计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：基于有向图模型的旅游领域命名实体识别
作者：崔丽平，古丽拉·阿东别克，王智悦
DOI：10.19678/j.issn.1000-3428.0060062
网络首发日期：2021-02-07
引用格式：崔丽平，古丽拉·阿东别克，王智悦．基于有向图模型的旅游领域命名实体识别．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0060062

计 算 机 工 程
Computer Engineering

文献标识码：A

基于有向图模型的旅游领域命名实体识别

崔丽平[1,2,3]　古丽拉·阿东别克[1,2,3]　王智悦[1]

[1]（新疆大学 信息科学与工程学院　乌鲁木齐 830046）

[2]（新疆多语种信息技术实验室　乌鲁木齐 830046）

[3]（国家语言资源监测与研究少数民族语言中心哈萨克和柯尔克孜语文基地 乌鲁木齐 830046）

摘　要　旅游领域命名实体识别是旅游知识图谱构建过程中的重要一环。旅游文本中的实体与通用领域中的实体相比，具有长度较长、一词多义、嵌套严重的特点。针对此类问题，提出了基于有向图模型的命名实体识别方法。该模型首先将预训练词向量通过卷积神经网络（CNN）提取丰富的字特征，然后利用词典构造句子的有向图，生成邻接矩阵融合字词信息，最终将包含局部特征的词向量和邻接矩阵输入图神经网络（GNN）中提取全局语义信息，并引入条件随机场（CRF）得到最优的标记序列。此外，针对旅游领域数据集缺乏的问题，构建了命名实体识别数据集。实验结果表明，该模型在 Tourism 和 Resume 数据集上的识别效果均优于现有的主流模型，F1 值分别达到了 86%和 95.02%。

关键词　知识图谱　命名实体识别　卷积神经网络　图神经网络　条件随机场

NAMED ENTITY RECOGNITION IN TOURISM BASED ON DIRECTED GRAPH MODEL

Cui Liping[1,2,3]　**Gulila Altenbek**[1,2,3]　**Wang Zhiyue**[1]

[1](*College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China*)

[2](*Xinjiang Laboratory of Multi-language Information Technology, Urumqi 830046, China*)

[3](*The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Center on Minority Languages, Urumqi 830046, China*)

【Abstract】 Named entity recognition in the field of tourism is an important part in the construction of tourism knowledge graph.Compared with entities in the general field,such entities have the characteristics of random naming, longer length and serious nesting. To solve this problem, a named entity recognition method based on directed graph model is presented. Convolutional Nerual Network (CNN) was used to extract rich character feature vectors, then the directed graph of the sentence is constructed using the lexicon to generate an adjacency matrix, and finally, the concatenated word vectors and adjacency matrices are trained in the graph neural network (GNN), and conditional random field (CRF) is introduced to label the entity recognition results. In addition, a dataset is being constructed to address the lack of corpus in the field of tourism. The experimental results on Tourism and Resume datasets show that the proposed model achieves 86% and 95.02%.

【Key words】Knowledge graph　Named entity recognition　CNN　GNN　CRF

DOI:10.19678/j.issn.1000-3428.0060062

基金项目：国家自然科学基金项目（62062062）；新疆大学基金项目（BS 180250）

作者简介：崔丽平(1994-)，女，硕士研究生，主研方向：自然语言处理；古丽拉·阿东别克(通讯作者)，女，教授，博士；王智悦（1995-）男，硕士研究生

E-mail：cui_lip@163.com

1 概述

随着信息化建设的加快，人们生活水平的不断提高，旅游逐渐成为人们休闲放松的重要方式，同时，旅游业作为当今主要发展的产业之一，备受关注。对于出游在外的游客们，运用智能化的应用软件去解决出行的问题是十分便利的手段。例如景点的智能线路推荐，关于景区的智能问答系统的实现等，旅游领域的命名实体识别作为智能化服务的基础任务亟需解决。

命名实体识别（Named entity recognition, NER）是自然语言处理中的一项基础研究任务，是信息检索、问答系统、机器翻译等诸多任务的基础。以往的命名实体识别任务大多针对通用领域，近年来，NER 在某些特定的领域开始新的尝试，在生物医学领域，王浩畅[1]用 SVM 进行蛋白质、基因、核糖核酸等实体识别；社交媒体领域中,李源等[2]对微博中的实体进行研究；罗凌等[3]对电子病历中的实体进行研究，此外，还有一些研究较少的实体，如化学实体[4]、古籍文本中的人名[5]等。

旅游领域的命名实体识别研究相对较少。薛征山等[6]提出基于 HMM 的旅游景点识别方法，该方法首次在旅游领域上进行命名实体识别任务，但是没有充分考虑到上下文信息，面对一词多义的问题表现欠佳。因为很多实体在不同的语境中会代表不同的意思，例如：“玉门关”，在其他的文本中，指的是地名，在旅游文本中，指的是旅游景点“玉门关”。郭剑毅等[7]提出使用层叠条件随机场识别景点名的方法，却过于依赖人工特征的建立，而且规则制定要耗费大量的人力，以致于不能广泛使用。刘小安等[8]提出了一种基于 CNN-BiLSTM-CRF 的网络模型，避免了人工特征的构建，但是该方法是基于字进行识别，未能充分利用词典信息。对于特定领域的命名实体识别任务，词典是十分重要的外部资源，尤其是旅游文本中存在许多较长的景点名，例如“阿尔金山自然保护区”，“巴音布鲁克天鹅湖”等，利用词典可以获取这类词汇的信息，进而提高命名实体识别的准确率。

