最好，这种现象是由语料的序列长度存在差异造成的。在DocRED语料中，序列的平均长度为198，而DocACE中的平均长度为66，在序列较短时，GCN层数的增加，会导致节点学到的信息冗余，不同节点携带的信息基本一致，不利于描述节点的特异性。

**表3 BiLSTM-GCN与BiLSTM模型的结果**

| 序号 | 模型 | 特征、GCN层数 | DocRED $F_1$/% | DocACE $F_1$/% |
|---|---|---|---|---|
| 1 | ① | Ⅰ、0层 | 50.35 | 40.06 |
| 2 | ② | Ⅲ+Ⅱ、1层 | 51.27 | 43.79 |
| 3 | ③ | Ⅲ+Ⅱ、2层 | 46.05 | 42.22 |
| 4 | ② | Ⅲ+Ⅱ、3层 | 51.73 | 41.46 |
| 5 | ② | Ⅲ+Ⅱ、4层 | 50.96 | 42.42 |

表4给出BiLSTM-GCN模型的消融实验的结果，通过采用不同特征进行建图，验证了不同特征对跨句子关系抽取任务的重要性。其中，建图特征表示在建图时利用2.2.1节提到的4种特征。

**表4 消融实验**

| 序号 | 特征 | DocRED $F_1$/% | DocACE $F_1$/% |
|---|---|---|---|
| 6 | 建图特征 | 51.27 | **43.79** |
| 7 | — 依存边 | **51.40** | 40.87 |
| 8 | —同指边 | 50.45 | 40.46 |
| 9 | — 相邻句 | 50.98 | 41.88 |

在句子级关系抽取中，类似于句法等局部依赖在任务中扮演重要的角色[19]。根据表4中的数据，在去掉同指依赖(8)和相邻句依赖(9)时，性能具有明显的下降趋势，表明在文档级的关系抽取任务中，同指等全局依赖的重要性逐渐凸显，在一定程度上反映了文档级与句子级的关系抽取任务之间的差别。

在两个语料中同指和相邻句的实验结果比较一致，而依存边则存在较大的差异。为了对此进行分析，我们统计了两个语料中的跨句子情况分布以及依存深度，DocRED中跨句子占比为77.57%，DocACE中跨句子占比为48.65%，语料分布差别较大。在依存深度方面，两个语料中均在深度为2处达到峰值，DocRED的平均深度为1.660，DocACE的平均深度为1.687，差别也不是那么明显。因此，依存深度不是造成该现象的主要原因。另一方面，我们通过斯坦福工具①获取自动句法，其中DocACE为新闻领域的语料，而DocRED包括但不局限于新闻领域。新闻领域的文本格式相对规范，句法分析的结果具有更高的可靠性。综上，我们推测两个语料上句法分析的性能、语料分布的差异对

① https://github.com/Lynten/stanford-corenlp

实验结果造成了一定的影响。

### 4.2　基于 BiLSTM-Multi-GCN 模型的性能

与 4.1 节的 BiLSTM-GCN 模型相比，BiLSTM-Multi-GCN 模型在前面基础上引入 Multi-head Attention，进而对不同的依赖特征加以区分，实验结果如表 5 所示。

**表 5　三种模型的实验结果**

| 序号 | 模型 | DocRED $F_1$/% | DocACE $F_1$/% |
|---|---|---|---|
| 10 | BiLSTM | 50.35 | 40.06 |
| 11 | BiLSTM-GCN | 51.27 | 43.79 |
| 12 | BiLSTM-Multi-GCN | **52.38** | **43.99** |

如表 5 所示，实验 10 与实验 11 的对比表明，图卷积模型可以捕获相对复杂的依赖信息，对基准模型中长距离依赖不足的问题加以弥补。实验 11 与实验 12 对比表明，对不同依赖特征给予同等关注的做法有失偏颇，在不同任务中甚至是同一任务的不同时期（句子级、文档级），对特征的依赖是变化的。通过对比 DocRED 语料上的实验结果，证明了本文提出的动态调整策略的有效性，这在其他任务中也是具备借鉴意义的。而 DocACE 上效果不明显可能是由语料规模决定的，随着 Multi-GCN 的引入，模型的参数逐渐增多，语料规模的局限性也更加明显。

### 4.3　BERT-GCN 与 BERT-Multi-GCN 模型的性能

与 BiLSTM 相比，BERT 能更好地捕获高层的语义信息，同时，BERT 采用的自注意力机制在捕获长距离依赖时更具优势。因此，本节用 BERT 替换基准模型中的 BiLSTM 进行实验，并给出在 BERT 基础上引入 GCN、Multi-GCN 的实验结果，如表 6 所示。

**表 6　BERT 及引入 GCN、Multi-GCN 的结果**

| 序号 | 模型 | DocRED $F_1$/% | DocACE $F_1$/% |
|---|---|---|---|
| 13 | BERT | 55.39 | 40.61 |
| 14 | BERT-GCN | 55.07 | 40.86 |
| 15 | BERT-Multi-GCN | 55.24 | 39.79 |
| 16 | BERT-GCN（依存边） | **55.43** | **42.08** |

从表 6 可以看出，采用 2.2.1 节的 4 种特征进行建图影响了性能，可能是因为 GCN 中的依存关系与 BERT 的中层编码的句法知识出现冗余。采用大规模语料进行预训练，虽然无监督的数据没有标签，但语言本身是有逻辑规律存在的。因此，BERT 对句法知识的把握或许更全面，这一点在实验 16 中得到了很好的印证。

