计算机科学
COMPUTER SCIENCE

CFGT:一种基于词典的中文地址要素解析模型

黄威，沈耀迪，陈松龄，傅湘玲

引用本文

黄威，沈耀迪，陈松龄，傅湘玲. CFGT:一种基于词典的中文地址要素解析模型[J]. 计算机科学，2024，51(9)：233-241.

HUANG Wei, SHEN Yaodi, CHEN Songling, FU Xiangling. CFGT:A Lexicon-based Chinese Address Element Parsing Model [J]. Computer Science, 2024, 51(9): 233-241.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900159

CFGT:一种基于词典的中文地址要素解析模型

黄 威 沈耀迪 陈松龄 傅湘玲
北京邮电大学计算机学院(国家示范性软件学院) 北京 100876
可信分布式计算与服务教育部重点实验室 北京 100876
(huangwei@bupt.edu.cn)

摘 要 地址要素解析作为地理编码过程中的关键环节,直接影响到地理编码的准确性。由于中文地址表达的多样性和复杂性,两段相似的地址文本在地理表示上却可能完全不同。传统的通过词典匹配进行地址要素解析的方法无法较好地应对歧义词,从而导致识别准确率欠佳。文中提出一种基于词典的中文地址要素解析模型(Collaborative Flat-Graph Transformer,CFGT),利用自匹配词、最近上下文等词汇信息增强地址文本字符序列表示,有效遏制了地址文本表达的歧义性。具体地,模型首先构建 Flat-Lattice 和 Flat-Shift 两种协作图,为地址字符捕获自匹配词和最近上下文词汇的知识,并设计融合层实现图之间的协作;其次,通过改进的相对位置编码,进一步强化词信息对地址文本字符序列的增强效果;最后,利用 Transformer 和条件随机场进行地址要素解析。在 Weibo 和 Resume 等多个公开数据集及 Address 私有数据集上开展的实验表明,CFGT 模型的性能优于已有的中文地址要素解析模型和中文命名实体识别模型。

关键词:中文地址识别;词典强化;外部信息;命名实体识别
中图分类号 TP391

CFGT:A Lexicon-based Chinese Address Element Parsing Model

HUANG Wei,SHEN Yaodi,CHEN Songling and FU Xiangling
School of Computer Science(National Pilot Software Engineering School),Beijing University of Posts and Telecommunications,Beijing 100876,China
Key Laboratory of Trustworthy Distributed Computing and Service(BUPT),Ministry of Education,Beijing 100876,China

Abstract As a key step in the geocoding process,address element parsing directly affects the accuracy of geocoding. Due to the diversity and complexity of Chinese address expressions,two similar address texts may be completely different in geographical representation. Traditional address element parsing based on dictionary matching cannot handle ambiguous words well,thus showing poor recognition accuracy. A lexicon-based Chinese address element parsing model CFGT:collaborative flat-graph transformer is proposed,which uses self-matched words,nearest contextual and other lexical information to enhance the character sequence representation of address text,effectively curbing the ambiguity of address text expression. Specifically,the model first constructs two collaboration graphs,flat-lattice and flat-shift,to capture the knowledge of self-matched words and nearest contextual words for address characters,and designs a fusion layer to implement collaboration between graphs. Secondly,with the help of the improved relative position encoding,the enhancing effect of word information on the address text character sequence is further strengthened. Finally,Transformer and conditional random fields are used to analyze address elements. Experiments are conducted on multiple public datasets such as Weibo and Resume,as well as the private dataset Address. Experimental results show that the performance of the CFGT is superior to previous Chinese address element parsing models and existing models in the field of Chinese named entity recognition.

Keywords Chinese address recognition,Lexicon enhancement,External information,Named entity recognition

1 引言

地址作为一种重要的文本数据,记录了社会生产活动及人的行为活动对应的地理空间信息,在地理商业智能、城市治理、金融风控等领域有着重要的应用价值[1-3]。随着互联网、大数据技术以及地理信息系统的快速发展,公众对位置信息的需求迅速增加[4]。据统计,世界上大约 70%的网页包含位置信息,其中大部分位置信息以非结构化文本的形式表达[5]。为了更好地向公众提供基于空间位置的大数据分析及其他服务,如何将这些非结构化的位置信息精准地转换为空间坐标,

到稿日期:2023-09-28 返修日期:2024-03-14
基金项目:国家自然科学基金(72274022)
This work was supported by the National Natural Science Foundation of China(72274022).
通信作者:傅湘玲(fuxiangling@bupt.edu.cn)

已经成为亟待解决的重要问题[6]。

地理编码是一种基于空间定位技术的编码方法，被视为在地址文本和空间坐标之间建立联系的一种最常用的有效方法[7]。它通过地址要素解析、地址标准化、地址匹配、空间定位实现非结构化地址文本到空间坐标的转换过程。其中，地址要素解析包括将非结构化地址文本拆分为地址元素，并对其分类的过程，是地理编码方法中最关键的步骤，其解析效果直接影响地理编码的准确性[8]。假设有中文地址文本“吉林省吉林市昌邑上海路”，地址要素解析首先将其拆分为若干个地址元素：“吉林省”“吉林市”“昌邑”“上海路”。然后将每个元素分到它对应的类别，即省份(Province)：吉林省；城市(City)：吉林市；区县(District)：昌邑区；街道(Road)：上海路。

然而，由于中文地址表达的多样性和复杂性，中文地址要素解析成为中文命名实体识别(Chinese Named Entity Recognition，CNER)任务中一个复杂而困难的研究问题。其难点主要表现在3个方面：(1)中国幅员辽阔，地理文化差异大，至今没有形成一个权威的、覆盖全国的统一命名标准；(2)中文地址数据结构复杂，随意性较强，多伴随地址要素的缺失或冗余，往往语义模糊且存在歧义[9]；(3)相比于其他大多数以空格为词与词之间分隔符的语言(尤其是英文)，中文的句子只能由标点符号作为词语之间的划分标志，这大大增加了识别难度。此外，中国城市的现代化发展迅速，地址更新迭代快[10]，简单地使用关键字匹配进行地址识别很难奏效。