基于上述问题，本文提出了融合词典信息的一种有向图神经网络模型用于旅游领域中的命名实体识别任务。首先使用具有多个卷积核的 CNN 提取字特征，对不同大小的词组信息进行卷积，充分获得局部依赖信息，并且基于词典构建每个句子的有向图并生成对应的邻接矩阵，通过边的连接融合词汇特征与字特征，之后将词向量和邻接矩阵一同输入图神经网络（GNN）进行全局语义信息的提取，图神经网络通过边和节点的计算，使每个字充分的学习到邻居节点的信息和与其构成词汇的其它节点信息，最后使用 CRF 解决标注偏置问题，获得最优序列。

本文的主要贡献有：1）由于缺乏公开的旅游领域数据集，构建了新疆旅游领域命名实体识别的数据集。2）构建了旅游领域的字典，可用于旅游领域的一些相关任务。3）首次将图神经网络用于旅游领域中的命名实体识别任务，提出了基于有向图的命名实体识别模型（L-CGNN），有效融合了词典信息，在实体命名不规范、一词多义现象严重的旅游文本中取得了较高的识别准确率，实验结果明显优于主流的机器学习方法和深度学习方法。

2 相关工作

2.1 命名实体识别

命名实体识别任务常用的方法有基于规则和词典的方法，基于统计机器学习的方法，以及基于深度学习的方法。基于规则和字典的方法需要考虑数据的结构和特点，在特定的语料上取得了较高的识别效果，但是依赖于大量规则的制定，手工编写规则又需耗费时间和精力。基于机器学习的方法则具有较好的移植性，对未登录词也具有较好的识别效果。常用的机器学习模型有支持向量机（Suppotr Vector Machine,SVM）[9],隐马尔科夫（Hidden Markov Model,HMM）[10]，条件随机场（Conditional Random Field,CRF）[11]，最大熵（Maximum Entropy,ME）[12]等，这些方法都被成功地用来进行命名实体的序列化标注，然而都需要从文本中选择对该项任务有影响的各种特征，并且将这些特征加入到词向量中，所以对语料库的依赖性很高。

随着深度学习在图像和语音领域的广泛应用，深度学习的众多方法也被越来越多的被应用到自然语言处理任务中。Collobert 等[13]首次提出基于神经网络的命名实体识别方法，该方法使用了具有固定大小的窗口，通过窗口在字符序列上滑动以提取特征。由于窗口的限制，不能考虑到长距离字符之间的有效信息。循环神经网络（Recurrent Neural Network,RNN）的优势在于它通过记忆单元存储序列信息。但是在实际的应用中，RNN 的记忆功能会随着距离的变长而衰减，从而导致其丧失学习远距离信息的能力。Hochreiter 等[14]在 RNN 的基础上，提出了长短时记忆神经网络（Long Short Term Memory，LSTM），该方法利用“门”结构解决了梯度消失的问题，然而 3 个“门”单元导致了计算量的增加。门循环单元（Gated Recurrent Unit，GRU）[15]只用了两个“门”保存和更新信息，减少了训练参数，缩短了训练的

时间。由于单向的 RNN 不能满足命名实体识别任务的需求，Graves 等[16]提出了双向 LSTM 模型（BiLSTM）用于序列标注任务，通过不同方向充分学习上下文特征。Huang 等[17]首次提出 BiLSTM 与 CRF 结合的模型，用 CRF 规范实体标签的顺序。至此，BiLSTM+CRF 的结构成为了命名实体识别任务中的主流模型[18-19]。

近年来，深度神经网络飞速发展，Vaswani 等[20]提出一种基于注意力机制(Attention）的机器翻译模型 Transformer，它摒弃了之前传统的 Encoder-Decoder 模型必须结合 RNN 或者 CNN 的固有模式，使用完全基于注意力机制的方式。由于具有强大的并行计算能力和长距离特征捕获能力，因此在机器翻译、预训练语言模型等语言理解任务中表现出色，逐渐取代了 RNN 结构成为特征提取的主流模型。但是，在命名实体识别任务上，基于 Self-Attention 的 Transformer 编码器的效果却远不如 LSTM，因为 Self-Attention 虽然可以进一步获得字词之间的相关关系，却无法捕捉字词间的顺序关系，并且经过 Self-Attention 计算后相对位置信息的特性会消失。位置信息的丢失和方向信息的缺失势必会影响命名实体识别的效果[21]。

目前在英文的命名实体识别任务上，主要使用基于词的方法，但是，在中文命名实体识别任务中，由于中文存在严重的边界模糊现象，基于词的方法会导致歧义问题的产生，进而影响命名实体识别结果。已经有大量的研究证明，基于字的方法比基于词的方法更适合中文命名实体识别任务[22-23]。然而基于字的方法也存在无法提取词汇信息的缺陷，这些潜在词的信息对命名实体识别任务十分重要。所以，越来越多研究者专注于将构造字词结合训练的方法[24-26]。

Zhang 等[27]于 2018 年提出的 Lattice LSTM 结构，使用了词典动态的将字词信息送入 LSTM 结构中计算，在多个数据集上取得了最好成绩。然而 RNN 的链式结构和缺乏全局语义的特点决定了单纯的基于 RNN 的模型容易产生歧义，如图 1 所示，“市长”和“长江”两个词都共同包含“长”字，RNN 会严格按照字和词汇出现的顺序进行信息的传递，所以“长”会优先被划分到左边的“市长”一词中[28]，这显然是错误的。为了解决这个问题，本文使用图神经网络进行信息的传递，在每一次计算的时候，每个节点都会同时的获得与其相连的节点的信息，以削弱字符语序和匹配词序对识别的影响。

