第47卷 第12期 计 算 机 工 程 2021年12月
Vol.47 No.12 Computer Engineering December 2021

·人工智能与模式识别· 文章编号:1000-3428(2021)12-0103-09 文献标志码:A 中图分类号:TP18

基于图卷积神经网络的中文实体关系联合抽取

张军莲[1,2,3],张一帆[1,2,3],汪鸣泉[1,3],黄永健[1,2,3]

(1.中国科学院上海高等研究院 碳数据与碳评估研究中心,上海 201210;2.中国科学院大学,北京 100049;
3.中国科学院低碳转化科学与工程重点实验室,上海 201210)

摘 要: 现有实体关系联合抽取方法未充分考虑中文句子中实体关系的复杂结构特征,为此,提出一种基于图卷积神经网络(GCN)的中文实体关系联合抽取方法。在双向长短时记忆网络抽取序列特征的基础上,利用GCN编码依存分析结果中的语法结构信息,借鉴改进的实体标注策略构建端到端的中文实体关系联合抽取模型。实验结果表明,该方法的F值可达61.4%,相比LSTM-LSTM模型提高了4.1%,GCN能有效编码文本的先验词间关系并提升实体关系抽取性能。

关键词: 信息抽取;关系抽取;联合抽取;图卷积神经网络;依存分析

开放科学(资源服务)标志码(OSID):

中文引用格式: 张军莲,张一帆,汪鸣泉,等.基于图卷积神经网络的中文实体关系联合抽取[J].计算机工程,2021,47(12):103-111.

英文引用格式: ZHANG J L,ZHANG Y F,WANG M Q,et al.Joint extraction of Chinese entity relations based on graph convolutional neural network[J].Computer Engineering,2021,47(12):103-111.

Joint Extraction of Chinese Entity Relations Based on Graph Convolutional Neural Network

ZHANG Junlian[1,2,3],ZHANG Yifan[1,2,3],WANG Mingquan[1,3],HUANG Yongjian[1,2,3]

(1.Shanghai Carbon Data Research Center,Shanghai Advanced Research Institute,Chinese Academy of Sciences,Shanghai 201210,China;
2.University of Chinese Academy of Sciences,Beijing 100049,China;
3.Key Laboratory of Low-Carbon Conversion Science and Engineering,Chinese Academy of Sciences,Shanghai 201210,China)

【Abstract】 The existing methods for extracting entity relations usually ignore the complex structural features of Chinese sentences.To address the problem,a Graph Convolutional neural Network(GCN)-based method is proposed for joint extraction of Chinese entity relations.Based on the sequence features extracted by the bidirectional long short term memory network,this method uses GCN to encode the grammatical structure information in dependency analysis results,and employs the idea of an improved entity tagging strategy to build an end-to-end model for the joint extraction of Chinese entity relations.Experimental results show that this method displays an F score of 61.4%,which is 4.1% higher than the LSTM-LSTM model.GCN can effectively encode the prior relations between words contained in the text,and effectively improve the performance of entity relation extraction.

【Key words】 information extraction;relation extraction;joint extraction;Graph Convolutional neural Network(GCN);dependency analysis

DOI: 10. 19678/j. issn. 1000-3428. 0059574

0 概述

实体关系抽取是信息抽取的下属子任务,信息抽取由美国国家标准技术研究院的自动内容抽取(Automatic Content Extraction,ACE)[1]定义。实体关系抽取任务的目标是从非结构化文本中抽取出实体关系三元组,即<实体1,关系,实体2>,其中,“实体1”和“实体2”是“关系”涉及的2个命名实体,“关系”指2个实体间的关系类型。实体关系抽取是语义理解中的关键技术,也是机器翻译、知识图谱构

基金项目: 国家自然科学基金面上项目(51778601);中国科学院青年创新促进会基金(2018327)。

作者简介: 张军莲(1996—),女,硕士研究生,主研方向为自然语言处理;张一帆,硕士研究生;汪鸣泉,副研究员、博士;黄永健(通信作者),工程师、硕士。

收稿日期: 2020-09-25 **修回日期:** 2020-12-10 **E-mail:** huangyj@sari.ac.cn

建、自动问答系统等应用研究的基础。

目前，实体关系抽取任务有2种主流研究框架：一是流水线方法，即在实体识别之后进行实体关系提取；二是联合抽取方法，即同时进行实体识别和关系抽取。流水线方法在命名实体识别的基础上进行关系提取，实体识别中所产生的错误会影响到关系预测结果，造成错误传播[2]。与流水线方法相比，联合抽取方法被认为具有更好的性能和潜力。2017年，ZHENG等[3]较早提出基于新标注策略的实体关系联合抽取方法，该方法把包含命名实体识别与关系分类2个任务的联合学习模型转变成序列标注问题，其取得了很好的效果。联合关系抽取虽然避免了流水线方法中的错误传播问题，但是其需要更复杂的模型结构以编码更丰富的语义信息。

依存分析的目的是通过分析句子中各个成分之间的依赖关系，从而揭示句子的句法结构。表征文本语法句法结构的依存分析信息可为联合关系抽取提供有效的先验文本结构化信息，帮助模型理清文本结构，从而提升实体关系抽取性能。文献[4]首先利用依存分析并结合中文语法启发式规则抽取关系表述，然后根据距离确定论元位置，最后输出三元组，由此避免了中文复杂的语法规则、灵活的表达方式、多样化的语义对关系抽取带来的限制。文献[5]在模型输入中加入基于最短依存路径的词序列，通过双向长短时记忆（Bidirectional Long Short Term Memory，Bi-LSTM）网络和卷积神经网络提取文本的语义信息，其在中文新闻语料上取得了较好效果。依存分析构建的是语法树结构，考虑到中文语法结构的复杂性，引入图的方法对依存分析中的结构信息进行编码，相比传统的树结构具有更高的灵活性和适用性。图卷积神经网络（Graph Convolutional neural Network，GCN）是卷积网络在图上的实现，可以提取拓扑图上的空间特征，能够有效聚合包含实体关系的实体节点，进而提升实体关系抽取的性能。为减少信息冗余，研究人员对依存分析图中的依赖关系进行裁剪，仅保留部分依赖关系[6-7]。

