计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于句法依存分析的图网络生物医学命名实体识别
作者：许力，李建华
收稿日期：2020-06-01
网络首发日期：2020-09-11
引用格式：许力，李建华．基于句法依存分析的图网络生物医学命名实体识别[J/OL]．计算机应用. https://kns.cnki.net/kcms/detail/51.1307.tp.20200910.1010.002.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(****)**-0000-00 doi:10.11772/j.issn.1001-9081.2020050738

基于句法依存分析的图网络生物医学命名实体识别

许力，李建华*

(华东理工大学 信息科学与工程学院，上海 200237)

(*通信作者电子邮箱 jhli@ecust.edu.cn)

摘要:现有的生物医学命名实体识别方法没有利用语料中的句法信息，准确率不高。针对这一问题，提出基于句法依存分析的图网络生物医学命名实体识别模型。首先利用卷积神经网络(CNN)生成字符向量并将其与词向量拼接，送入双向长短期记忆网络(BiLSTM)进行训练；其次，以句子为单位对语料进行句法依存分析，构建邻接矩阵；最后将 BiLSTM 的输出和经过句法依存分析构建的邻接矩阵送入图卷积神经网络进行训练，并引入图注意力机制优化邻接节点的特征权重得到模型输出。模型在 JNLPBA 数据集和 NCBI-disease 数据集上分别达到了 76.91%和 87.80%的 F1 值，相比于基准模型分别提升了 2.62 和 1.66 个百分点。实验结果证明，提出的方法能有效提升模型在生物医学命名实体识别任务的表现。

关键词: 生物医学；命名实体识别；图卷积神经网络；句法依存分析；图注意力机制

中图分类号:TP391.1 **文献标志码**：A

Biomedical named entity recognition with graph network based on syntactic dependency parsing

XU Li, LI Jianhua*

(School of Information Science and Engineering, East University of Science and Technology,Shanghai 200237, China)

Abastract: The existing biomedical named entity recognition methods did not use the syntactic information in the corpus, achieved low precision. To solve this problem, a biomedical named entity recognition model with graph network based on syntactic dependency parsing was proposed. Firstly, Convolutional Nerual Network (CNN) was used to generate character vector, which was concatenated with word embedding and sent to Bidirectional Long Short Term Memory (BiLSTM). Secondly, syntactic dependency parsing based on sentences was conducted and adjacency matrix was constructed. Finally, the output of BiLSTM and the adjacency matrix constructed by syntactic dependency parsing were sent to graph convolution network for training, and the graph attention mechanism was introduced to optimize the feature weight of adjacency nodes to get the model output. In JNLPBA dataset and NCBI-disease dataset, the model reached 76.91% and 87.80% of F1 value respectively, which was 2.62 and 1.66 percentage points higher than the baseline model. Experimental results show that the proposed method can effectively improve the performance of the model in the biomedical named entity recognition task.

Keywords: biomedicine; named entity recognition; graph convolution network ; syntactic dependency parsing; graph attention mechanism

0 引言

在生物医学领域，每年都会新增大量的专利、期刊和报告等文献。这些文献中包含的实体信息可应用于药物设计和临床医疗，对于生物医学研究具有重大意义。随着自然语言处理技术的发展，生物医学命名实体识别逐渐成为研究热点。广义的命名实体识别针对文本中的特定实体，如人名、地理位置和组织名称等。而在生物医学领域，实体则特指药物名称、蛋白质名称、疾病名称和基因名称等。相比于通用领域的命名实体识别任务，生物医学命名实体识别由于命名规则多样、实体名称较长且包含关系复杂，面临着更多困难。

现有的生物医学命名实体识别方法主要分为：基于规则和词典的方法、基于传统机器学习的方法和基于深度学习的方法。基于规则和字典的方法常见于早期命名实体识别研究，主要通过设计规则模板，构建实体字典来识别实体，如 Krauthammer 等人[1]提出基本局部比对检索工具(Basic Local Alignment Search Tool, BLAST)，Hanisch 等 [2]使用基于规则的方法识别基因和蛋白质实体等。这种方法虽然简单实用，

收稿日期: 2020-06-01; **修回日期**:2020-07-29; **录用日期**: 2020-08-03。

基金项目：国家重大新药创制(2018ZX09735002)；国家重点研发计划项目(2016YFA0502304)

作者简介:许力(1997—)，男，安徽合肥人，硕士研究生，主要研究方向：自然语言处理；李建华(1977—)，男，安徽广德人，博士，副教授 CCF 会员，主要研究方向：计算机辅助设计、药物数据挖掘、生物信息学。