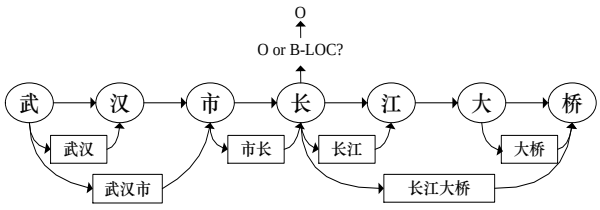

图 1 Lattice 结构

Fig. 1 Structure of lattice model

2.2 图神经网络

图是由一系列对象（节点）和关系类型（边）组成的结构化数据。图神经网络的概念最早由 Gori 等[29]在 2005 年提出。受到卷积网络在计算机视觉领域所获巨大成功的激励，近来出现了很多为图数据重新定义卷积概念的方法。Bruna 等[30]于 2013 年提出了关于图卷积网络的第一项重要研究，基于谱图论（spectral graph theory）的一种图卷积的变体。自此，图卷积网络不断改进、拓展、进阶，不断有新的方法被提出，包括图注意力网络[31]（GAT）、图生成网络[32]等。最近，图神经网络在自然语言处理领域的应用也逐渐成为一大热点，Yao 等[33]提出将图卷积神经网络（GCN）用于文本分类，Zhang 等[34]利用依存句法分析构建图神经网络用于关系抽取。

3 L-CGNN 旅游领域命名实体识别模型

模型的整体结构分为 3 个部分。特征表示层，GGNN 层，CRF 层。特征表示层的主要任务有两个，第一个任务是获取预训练词向量并使用具有不同卷积核的 CNN 进行局部特征的提取，充分获得每个字的局部特征，第二个任务是通过词典匹配句子中的词汇信息，构建句子的有向图结构，得到相应的邻接矩阵用来表示字与词汇的关系。GGNN 层接收特征表示层传入的词向量矩阵和邻接矩阵，动态地融合字词信息，获得全局的语义表示。最后通过 CRF 层进行解码，获得最优标签序列。完整的模型结构如图 2 所示。

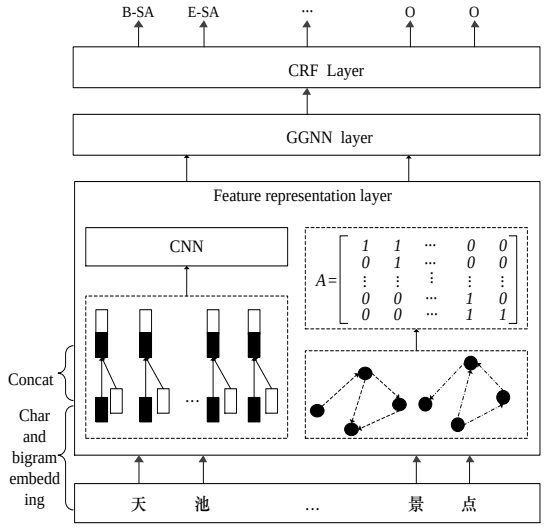

图 2 L-CGNN 完整模型

Fig. 2 The whole achitecture of L-CGNN

3.1 特征表示层

特征表示层首先对文本进行词向量表示，然后构建文本的图结构。

（1）词向量

神经网络的输入是向量矩阵,所以先将字转换成向量矩阵形式。给定包含 n 个字的句子 $S=\{c_1,\cdots,c_n\}$，其中 c_i 是第 i 个字，每个字通过查询预训练字向量表，转换为基于字的词向量:

$$x_i = E^c(c_i) \tag{1}$$

E^c 是预训练词向量表，此外，bigram 的引入，也已经被证明可以提高命名实体识别效果[35-36]。所以加入 bigram 特征后得到的词向量由三部分组成，基于字的向量，前向 bigram 词向量,后向 bigram 词向量，E^b 为预先训练的 bigram 向量矩阵。

$$x_i = \left[E^c(c_i); E^b(c_i,c_{i+1}); E^b(c_{i-1},c_i)\right] \tag{2}$$

对于旅游文本，因为实体名通常较长，并且嵌套现象严重，字向量和 bigram 向量并不能很好的表示局部信息。例如“天山大峡谷是新疆著名景点”，对于“山”字，除了字向量特征外，只能获取到“天山”和“山天”的信息，这就导致了“天山”很可被当作单独的一个景点名被识别，然而这里的“天山大峡谷”是一个完整的景点名，这就需要更多的信息辅助识别。

卷积神经网络(CNN)最初被应用于图像处理任务中，随后，逐渐被用于自然语言任务中局部特征的提取。通用的 CNN 结构包含卷积层、激活层、池化层，由于池化层会削弱位置特征的表达，而位置特征对于序列标注任务十分重要，所以本文没有使用池化操作，而是使用了 3 个不同大小的卷积核提取特征，为了获得相同维度的表示，进行了填充操作。三个卷积核的大小为 $k\times w$，其中 k 依次取 $1,3,5$，对应 w 依次取 $d, d+2, d+4$，d 为词向量 x_i 的维度，局部特征提取过程如图 3.