本文优化ZHENG等所提的新标注策略[3]，提出一种基于GCN的中文实体关系联合抽取方法，并构建融合Bi-LSTM网络和GCN的端到端实体关系联合抽取模型LSTM-GCN-LSTM。借鉴新标注策略的思路，优化标注模式，以标注中文文本中的分词，利用端到端序列标注模型实现中文实体关系联合抽取。通过GCN编码文本依存分析的图结构特征，从而表征文本所蕴含的先验词间关系并构建包含文本序列特征和图结构特征的模型。

1 相关工作

1.1 共享模型参数的联合抽取

文献[8]将神经网络方法用于实体与关系的联合表示，建立用双向序列和双向树结构的LSTM-RNNs表示词序列和依赖树结构的端到端关系提取模型，使实体识别与关系分类共享编码层的Bi-LSTM表示。该模型在数据集ACE2004和ACE2005上的表现优于对比模型，为共享参数的联合学习模型研究奠定基础。文献[9]不依赖依存树与词序列特征，仅将词向量作为模型的输入特征，利用多层Bi-LSTM识别实体，同时借助Attention机制[10]计算当前位置上识别出的实体与已知实体的相似度，从而识别实体之间的关系。

在针对中文语料的研究中，文献[11]人工标注某医院临床医学记录，将Bi-LSTM-CRF和Bi-LSTM组合到统一的框架中，在实体属性的关系中引入关系约束以限制关系的预测结果，并通过组合系数，利用实体或属性识别、实体属性关系2个子任务模块的信息实现关系联合抽取。文献[12]在2个子任务之间引入反馈机制，使用混合神经网络模型来实现联合抽取，在从百度百科和专利文本中爬取到的26 399句资源描述文本中，得到相比其他模型更高的F值。

1.2 基于新标注策略的联合抽取

共享模型参数的联合抽取方法改善了传统流水线方法中忽视2个子任务之间依赖关系的问题，但其在训练时需要先识别出实体，再根据实体信息对实体进行匹配以完成关系分类子任务，该过程中会产生没有关系的实体，出现实体冗余现象。为解决该问题，基于新标注策略的实体关系联合抽取方法应运而生。

2017年，ZHENG等[3]提出基于新标注策略的实体关系联合抽取方法，其把包含命名实体识别与关系分类的联合学习模型转变成序列标注问题。该模型使用Bi-LSTM对句子进行编码，利用LSTM对其进行解码，最后输出实体关系三元组，其解决了共享模型参数的联合抽取方法带来的实体冗余问题。文献[13]基于新标注策略，通过预训练实体识别模型中隐藏层向量得到实体特征，将其作为联合模型的特征，引入Attention机制选择对关系预测影响更大的句子成分。该模型有效提升了NYT（New York Times）数据集上的实体关系提取性能。文献[14]为解决关系重叠问题，添加象征该词所在实体参与多个关系类别的M标签，并改进实体与关系的匹配策略，改进后的实体关系联合抽取模型性能优于使用相同模型的流水线方法，在药物-药物交互作用（Drug-Drug Interactions，DDI）数据集上，实体识别F值为89.9%，关系抽取F值为67.3%。文献[15]借鉴该标注策略，在模型中引入Attention机制以增强对文本中更能体现关系的词语的编码能力，在模型训练中使用对抗训练，该文所提出的LSTM-LSTM-ATT-Bias端到端模型在NYT数据集上，实体1识别F值为53.4%，实体2识别F值为51.9%，实体关系抽取F值为53%。

1.3 基于图的信息抽取

语言是按照复杂的句法语法规则进行组词成句的，多数传统方法仅提取文本中的序列特征，不足以表征文本的复杂语义。利用图结构特征将不同类

型、不同结构的分词通过边的形式连接起来[16],可以更全面地表达句中的语法关系,因此,该方法被广泛应用于信息抽取、关系抽取等领域。

在信息抽取领域:文献[17]为了突破多数信息抽取系统仅基于序列特征而实现的局限性,提出一种基于文本底层结构且针对特定任务的在图形拓扑上学习局部和全局表示的信息提取框架GraphIE(Graph Information Extraction),该框架联合单词的节点表示或句子的节点表示及其互相依赖关系;文献[18]提出实现信息抽取多任务的动态跨度图框架DYGIE(Dynamic Graph Information Extraction),利用动态跨度图方法,将文本跨距视为图形结构中的节点,根据预测的节点间相互参照关系以及与图中其他节点的关联关系,为每个节点构造加权弧。

在关系抽取领域:文献[19]将实体及其关系转换为有向图,并使用基于神经转换的解析系统实现求解,不仅对实体与关系之间的依赖关系进行建模,而且对不同关系之间的依赖关系进行建模,从而实现实体和关系的联合抽取;文献[20]提出基于图LSTM的通用框架,将句中关系抽取任务扩展为跨句子的多元关系抽取。

图卷积神经网络是为了实现图结构数据编码,在卷积神经网络的基础上改编得到的一种网络[21]。文献[21]在每个节点周围的一阶邻域上操作限制滤波器,产生局部图结构和节点特征的编码表示,从而简化文献[22-23]提出的图神经网络。文献[6-7]将GCN与Bi-LSTM等递归网络相结合,提取文本中的语境化信息和句法知识,针对依存图的信息冗余问题,分别提出以最近公共祖先为中心的剪枝技术和基于Attention的剪枝策略,以忽略无关信息并降低计算复杂度。

2 本文方法

本文借鉴ZHENG等所提的新标注策略[3]对文本标注其所蕴含的实体与关系信息,利用Bi-LSTM提取文本序列特征和GCN编码文本中的先验词间关系,通过分类网络得到文本的标签预测结果,最后按照关系提取规则从文本中抽取出其蕴含的实体关系三元组信息。

2.1 标注模式及提取规则

英文分词以空格作为分词标志,实体名多由2个以上分词组成,中文虽然无明显的分词标志,但利用分词工具得到的分词结果大多可直接表达实体名。本文针对中文分词的这一特点,在新标注策略的基础上对标注模式进行优化,采用更简易的“BIO”标注方案。另外,本文将关系三元组中2个实体的实体类别也标记在关系标签中。在本文的标注模式下,文本的标注结果如图1所示。