目前中文地址要素解析任务的研究发展情况可以概括为3个阶段：词典匹配、机器学习以及深度学习方法。基于词典匹配的方法依托语料、词典等数据对地址进行解析拆分，具有简单易构造的优点。Zhang等[11]基于大规模地名词典和地址数据库，提出了中文地址的数字表达方法。Zhao等[12]基于整词二分分词词典，采用FMM算法实现了对地名地址串的有效拆分。机器学习方法针对词典匹配方法存在的未登录词和歧义识别等扩展性差的问题，进一步提升了地址识别能力。Duan等[13]通过构建语料特征模板，基于条件随机场(Conditional Random Field，CRF)实现了中文地址行政区划提取，地址信息解析的准确率达到了89.93%。Wang等[14]引入隐马尔可夫模型(Hidden Markov Model，HMM)，构建了一种地名地址语义解析及地址空间化的方法，该方法达到了较高的识别精度。然而，机器学习方法的解析效果受到特征设定的限制，其识别性能仍然不够理想。因此，深度学习方法被用于进一步改进和提升中文地址要素解析的效率和计算性能。Cheng等[15]提出了融合双向长短时记忆网络(Bidirectional Long Short-Term Memory，BiLSTM)和CRF的中文层级地址分词模型，利用BiLSTM获取地址字符序列的上下文信息，通过CRF的转移概率矩阵控制地址标注的输出，实现了较高的识别精度。Li等[16]首先利用Jieba分词工具获得地址字符序列的分词结果，再基于双向门控循环单元(Bi-Gated Recurrent Unit Recurrent Neural Networks，Bi-GRU)生成基于中文地址分词的标签特征，最终通过CRF推断地址分词的标签。尽管Bi-GRU在识别效果上优于BiLSTM，但是错误的地址分词会传递给Bi-GRU，影响标签推断的准确性，模型性能很大程度上受限于中文分词工具的性能。Liu等[17]利用卷积神经网络(CNN)提取地址文本的局部信息特征，并结合BiLSTM-CRF的网络结构获取地址文本的上下文信息，对中文景点的识别F1值达到了93.9%。进一步地，基于Transformer的预训练模型，如BERT[18]，在CNER领域表现出了优异的性能。Zhang等[19]在BERT之上增加了一个CRF层，在其地址数据集上达到了95%的F1值。Sun等[20]使用BERT获取中文地址的抽象特征，结合BiLSTM-CRF，根据文本特征进行地址标签识别，同样取得了良好的结果。

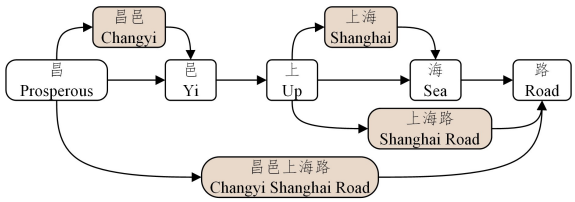


图1 词增强的Lattice结构

Fig. 1 Lexicon-enhanced Lattice structure

虽然上述方法在地址识别任务中取得了显著进步，但它们仍然在融合地址文本的字符信息和词信息方面存在不足。(1)早期简单利用词典进行模糊匹配的方法，其匹配精度以及分类准确率较低，已经不适用于现阶段的应用场景；而先分词再推断标签的方法，其性能非常依赖于分词效果的好坏。(2)基于字符的中文地址识别方法没有利用词信息，而Zhang等[21]的研究表明，词信息融入对基于字符的CNER方法具有重要价值。如图1所示，“上海路”及“昌邑上海路”能够用来消除上下文中潜在的地址实体的歧义，避免将吉林市的昌邑区上海路识别为昌邑市或上海市。(3)基于Transformer进行词增强的方法没有考虑最近的上下文词汇信息对字符的增强作用。字符的最近上下文词汇表示与该字符在句子中最近的前或后子序列相匹配的词汇，例如，字符“上”的最近上下文词汇是词“昌邑”，通过“昌邑”对“上”的增强，模型会预测“上”的标签为“B-road”而不是“B-city”。然而，以Lattice LSTM[21]以及Flat-Lattice Transformer(FLAT)[22]为代表的模型仅通过对上下文的学习隐式地整合最近上下文词汇的知识，该过程仍易受到其他信息干扰。

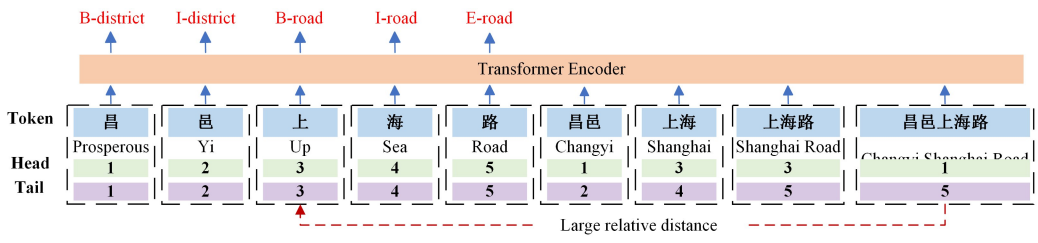


图2 词增强的Flat-Lattice结构

Fig. 2 Lexicon-enhanced Flat-Lattice structure

为了解决已有方法在字符信息、词信息融合方面表现不佳的问题，本文提出一种基于词典的中文地址要素解析模型CFGT。模型一方面构建了中文地址文本的Flat-Lattice图结构，为字符捕获自匹配词和最近上下文词汇的部分信息；另一方面，模型创新性地提出Flat-Shift图结构，建立起字符和最近上下文词汇之间的直接联系。Flat-Lattice图和Flat-Shift图的构建过程中没有利用外部分词工具，避免了错误的传播。同时，这些图是互补的，因此模型设计了一个融合层实现图之间的协作。此外，如图2所示，由于Lattice结构的限制，Lattice LSTM只能利用词“昌邑上海路”增强词尾字符“路”。FLAT通过设计相对位置编码改进了自匹配词对词中字符的增强效果，但仍然存在无法有效增强词中非首尾字符的问题。比如，词“昌邑上海路”与该词中非首尾字符“上”具有较大的相对距离（相对距离越大，相关性越小）。针对这一问题，CFGT提出改进的相对位置编码，用于有效调控字符及相应匹配词汇间的相对距离，保证字符能够被其自匹配词有效增强。最后，进一步探索了汉字字形及拼音信息对模型性能的影响。本文的主要贡献如下：

（1）提出了一种基于词典的中文地址要素解析方法CFGT。该方法通过融合协作平面图（Collaborative Flat-Graph）实现了词信息对地址文本字符序列的增强。据调研，本文是首个使用词信息增强基于字符的中文地址要素解析方法。

（2）提出了一种改进的相对位置编码，解决了FLAT不能有效利用匹配词增强词中非首尾字符的问题。使用地址数据集以及CNER领域的部分数据集对模型进行测试，结果表明所提模型优于已有的中文地址要素解析模型和CNER领域现存的先进（State Of The Art，SOTA）模型。

（3）探索了汉字字形及拼音信息对地址识别性能的影响，提出了一种融合词、字形及拼音信息的地址识别改进模型。使用CNER领域的部分数据集和构建的虚假地址数据集对模型进行测试，实验结果证明汉字字形和拼音信息对提升中文地址识别性能具有积极作用。