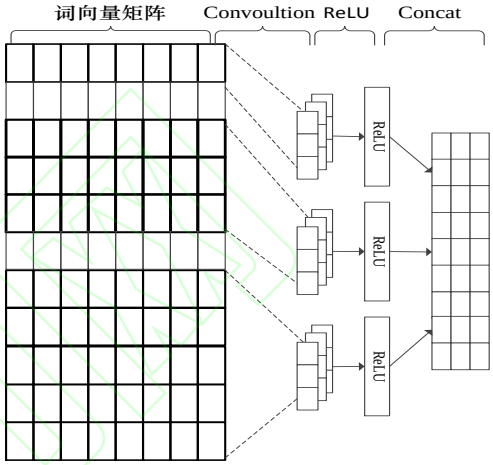

图 3 CNN 提取局部特征

Fig. 3 Schematic diagram of local feature extraction from CNN model

$$h_i^k = f(W^{conv}\cdot x_{i:i+k-1} + b) \tag{3}$$

$$h_i = h_i^1 \oplus h_i^3 \oplus h_i^5 \tag{4}$$

其中 $W^{conv}\in R^{k\times w}$，$f$ 是线性修正单元（Relu），b 是偏置项，将不同卷积核提取的局部特征拼接，得到最终的特征表示。

（2）文本图结构

对于一个有 n 个节点的图，可以用形状为 $n\times n$ 的邻接矩阵表示。本文中图结构的构建主要分为两个步骤。给定包含 n 个汉字的句子 $S=\{c_1,\cdots,c_n\}$，将句子中的每一个字作为图的节点。首先是所有相邻节点的连接，因为信息传递的方向性对于序列标注任务具有重要意义，所以在句子的第 i 个字和第 $i+1$ 个字之间，即 c_i，c_{i+1} 之间，都连接两条方向相反的边。其次，是词汇边的连接，若 i 和 j 是第 i 个字从字典中匹配到的词的开始节点和结束节点，我们在这两个节点之间连接两条方向相反的边，即令 $A_{i,j}=1, A_{j,i}=1$。如图 4 所示。

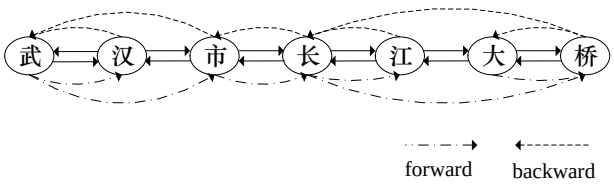

图 4 字词结合的有向图

Fig. 4 Word-Character containing directed graph

从上述有向图结构中可以看出，如果一个节点在字典中匹配到词汇数不止一个，则该节点和与之构成词汇的所有节点之间都存在相应的边，这样在后续的传递的过程中，可以同时学习所有词汇与字的信息，有效克服字或词汇的固有序列的影响。

3.2 基于门控机制图神经网络（GGNN）

门控图神经网络（GGNN）是一种基于 GRU 的经典空间域消息传递模型[37],与 GCN 等其他图神经结构相比，GGNN 在捕捉长距离依赖信方面优于 GCN，更适合于中文的命名实体识别任务。将特征表示层得到的词向量和邻接矩阵传入 GGNN 进行上下文语义的学习。信息传递过程如下:

$$h_i^{(0)} = h_i \tag{5}$$

$$a_i^{(t)} = A_{i:}^T [h_1^{(t-1)T} \cdots h_{|i|}^{(t-1)T}]^T + b \tag{6}$$

$$z_i^t = \sigma(W^z a_i^{(t)} + U^z h_i^{t+1}) \tag{7}$$

$$r_i^t = \sigma(W^r a_i^{(t)} + U^r h_i^{(t-1)}) \tag{8}$$

$$\tilde{h}_i^{(t)} = tanh(W a_i^{(t)} + U(r_i^{(t)} \odot h_i^{(t-1)})) \tag{9}$$

$$h_i^{(t)} = (1 - z_i^t) \odot h_i^{(t-1)} + z_i^t \odot \tilde{h}_i^{(t)} \tag{10}$$

$h_i^{(0)}$ 代表特征表示层获得的词向量矩阵，$A_{i:}$ 代表从邻接矩阵中，选定节点 i 对应的行向量，$h_i^{(t-1)}$ 代表 $t-1$ 时刻节点 i 的信息，$a_i^{(t)}$ 代表节点 i 在 t 时刻的状态信息，图中每个节点都通过该节点对应的传入边和传出边进行信息的传递。(7)- (9)则是普通的 GRU 更新信息过程。

3.3 CRF 层

条件随机模型可以看成是一个无向图模型或马尔科夫随机场，用来学习标签的约束，解决了标签偏置问题。对于给定的观察列，通过计算其整个标记序列的联合概率的方法获得最优标记序列。随机变量 $X = \{x_1, x_2, \cdots, x_n\}$ 表示观察序列，随机变量 $Y = \{y_1, y_2, \cdots, y_n\}$，表示相应的标记序列，$P(Y|X)$ 表示在给定 X 的条件下 Y 的条件概率分布，则条件随机场计算可表示为:

$$P(y|x) = \frac{\sum_{t=1}^{T} e^{f(y_{t-1}, y_t, x)}}{\sum_{y'}^{Y(x)} \sum_{t=1}^{T} e^{f(y'_{t-1}, y'_t, x)}} \tag{11}$$