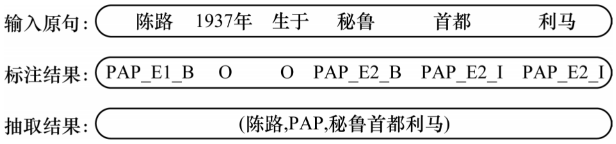

图1 本文标注模式下的中文文本标注结果

Fig.1 Annotation results of Chinese texts under annotation mode in this paper

文本的每个分词都被标注一个标签,标签中包含实体关系三元组信息。文本的标注结果包括3个组成部分:

1)关系类型,即关系三元组中的关系,本文将数据集中预定义的关系和实体类别拼接构成关系类型。

2)实体角色,即分词在关系三元组中的角色信息,用“E1”表示该分词属于首实体的组成之一,用“E2”表示该分词属于尾实体的组成之一。

3)分词位置,即分词在实体名称中的位置信息。

本文采用“BIO”标注方案标注分词位置信息。若实体仅由一个分词构成,用“B(Begin)”标注该分词;若实体由多个分词构成,用“B(Begin)”标注第一个分词,用“I(Inside)”标注其后所有分词。文本中包含在三元组中的分词,其标注结果由以上3个部分拼接形成,而对于不包含在三元组中的分词,本文用“O(Other)”做标注。

在图1的示例中,原句包含实体关系三元组<陈路,PAP,秘鲁首都利马>,其中,“PAP”表示“人物/祖籍/地点(Person/Ancestor/Place)”,是由“陈路”的实体类别“人物”、“秘鲁首都利马”的实体类别“地点”以及预定义的关系“祖籍”这3个信息拼接而成的关系类型。首实体仅含有一个分词“陈路”,按照上文所述的标注模式,其被标注为“PAP_E1_B”;尾实体“秘鲁首都利马”含有“秘鲁”“首都”“利马”3个分词,根据其在尾实体中的位置,分别被标注为“PAP_E2_B”“PAP_E2_I”“PAP_E2_I”。分词“1937年”和“生于”因没有包含于三元组中而被标注为“O”。

分词标注结果指明实体关系三元组中首尾实体的分词信息和所属的关系类型。在模型预测出句子中分词的标注结果后,将标注相同关系类型的分词相结合,根据实体角色和分词位置将分词组合起来得到首尾实体名称,最终获取<实体1,关系,实体2>三元组。

上述介绍的分词标注模式将实体关系三元组的抽取问题转化为端到端的序列标注问题。本文考虑一个实体仅属于一个三元组的情况。在预测文本包含的实体关系时,若预测标注结果中包含多于一个具有相同关系类型的三元组,本文按照最邻近原则将最近的2个实体相组合形成三元组,并作为实体关系的预测结果。

2.2 模型总体框架

本文的实体关系联合抽取模型包含4个组成部分,分别为表示层、Bi-LSTM与GCN编码层、LSTM解码层、Softmax层。总体框架如图2所示。

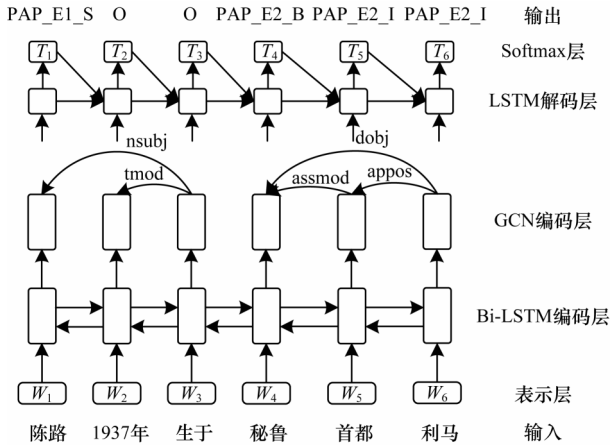

图2 模型框架

Fig.2 Model framework

2.3 表示层

通过词向量表将中文分词转换成表征分词信息的低维向量，作为下一层的输入向量。本文使用百度百科语料库训练语言模型得到词向量表（https://github.com/Embedding/Chinese-Word-Vectors），该词向量表包含语料库中所有分词通过语言模型训练得到的向量表示。检索词向量表得到分词的向量表示的过程具体如下：对于包含n个分词的输入句子S，$s=\{t_1,t_2,\cdots,t_n\}$，句中的每个分词为t_i，从词向量表中检索到其对应的词向量表示$\boldsymbol{x}_i$，最终，句子S转换成其分词的向量表示序列：$\boldsymbol{s}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$。

2.4 编码层

编码层中使用Bi-LSTM提取文本中的序列特征，再利用GCN编码文本中基于依存分析图的局部依赖特征以及先验词间关系。

2.4.1 Bi-LSTM编码

Bi-LSTM编码层由2个平行的LSTM层组成，即前向LSTM层和反向LSTM层[24]。Bi-LSTM中前向网络的神经元结构如图3所示。

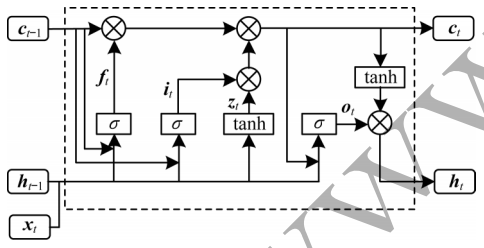

图3 Bi-LSTM编码层中的前向网络神经元结构

Fig.3 Forward network neuron structure in Bi-LSTM coding layer

LSTM通过遗忘门、输入门和输出门来对输入信息进行保护和控制。在前向网络中，每次新输入一个分词特征向量$\boldsymbol{x}_t$，并与上一时刻状态$\boldsymbol{h}_{t-1}$共同产生下一时刻的状态$\boldsymbol{h}_t$，其中，t代表时间步长。隐藏状态$\boldsymbol{h}_t$的计算如下所示[25]：