但规则和词典的设计过程复杂耗时且容易产生错误。基于传统机器学习的方法对于特征选取的要求较高，需选取例如前后缀、大小写等人工特征来训练模型。Leaman 等人[3]提出了 tmchem 模型，该模型融合了多种人工特征，在化学命名实体识别取得了较好效果。Li 等人[4]在条件随机场(Conditional Random Field, CRF)中融入词频和共现信息识别基因实体，进一步提升了模型表现。基于传统机器学习的方法进一步提高了实体识别的准确率，但由于过于依赖特征选取且识别策略单一，导致模型鲁棒性和泛化性能较差。基于深度学习的方法主要通过神经网络提取文本特征，在命名实体识别任务中获得了广泛应用。Rocktäschel 等人[5]提出 ChemSpot 模型，将 CRF 与词典结合，学习化学品名称的多种形式。Huang 等人[6]提出了 BiLSTM-CRF 模型，使用双向长短期记忆网络(Bidirectional Long Short Term Memory, BiLSTM)提取上下文特征，并通过 CRF 学习实体标签间的转换概率以修正 BiLSTM 的输出。Guo 等人[7]提出了 BiLSTM-CNN-CRF 模型，在 BiLSTM-CRF 基础上使用卷积神经网络(Convolutional Nerual Network, CNN)提取字符级别特征，模型在生物医学命名实体识别中获得了良好的效果。Dang 等人[8]提出了 D3NER 模型，在 BiLSTM-CRF 模型中引入语言学特征，对词向量进行优化。Crichton 等人[9]在 CNN 模型中融入多任务学习思想，使用多种不同标准和不同实体类型的数据集训练模型，提升了模型的泛化性能。Cho 等人[10]通过 N-Gram 模型在词向量中融入上下文信息，进一步增强了模型表现。基于深度学习的方法在不依赖人工特征的情况下，在命名实体识别任务中获得了更好的表现。以上这些方法虽然均在一定程度上增强了模型在任务中的表现，但其改进的思路大多是通过增加特征，优化词向量来丰富词与词之间的关联，没有从句法的角度考虑词与词之间的联系，效果有所局限。

句法依存分析是获取文本句法信息的重要方法。它以句子为单位构建依存分析图，揭示了词与词之间的依存关系，提升了模型在自然语言处理任务中的表现。现有的句法依存分析方法大多采用线性方式对句法依存分析图进行编码[11][12]，不能充分地利用依存分析图中的句法信息。随着图网络技术的发展，将图卷积网络(Graph Convolutional Network,GCN)应用于句法依存分析的研究已经出现，例如 Bastings 等人[13]利用图卷积神经网络编码句法依存信息用于机器翻译，Marcheggiani 等人[14]使用图卷积网络学习文本句法结构，并将其与 BiLSTM 结合应用于语义角色标注等。虽然句法依存分析在自然语言处理的多个领域已有应用，但在生物医学命名实体识别上的研究还未出现。

针对以上问题，本文提出一种基于句法依存分析的图网络生物医学命名实体识别模型。模型利用 CNN 生成字符向量并将其与词向量拼接，送入 BiLSTM 进行训练。其次，对语料进行句法依存分析，使用图卷积网络对句法依存分析图进行编码并引入图注意力机制(Graph Attention Network,GAT)优化邻接节点的特征权重，得到模型输出。实验结果表明，本文提出的方法有效地提升了生物医学命名实体识别的准确率。

本文首次提出将图卷积神经网络应用于生物医学命名实体识别领域，通过图网络学习文本句法信息并使用图注意力机制增强实体识别效果，为后续的生物命名实体识别研究拓展了思路。本文的贡献主要包含以下两点：

1）现有的生物医学命名实体识别方法没有利用语料中的句法信息，效果有所局限。本文对语料进行句法依存分析，并使用图卷积网络对句法依存分析图进行编码，充分利用了文本中的句法信息。

2）引入图注意力机制优化句法依存分析图中邻接节点的特征权重，更好地聚合了邻接节点的特征，提升了模型的识别效果。

1 基于句法依存分析的图网络生物医学命名实体识别

1.1 句法依存分析

句法依存分析通过分析句子内的依存关系来获取文本句法结构，主张句子中核心动词是支配其它成分的中心成分，而它本身却不受其它任何成分的支配，所有受支配成分都以某种依存关系从属于支配者[15]。依存关系用依存弧表示，方向由从属词指向支配词。每个依存弧上有个标记，称为关系类型，表示该依存对上的两个词之间存在什么样的依存关系[16]。常见的依存关系有主谓关系(SBV)，动宾关系(VOB)和状中关系(ADV)等。