其中，$Y(x)$ 表示所有可能的标签序列，$f(y_{t-1}, y_t, x)$ 用来计算 y_{t-1} 到 y_t 的转移分数和 y_t 的分数。

$$Y^* = argmax_y P(y|x) \tag{12}$$

最后使得 $P(y|x)$ 分数最大的标记序列 y，即句子所对应的实体标签序列。

4 实验

4.1 数据集

（1）旅游数据集（Tourism），由于目前还没有公认度较高的旅游领域数据集，本文从去哪儿网、携程、马蜂窝等旅游网站爬取有关新疆的旅游攻略，经过去除空白行、空格、非文本相关内容等预处理操作，得到旅游领域文本 1200 余篇。使用 NLTK 工具对预处理后的语料进行半自动化标注，之后进行人工校对、标注，构建了用于旅游领域实体识别的训练集，评估集和测试集，并且通过高德地图旅游景点数据和旅游网站检索构造了旅游景点词典。

针对旅游领域实体类型的定义，本文参考了郭剑毅[7]等人的分类标准，该文将旅游领域实体分为地名、景点名、特色美食三大类。考虑到新疆地域的特点，本文新增了人名、民族两种实体类型。采用 BIOES 标注体系进行实体的标注。例如:“天山大峡谷位于乌鲁木齐县境内”，按照采用的标注体系，可以标记为“天/B-SA 山/I-SA 大/I-SA 峡/I-SA 谷/E-SA 位/O 于/O 乌/B-LOC 鲁/I-LOC 木/I-LOC 齐/I-LOC 县/E-LOC 境/O 内/O ”。具体的数据统计见表 1。

表 1 Tourism 数据集详细统计

Table 1 Detailed statistics of tourism dataset

Type	Train	Dev	Test
SA(景点名)	2055	288	262
LOC(地名)	1571	177	184
SC(特色小吃)	139	13	26
NA(民族)	218	31	28
PER(人名)	193	32	40
合计	4176	541	540

（2）简历数据集(Resume)，由 Zhang 等[27]提出。该数据集共有 8 种不同的实体类型，CONT(country), EDU(educational institution), LOC, PER,ORG, PRO(profession), RACE(ethics background), 和 TITLE(job title)，两个数据的详细统计见表 2。

表 2 两种数据集的统计

Table 2 Statistics of two datasets

Dataset	Type	Train	Dev	Test
Resume	Sentences	3.8k	0.48k	0.46k
	Entities	1.34k	0.15k	0.16k
Tourism	Sentences	4.4k	0.5k	0.5k
	Entities	4.2k	0.54k	0.54k

实验使用的预训练词向量表来自于[38]，通用的词典来自于[27]，该字典包含 704.4k 个词，其中单个字有 5.7k，两个字构成的词有 291.5k，三个字构成的词有 278.1k，其它 129.1k。

此外，专门构建了旅游领域景点名词典。

4.2 对比模型

为了证明本文模型的有效性，使用了现有的应用于旅游领域命名实体识别任务的机器学习方法和几种主流的深度学习模型进行对比。

（1）**HMM[6]:**以 HMM 算法为原理，用于旅游领域命名实体识别任务。

（2）**CRF[7]:**使用层叠条件随机场方法解决旅游实体嵌套问题。

（3）**BiLSTM+CRF:**模型是命名实体识别任务的经典模型。

（4）**BiLSTM+CRF（+bigram）:**为了验证 bigram 对命名实体识别任务的作用，设计了包含 bigram 特征的 BiLSTM+CRF 模型进行对比分析。

（5）**Transformer+CRF[21]:**Transformer 由于强大的特征提取能力，在很多的自然语言处理任务中逐渐取代了 RNN 模型，所以本文加入了该模型的对比。

（6）**ID-CNN+CRF[24]:**膨胀卷积、空洞卷积，主要是通过扩大感受域的方法获得更广泛的序列信息。在英文命名实体识别任务上曾取得最佳成绩。

（7）**Lattice LSTM[27]:**该模型是字词结合训练的代表性方法，创造性地将字符和词汇通过网格的方法融合在一起，并且在 MSRA、Weibo、OntoNotes4、Resume 四个数据集上取得了最好成绩。

（8）**Bert+CRF:**Bert 作为一种预训练模型，在自然语言处理的多项任务中逐渐成为主流模型。

4.3 实验环境及参数设置

本文模型使用的 GPU 为 GeForce GTX 1080Ti，操作系统为 Ubuntu18.04，编程语言为 Python3.6，框架为 PyTorch 1.1.0。为了实体识别算法的一致性，设置了初始化参数，预训练词向量维度为 300，GGNN 神经元个数为 200，丢码率为 0.5，初始学习率为 0.001，衰减率为 0.05。

4.4 评价指标

评价指标采使用准确率（P），召回率（R）和 $F1$ 值，公式参数定义如下，TP 为正确识别的实体个数，FP 是识别出的不相关的实体个数，FN 是数据集中存在且未被识别出来的实体个数

$$P = \frac{TP}{TP + FP} \times 100\% \tag{13}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{14}$$

通常精确率和召回率的数值越高，代表实验的效果好。然而一般精确率和召回率会出现矛盾的情况，即精确率越高，召回率越低。所以需要综合考量他们的加权调和平均值，也就是 $F1$ 值，$F1$ 值定义如下：

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

4.5 实验结果与分析

在旅游领域命名实体识别数据集上，选择了 HMM、CRF 、 BiLSTM+CRF 、 BiLSTM+CRF(+bigram) 、Transformer+CRF、ID-CNN+CRF、Lattice LSTM、Bert+CRF 等模型进行实验，实验结果见表 3。

表 3 Tourism 上的实验结果（*Dic 为自建词典）

Table 3 Main results on Tourism (* Dic is tourism domain lexicon)