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{x_i}\boldsymbol{x}_t+\boldsymbol{W}_{h_i}\boldsymbol{h}_{t-1}+\boldsymbol{W}_{c_i}\boldsymbol{c}_{t-1}+\boldsymbol{b}_i) \tag{1}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{x_f}\boldsymbol{x}_t+\boldsymbol{W}_{h_f}\boldsymbol{h}_{t-1}+\boldsymbol{W}_{c_f}\boldsymbol{c}_{t-1}+\boldsymbol{b}_f) \tag{2}$$

$$\boldsymbol{z}_t=\tanh(\boldsymbol{W}_{x_c}\boldsymbol{x}_t+\boldsymbol{W}_{h_c}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c) \tag{3}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\boldsymbol{z}_t \tag{4}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{x_o}\boldsymbol{x}_t+\boldsymbol{W}_{h_o}\boldsymbol{h}_{t-1}+\boldsymbol{W}_{c_o}\boldsymbol{c}_t+\boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\tanh(\boldsymbol{c}_t) \tag{6}$$

$$\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}}_t\oplus\overleftarrow{\boldsymbol{h}}_t]\in\mathbb{R}^{2d_e} \tag{7}$$

$$\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)\in\mathbb{R}^{2d_e\times n} \tag{8}$$

其中：$\boldsymbol{i}$、$\boldsymbol{f}$、$\boldsymbol{o}$分别为输入门、遗忘门、输出门；$\boldsymbol{b}$是偏置项；$\boldsymbol{W}$为参数矩阵。前向LSTM层通过从分词向量$\boldsymbol{x}_1$到$\boldsymbol{x}_t$，考虑$\boldsymbol{x}_t$的前文信息来编码$\boldsymbol{x}_t$，输出记为$\overrightarrow{\boldsymbol{h}}_t$。类似地，反向LSTM层从分词向量$\boldsymbol{x}_n$到$\boldsymbol{x}_t$，考虑$\boldsymbol{x}_t$的后文信息来编码$\boldsymbol{x}_t$，输出记为$\overleftarrow{\boldsymbol{h}}_t$。最后，级联$\overrightarrow{\boldsymbol{h}}_t$和$\overleftarrow{\boldsymbol{h}}_t$来表示第$t$个分词编码后的信息，如式(7)所示，⊕表示向量级联，d_e为单向LSTM网络维度。对于输入的S，该层的输出如式(8)所示，$\boldsymbol{h}$输出到下一层作为输入。

2.4.2 GCN编码

GCN是一种简单有效的基于图的卷积神经网络，其能够通过图节点间的信息传递来有效捕捉数据之间的依赖性，因此，经常被用来处理对象间关系丰富且存在相互依赖关系的数据。GCN被直接作用于图上[26]，网络的输入是图的结构和图中节点的特征表示。对于图中的每个节点，GCN通过该节点附近其他节点的性质融合归纳得到该节点的特征表示向量。

不同于GCN在图像领域中直观地将图像中的每个像素点作为图中的节点，本文借助文本的依存分析结果，将文本的每个分词经Bi-LSTM生成的特征向量表示作为图中的节点，依存分析结果中不同节点之间的关系作为图中的边，构成图卷积神经网络的基本图结构。依存分析图展示的是文本分词之间的依赖关系，在依存分析图中，root是虚拟根节点，有且仅有一个节点依赖于根节点，边表示分词之间的依赖关系。图4所示为“公司于2015年02月27日在海淀分局登记成立”的依存分析图：“公司”和“登记”之间是名词性主语和动词之间的关系，该关系属于nsubj关系；“登记”和“02月27日”之间是动词和名词组成的非核心依赖关系，该关系属于nmod关系；“2015年”和“02月27日”之间是2个名词之间的补语关系；“于”和“02月27日”之间则为介词与其所依赖的名词之间的关系，属于case关系。连接“公司”和“02月27日”的“登记”是表征公司成立日期关系的关键分词，在依存分析图中可以通过词节点与边将2个实体联系起来。

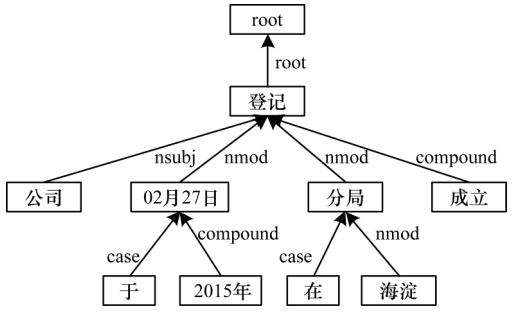

图4 依存分析图示例

Fig.4 Dependency analysis diagram example

基于依存分析图结构的GCN编码层利用前面的Bi-LSTM生成的分词特征向量表示，将每个节点邻域内的相关信息编码为一个新的表示向量。

对于一个有 n 个节点的依存分析图，本文使用 $n\times n$ 的邻接矩阵 $\boldsymbol{A}_{ij}$ 表示其图结构，通常使 $\boldsymbol{A}_{ij}=1$ 代表节点 i 到节点 j 之间存在边。因为依存分析图的边可能存在不同的依赖关系，本文对表征节点 i 与节点 j 之间边的 $\boldsymbol{A}_{ij}$ 赋予不同的数值，以区别不同的依赖关系。表征图4所示文本依存分析图的邻接矩阵如图5(a)所示。

	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	3	4	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	3	4	0	0	0
8	0	2	0	0	5	0	0	5	0	4
9	0	0	0	0	0	0	0	0	0	0

(a)未改进邻接矩阵

	0	1	2	3	4	5	6	7	8	9
0	1	0	0	0	0	0	0	0	0.03	0
1	0	1	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0
4	0	0	0.10	0.13	1	0	0	0	0	0
5	0	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0.10	0.13	1	0	0
8	0	0.06	0	0	0.16	0	0	0.16	1	0.13
9	0	0	0	0	0	0	0	0	0	1

(b)改进后邻接矩阵

图5　邻接矩阵

Fig.5　Adjacency matrix

在 L 层GCN中，$\boldsymbol{h}_i^{(l-1)}$ 表示输入向量，$\boldsymbol{h}_i^{(l)}$ 表示节点 i 在第 l 层的输出向量，一个图卷积操作如下所示：

$$\boldsymbol{h}_i^{(l)}=\sigma\left(\sum_{j=1}^{n}\boldsymbol{A}_{ij}\boldsymbol{W}^{(l)}\boldsymbol{h}_j^{(l-1)}+\boldsymbol{b}^{(l)}\right) \tag{9}$$