### 4.4　实验参数设置

在本文设计的实验中，词向量维度、隐层数目、学习率与优化器都选用与基准模型相同的设置。为了验证多头注意力机制中头（head）的数目对实验结果的影响，我们采用 BiLSTM-Multi-GCN 在 DocRED 语料上设置了对比实验（表 7）。

**表 7　不同 head 的实验结果**

| head 数目 | DocRED $F_1$/% |
|---|---|
| 1 | 51.25 |
| 2 | **52.38** |
| 4 | 51.43 |

上述结果表明，在 head 数目为 2 时，实验效果最好。

## 5　总结和展望

本文基于文档级关系抽取语料，针对基准模型 BiLSTM 存在的某些问题，对编码部分进行改进，实验结果印证了方法的有效性。本文的主要贡献如下：

（1）借助同指信息，对 ACE2005 语料进行扩展，与 DocRED 语料上的实验结果进行对比分析；

（2）针对基准模型中存在长句子依赖与句法信息捕获不足的问题，本文尝试引入图信息，并借助 GCN 模型捕获邻域信息。借助消融实验，分析不同的特征组合对关系抽取任务的影响，进一步引发对任务本身的一些思考；

（3）为了区分不同特征对文档级关系抽取任务的贡献度，本文将 GCN 模型与 Multi-head Attention 进行有机结合，形成 Multi-GCN。一方面，可以从多层面、多角度获取信息；另一方面，通过调整不同特征的比重实现动态的特征优选。

目前的关系抽取任务多集中在句子级别，对文档级关系抽取的关注相对缺乏。与句内关系抽取相比，文档级关系抽取将面临更多的挑战。随着抽取范围的扩大，候选实体对的数目激增，而经验与统计数据表明，大部分情况下的候选实体对之间是没有

关系的。因此,对候选实体对的筛选可以引入以后的工作中。另一方面,对句间依赖的处理方式过于简单,目前只通过连接相邻句子中依存句法的根节点实现,下一步考虑将篇章关系引入,使用句子间的逻辑修辞关系对句间依赖做进一步补充,挖掘篇章的内在逻辑关系。

## 参考文献


[1] 车万翔, 刘挺, 李生. 实体关系自动抽取[J]. 中文信息学报, 2005, 19(2): 2-7.

[2] Yu M, Yin W, Hasan K S, et al. Improved neural relation detection for knowledge base question answering [J].arXiv preprint arXiv:1704.06194, 2017.

[3] 郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报, 2014, 28(6): 183-189.

[4] Jiang X, Wang Q, Qi B, et al. Attentive path combination for knowledge graph completion[C]//Proceedings of the Asian Conference on Machine Learning, 2017: 590-605.

[5] Zhang M, Zhang J, Su J, et al. A composite kernel to extract relations between entities with both flat and structured features[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, 2006: 825-832.

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

[7] Li Q, Ji H. Incremental joint extraction of entity mentions and relations[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 402-412.

[8] Miwa M, Bansal M. End-to-end relation extraction using LSTMS on sequences and tree structures[J]. arXiv preprint arXiv:1601.00770, 2016.

[9] Zhang M, Zhang Y, Fu G. End-to-end neural relation extraction with global optimization[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 1730-1740.

[10] Swampillai K, Stevenson M. Inter-sentential relations in information extraction corpora[C]//Proceedings of the LREC, 2010.

[11] Swampillai K, Stevenson M. Extracting relations within and across sentences[C]//Proceedings of the International Conference Recent Advances in Natural Language Processing, 2011: 25-32.

[12] Peng N, Poon H, Quirk C, et al. Cross-sentence n-ary relation extraction with graph LSTMS[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 101-115.

[13] Yao Y, Ye D, Li P, et al.DocRED: A large-scale document-level relation extraction dataset[J]. arXiv preprint arXiv:1906.06127, 2019.

[14] Quirk C, Poon H. Distant supervision for relation extraction beyond the sentence boundary[J].arXiv preprint arXiv:1609.04873, 2016.

[15] Song L, Zhang Y, Wang Z, et al. N-ary relation extraction using graph state LSTM[J]. arXiv preprint arXiv:1808.09101, 2018.

[16] Gupta P, Rajaram S,Schütze H, et al. Neural relation extraction within and across sentence boundaries [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 6513-6520.

[17] Verga P,Strubell E, McCallum A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction[J]. arXiv preprint arXiv:1802.10569, 2018.

[18] Sahu S K, Christopoulou F, Miwa M, et al. Inter-sentence relation extraction with document-level graph convolutional neural network[J]. arXiv preprint arXiv:1906.04684, 2019.

[19] Zhang Y, Qi P, Manning C D. Graph convolution over pruned dependency trees improves relation extraction[J].arXiv preprint arXiv:1809.10185, 2018.

[20] Vashishth S, Dasgupta S S, Ray S N, et al. Dating documents using graph convolution networks[J]. arXiv preprint arXiv:1902.00175, 2019.

[21] Jawahar G, Sagot B, Seddah D, et al. What does BERT learn about the structure of language? [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[22] Li Q, Ji H. Incremental joint extraction of entity mentions and relations[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 402-412.



吴婷(1994—),硕士研究生,主要研究领域为自然语言处理、关系抽取。

E-mail: 20174227042@stu.suda.edu.cn



孔芳(1977—),通信作者,博士,教授,主要研究领域为机器学习、自然语言处理、篇章分析。

E-mail: kongfang@suda.edu.cn