1.2 图卷积神经网络

图卷积网络是卷积神经网络的一种，可用于编码与图相关的信息。给定一个具有 n 个节点的图，可以使用一个 n´ n 的邻接矩阵 $\boldsymbol{A}$ 来表示图结构，若节点 i 到节点 j 有边，则令 $\boldsymbol{A}_{ij}=1$。为了聚合节点自身的特征，需对邻接矩阵 $\boldsymbol{A}$ 进行自环操作，即令 $\boldsymbol{A}$ 对角线上所有元素都为 1。其次，对 $\boldsymbol{A}$ 进行归一化处理，引入邻接矩阵 $\boldsymbol{A}$ 的度矩阵 $\boldsymbol{D}$，其中 $\boldsymbol{D}_{ii}=\sum_{j}\boldsymbol{A}_{ij}$，令 $\hat{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$。在一个具有 l 层的图卷积网络中，定义 $\boldsymbol{h}_i^{(l-1)}$ 为输入向量，$\boldsymbol{h}_i^{(l)}$ 为节点 i 在第 l 层的输出向量，图卷积的操作如下所示：

$$\boldsymbol{h}_i^{(l)}=\sigma\left(\sum_{j=1}^{n}\hat{\boldsymbol{A}}_{ij}\boldsymbol{W}^{(l)}\boldsymbol{h}_j^{(l-1)}+\boldsymbol{b}^{(l)}\right) \tag{1}$$

其中，$\boldsymbol{W}^{(l)}$ 是权重矩阵，$\boldsymbol{b}^{(l)}$ 是偏置项，σ 是非线性映射函数。每个节点可以通过邻接矩阵聚合邻接节点的特征信息，并传入下一层作为输入。

1.3 模型架构

本文提出的模型框架如图 1 所示，模型由词表示层、BiLSTM 层和图卷积层组成。首先利用 CNN 生成字符向量，将其与词向量拼接，送入 BiLSTM 层训练；其次，以句子为单位对语料进行句法依存分析，并构建邻接矩阵；最后将 BiLSTM 的输出和构建好的邻接矩阵送入图卷积神经网络训练，并引入图注意力机制优化邻接节点的特征权重，得到模型输出。

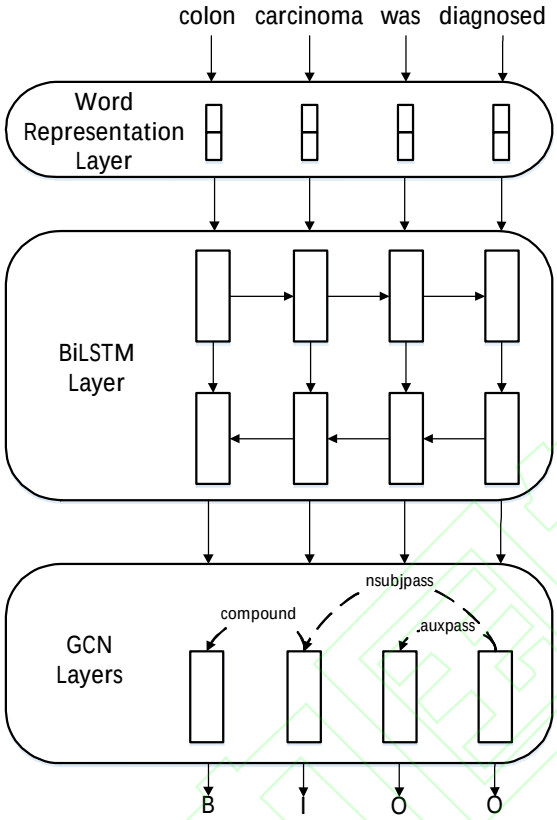


图1 模型整体结构图

Fig. 1 Overall structure of model

1.3.1 词表示层

词表示层的作用是将原始文本转换为向量形式，其示意图如图 2 所示。模型使用 Glove(Global Vectors For Word Representation,Glove)将单词转化为向量，得到长度为 200 维的词向量 $\boldsymbol{X}_1$。将原始的单词按字母拆解，填充为长度为 52 的字母序列后送入 CNN 学习，经过卷积，最大池化操作后，映射成为 30 维的字符向量 $\boldsymbol{X}_2$。将这两部分向量拼接后，构成模型输入 $\boldsymbol{X}=\boldsymbol{X}_1 \oplus \boldsymbol{X}_2$。

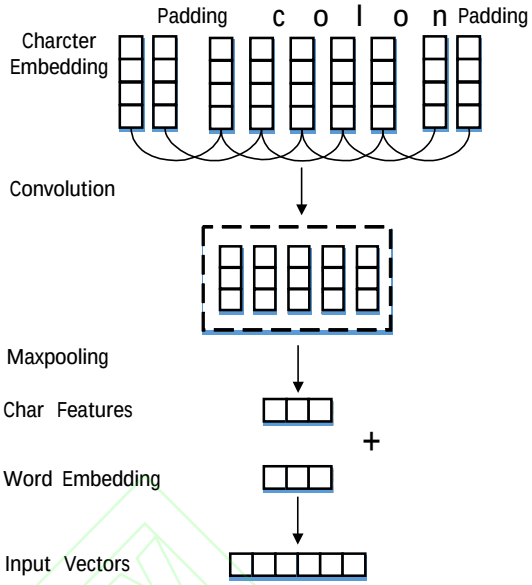


图2 词表示层

Fig. 2 The word representation layer

1.3.2 BiLSTM 层

BiLSTM 层由前向和后向 LSTM 组成，主要用于提取文本中的上下文特征。Marcheggiani 等人[14]的工作指出，图卷积网络的主要问题在于难以捕捉长距离节点之间的依存关系，将其与 LSTM 结合后可以很好地避免这一问题。因此，本文沿用这一思路，将字符向量和词向量进行拼接后，加入到 BiLSTM 中进行编码。LSTM 的主要结构可以表示为：