其中：$\boldsymbol{W}^{(l)}$ 是线性转换；$\boldsymbol{b}^{(l)}$ 是偏置项；σ 是非线性函数(如ReLU)；$\boldsymbol{A}_{ij}$ 是邻接矩阵。在每一次图卷积计算中，各节点汇集图中其相邻节点上的信息。

直接使用式(9)会出现不同节点表示之间量级差距过大的现象，导致句子的特征表示不考虑节点中包含的信息内容，仅仅偏向于高阶节点，为此，在实际使用中需要对邻接矩阵 $\boldsymbol{A}_{ij}$ 进行归一化处理。此外，式(9)依赖树中的节点永远不会再连接到自身，即 $\boldsymbol{h}_i^{(l-1)}$ 中的信息永远不会传递给 $\boldsymbol{h}_i^{(l)}$，因此，本文为图中的每个节点添加自循环，将归一化后的 $\boldsymbol{A}_{ij}$ 对角线元素设为1，形成改进后的邻接矩阵 $\tilde{\boldsymbol{A}}_{ij}$，最后将其通过非线性函数反馈给GCN。上述改进使图中主要特征仍是节点本身，符合特征提取原则。表征图4所示文本依存分析图的改进邻接矩阵如图5(b)所示。式(9)改进如下：

$$\boldsymbol{h}_i^{(l)}=\sigma\left(\sum_{j=1}^{n}\tilde{\boldsymbol{A}}_{ij}\boldsymbol{W}^{(l)}\boldsymbol{h}_j^{(l-1)}+\boldsymbol{b}^{(l)}\right) \tag{10}$$

2.5　LSTM解码层

本文使用LSTM结构对编码层基于图结构的编码输出进行解码。在基于依存分析图的编码结果中，根据代表文本分词的节点特征向量，将图结构的特征表示转换成序列结构的特征向量。解码层采用一个单向的LSTM层，结构如图6所示。

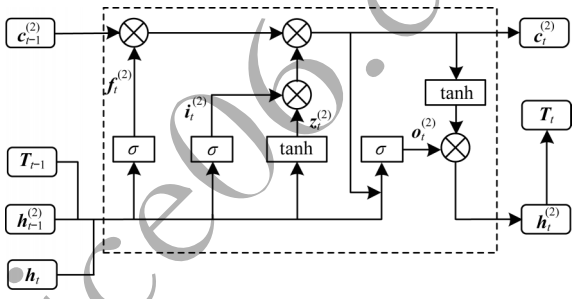

图6　LSTM解码层神经元结构

Fig.6　LSTM decoding layer neuron structure

在检测分词 $\boldsymbol{x}_t$ 的标签时，解码层的输入是从编码层获得的分词 $\boldsymbol{x}_t$ 的上下文表示向量 $\boldsymbol{h}_t$，前一神经元的预测标签表示为 $\boldsymbol{T}_{t-1}$，前一神经元值为 $\boldsymbol{c}_{t-1}^{(2)}$，解码层前一隐层向量为 $\boldsymbol{h}_{t-1}^{(2)}$，最终输出是 $\boldsymbol{x}_t$ 预测标签的向量表示 $\boldsymbol{T}_t$，解码层表示为[3]：

$$\boldsymbol{i}_t^{(2)}=\sigma(\boldsymbol{W}_{x_i}^{(2)}\boldsymbol{h}_t+\boldsymbol{W}_{h_i}^{(2)}\boldsymbol{h}_{t-1}^{(2)}+\boldsymbol{W}_{t_i}\boldsymbol{T}_{t-1}+\boldsymbol{b}_i^{(2)}) \tag{11}$$

$$\boldsymbol{f}_t^{(2)}=\sigma(\boldsymbol{W}_{x_f}^{(2)}\boldsymbol{h}_t+\boldsymbol{W}_{h_f}^{(2)}\boldsymbol{h}_{t-1}^{(2)}+\boldsymbol{W}_{t_f}\boldsymbol{T}_{t-1}+\boldsymbol{b}_f^{(2)}) \tag{12}$$

$$\boldsymbol{z}_t^{(2)}=\tanh(\boldsymbol{W}_{x_c}^{(2)}\boldsymbol{h}_t+\boldsymbol{W}_{h_c}^{(2)}\boldsymbol{h}_{t-1}^{(2)}+\boldsymbol{W}_{t_c}\boldsymbol{T}_{t-1}+\boldsymbol{b}_c^{(2)}) \tag{13}$$

$$\boldsymbol{c}_t^{(2)}=\boldsymbol{f}_t^{(2)}\boldsymbol{c}_{t-1}^{(2)}+\boldsymbol{i}_t^{(2)}\boldsymbol{z}_t^{(2)} \tag{14}$$

$$\boldsymbol{o}_t^{(2)}=\sigma(\boldsymbol{W}_{x_o}^{(2)}\boldsymbol{h}_t+\boldsymbol{W}_{h_o}^{(2)}\boldsymbol{h}_{t-1}^{(2)}+\boldsymbol{W}_{c_o}^{(2)}\boldsymbol{c}_t^{(2)}+\boldsymbol{b}_o^{(2)}) \tag{15}$$

$$\boldsymbol{h}_t^{(2)}=\boldsymbol{o}_t^{(2)}\tanh\left(\boldsymbol{c}_t^{(2)}\right) \tag{16}$$

$$\boldsymbol{T}_t=\boldsymbol{W}_{t_s}\boldsymbol{h}_t^{(2)}+\boldsymbol{b}_{t_s} \tag{17}$$

$$\boldsymbol{T}=(\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n)\in\mathbb{R}^{d_d\times n} \tag{18}$$

其中：$\boldsymbol{i}$、$\boldsymbol{f}$、$\boldsymbol{o}$ 分别为输入门、遗忘门、输出门；$\boldsymbol{b}$ 是偏置项；$\boldsymbol{W}$ 为参数矩阵。对于输入的 $\boldsymbol{h}_t$，该层的输出为预测标签的向量序列，如式(18)所示，d_d 为编码层的网络维度。