（4）在中文地址领域发布了一个新数据集Address，其中包含了17 450条来自全国各省市的带标注地址数据。

2　相关工作

中文地址要素解析与CNER方法密切相关。近年来，增强基于字符的CNER方法主要概括为两类：（1）将词信息集成到基于字符的序列编码器，以明确地建模词特征；（2）集成大规模的预训练上下文嵌入表示，例如BERT，它已被证明能捕获隐含的词级别的语法和语义知识[23]。此外，还有一些研究综合考虑了这两方面。

2.1　基于词增强的方法

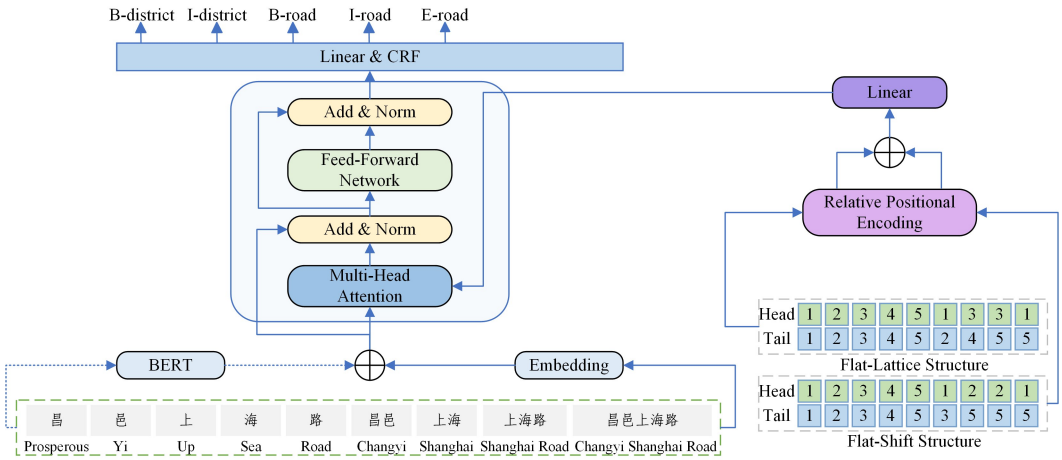
基于词增强的方法指利用词信息增强基于字符的模型。Zhang等[21]首先提出了使用词信息增强字符序列的Lattice LSTM，其中，使用额外的词cell单元编码自匹配词，并使用注意力机制将可变数目的自匹配词与词尾字符融合。然而，基于RNN的模型很难对长距离关系建模，且Lattice结构本身的局限性使得自匹配词只能对词尾字符增强。针对Lattice的不足，后续的工作在上下文信息学习[22]、图结构[24-25]、词增强模型向基于词的模型的退化[26]等方面做出了改进。其中，FLAT[22]基于位置信息将Lattice结构转化为Flat-Lattice结构，并通过自注意力机制的一种变体[27]来利用相对位置编码。这项工作是2020年CNER在4个主要公开数据集上的SOTA模型。

2.2　基于预训练模型的方法

基于预训练模型的方法考虑了对大规模预训练上下文信息的集成，这类方法以BERT为代表。Hu等[28]证明使用BERT的CNER模型优于使用静态嵌入的方法。最近的工作试图整合词增强方法和预训练模型。Ma等[29]拼接字符特征、BERT表示和词信息，并将其输入到LSTM中进行特征融合。Liu等[30]在BERT的Transformer层之间融入词信息，这项工作是2021年CNER在4个主要公开数据集上的SOTA模型。

3　中文地址要素解析模型

CFGT模型的主要结构如图3所示。



注：虚线的部分表示加入BERT的上下文表示。

图3　CFGT模型结构

Fig. 3　CFGT model structure

与FLAT[22]相比，CFGT主要在3个方面进行了改进。首先，使用了改进的相对位置编码来实现自匹配词对词中非首尾字符的有效增强。其次，构建Flat-Shift结构，帮助字符捕获其最近上下文词汇的语义信息。最后，将Flat-Shift结构

对应的相对位置编码与 Flat-Lattice 结构的相对位置编码进行融合，有效集成词汇信息。具体地，对于给定的中文地址文本，模型希望获取自匹配词汇（Self-Matched Words，SMWs）和最近上下文词汇（Nearest Contextual Words，NCWs）的相关信息。构建 Flat-SMWs（即图中的 Flat-Lattice 结构），建模汉字字符与其自匹配词汇的直接信息交互，以获取字符的自匹配词的语义信息和最近上下文词的部分信息；构建 Flat-NCWs（即图中的 Flat-Shift 结构），建模汉字字符及其最近上下文词汇的直接信息交互，最大限度地利用地址实体的级联特性。然后，CFGT 通过一个融合层来实现这两个协同平面图之间的协作。此外，针对非首尾字符不能被其自匹配词汇有效增强的问题，CFGT 设计了一种新的自适应的相对位置编码方案，以自适应地调节字符与其自匹配词汇之间的相对距离，确保字符得到充分增强。

3.1 改进的相对位置编码

Flat-Graph 结构由不同长度的跨度组成，各跨度包含了字符或词本身的位置信息。Flat-Graph 结构中的两个不同跨度之间存在着 3 种相对位置关系：相交、包含和分离[22]。为了对跨度之间的相对位置关系进行编码，并避免出现 FLAT 的相对位置编码中无法有效增强词中非首尾字符的问题，CFGT 使用密集向量来建模两个跨度之间的相对位置关系，以期更有效地利用自匹配词增强词中字符。具体地，该密集向量由不同跨度的头尾信息经过简单非线性变换得到。令 $head[i]$ 和 $tail[i]$ 表示跨度 x_i 的头和尾位置，两个跨度 x_i 和 x_j 之间的相对位置关系可以由 4 种相对距离来表示：

$$d_{ij}^{(hh)}=head[i]-head[j] \tag{1}$$

$$d_{ij}^{(ht)}=head[i]-tail[j] \tag{2}$$

$$d_{ij}^{(th)}=tail[i]-head[j] \tag{3}$$

$$d_{ij}^{(tt)}=tail[i]-tail[j] \tag{4}$$

$$[d_{ij}^{(hh)},d_{ij}^{(ht)},d_{ij}^{(th)},d_{ij}^{(tt)}]=e^{w}[d_{ij}^{(hh)},d_{ij}^{(ht)},d_{ij}^{(th)},d_{ij}^{(tt)}](w\leqslant 0) \tag{5}$$

$$w=d_{ij}^{(hh)}\cdot d_{ij}^{(tt)} \tag{6}$$

其中，$d_{ij}^{(hh)}$ 表示 x_i 的头部和 x_j 的头部之间的距离，$d_{ij}^{(ht)}$，$d_{ij}^{(th)}$，$d_{ij}^{(tt)}$ 的含义类似；w 是 $d_{ij}^{(hh)}$ 与 $d_{ij}^{(tt)}$ 的乘积。为了实现对词中非首尾字符的有效增强，对于具有包含关系（$w\leqslant 0$）的跨度 x_i 和 x_j，将它们的相对距离同时乘 e^{w} 来进行缩放。w 越小，缩放的效果越明显。最后，对 4 种相对距离做非线性变换，得到最终的相对位置编码：

$$\boldsymbol{R}_{ij}=\mathrm{ReLU}(W_r(PE_{d_{ij}^{(hh)}}\oplus PE_{d_{ij}^{(ht)}}\oplus PE_{d_{ij}^{(th)}}\oplus PE_{d_{ij}^{(tt)}})) \tag{7}$$