Models	P	R	F1
HMM[6]	68.37	67.77	67.93
CRF[7]	79.55	75.53	77.49
BiLSTM+CRF	83.30	77.10	80.08
BiLSTM+CRF(+bigram)	82.23	81.75	81.99
Transformer+CRF[21]	74.91	82.91	78.71
ID-CNN+CRF[24]	80.27	76.71	79.59
Lattice LSTM[27]	84.31	83.50	83.90
Bert+CRF	85.94	86.61	86.27
L-CGNN(ours)	**85.30**	**85.63**	**85.47**
L-CGNN(*Dic)	**85.80**	**87.96**	**86.86**

由表 3 可知，HMM 和 CRF 模型在旅游领域命名实体识别任务上的 P、R、F1 三个数值都低于其他深度学习模型，HMM 算法的原理决定其只依赖于当前状态和其对应的观察对象，序列标注问题不仅和单个词相关，同时与观察序列的

长度，单词的上下文等相关。CRF 模型解决了标注偏置问题，所以识别效果相较于 HMM 有很大程度的提高，但是因为不能充分捕捉上下文语义信息，尤其是在不规范的旅游文本上，所以识别效果同样不佳。

通过 ID-CNN+CRF 与比 BiLSTM+CRF 的对比，后者识别效果略好，原因是 BiLSTM 模型可以获得长距离依赖关系，加强对语义的理解，而 IDCNN 虽然通过扩大感受域的方法加强了距离关系的捕捉，但是还存在不足。BiLSTM+CRF 融合 bigram 特征后，对实体识别的效果略有提升，证明了加入 bigram 特征是可以提高命名实体识别效果的。

对比 Transformer+CRF 模型和 BiLSTM+CRF,可以看出，Transformer 在命名实体识别效果上，并不如 BiLSTM+CRF。因为 Transformer 在方向性、相对位置、稀疏性方面不适合 NER 任务。尽管引入了绝对位置编码的方式来弥补这一缺失。但是，在命名实体识别任务上，效果仍然不理想。

Lattice LSTM 模型是通过字典的方式将词汇信息与字符信息融合以提升命名实体识别效果，但是，由于其严格的序列学习的特性，每次都会按照匹配词出现的顺序学习，这就会 X 导致歧义现象的产生，所以实验效果不如 L-CGNN。

Bert+CRF 模型在该任务上的结果优于其他模型，是因为 Bert 利用 transformer 的编码器，提高了特征提取能力，获得了充分的上下文信息。然而，对于旅游领域，词典是非常重要的外部资源，它对于命名实体识别等任务具有十分重要的意义，所以，L-CGNN（*Dic）模型在旅游数据集上识别效果显然优于 Bert+CRF 模型。

本文提出的模型（L-CGNN）明显优于前几种模型，因为 L-CGNN 模型通过词典构建有向图结构，然后通过图神经网络获得语义信息，这样不仅融合了字符与词汇信息，还可以利用图特殊的结构传递，在每一次计算时，同时将节点匹配到的所有词汇信息融合，从而减少了词序导致的歧义现象。

为证明 L-CGNN 模型确实可以解决匹配词先后顺序对命名实体识别效果的影响，我们在公开的 Resume 数据集上进一步进行了实验，实验结果见表 4。

表 4 Resume 数据集上实验结果

Table 4 Main results on Resume

Models	P	R	F1
BiLSTM+CRF	93.71	93.74	93.73
BiLSTM+CRF（+bigram）	93.72	94.36	94.04
文献[24]	92.69	94.85	93.75
文献[21]	-	-	93.43
文献[27]	94.81	94.29	94.46
Bert+CRF	95.28	96.24	95.76
L-CGNN(ours)	**95.14**	**94.91**	**95.02**

表 4 的实验结果中，文献[12]中的 P、R 没有公布，所以未能获取，通过以上几组实验可以看出，L-CGNN 比其他的模型在 P、R、F1 三个值上的分数高于以往模型的模型。本文的模型略低于 BERT+CRF,主要是因为有向图结果依赖于字典的质量，通用的词典质量低于专有领域词典，未能取得像旅游领域上那样高于 BERT+CRF 的成绩。这组实验也进一步表明了本文提出的模型具有一定的泛化能力。

4.6 消融实验

为了探讨不同特征对实验效果的影响，我们设计了进一步的对比实验。分别去除某些特征，进行命名实体的识别，实验结果见表 5。

表 5.不同特征的有效性验证，其中 W/O 代表去除该特征，例如“W/O lexicon”代表去除字典信息。

Table 5. Validation of different features for experiments, where W / O means without the feature, e.g. "w / O lexicon" stands for removing lexicon feature.

Models	Tourism	Resume
Complete model(L-CGNN)	85.47	95.02
W/O lexicon	82.56	94.41
W/O bigram	82.69	94.56
W/O bigram+lexicon	81.38	94.39

从表 5 的结果可以看出，两个数据集上，如果去除字典特征，最终的识别效果较差，可见字典信息的重要性。同样的，在去除 bigram 特征的情况下，识别效果也会被削弱。尤其是在同时去除字典和 bigram 两个特征后，F1 值有了很大程度的降低，这就证明了加入的特征都是可以提升最终识别效果的。