2.6 Softmax层

本文模型的分类层使用Softmax分类器进行标签分类。通过Softmax层运算得到条件概率$p_t^i(\theta)$,即分词t为标签i的概率,如下:

$$p_t^i(\theta)=\frac{\exp(y_t^i)}{\sum_{j=1}^{N_t}\exp(y_t^j)} \tag{19}$$

其中:θ为模型需要用到的所有参数;N_t表示总的标签数目;y_t^i表示y_t中第i个元素。$\boldsymbol{y}_t$是模型对分词t在所有标签类型上的评分,其定义为:

$$\boldsymbol{y}_t=\boldsymbol{W}_y\boldsymbol{T}_t+\boldsymbol{b}_y \tag{20}$$

其中:$\boldsymbol{W}_y\in\mathbb{R}^{N_t d}$是参数矩阵;$\boldsymbol{b}_y\in\mathbb{R}^{N_t}$是偏置项。

在测试阶段,将所学习到的标签特征$\boldsymbol{T}_t$乘以概率p得到$\boldsymbol{T}_t'=p\boldsymbol{T}_t$,用$\boldsymbol{T}_t'$进行标签预测。最终,得到分词$t$具有如下标签:

$$\hat{t}=\operatorname{argmax} p_t^i(\theta) \tag{21}$$

3 实验结果与分析

3.1 数据集和实验设置

本文在2019年百度语言与智能技术竞赛的关系抽取任务所提供数据集基础上进行数据清洗与筛选,从而形成本文实验数据集,每个文本中仅包含一个目标提取三元组。实验所用标记数据集共包含50种实体关系类别,分布在132 952个句子中,其中,训练集包含118 121句,测试集包含14 831句。

准确率(Precision)、召回率(Recall)、F值是目前实体关系抽取研究领域中通用的性能评测指标,其中,F值是考虑准确率和召回率的综合性指标。在实际的模型训练中,具体的超参数值如表1所示。

表1 模型超参数设置

Table 1 Model hyper parameters setting

超参数名称	超参数值
词向量维数	300
Bi-LSTM编码层维数	300
Bi-LSTM编码层层数	1
GCN层数	1
LSTM解码层维数	600
LSTM解码层层数	1
Batch大小	20
学习率	0.000 8
优化算法	Adam
权重衰减速率	0.000 1

3.2 基线模型与评估方案

为验证GCN可以有效编码词间先验关系,并评估所提2个模型(LSTM-GCN-CRF、LSTM-GCN-LSTM)在中文实体关系抽取中的性能,本文选择经典模型LSTM-CRF和LSTM-LSTM以及LSTM-LSTM-Bias、LSTM-GCN-Pruned作为基线模型,分别进行中文实体关系抽取实验,并对比分析各个模型的评价指标结果。

本文为证明GCN编码层能有效编码词间先验关系,在LSTM-CRF和LSTM-LSTM中加入GCN编码层,在文本序列特征的基础上提取图结构特征,相应地生成LSTM-GCN-CRF、LSTM-GCN-LSTM这2种模型。LSTM-CRF[27]采用LSTM编码文本进行实体识别,通过简单的条件随机场架构对输出标签进行建模,预测实体标签序列。LSTM-LSTM[28]则使用LSTM对通过之前网络学习的信息进行解码,实现实体标签序列预测。LSTM-GCN-Pruned[6]在LSTM编码之后使用GCN编码,使词向量融合上下文信息,同时提出以最近公共祖先为中心的剪枝技术,以去除依存分析图中的无关依赖信息。

本文为验证LSTM-GCN-CRF、LSTM-GCN-LSTM模型对中文实体关系抽取性能的提升作用,选择同样基于标注策略的实体关系联合抽取模型,即ZHENG等提出的LSTM-LSTM-Bias作为对比基线模型,在中文实体关系数据集上训练模型,从而预测实体关系。上述模型以不同方式增强实体间的联系:LSTM-LSTM-Bias在模型训练时使用增加了偏置的目标函数进行训练,优化模型参数,增加文本中的实体标签对损失函数的影响,同时减少非实体标签对损失函数的影响,以此增强实体之间的联系;本文所提模型通过GCN编码层提取文本中的图结构特征,通过依存分析图中节点和不同关系的边强调实体之间的连接。

3.3 结果分析

本文按照3.2节的评估方案,在中文实体关系数据集上,训练本文所提模型(LSTM-GCN-CRF、LSTM-GCN-LSTM)、经典实体关系抽取模型(LSTM-CRF、LSTM-LSTM)、对依存分析图进行裁剪去除冗余信息的LSTM-GCN-Pruned模型以及LSTM-LSTM-Bias模型。

在中文数据集上,不同基线模型的实体关系抽取准确率、召回率与F值结果如表2所示。从表2可以看出:加入GCN编码层后的LSTM-GCN-CRF和LSTM-GCN-LSTM这2种模型的F值分别达到61.4%、61.2%,相比只提取序列特征的LSTM-CRF和LSTM-LSTM,F值分别提升3.0%、4.1%;LSTM-GCN-Pruned模型的3项指标均高于未采用GCN编码的经典模型,GCN通过分词节点和边关系信息充分学习中文文本中蕴含的复杂

句法信息，能表征更丰富的语义信息，GCN编码之后的2个模型都取得更高的召回率和F值，说明GCN编码层可改善实体关系抽取性能；相较于LSTM-GCN-CRF、LSTM-GCN-LSTM模型，LSTM-GCN-Pruned虽然没有使用文本依存分析图中的全部依赖关系，但是其实体关系抽取性能并未因此而降低，表2中的3项评价指标略高于其他模型，这是因为大多数与关系相关的信息通常包含在以2个实体的最近公共祖先为根的子树中，LSTM-GCN-Pruned模型采用剪枝技术仅保留所有直接连接到依赖路径上的节点，从而保留了大部分关键信息。

表2　不同模型的关系三元组预测性能比较

Table 2　Comparison of relational triple prediction performance of different models　%

模型	准确率	召回率	F值
LSTM-CRF	63.8	56.1	59.6
LSTM-LSTM	64.9	53.8	58.8
LSTM-GCN-CRF	65.1	58.1	61.4
LSTM-GCN-LSTM	64.7	58.3	61.2
LSTM-GCN-Pruned	65.1	58.5	61.8