其中，W_r 是可学习的参数，$\oplus$ 代表张量的连接操作，PE_d 的计算方式与 Vaswani 等[31]的方法一致：

$$PE_d^{(2i)}=\sin(d/10\,000^{2i/d_{\text{model}}}) \tag{8}$$

$$PE_d^{(2i+1)}=\cos(d/10\,000^{2i/d_{\text{model}}}) \tag{9}$$

其中，PE_d 的维度与字符嵌入的维度 d_{model} 保持一致，i 表示位置编码维度的索引。

3.2 Flat-Shift 结构

Flat-Graph 结构可由有向无环图结构转换得到，因此，每一个 Flat-Graph 结构都对应了一种图结构，并且两者之间能够相互转换。Sui 等[25]通过多种图结构的融合来整合词汇信息，其中的 T-graph（Word-Character-Transition graph）建图方式能够有效地帮助字符捕获最近上下文词汇的语义信息。受此启发，本文提出了 Flat-Shift 结构来融合最近上下文词汇信息。

Flat-Lattice 结构可以有效捕捉自匹配词的边界和语义信息来对字符进行增强。此外，中文字符在句子中所对应的最近上下文词汇信息也有助于理解该字符。地址“四川攀枝花仁和区”的部分 T-graph 结构如图 4 所示。字符“花”属于“攀枝花”实体，但由于“攀枝花”既可能代表一个市，也可能代表一种花，因此字符“花”对应的标签可能为“E-city”或“O”。在加入了其最近上下文词汇“仁和区”的语义信息后，标签“E-city”将会取代“O”，因为区级实体前为市级实体的可能性更大。Flat-Shift 结构通过对 T-graph 结构的转换来帮助字符捕获其最近上下文词汇的语义信息，由图结构到对应的 Flat-Shift 结构的转换过程如图 5 所示。首先，将头尾位置相同的单个字符构成字符跨度序列，然后使用词语以及它们的头尾位置信息构建词语跨度序列，最后将两部分拼接起来得到 Flat-Shift 结构。

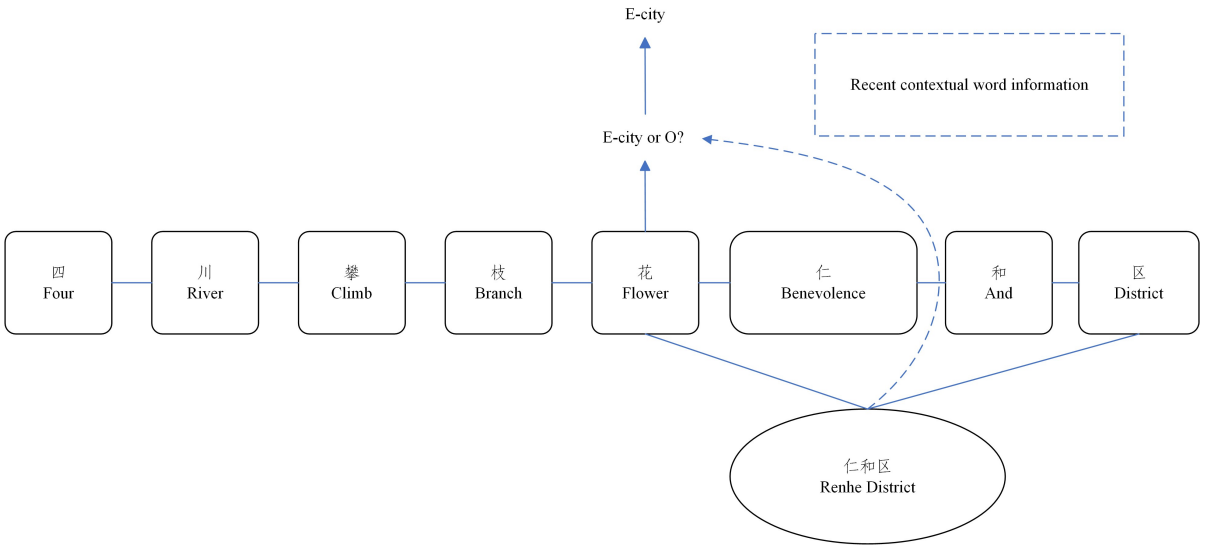


图 4 最近上下文词汇对字符的增强

Fig. 4 Character enhancement by nearest contextual words

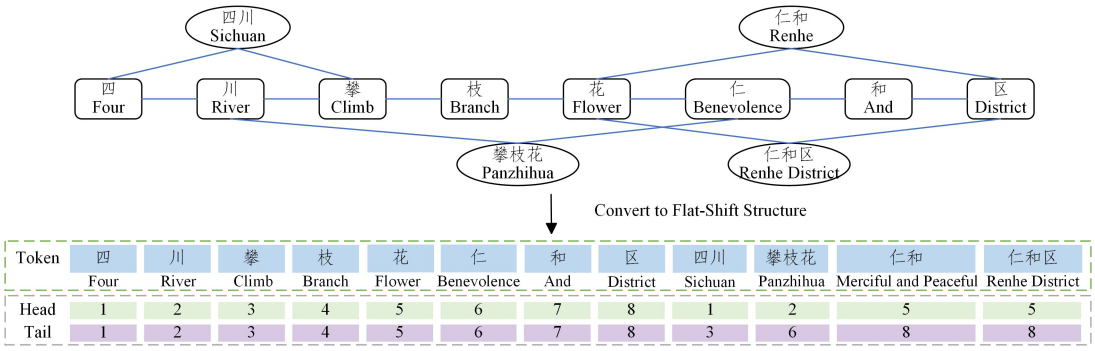


图 5　图结构转换为 Flat-Shift 结构

Fig. 5　Convert graph structure to Flat-Shift structure

3.3　相对位置编码融合

Flat-Lattice 与 Flat-Shift 结构包含了自匹配词和最近上下文词的边界和语义信息,因此我们通过融合这两种 Flat-Graph 结构的相对位置编码来集成两种词汇信息。