$$i_t = \sigma\left(\boldsymbol{x}_t \times \boldsymbol{w}_{xh}^{i} + \boldsymbol{h}_{t-1} \times \boldsymbol{w}_{hi}^{j} + \boldsymbol{b}_h^{i}\right) \tag{2}$$

$$f_t = \sigma\left(\boldsymbol{x}_t \times \boldsymbol{w}_{xh}^{f} + \boldsymbol{h}_{t-1} \times \boldsymbol{w}_{hh^{\text{¢}}}^{f} + \boldsymbol{b}_h^{f}\right) \tag{3}$$

$$o_t = \sigma\left(\boldsymbol{x}_t \times \boldsymbol{w}_{xh}^{o} + \boldsymbol{h}_{t-1} \times \boldsymbol{w}_{hh^{\text{¢}}}^{o} + \boldsymbol{b}_h^{o}\right) \tag{4}$$

$$\text{¢}_t = tanh\left(\boldsymbol{x}_t \times \boldsymbol{w}_{xh}^{c} + \boldsymbol{h}_{t-1} \times \boldsymbol{w}_{hh^{\text{¢}}}^{c} + \boldsymbol{b}_h^{c}\right) \tag{5}$$

$$\boldsymbol{c}_t = i_t \text{Ä} \text{¢}_t + f_t \text{Ä} \boldsymbol{c}_{t-1} \tag{6}$$

$$\boldsymbol{h}_t = O_t \text{Ä} tanh\left(\boldsymbol{c}_t\right) \tag{7}$$

其中 σ 是 sigmoid 函数，i，f，o 和 $\boldsymbol{c}$ 分别是输入门，遗忘门，输出门和细胞向量；Ä是点乘运算，$\boldsymbol{w}$，$\boldsymbol{b}$ 代表输入门、忘记门、输出门的权重矩阵和偏置向量。

1.3.3 图卷积层

图卷积层的作用是对语料中的句法依存分析信息进行编码，其输入由两部分构成，一部分是 BiLSTM 的输出，另一部分是根据句法依存分析构建好的邻接矩阵。本文使用 Spacy 工具库对数据集进行句法依存分析，以句子为单位构建依存分析图。将句子中的单词看作图中的节点，将单词与单词之间的依存关系看作图中的边。Spacy 以 child 和 head

描述单词与单词之间的依存关系，依存关系弧从 head 指向 child。得到一个句子的依存分析图之后，我们将其构建成邻接矩阵。假设一个句子由 n 个单词构成，我们需要构建一个大小为 n´n 的邻接矩阵。若单词 i 与单词 j 之间存在依存关系，且依存弧由 i 指向 j，则相应的令 $\boldsymbol{A}_{ij}=1$。邻接矩阵构建好后，将会作为图卷积网络的部分输入。以句子‘China's achievements attract the world's attention’为例，其依存分析图如图 3 所示。

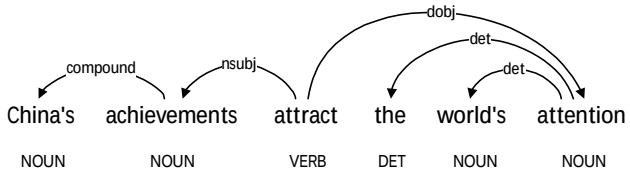


图3 依存分析示意图

Fig. 3 Syntactic dependency parsing

根据依存分析图得到的邻接矩阵，经过自环操作后转为：

$$\begin{pmatrix}1&0&0&0&0&0\\1&1&0&0&0&0\\0&1&1&0&0&1\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&1&1&1\end{pmatrix}$$

经过归一化后，邻接矩阵转换为：

$$\begin{pmatrix}1&0&0&0&0&0\\0.5&0.5&0&0&0&0\\0&1/3&1/3&0&0&1/3\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&1/3&1/3&1/3\end{pmatrix}$$

从图中矩阵可以看出，每个节点的出度，经过归一化之后得到相同的值。这种操作为节点的邻接节点分配了相同的权重，忽视了节点之间依存关系的差异性。因此，本文决定引入图注意力机制[17]，在 GCN 的基础上，对邻接节点的特征聚和做出调整。其操作如下所示：

$$\boldsymbol{z}_i^{(l)}=\boldsymbol{W}^{(l)}\boldsymbol{h}_i^{(l)} \tag{8}$$

$$e_{ij}^{(l)}=\mathrm{LeakyReLU}\left(\vec{\boldsymbol{a}}^{(l)^T}\left(\boldsymbol{z}_i^{(l)}\|\,\boldsymbol{z}_j^{(l)}\right)\right) \tag{9}$$