如表3所示，LSTM-LSTM-Bias在英文实体关系抽取数据集（NYT）上的性能表现较好，但其直接用于中文数据集时，3项指标明显降低，F值仅有41.2%：一方面是因为中文在组词、句法语法规则上更加灵活，更容易对文本内容产生语义理解分歧；另一方面是因为LSTM-LSTM-Bias仅用Bi-LSTM提取文本的长距离依赖关系序列特征，不足以表征中文文本中复杂的句法信息。本文所提LSTM-GCN-CRF、LSTM-GCN-LSTM模型的F值分别达到61.4%、61.2%，相比LSTM-LSTM-Bias模型分别提高了49.0%、48.5%，由此说明本文LSTM-GCN-CRF、LSTM-GCN-LSTM模型可有效提升中文实体关系抽取性能。

表3　LSTM-LSTM-Bias模型的预测性能

Table 3　Prediction performance of LSTM-LSTM-Bias model　%

语言	准确率	召回率	F值
English	61.5	41.4	49.5
Chinese	53.5	40.1	41.2

3.4　GCN分析

上文中经过不同模型指标数据的对比分析，证明了GCN编码层的加入可有效提升实体关系抽取性能。本文统计测试集中实体1、实体2被正确预测的句子数，进一步验证GCN编码层对实体关系抽取结果的改善作用。实验结果如表4所示，其中：E1T_E2T表示实体1和实体2均预测正确；E1F_E2F表示实体1和实体2均预测错误；E1T_E2F表示实体1预测正确、实体2预测错误；E1F_E2T表示实体1预测错误、实体2预测正确。

表4　实体1、实体2被正确预测的句子数

Table 4　Number of sentences for entity 1 and entity 2 which are predicted correctly

类别	LSTM-CRF	LSTM-LSTM	LSTM-GCN-CRF	LSTM-GCN-LSTM
E1T_E2T	8 005	7 647	8 276	8 415
E1F_E2F	956	1 279	797	850
E1T_E2F	4 721	4 711	4 742	4 518
E1F_E2T	1 149	1 194	1 016	1 048

实体三元组包含首尾2个实体以及两者之间的关系。本文在观察测试集中所有句子的实体关系三元组抽取结果时发现，存在实体1、实体2其中一个抽取错误的现象，因此，统计LSTM-CRF、LSTM-LSTM和LSTM-GCN-CRF、LSTM-GCN-LSTM这4个模型在包含14 831个句子的测试集上实体1、实体2被正确预测的句子数情况。从表4可以看出，加入GCN编码层的LSTM-GCN-CRF、LSTM-GCN-LSTM模型将实体1、实体2同时预测正确的句子数多于原始模型，这是因为GCN基于依存分析图的图结构提取文本语义信息，在依存分析图上三元组中2个实体通过携带句中分词关系类别的边而更加紧密地联系起来，增加了2个实体同时被提取出来作为同种关系涉及的实体对的可能性，从而提高了实体三元组的提取完整性。此外，4个模型抽取结果中单个实体1预测正确的句子数普遍多于单个实体2，这是因为数据集里大多数中文文本的语言表达按照主语谓语宾语的语法顺序，根据数据集中关系和实体的标注规则可知，实体1是主语，多位于句子靠前的位置，而实体2是宾语，多位于句子靠后的位置，主语被作为实体关系三元组中的实体被抽取出来的可能性更大，而实体2被抽取出来需要依靠句子更丰富的语义信息。

3.5　实例分析

为了更直观地体现GCN编码层在中文文本实体关系抽取中的效果，本文列出2个典型实体关系抽取结果实例，如图7所示，其中加粗表示预测错误的标签。图中展示出关于实例的4行信息，从上至下依次为原句、正确的实体关系抽取结果、LSTM-LSTM的抽取结果以及LSTM-GCN-LSTM的抽取结果。

图7 实体关系抽取结果

Fig.7 Entity relationship extraction results

原句1中存在可能混淆抽取结果的其他实体名。对比LSTM-LSTM和LSTM-GCN-LSTM模型的抽取结果可以发现,LSTM-GCN-LSTM通过GCN编码层获取到句子依存信息,增强了“赵灵儿”与“李忆如”之间“母亲”关系的连接,从而提取出正确的实体关系;而LSTM-LSTM则误将“抚养”关系当成“母亲”关系,提取出了错误的实体关系。

在原句2中,目标实体关系三元组中的某个实体同时存在于其他关系三元组中。LSTM-LSTM仅提取出一个实体,无法构成三元组,且错误地将“谁偷了谁的忧伤”预测为“人物/作者/图书作品”实体关系的实体;LSTM-GCN-LSTM虽然提取出2个实体并正确提取出“晋江文学城”的实体分类结果,却将实体1“谁偷了谁的忧伤”归类到“人物/作者/图书作品”实体关系中,同时也未将“玲小旭”预测出来,造成实体三元组信息的不完整。从句子内容来看,本句的后半部分确实提及本书的作者,存在2个实体三元组,这说明LSTM-GCN-LSTM在处理多个实体之间存在2个实体关系三元组的实体重叠问题时仍有不足。

在含有数字的文本中,包含数字的实体关系三元组中数字实体重叠现象较为普遍,如图8所示,例句中的数字“50”存在于4个待提取的实体关系三元组中。数字作为特定领域(如能源领域)文本中的关键信息,提取其所描述的具体信息非常有必要。因此,实体重叠是后续工作中需要解决的重要问题。

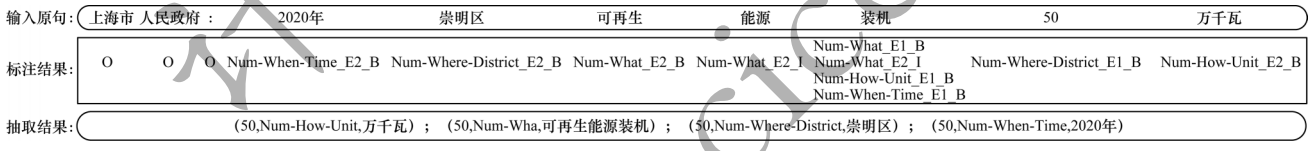

图8 含有数字实体关系的中文文本标注结果

Fig.8 Chinese text annotation results with digital entity relationship

4 结束语

本文提出一种基于GCN的中文实体关系联合抽取方法,利用GCN编码依存分析图中的先验词间关系信息,通过改进的标注策略标记实体关系,将实体关系联合抽取问题转化为序列标注问题,最终输出实体关系三元组。实验结果表明,GCN具有编码局部特征和先验词间关系的能力,联合抽取模型在加入GCN编码的信息后能够提高三元组中2个实体均被正确抽取的概率,从而提升网络性能。下一步尝试利用图网络在非欧空间上对拓扑关系的编码能力来解决实体重叠问题,从而提升模型的适用性。

参考文献

[1] LEE H J, WANG J S. Design of a mathematical expression understanding system [J]. Pattern Recognition Letters, 1997, 18(3): 289-298.