假设 R_{ij}^{l} 和 R_{ij}^{s} 分别是 Flat-Lattice 结构与 Flat-Shift 结构对应的相对位置编码,两者的融合过程如下:

$$\boldsymbol{R}_{ij}^{*}=W_t(\boldsymbol{R}_{ij}^{l}\oplus\boldsymbol{R}_{ij}^{s}) \tag{10}$$

其中,W_t 是可学习的参数。在自注意力机制的变体中,利用融合后的相对位置编码信息得到不同跨度之间的注意力权重:

$$Att(\boldsymbol{A},\boldsymbol{V})=\mathrm{softmax}(\boldsymbol{A})\boldsymbol{V} \tag{11}$$

$$\boldsymbol{A}_{ij}=\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{K}_j+\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{W}_R\boldsymbol{R}_{ij}^{*}+\boldsymbol{u}^{\mathrm{T}}\boldsymbol{K}_j+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{W}_R\boldsymbol{R}_{ij}^{*} \tag{12}$$

$$[\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}]=E_x[\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V] \tag{13}$$

其中,$\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V,\boldsymbol{W}_R\in\mathbb{R}^{d_{\mathrm{model}}\times d_{\mathrm{head}}}$ 和 $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^{d_{\mathrm{head}}}$ 都是可学习的参数。

将经过 Transformer Encoder 得到的新字符表示作为输出层的输入,并使用条件随机场(CRF)提高最终预测标签序列的可靠性。

3.4　增强模型 CFGT-gp

本节在 CFGT 模型的基础上,利用汉字的字形信息和拼音信息增强基于字符的中文地址识别能力,提出融合词、字形及拼音增强学习的中文地址识别改进模型 CFGT-gp(Collaborative Flat-Graph Transformer with glyph and pinyin)。模型的整体架构如图 6 所示。该模型共包含 7 个子模块:协同平面图构建模块(Collaborative Flat-Graph Construction Module)、自适应相对位置编码模块(Inner-Character Augmented Positional Encoding Module)、相对位置编码前融合模块(Positional Encoding Fusion Module)、字形提取模块(Glyph Module)、拼音提取模块(Pinyin Module)、特征融合模块(Embedding Module)和模型训练和输出模块(Tranformer Predictor Module)。其中,字形提取模块和拼音提取模块是 CFGT-gp 相较于 CFGT 增加的部分,它们分别利用全连接层和卷积神经网络实现对字形及拼音信息的提取。

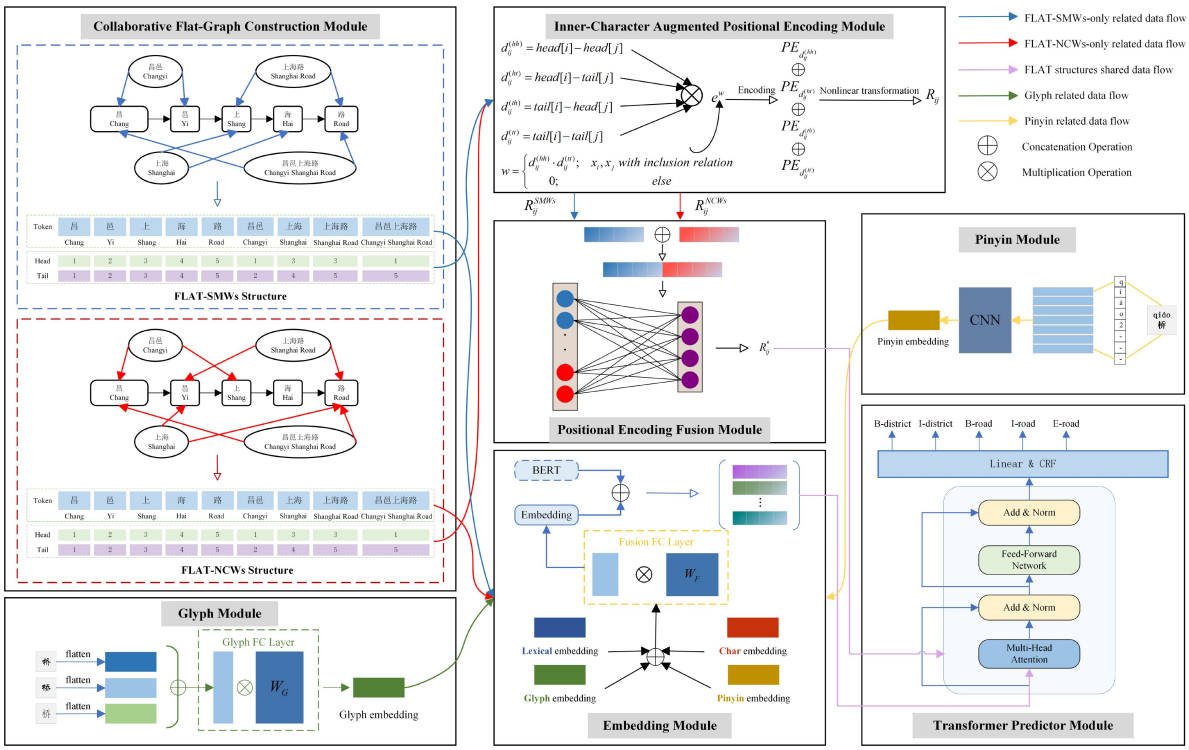


图 6　CFGT-gp 模型架构图

Fig. 6　CFGT-gp architecture

以输入“昌邑上海路”为例，模型首先利用协同平面图构建模块（Flat-SMWs 和 Flat-NCWs）对字词交互图进行网格化，构建两种类型的平面图：Flat-Lattice 以及 Flat-Shift。其次，利用自适应相对位置编码模块和相对位置编码前融合模块，得到位置敏感的自适应的相对位置编码，即 $\boldsymbol{R}_{ij}$。然后，特征融合模块接收字形提取模块和拼音提取模块输出的字形及拼音嵌入，与字符的原始嵌入和词嵌入融合得到最终的字符表征。最后，将相对位置编码和特征融合模块得到的表征作为 Tranformer-xl[27] 的输入，训练模型对中文地址进行分类，输出最终结果。

3.4.1 汉字字形特征提取

字形特征提取模块利用仿宋、行楷及隶书 3 种字体的字形增强文字信息。具体来说，实验首先将 3 种字体实例化为具有 0～255 浮点像素的 24×24 图像，然后使用全连接层将 24×24×3 维的特征向量转化为一维向量并降维，得到汉字字符的字形嵌入。字形特征提取过程如图 6 中的 Glyph Module 部分所示。