其中 $\boldsymbol{z}_i^{(l)}$ 表示对 l 层第 i 个节点的特征进行线性转换，$e_{ij}^{(l)}$ 代表节点 j 的特征对于节点 i 的重要性，$\|$是拼接操作，$\vec{\boldsymbol{a}}^{(l)^T}$ 代表一个可以学习的权重向量。将 $\boldsymbol{z}_i^{(l)}$ 与 $\boldsymbol{z}_j^{(l)}$ 拼接后，使用 LeakyReLu 函数进行非线性映射。为了使权重系数在不同节点间方便比较，使用 softmax 函数对 $e_{ij}^{(l)}$ 进行归一化操作，$\mathcal{N}(i)$代表节点 i 的邻接节点集合

$$a_{ij}^{(l)}=\frac{\exp\left(e_{ij}^{(l)}\right)}{\sum_{k\in N(i)}\exp\left(e_{ik}^{(l)}\right)} \tag{10}$$

得到节点 i 的邻接节点注意力权重后，对邻接节点特征进行加权求和

$$\boldsymbol{h}_i^{(l+1)}=\sigma\left(\sum_{j\in N(i)}a_{ij}^{(l)}\boldsymbol{z}_j^{(l)}\right) \tag{11}$$

为了均衡注意力机制的输出，采用多头注意力机制执行相同的操作,拼接后得到输出，K 代表注意力头数

$$\vec{\boldsymbol{h}}_i=\Big\|_{k=1}^{K}\sigma\left(\sum_{j\in N_i}a_{ij}^{k}\boldsymbol{W}^{k}\vec{\boldsymbol{h}}_j\right) \tag{12}$$

获得图卷积层输出后，使用 softmax 函数得到模型分类结果。

2 实验过程与评估

2.1 数据集

本文在 JNLPBA[18]，NCBI-disease[19]数据集中训练模型，数据集分布如表 1 所示。NCBI-disease 数据集来源于 793 篇生物医学领域的摘要，主要包含疾病实体。JNLPBA 数据集包含 DNA 、RNA 、Cell_line 、Cell_type 、Protein 五种实体，模型除了识别出实体，还需要给出实体的具体分类。数据集使用 BIO 标注方案，B 代表 Beginning，标注一个实体的开始部分，I 代表 Inside，标注组成实体的中间部分，O 代表 Outside，标注与实体无关的信息。

表1 实验数据集详细信息

Tab. 1 Details of datasets used in the experiment

数据集	实体类型	数据子集	实体数量
NCBI-disease	Disease	Training	5,423
		Validation	922
		Test	939
JNLPBA	Gene/Protein	Training	14,690
		Validation	3,856
		Test	3,856

2.2 评估标准

生物医学命名实体识别任务需要模型正确判定实体的边界并输出正确的实体类别。为了精确衡量实体识别效果，本文采用准确率(Precision),召回率(Recall)和 F1 值三种评测指标，具体公式如下：

$$P=\frac{T_p}{T_p+F_p}\times 100\% \tag{13}$$

$$R=\frac{T_p}{T_p+F_n}\times 100\% \tag{14}$$

$$F1=\frac{2P\times R}{P+R}\times 100\% \tag{15}$$

其中 T_p 为正确识别的实体个数，F_p 代表实体边界判定错误或类别分类错误的实体个数，F_n 代表未识别出的实体个数。

表2 模型在数据集上的表现对比

Tab. 2 Comprison of model performance on datasets

Methods	NCBI-disease			JNLPBA		
	P	R	F1	P	R	F1
Glove+BiLSTM	80.65	76.24	78.38	70.58	72.81	71.68
Glove+char+BiLSTM	82.64	81.72	82.18	71.42	74.97	73.15
Glove+char+GCN+BiLSTM	84.83	85.51	85.17	76.83	74.25	75.52
Glove+char+GAT+BiLSTM(ours)	**87.29**	**88.31**	**87.80**	**77.56**	**76.28**	**76.91**

2.3 实验设置

本文使用 Glove 将单词转换为 200 维词向量，利用 CNN 生成 30 维字符向量，将两部分向量拼接后形成 230 维特征表示。BiLSTM 隐藏层单元数设为 200，Dropout 率设为 0.5。图卷积层单元数设为 50，图注意力机制头数设置为 6。获得图卷积层输出后，通过 softmax 函数得到最终分类结果。实验运行环境为 Keras 2.2.4，模型学习率设为 5e-4，损失函数定义为多类交叉熵损失函数，优化器采用 Nadam 算法，模型经过 100 轮训练后达到收敛。