4.7 收敛速率与资源消耗对比

（1）为了进一步说明本文所提模型的性能，本文对比了 BiLSTM+CRF、Lattice LSTM、和 L-CGNN 三种模型的收敛速度，收敛曲线见图 4。

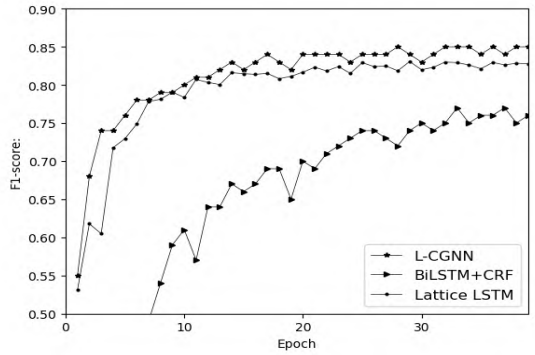

图 5 收敛速率对比图

Fig. 5 Comparison of convergence rate

通过图 5 的曲线，可以看出，在收敛速度上，L-CGNN 模型也优于其他模型，主要原因在于 BiLSTM+CRF 是通过双向 LSTM 学习，这就使得信息的更新比较慢，并且没有包含任何词汇特征，所以识别速率提升较慢。而 Lattice 和 L-CGNN 都包含字典外部信息，识别效果就相对较好，在一段时间后，L-CGNN 模型识别效果明显优于 Lattice LSTM，也说明本文模型在融合词汇方面更优。

（2）在资源消耗方面，本文从训练时间上分别在 HMM、CRF、Bilstm+CRF、Lattice LSTM、和 L-CGNN 等模型进行了对比实验，结果见表 6。

表 6　Tourism 数据集上时间对比

Table 6 Time per epoch of models on Tourism

Model	Time(s)/epoch
HMM[6]	1.54
CRF[7]	19
BiLSTM+CRF	150
Lattice LSTM[27]	615
L-CGNN	214

从表 6 中可以看出，HMM 和 CRF 是基于机器学习的方法，所以训练速度很快，但识别效果欠佳。对比 L-CGNN 与 BiLSTM+CRF，由于要进行邻接矩阵的构建，所以 L-CGNN 在训练上的时间消耗略大，对比融合词典的 Lattice LSTM 模型，本文模型的时间消耗远低于 Lattice LSTM，且取得了最优的识别效果。

5　结束语

本文针对旅游领域的命名实体识别任务，提出了基于字典构建文本的有向图结构的方法，该方法通过节点的连接，融合了字词信息，在图神经网络上动态地更新信息，有效的提高了在命名实体识别任务上的效果。此外，本文收集、整理了新疆旅游领域文本，建立了新疆旅游领域命名实体识别数据集。通过多组实验，证明了本文所提模型的有效性。下一步将继续探究更有效的图神经网络构建方法用于命名实体识别任务。

参考文献

[1] WANG Haochang,ZHAO Tiejun.SVM-based biomedical named entity recognition[J].Journal of Harbin Engineering University, 2006, 027(B07): 570-574.(in Chinese)

王浩畅,赵铁军.基于SVM的生物医学命名实体的识别[J].哈尔滨工程大学学报,2006,027(B07):570-574.

[2] LUO Ling, YANG Zhihao, SONG Yawen, LI Nan,et al. Research on naming entity recognition of Chinese electronic medical records based on stroke ELMo and multi-task learning [J].Chinese Journal of Computers, 2020, 43(10):1943-1957.(in Chinese)

罗凌,杨志豪,宋雅文,李楠,等.基于笔画ＥＬＭｏ和多任务学习的中文电子病历命名实体识别研究[J].计算机学报,2020,43(10):1943-1957.

[3] LI Yuan, MA Lei, SHAO Dangguo,et al. Chinese named entity recognition for social media [J]. Chinese Information Journal,2020,34(8):61-69.(in Chinese)

李源,马磊,邵党国,等.用于社交媒体的中文命名实体识别[J].中文信息学报,2020,34(8):61-69.

[4] ROBERT, LEAMAN, CHIH-HSUAN, et al. tmChem: a high performance approach for chemical named entity recognition and normalization[J].Journal of Cheminformatics,2015,7(S-1):S3.

[5] TANG Yafen.Study on automatic recognition of names in ancient Chinese classics before Qin Dynasty[J].Modern Library and Information Technology,2013,29(7-8):63-68.(in Chinese)

汤亚芬.先秦古汉语典籍中的人名自动识别研究[J].现代图书情报技术,2013,29(7-8):63-68.

[6] XUE Zhengshan, GUO Jianyi, YU Zhengtao, et al. Recognition of Chinese tourist attractions based on HMM[J]. Journal of Kunming University of Science and Technology (Science and Technology Edition). 2009, 34(6):44-48.(in Chinese)

薛征山,郭剑毅,余正涛,等.基于HMM 的中文旅游景点的识别[J].昆明理工大学学报(理工版) 2009,34(6):44－48.

[7] GUO Jianyi,XUE Zhengshan,YU Zhengtao,et al. Recognition of named entities in the tourism field based on stacked conditional random fields[J].

Journal of Chinese Information Processing, 2009,23(5):47-53.(in Chinese)
郭剑毅,薛征山,余正涛,等.基于层叠条件随机场的旅游领域命名实体识别[J]. 中文信息学报,2009,23(5):47-53.

[8] LIU Xiaoan, PENG Tao. Research on Chinese scenic spot recognition based on convolutional neural network[J]. Computer Engineering and Applications,2020,056(004):140-145.(in Chinese)
刘小安,彭涛.基于卷积神经网络的中文景点识别研究[J]. 计算机工程与应用,2020,056(004):140-145.

[9] EKBAL A,BANDYOPADHYAY S.Named entity recognition using support vector machine: A language independent approach[J].International Journal of Computer Systems ence & Engineering, 2010,4(3):589-604.