3.4.2 汉字拼音特征提取

拼音特征提取模块利用开源包 pypinyin 生成汉字的拼音序列，以数字 1—4 的形式表示汉字的声调并将其附加到拼音序列末尾，包括：第一声阴平，表示为“-”；第二声阳平，表示为“ˊ”；第三声上声，表示为“ˇ”；第四声去声，表示为“ˋ”。拼音特征提取过程通过 CNN 学习实现，如图 6 中的 Pinyin Module 部分所示。

4 实验与分析

4.1 数据集

实验在囊括了粗粒度和细粒度 CNER 任务的 5 个数据集上开展。表 1 列出了各数据集的统计数据，包含每个数据集中训练集、验证集、测试集的句子和字符个数。

表 1 数据集统计情况

Table 1 Statistics of datasets

Dataset	Type	Train	Dev	Test
Address	Sentence	10500	3500	3500
	Char	213500	71100	15100
Weibo	Sentence	1400	270	270
	Char	73800	14500	14800
Ontonotes	Sentence	15700	4300	4300
	Char	491900	200500	208100
MSRA	Sentence	46400	—	4400
	Char	2169900	—	172600
Resume	Sentence	3800	460	480
	Char	124100	13900	15100

4.1.1 Address 数据集

由于中文地址要素的类别繁多并且要求标注者对地址所涉及的区域有一定了解，因此，中文地址数据的标注工作较为困难，进而导致中文地址公开数据集相对匮乏。为了评估所提模型在地址数据以及细粒度 NER 任务上的有效性，同时为其他研究者多提供一个中文地址数据选择，本文构建并发布了一个新的中文地址数据集 Address。基于网络爬取的 20027 条全国各地的地址数据，进行 3 轮人工标注和复查，筛选过滤了 2577 条存在问题（如地址不完整、地址存在套叠现象等）的地址数据，最终得到了 17450 条正确标注的全国地址数据。数据集中共包含了 20 个地址要素实体类型。为了使验证集和测试集样本数据相对丰富，将地址数据按 6∶2∶2 的比例随机划分为训练集、验证集和测试集[1)]。

4.1.2 NER 基准数据集

为了评估模型在不同领域的粗粒度 NER 数据集上的表现，在 4 个命名实体识别的基准数据集上进行了实验，包括 Weibo[32-33]，Resume(Zhang Y and Yang J, 2018[21])，MSRA[34]，Ontonotes4.0[35]。Weibo 数据集的语料来自社交媒体，其中包含 4 类实体：PER，ORG，LOC 和 GPE。Resume 数据集的语料来自新浪财经的简历数据，其中包含 CONT，EDU，LOC，NAME，ORG，PRO，RACE 和 TITLE 8 种实体。MSRA 和 Ontonotes4.0 的语料来自于新闻数据，其中 MSRA 包含 ORG，PER，LOC 3 类实体，OntoNotes4.0 包含 PER，ORG，LOC，GPE 4 类实体。

4.2 实验设置

4.2.1 基准模型

为了评估 CFGT 模型在中文地址要素解析任务上的有效性，将 HMM 与 BiLSTM-CRF 作为基准模型进行了对比实验。其中，HMM 属于机器学习方法，而基于深度学习的 BiLSTM-CRF 模型是目前中文地址要素解析领域的 SOTA 模型。由于 HMM 与 BiLSTM-CRF 并未引入外部的词信息进行增强，因此实验中还将 CFGT 与 CNER 领域最近几年提出的基于词增强的方法进行对比，如 Lattice LSTM[21]，CGN[25]，FLAT[22] 等，其中 FLAT 是 2020 年 CNER 领域的 SOTA 模型。此外，为了证明 CFGT 在 CNER 领域具有同样优秀的表现，引入了目前该领域的 SOTA 模型 LEBERT 作为基准模型进行对比实验。LEBERT 通过词典适配层将外部的词汇知识集成到预训练模型 BERT 的底层中，实现了字符信息与词汇信息的深度融合。本文将 BERT＋CFGT 模型与 LEBERT 进行比较，最后将 BERT＋CFGT 与 BERT 以及 BERT＋FLAT 比较，来验证模型与 BERT 的兼容性。实验中所有利用词增强方法的模型使用的词典和字、词向量与文献[21]相同。

4.2.2 超参数

模型使用 SGD 优化器执行梯度下降，其中，学习率为 1×10^{-3}，动量为 0.9；最大训练轮次设置为 100，批次大小为 10；另外，仅使用了单层的 Transformer Encoder，在多头自注意力层中设置了 8 个注意力头。这些参数在所有数据集上基本一致，它们的设置参考了 Li 等[22] 的工作。由于受到机器硬件配置的影响，批次大小在 Ontonotes 和 MSRA 数据集上取值为 4。

4.3 实验结果

表 2 列出了不同模型在 Address 数据集上的实验结果。结果表明，使用词信息增强的模型的 F1 分数大多高于不使用词信息的模型，这充分证明了词信息能提升模型表现，对于中文地址要素解析任务具有重要作用。CFGT 模型在

1) https://github.com/hoppyNaut/Address_Dataset

Address 数据集上获得了最高的准确率(Precision)、召回率(Recall)以及 F1 分数，相比于在中文地址要素解析领域的 SOTA 模型 BiLSTM-CRF 分别提高了 0.26%，0.39%，0.32%。本文模型在 Address 数据集上的指标并没有获得较为明显的提升，这可能是由于 Address 数据集属于小型精标注数据集，其中的噪声较少，因此各种模型的 F1 分数普遍较高。

表 2　不同模型在 Addrress 数据集上的实验结果

Table 2　Experimental results of different models on Address dataset

Models	*Precision*	*Recall*	*F*1
HMM	0.9337	0.9343	0.9332
BiLSTM-CRF	0.9680	0.9681	0.9680
Lattice LSTM	0.9704	0.9715	0.9709
LGN	0.9678	0.9679	0.9678
CGN	0.9698	0.9713	0.9706
FLAT	0.9697	0.9711	0.9704
CFGT	**0.9706**	**0.9720**	**0.9712**
BERT	—	—	0.9712
BERT-FLAT	0.9731	0.9746	0.9738
LEBERT	0.9725	0.9746	0.9735
BERT+CFGT	**0.9748**	**0.9751**	**0.9749**

表 3 列出了不同模型在 4 个 CNER 基准数据集上的 F1 分数，其中 * 表示当前结果是使用文献[21]的词典运行模型开源代码得到的。从表中数据可以看出，CFGT 模型在 4 个 CNER 基准数据集上比 BiLSTM-CRF 以及其他基于词增强的模型表现更好。相较于 BiLSTM-CRF 以及 Lattice LSTM，该模型在 4 个数据集上的平均 F1 分数分别有 3.10% 和 1.74%的提升。而相较于 FLAT，该模型在 Ontonotes，MSRA，Resume 数据集上的 F1 分数有轻微提升，在 Weibo 数据集上获得 1.26%的明显提升。此外，BERT+CFGT 在 4 个数据集上的表现都超过了 LEBERT。上述实验结果充分验证了 CFGT 模型在粗粒度和细粒度 CNER 任务上的有效性，并且 BERT+CFGT 在 5 个数据集上的优异表现也证明了 CFGT 模型与 BERT 有很好的兼容性。