2.4 实验结果与参数分析

2.4.1 对比实验分析

为了验证提出方法的有效性，本文进行对比实验说明各个模块作用，实验结果如表 2 所示。

为了说明字符向量在向量表征方面的优势，本文选择 BiLSTM 进行对比实验。从表中可以看出，Glove+char+BiLSTM 模型相比于 Glove+BiLSTM 的方法，其准确率平均提升了 1.42 个百分点，F1 值平均提升了 2.64 个百分点。生物医学实体名称大多包含大小写和特殊字符，使用 CNN 提取字符特征，能够帮助模型区分实体，提高实体识别率。实验结果证明，使用 CNN 提取字符特征生成的字符向量，在与词向量结合后，提升了模型的准确率。

传统的生物医学命名实体识别方法不能很好地利用语料中的句法信息，而句法信息可以丰富实体与实体之间的关联，对于生物医学命名实体识别有重要意义。本文提出一种方法，对语料进行句法依存分析，并使用图卷积网络利用这部分信息进行训练。对比 Glove+char+BiLSTM 模型和 Glove+char+GCN+BiLSTM 模型的实验结果可以发现，后者相比于前者，准确率平均提升了 3.80 个百分点，F1 值平均提升了 2.68 个百分点。实验结果证明，使用 GCN 可以很好地利用句法信息，增强模型表现。

GCN 使用归一化操作处理邻接矩阵，为每个节点的邻接节点赋予相同的权重，忽视了节点之间依存关系的差异性。本文在 GCN 的基础上，引入图注意力机制改善这一问题。对比 Glove+char+GCN+BiLSTM 模型和 Glove+char+GAT+BiLSTM 模型的实验结果可以发现，后者相比于前者，准确率平均提升了 1.60 个百分点，F1 值平均提升了 2.01 个百分点。实验结果证明，引入图注意力机制可以更好地聚合邻接节点的特征，提升模型性能。

2.4.2 与现有其他方法对比

表3 模型在 JNLPBA 数据集上的表现对比

Tab. 3 Comprison of Methods on JNLPBA dataset

Model	Precision	Recall	F1 socre
Tang et al.[20](2014)	70.78	72.00	71.39
Li et al.[21](2016)	74.77	70.85	72.76
Wei et al.[22](2019)	71.57	75.55	73.50
Dai et al.[23](2019)	-	-	74.29
Our model	**77.56**	**76.28**	**76.91**

模型在 JNLPBA 数据集上的表现对比如表 3 所示。Tang 等人[20]在生物医学命名实体识别中融入词表示特征，Li 等人

[21]提出在模型中结合使用句子向量和双词向量；Wei 等人[22]在 BiLSTM-CRF 模型中引入注意力机制，获得了 71.57%的准确率和 73.50%的 F1 值；Dai 等人[23]在 BiLSTM-CRF 模型中使用预训练词向量模型 ELMO(Embeddings From Language Models,ELMO)，获得了 74.29%的 F1 值。以上方法均没有利用语料中的句法信息，效果有所局限。本文提出的模型，使用 CNN 提取字符特征，帮助模型识别包含特殊字符的实体，在不使用人工设计特征的情况下，提升了实体识别率；其次，本文使用图卷积网络学习语料中的句法依存分析信息，增强了词与词之间的关联。图卷积网络在归一化操作后会为每个节点的邻接节点赋予相同的权重，忽视了节点之间依存关系的差异性。针对这一问题，本文引入图注意力机制优化邻接节点的特征权重，让模型更好地聚合邻接节点特征。实验结果表明，本文模型的 F1 值比 Dai 等人提出的模型高出了 2.62 个百分点，相比于 Wei 等人提出的模型，其准确率提高了 5.99 个百分点，F1 值提高了 3.41 个百分点，获得了更好表现。

表4 模型在 NCBI 数据集上的表现对比

Tab. 4 Comprison of Methods on NCBI dataset

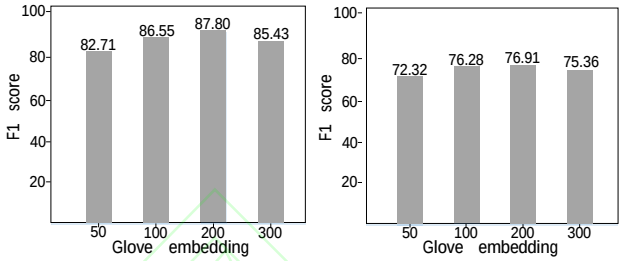
Model	Precision	Recall	F1 socre
Leaman et al.[24] (2013)	82.20	77.50	79.80
Lu et al.[25](2016)	83.50	79.60	81.50
Dang et al.[8](2018)	85.03	83.80	84.41
Wang et al.[26](2019)	85.86	86.42	86.14
Our model	**87.29**	**88.31**	**87.80**