[10] SAITO K,NAGATA M.Multi-Language Named-Entity Recognition System based on HMM[C]//Proceedings of the Workshop on Multilingual and Mixed-language Named Entity Recognition(NER@ACL).Sapporo,Japan:ACM Press, 2003:41-48.

[11] LAFFERTY J,MCCALLUM A,PEREIRA F.Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C] //Proceedings of the Eighteenth International Conference on Machine Learning.MA,USA:ACM Press,2001:282-289.

[12] VARGA D,SIMON E.Hungarian named entity recognition with a maximum entropy approach[J]. Acta Cybernetica,2007,18(2):293-301.

[13] COLLOBERT R,WESTON J,BOTTOU,LÉON,et al.Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research, 2011,12(1):2493-2537.

[14] HOCHREITER S, SCHMIDHUBER J.Long short-term memory[J]. Neural Computation, 1997,9(8):1735-1780.

[15] CHO K,VAN MERRIENBOER B,GULCEHRE C,et al.Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.Doha,Qatar,IEEE Press:2014:1724-1734.

[16] GRAVES A, JÜRGEN SCHMIDHUBER.Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J].Neural Networks, 2005, 18(5-6):602-610.

[17] HUANG Z H,WEI X and KAI Y. Bidirectional LSTM-CRF Models for Sequence Tagging [EB/OL]. [2015-09-15]. https://arxiv.org/pdf/1508.01991.pdf.

[18] MA X, HOVY E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.Berlin,Germany: ACL, 2016: 1064-1074.

[19] HABIBI M, WEBER L, NEVES M L, et al. Deep learning with word embeddings improves biomedical named entity recognition[J]. Bioinformatics,2017,33(14):i37-i48.

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[C]//Proceedings of Advances in Neural Information Processing Systems.Long Beach,CA USA:ACM Press, 2017:6000-6010

[21] YAN H, DENG B, LI X, et al. TENER: Adapting Transformer Encoder for Named Entity

Recognition[EB/OL].[2019-12-02] .https://arxiv.org/pdf/1911.04474.pdf.

[22] HE J Z and WANG H F. Chinese Named Entity Recognition and Word Segmentation Based on Character[C]//Proceedings of the International Joint Conference on Natural Language Processing. Hyderabad,India:ACM Press,2008:128-132.

[23] LIU Z, ZHU C, ZHAO T. Chinese Named Entity Recognition with a Sequence Labeling Approach: Based on Characters, or Based on Words?[C] //Proceedings of Third International Joint Conference on Natural Language Processing. Changsha,China:Springer,2010: 128-132.

[24] GUI T, MA R, ZHANG Q,et al. CNN-Based Chinese NER with Lexicon Rethinking[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.Macao,China:ijcai.org, 2019:4982-4988.

[25] YAN X,YINING W,TIANREN L,et al. Joint segmentation and named entity recognition[J]. Journal of the American Medical Informatics Association,2013,21(e1): e84-e92.

[26] WU F,LIU J,WU C,et al. Neural Chinese Named Entity Recognition via CNN-LSTM-CRF and Joint Training with Word Segmentation[C]//Proceedings of the World Wide Web Conference.San Francisco,CA,USA:ACM Press,2019:3342 – 3348.

[27] ZHANG Y, YANG J. Chinese NER Using Lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia:ACM Press,2018:1554-1564.

[28] CHARLES A P, HAITAN L. The constituency model of Chinese word Identification[J],Reading Chinese Script: A Cognitive Analysis. 1999:115-134.

[29] GORI M, MONFARDINI G, SCARSELLI F, et al. A new model for learning in graph domains[C]// Proceedings of 2005 IEEE International Joint Conference on Neural Networks.Montreal, Que, Canada:IEEE Press.2005:729-734.

[30] BRUNA J , ZAREMBA W , SZLAM A , et al. Spectral Networks and Locally Connected Networks on Graphs[C]//Proceedings of 2nd International Conference on Learning Representations.Banff,AB,Canada:OALib Press.2014

[31] VELICKOVIC P,CUCURULL G,CASANOVA A, et al.Graph Attention Networks[EB/OL]. [2020-02-04]. https://arxiv.org/pdf/1710.10903.pdf.

[32] YOU J,REX,REN X,et al. GraphRNN: A Deep Generative Model for Graphs [EB/OL]. [2020-06-23]https://arxiv.org/pdf/1802.08773.pdf.

[33] YAO L, MAO C, LUO Y, et al. Graph Convolutional Networks for Text Classification[C]//Proceedings of national conference on artificial intelligence. Honolulu, Hawaii:AAAI Press,2019:7370-7377.

[34] ZHANG Y, QI P, MANNING C D, et al. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL,2018:2205-2215.

[35] CHEN X, QIU X, ZHU C,et al. Long Short-Term Memory Neural Networks for Chinese Word Segmentation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.Lisbon, Portugal:ACL.2015:1197-1206.

[36] LIN C Y, XUE N, ZHAO D, et al. Character-Based {LSTM-CRF} with Radical-Level Features for Chinese Named Entity Recognition[C] //Proceedings of Natural Language Understanding and Intelligent Applications - 5th Conference on Natural Language Processing and Chinese Computing and 24th International Conference on Computer Processing.Kunming,China:Springer. 2016:239-250

[37] LI Y J, TARLOW D, BROCKSCHMIDT M,et al. Gated Graph Sequence Neural Networks[EB/OL].[2020-09-22]. https://arxiv.org/pdf/1511.05493.pdf.

[38] LI S, ZHAO Z, HU R,et al. Analogical Reasoning on Chinese Morphological and Semantic Relations[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia:ACL,2018:138-143.