表 3　在 4 个 CNER 基准数据集上的结果(*F*1)

Table 3　Results on four CNER benchmark datasets(*F*1)

Models	Weibo	Ontonotes	MSRA	Resume
BiLSTM-CRF	0.5675	0.7181	0.9187	0.9441
Lattice LSTM	0.5879	0.7388	0.9318	0.9446
CGN	0.5966*	0.7479	0.9347	0.9412
LGN	0.6015	0.7485	0.9363	0.9541
FLAT	0.6032	0.7645	0.9371	0.9545
CFGT	**0.6148**	**0.7652**	**0.9376**	**0.9550**
BERT	0.6820	0.8014	0.9495	0.9553
BERT-FLAT	0.6855	0.8141	0.9546	0.9586
LEBERT	0.6878*	0.8087*	0.9541*	0.9609*
BERT+CFGT	**0.6970**	**0.8201**	**0.9549**	**0.9658**

4.4　消融实验

本节通过消融实验来验证两种 Flat-Graph 结构与改进的相对位置编码的有效性。

4.4.1　实验设置

设计了以下对比模型来进行消融实验：(1) w/o FL：不包含 Flat-Lattice 结构；(2) w/o FS：不包含 Flat-Shift 结构；(3) w/ ORPE(Original Relative Positional Encoding)：使用未改进的相对位置编码；(4) w/ output-fusion：分别将两个 Flat-Graph 结构的相对位置编码输入 Transformer Encoder 中，再将两个输出结果进行融合。选取了 Address 数据集以及 Weibo 数据集来评估各模型的表现。

4.4.2　实验结果

表 4 列出了模型的消融实验结果。实验结果表明，无论是在 Address 还是 Weibo 数据集上，移除两种 Flat-Graph 结构的任一种都会导致模型性能的下降，但两种结构的重要程度有所不同。在 Weibo 数据集上，移除 Flat-Lattice 后模型 F1 值下降了 1.07%，而移除 Flat-Shift 后模型的 F1 值下降了 0.58%。此时，Flat-Lattice 的重要程度高于 Flat-Shift，这是因为 Flat-Shift 结构在词信息增强中发挥的是辅助作用。Flat-Shift 结构主要用于捕获最近上下文词汇的语义信息，当它与 Flat-Lattice 结构协作时，辅助融入了额外的最近上下文词信息，从而有效地提高了词增强的效果。但是 Flat-Lattice 结构仍然在词信息增强中占据主导地位，因此仅仅依靠 Flat-Shift 的效果略弱于单独使用 Flat- Lattice 结构的效果。在 Address 数据集上，两种 Flat-Graph 结构的表现则相差无几，模型的 F1 值相较于完整模型仅有轻微的降低。其次，改进的相对位置编码也有效提高了模型的性能，相对于原始的相对位置编码，在 Weibo 数据集上 F1 值提升了 0.58%。最后，当使用对两种 Flat-Graph 结构经过 Transformer-Encoder 的输出进行融合代替相对位置编码融合时，模型在两个数据集上的 F1 值出现了较大幅度的下降，在 Weibo 数据集上表现得尤其明显。这是因为对输出进行融合的方式属于浅层融合；而相对位置编码融合的方法属于深层的融合，可以充分利用多头自注意力机制的表示能力，并对两类词汇信息进行融合。

表 4　消融实验结果

Table 4　Ablation experimental results

Models	Address			Weibo		
	P	*R*	*F*1	*P*	*R*	*F*1
CFGT	0.9702	0.9722	**0.9712**	0.6464	0.5841	**0.6148**
w/o FL	0.9703	0.9719	0.9711	0.6555	0.5598	0.6041
w/o FS	0.9703	0.9715	0.9709	0.6610	0.5646	0.6090
w/ ORPE	0.9702	0.9720	0.9710	0.6033	0.6148	0.6090
w/ output-fusion	0.9693	0.9714	0.9703	0.6299	0.5741	0.6007

4.5　CFGT-gp 模型实验结果

本节验证字形信息和拼音信息对模型性能的影响，验证方法包括：CFGT(无字形及拼音信息)、CFGT-p(移除字形信息，仅添加拼音信息)、CFGT-gp(包含字形及拼音信息)。实验基于体量较小的 NER 基准数据集 Weibo 和 Resume 进行，结果如表 5 所列。

表 5　CFGT-gp 模型的消融实验结果

Table 5　Ablation experimental results of CFGT-gp

Models	Weibo	Resume
CFGT	0.6148	0.9550
CFGT-p	0.6222	0.9559
CFGT-gp	**0.6248**	**0.9584**

实验结果表明，CFGT-p 的性能表现比 CFGT 更好，证明了拼音信息添加的有效性；而 CFGT-gp 在融合字形和拼音信息后，表现优于 CFGT-p，在两个测试数据集上均获得了最高的 F1 得分，证明了字形信息添加的有效性。

5 讨论

本文提出了一种基于词典的中文地址要素解析方法，并且通过实验验证了该方法的有效性。此外，该方法能够为以下方面做出实际贡献。

首先，在物流、金融等领域，用户胡乱填写虚假地址的现象层出不穷，给相关公司造成了巨大的利益损失。目前主流的虚假地址识别方法大多依赖于中文地址要素解析的结果，因此中文地址要素解析模型表现的提升有助于提高虚假地址识别的准确率。另外，中文地址要素解析是中文地址编码的核心技术之一，中文地址编码是指将自然语言描述的地址信息，根据地址模型和编码规则进行智能语义解析，并通过与数据库的匹配来建立与对应的坐标空间信息和地理编码关联的过程。利用地址编码技术可以使大量的地址数据具有空间定位的性质，从而大大促进地理信息系统(GIS)的应用。本文提出的方法能够为地理编码关联过程提供更加可靠的智能语义解析结果作为输入，从而减少地址与地理编码产生误匹配的情况。

结束语 本文利用词信息增强地址文本字符序列，提出了一种基于词典的中文地址要素解析模型 CFGT。该模型通过构建、融合协作的平面图结构，建立了字符与其自匹配词、最近上下文词汇的直接联系，有效增强了字符序列的表示。同时，模型使用改进的相对位置编码调整字符和匹配词之间的相对位置，强化了匹配词对词中非首尾字符的增强能力。大量实验结果证明了 CFGT 模型在中文地址要素解析与 CNER 领域中的有效性。未来计划对结合字形、拼音等信息的增强模型做进一步探索。此外，如何构建一种能同时处理不同 NER 问题的统一模型，也是探索的方向之一。

参 考 文 献

[1] GOLDBERG D W, WILSON J P, KNOBLOCK C A. From text to geographic coordinates: the current state of geocoding[J]. URISA Journal, 2007, 19(1): 33-46.