模型在 NCBI-disease 数据集上的表现对比如表 4 所示。Leaman 等人[24]提出 Dnorm 模型，使用 CRF 结合多种人工设计特征；Lu 等人[25]提出 TaggerOne 模型，通过正则化方法降低实体识别错误率；Dang 等人[8]提出 D3NER 模型，在 BiLSTM-CRF 模型中引入语言学特征对词向量进行优化，取得了 85.03%的准确率和 84.41%的 F1 值；Wang 等人[26]提出在 BiLSTM-CRF 模型融入多任务学习思想，同时学习多种生物医学实体类型，取得了 85.86%的准确率和 86.14%的 F1 值。以上这些方法的思路大多是通过融合特征或优化词向量来提升模型效果，没有从句法的角度考虑词与词之间的关联。本文提出的模型，利用图卷积网络学习文本的句法依存信息，并引入图注意力机制优化邻接节点的特征权重，充分地利用了语料中的句法信息。实验结果表明，相比于 Wang 等人提出的模型，本文模型的准确率提升了 1.43 个百分点，F1 值提升了 1.66 个百分点；相比于 D3NER 模型，其准确率提升了 2.26 个百分点，F1 值提升了 3.39 个百分点，取得了良好的效果。

综上，实验结果证明，本文提出的模型通过图卷积网络编码句法依存分析图，充分地利用了语料中的句法信息，在生物医学命名实体识别任务中获得了更好的表现。

2.4.3 实验参数分析

图 1 为模型在不同词向量嵌入维度中的性能表现。从图中可以看出，低维度的词嵌入模型准确率较低，因为过低的嵌入维度不能很好的区分单词与单词间的语义，而过高的词向量维度会导致图卷积层参数过多，难以训练，降低模型性能。经过实验后，确定合适的词向量维度为 200 维。

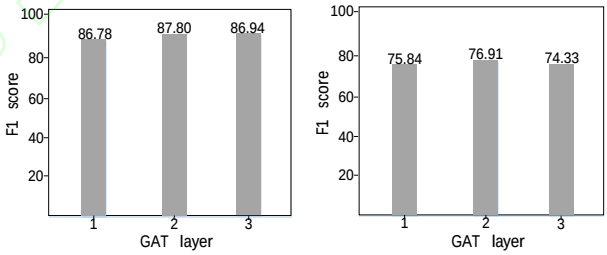


(a)JNLPBA 数据集 (b)NCBI-disease 数据集

图4 不同嵌入维度的 F1 值

Fig. 4 F1 score of varing embedding dimensions

图 5 为模型在不同图卷积层数下的性能表现。可以看出，图卷积层数为 2 时，模型准确率最高，相比于 1 层图卷积，其准确率和 F1 值均得到了一定的提升。图卷积层数为 3 后，准确率和 F1 值反而有所降低。这是因为图卷积层的堆叠层数过多时，会出现过平滑问题[27]，使得邻接节点的特征表示越来越趋同，实体识别率下降。



(a)JNLPBA 数据集 (b)NCBI-disease 数据集

图5 不同图网络层数的 F1 值

Fig. 5 F1 score of varing GCN layers

3 结语

本文提出一种基于句法依存分析的图网络生物医学命名实体识别方法。模型以句子为单位对语料进行句法依存分析，通过图卷积网络学习语料中的句法依存信息并引入图注意力机制优化邻接节点特征权重。实验结果表明，本文提出方法有效地利用了语料中的句法信息，提升了生物医学命名实体识别的准确率。未来会根据句法依存分析图的特点，尝试多种图卷积网络的变型，作出进一步改进。

参考文献

[1] KRAUTHAMMER M, RZHETSKY A , MOROZOV P , et al. Using BLAST for identifying gene and protein names in journal articles[J]. Gene, 2001, 259(1-2):245-252.

[2] HANISCH D, FUNDEL K, MEVISSEN H, et al. ProMiner: Rule-based Protein and Gene Entity Recognition[J]. BMC bioinformatics, 2005, 6 (S1):1633–1640.

[3] LEAMAN R, WEI C H , LU Z Y. tmChem: a high performance approach for chemical named entity recognition and normalization[J]. Journal of Cheminformatics, 2015, 7(S1):1746-1758.

[4] LI Y, LIN H, YANG Z. Incorporating rich background knowledge for gene named entity classification and recognition[J]. BMC bioinformatics, 2009, 10(1):223-240.

[5] ROCKTÄSCHEL T, WEIDLICH M, LESER U. ChemSpot: a hybrid system for chemical named entity recognition.[J]. Bioinformatics (Oxford, England), 2012, 28(12):82-91.

[6] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. [2015-08-09]. https://arxiv.org/abs/1508.01991.

[7] 李丽双,郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别[J]. 中文信息学报, 2018(1):116-122. (LI L S,GUO Y K. Biomedical Named Entity Recognition with CNN-BLSTM-CRF[J]. Journal of Chinese Information Processing, 2018(1):116-122.)