[2] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述[J]. 软件学报,2019,30(6):1793-1818.
E H H, ZHANG W J, XIAO S Q, et al. Survey of entity relationship extraction based on deep learning [J]. Journal of Software, 2019, 30(6): 1793-1818. (in Chinese)

[3] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme [M]. Stroudsburg, USA: Association for Computational Linguistics, 2017.

[4] 李明耀,杨静. 基于依存分析的开放式中文实体关系抽取方法[J]. 计算机工程,2016,42(6):201-207.
LI M Y, YANG J. Open Chinese entity relation extraction method based on dependency parsing [J]. Computer Engineering, 2016, 42(6): 201-207. (in Chinese)

[5] 孙紫阳,顾君忠,杨静. 基于深度学习的中文实体关系抽取方法[J]. 计算机工程,2018,44(9):164-170.
SUN Z Y, GU J Z, YANG J. Chinese entity relation extraction method based on deep learning [J]. Computer Engineering, 2018, 44(9): 164-170. (in Chinese)

[6] ZHANG Y H, QI P, MANNING C D. Graph convolution over pruned dependency trees improves relation extraction [C]//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2018: 2205-2215.

[7] PARK C, PARK J, PARK S. AGCN: attention-based graph convolutional networks for drug-drug interaction extraction [J]. Expert Systems with Applications, 2020, 159: 113538-113550.

[8] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures [M]. Stroudsburg, USA: Association for Computational Linguistics, 2016.

[9] KATIYAR A, CARDIE C. Going out on a limb: joint extraction of entity mentions and relations without dependency trees [M]. Stroudsburg, USA: Association for Computational Linguistics, 2017.

[10] 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. Long Beach, USA: Curran Associates Inc., 2017: 6000-6010.

[11] SHI X, YI Y, XIONG Y, et al. Extracting entities with attributes in clinical text via joint deep learning [J]. Journal of the American Medical Informatics Association, 2019, 26(12): 1584-1591.

[12] 马建红，李振振，朱怀忠，等. 反馈机制的实体及关系联合抽取方法[J]. 计算机科学，2019，46(12)：242-249.
MA J H, LI Z Z, ZHU H Z, et al. Entity and relationship joint extraction method of feedback mechanism [J]. Computer Science, 2019, 46(12): 242-249. (in Chinese)

[13] YAN Z, HUANG L T, GUO T, et al. An attention-based model for joint extraction of entities and relations with implicit entity features [M]. New York, USA: Assoc Computing Machinery, 2019.

[14] 曹明宇，杨志豪，罗凌，等. 基于神经网络的药物实体与关系联合抽取 [J]. 计算机研究与发展，2019，56(7)：1432-1440.
CAO M Y, YANG Z H, LUO L, et al. Joint drug entities and relations extraction based on neural networks [J]. Journal of Computer Research and Development, 2019, 56(7): 1432-1440. (in Chinese)

[15] 黄培馨，赵翔，方阳，等. 融合对抗训练的端到端知识三元组联合抽取 [J]. 计算机研究与发展，2019，56(12)：2536-2548.
HUANG P X, ZHAO X, FANG Y, et al. End-to-end knowledge triplet extraction combined with adversarial training [J]. Journal of Computer Research and Development, 2019, 56(12): 2536-2548. (in Chinese)

[16] HONG Y, LIU Y X, YANG S Z, et al. Improving graph convolutional networks based on relation-aware attention for end-to-end relation extraction [J]. IEEE Access, 2020, 8: 51315-51323.

[17] QIAN Y J, SANTUS E, JIN Z J, et al. GraphIE: a graph-based framework for information extraction [C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics, 2019: 751-761.

[18] YI L, WADDEN D, HE L H, et al. A general framework for information extraction using dynamic span graphs [C]// Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics, 2019: 3036-3046.

[19] WANG S L, ZHANG Y, CHE W X, et al. Joint extraction of entities and relations based on a novel graph scheme [C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Washington D. C., USA: IEEE Press, 2018: 4461-4467.

[20] PENG N Y, 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.

[21] KIPF T, WELLING M. Semi-supervised classification with graph convolutional networks [C]//Proceedings of the 5th International Conference on Learning Representations. Washington D. C., USA: IEEE Press, 2017: 1-14.

[22] GORI M, MONFARDINI G, SCARSELLI F. A new model for learning in graph domains [C]//Proceedings of 2005 IEEE International Joint Conference on Neural Networks. Washington D. C., USA: IEEE Press, 2005: 729-734.

[23] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model [J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61-80.

[24] 李自荐，迟呈英，战学刚. 基于Bi-LSTM与CRF的泰语句子切分模型[J]. 计算机工程，2020，46(10)：294-300.
LI Z J, CHI C Y, ZHAN X G. Thai sentence segmentation model based on Bi-LSTM and CRF[J]. Computer Engineering, 2020, 46(10): 294-300. (in Chinese)

[25] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2013: 6645-6649.

[26] DUVENAUD D, MACLAURIN D, AGUILERA I J, et al. Convolutional networks on graphs for learning molecular fingerprints [EB/OL]. [2020-08-25]. https://dash.harvard.edu/bitstream/handle/1/24873720/Convolutional;jsessionid=E32E6592300CFC9DF22D39C1CC3A3EF0? sequence=1.

[27] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [C]// Proceedings of NAACL International Conference. San Diego, USA: Association for Computational Linguistics, 2016: 260-270.

[28] VASWANI A, BISK Y, SAGAE K, et al. Supertagging with LSTMs[C]//Proceedings of NAACL International Conference. San Diego, USA: Association for Computational Linguistics, 2016: 232-237.

编辑　吴云芳