[2] GOLDBERG D W. Advances in geocoding research and practice [J]. Transactions in GIS, 2011, 15(6): 727-733.

[3] KARIMI H A, SHARKER M H, ROONGPIBOONSOPIT D. Geocoding recommender: an algorithm to recommend optimal online geocoding services for applications[J]. Transactions in GIS, 2011, 15(6): 869-886.

[4] DHAR S, VARSHNEY U. Challenges and business models for mobile location-based services and advertising[J]. Communications of the ACM, 2011, 54(5): 121-128.

[5] CONG G, JENSEN C S. Querying geo-textual data: Spatial keyword queries and beyond[C]// Proceedings of the 2016 International Conference on Management of Data. New York: Association for Computing Machinery, 2016: 2207-2212.

[6] LI P, LUO A, LIU J, et al. Bidirectional gated recurrent unit neural network for chinese address element segmentation[J]. ISPRS International Journal of Geo-Information, 2020, 9(11): 635.

[7] MELO F, MARTINS B. Automated geocoding of textual documents: A survey of current approaches[J]. Transactions in GIS, 2017, 21(1): 3-38.

[8] KUAI X, GUO R, ZHANG Z, et al. Spatial context-based local toponym extraction and chinese textual address segmentation from urban poi data[J]. ISPRS International Journal of Geo-Information, 2020, 9(3): 147.

[9] LI X, ZHANG Y, LI L. A Chinese address recognition method based on address semantics [J]. Computer Engineering & Science, 2019, 41(3): 171-178.

[10] LIN Y, KANG M, HE B. Spatial pattern analysis of address quality: A study on the impact of rapid urban expansion in china [J]. Environment and Planning B: Urban Analytics and City Science, 2021, 48(4): 724-740.

[11] ZHANG X, LV G, LI B, et al. Rule-based Approach to Semantic Resolution of Chinese Address[J]. Journal of Geo-information Science, 2010(1): 9-16.

[12] ZHAO Y, WANG L, QIU A. An improved algorithm for address segmentation[J]. Science of Surveying and Mapping, 2013, 38(5): 74-76.

[13] DUAN Y, LI X, HUANG S. Extraction of administrative division of Chinese address based on conditional random fields[J]. Journal of Wuhan Institute of Technology, 2015(11): 47-51.

[14] WANG Y, ZHOU S, XING C. The address spatiotemporal data engine building method based on HMM[J]. Science of Surveying and Mapping, 2020, 45(10): 7.

[15] CHENG B, LI W, TONG H. Chinese Address Segmentation based on BiLSTM-CRF[J]. Journal of Geo-information Science, 2019, 21(8): 1143-1151.

[16] LI P, LUO A, LIU J, et al. Bidirectional gated recurrent unit neural network for chinese address element segmentation[J]. International Journal of Geo-Information, 2020, 9(11): 635.

[17] LIU X, PENG T. Research on Chinese Scenic Spot Named Entity Recognition Based on Convolutional Neural Network[J]. Computer Engineering & Science, 2020, 56(4): 145-150.

[18] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of NAACL-HLT. Stroudsburg: Assoc Computational Linguistics-ACL, 2019: 4171-4186.

[19] ZHANG H, REN F, LI H, et al. Recognition method of new address elements in chinese address matching based on deep learning[J]. ISPRS International Journal of Geo-Information, 2020, 9(12): 745.

[20] SUN S, TANG K. Chinese address segment method based on BERT[J]. Electronic Design Engineering, 2021, 29(9): 155-159.

[21] ZHANG Y, YANG J. Chinese ner using lattice lstm[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Assoc Computational Linguistics, 2018: 1554-1564.

[22] LI X, YAN H, QIU X, et al. FLAT: Chinese NER Using Flat-Lattice Transformer [C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Assoc Computational Linguistics, 2020: 6836-6842.

[23] HEWITT J,MANNING C D. A structural probe for finding syntax in word representations[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Stroudsburg: Assoc Computational Linguistics, 2019: 4129-4138.

[24] DING R,XIE P,ZHANG X,et al. A neural multi-digraph model for chinese ner with gazetteers[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg:Assoc Computational Linguistics,2019:1462-1467.

[25] SUI D,CHEN Y,LIU K,et al. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). Stroudsburg: Assoc Computational Linguistics, 2019:3830- 3840.

[26] LIU W,XU T,XU Q,et al. An encoding strategy based word-character lstm for chinese ner[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Stroudsburg: Assoc Computational Linguistics, 2019: 2379-2389.

[27] DAI Z,YANG Z,YANG Y,et al. Transformer-xl:Attentive language models beyond a fixed-length context[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. Stroudsburg: Assoc Computational Linguistics, 2019: 2978-2988.

[28] HU Y,VERBERNE S. Named entity recognition for Chinese biomedical patents[C]// Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg: Assoc Computational Linguistics,2020:627-637.

[29] MA R,PENG M,ZHANG Q,et al. Simplify the usage of lexicon in chinese ner [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg:Assoc Computational Linguistics,2020:5951-5960.

[30] LIU W,FU X,ZHANG Y,et al. Lexicon enhanced chinese sequence labelling using bert adapter[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Assoc Computational Linguistics,2021:5847-5858.

[31] 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. California: Neural Information Processing Systems (NIPS), 2017: 6000-6010.

[32] PENG N,DREDZE M. Named entity recognition for chinese social media with jointly trained embeddings[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Assoc Computational Linguistics, 2015:548-554.

[33] HE H,SUN X. F-score driven max margin neural network for named entity recognition in chinese social media[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Assoc Computational Linguistics,2017:713-718.

[34] LEVOW G A. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition [C]// Proceedings of the Fifth SIGHAN workshop on Chinese language processing. Stroudsburg: Assoc Computational Linguistics,2006:108-117.

[35] WEISCHEDEL R,PARADHAN S,RAMSHAW L,et al. Ontonotes release 4. 0[DB/OL]. http://catalog. ldc. upenn. edu. LDC2011T03.

HUANG Wei, born in 1998, postgraduate. His main research interests include data mining and anomaly detection.

FU Xiangling, born in 1975, Ph.D, professor, Ph.D supervisor. Her main research interests include natural language processing, smart finance and smart healthcare.

(责任编辑:柯颖)