[8] DANG T H, LE H Q, NGUYEN T M. D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information.[J]. Bioinformatics (Oxford, England), 2018, 34(20): 3539–3546.

[9] CRICHTON G, PYYSALO S, CHIU B. A neural network multi-task learning approach to biomedical named entity recognition.[J]. BMC bioinformatics, 2017, 18(1):3973–3981.

[10] CHO H, LEE H. Biomedical named entity recognition using deep neural networks with contextual information.[J]. BMC bioinformatics, 2019, 20(1):1745–1752.

[11] LEVY O, GOLDBERG Y. Dependency-Based Word Embeddings[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, md : Association for Computational Linguistics, 2014: 302-308.

[12] JIE Z M, LU W. Dependency-Guided LSTM-CRF for Named Entity Recognition[EB/OL]. [2019-11-16]. http://www.statnlp.org/research/ie/jie2019depner.pdf.

[13] BASTINGS J, TITOV I, AZIZ W. Graph convolutional encoders for syntax-aware neural machine translation[EB/OL]. [2017-06-18]. https: //arxiv.org/ pdf/ 1704. 04675.pdf.

[14] MARCHEGGIANI D , TITOV I . Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling[EB/OL]. [2017-07-30]. https://arxiv.org/pdf/1703.04826.pdf.

[15] 宋晓思.词项语法句法学中的依存关系探析[J].边疆经济与文化,2013(03):144-145.(Song X S.An analysis of the dependency relationship in lexical grammar and syntax[J].The Border Economy and Culture, 2013(03): 144-145.)

[16] 冯时,付永陈,阳锋,等.基于依存句法的博文情感倾向分析研究[J].计算机研究与发展,2012,49(11):2395-2406.(FENG S,FU Y C,YANG F,et al.Research on sentiment tendency analysis of blog posts based on dependency syntax[J].Journal of Computer Research and Development, 2012, 49(11): 2395-2406.)

[17] VELIKOVI, PETAR, CUCURULL G , CASANOVA A , et al. Graph Attention Networks[C]// Proceedings of the 2018 International Conference on Learning Representations. Vancouver, BC: ICLR, 2018: 1469-1477.

[18] KIM, JIN D, et al. Introduction to the bio-entity recognition task at JNLPBA.[EB/OL]. [2004-05-18]. https:// dl.acm. org/ft_gateway. cfm ? id=1567610&type=pdf

[19] DOĞAN R I, LEAMAN R, LU Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. [EB/OL]. [2014-12-01]. https:// www. researchgate. net/publication/ 259606335 _NCBI_Disease_Corpus_A_Resource_for_Disease_Name_Recognitio n_and_Concept_Normalization.

[20] TANG B ,CAO H ,WANG X ,et al. Evaluating word representation features in biomedical named entity recognition tasks[EB/OL]. [2014-04-16]. http:// downloads. hindawi. com/journals/ bmri/ 2014/240403.pdf.

[21] LI L, JIN L, JIANG Y, et al .Recognizing biomedical named entities based on the sentence vector / twinword embeddings conditioned bidirectional LSTM[C]// Proceedings of China National Conference on Chinese Computational Linguistics. Cham: Springer, 2016: 165-176 .

[22] WEI H , GAO M , ZHOU A , et al. Named Entity Recognition from Biomedical Texts Using an Fusion Attention-based BiLSTM-CRF. [EB/OL]. [2019-03-26]. https://ieeexplore.ieee.org/ stamp/ stamp. jsp?tp=&arnumber=8730360

[23] DAI X , KARIMI S , HACHEY B , et al. Using Similarity Measures to Select Pretraining Data for NER[J]. Association for Computational Linguistics, 2019,10(9):1460–1470.

[24] LEAMAN R,ISLAMAJ D R,LU Z Y. DNorm: disease name normalization with pairwise learning to rank.[J]. Bioinformatics (Oxford, England),2013,29(22):2909–2917.

[25] LEAMAN R,LU Z Y. TaggerOne: joint named entity recognition and normalization with semi-Markov Models.[J]. Bioinformatics (Oxford, England),2016,32(18):2839–2846.

[26] WANG X, ZHANG Y, REN X, et al. Cross-type biomedical named entity recognition with deep multi-task learning[J]. Bioinformatics,2019, 35(10): 1745-1752.

[27] LI Q, HAN Z, WU X M. Deeper insights into graph convolutional networks for semi-supervised learning[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 3538-3545.

This work is partially supported by the National Key Research and Development program of China(2016YFA0502304), the National Major Scientific and Technological Special Project for “Significant new Drugs Development”(2018ZX09735002).

Xu Li, born in 1997, master student.His research interests include natural language processing,data mining.

Li JianHua, born in 1977, Ph. D, professor. His research interests include computer aided design,drug data mining, bioinformatics.