Computer Science and Application 计算机科学与应用, 2023, 13(4), 808-818
Published Online April 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.134080

Hans汉斯

基于LEBERT-CRF和知识图谱的中文地址修正补全方法

王钦民，刘　鹏，邓国威

暨南大学信息科学技术学院，广东 广州

收稿日期：2023年3月18日；录用日期：2023年4月17日；发布日期：2023年4月24日

摘　要

为解决人工中文地址因输入不准确造成的地址解析错误问题，本文首先结合词汇增强的基于Transformer的双向编码表征模型(LEBERT)与条件随机场(CRF)，提出了LEBERT-CRF模型，相较BERT-长短期记忆-CRF模型(BERT-BiLSTM-CRF)在分词准确率、召回率以及F值上分别提升了1.45%、1.89%和1.67%。然后，通过标准层级地址数据，并引入别名、旧名等地址信息构建了地址知识图谱库。最终，利用经过分词处理的地址数据，并根据地址数据存在的几种可能错误类型，设计出一种基于地址知识图谱库的匹配算法，对分词完的地址数据进行匹配修正并得到准确地址信息，相较于中文省份城市地区匹配器(CPCA)，地址解析在一级地址、二级地址、三级地址上解析准确率分别提升了2.12%、2.36%和1.12%。

关键词

中文地址分词，中文地址匹配，LEBERT，CRF，知识图谱

Chinese Address Correction Completion Method Based on LEBERT-CRF and Knowledge Graph

Qinmin Wang, Peng Liu, Guowei Deng

College of Information Science and Technology, Jinan University, Guangzhou Guangdong

Received: Mar. 18th, 2023; accepted: Apr. 17th, 2023; published: Apr. 24th, 2023

Abstract

In order to solve the problem of address resolution errors caused by inaccurate input of manual Chinese addresses, in this paper, we first propose a LEBERT-CRF model which is based on the combi-

文章引用: 王钦民, 刘鹏, 邓国威. 基于 LEBERT-CRF 和知识图谱的中文地址修正补全方法[J]. 计算机科学与应用, 2023, 13(4): 808-818. DOI: 10.12677/csa.2023.134080

nation of the word-enhanced deep learning model Lexicon Enhanced Bidirectional Encoder Representations from Transformers (LEBERT) and Conditional Random Fields (CRF). Compared with BERT-Bidirectional Long Short Term Memory-CRF (BERT-BiLSTM-CRF) model, the segmentation accuracy, recall rate and F-score were increased by 1.45%, 1.89% and 1.67%, respectively. Then, based on the standard multi-level address data, an address knowledge graph database is constructed with address information such as aliases and old names. Finally, a matching algorithm based on the address knowledge graph database is designed based on the address data processed by word segmentation and several possible error types exist in the address data. The address data after word segmentation is matched and corrected and accurate address information is obtained. Compared to the Chinese Province City Area mapper (CPCA), the resolution accuracy of 1st-level address, 2nd-level address and 3rd-level address is improved by 2.12%, 2.36% and 1.12%, respectively.

Keywords

Chinese Address Segmentation, Chinese Address Matching, LEBERT, CRF, Knowledge Graph

1. 引言

地理信息化建设是国家近年来的重大建设任务，而中文地址解析是其中的重要组成部分。发展中文地址解析技术，有利于构建更健壮的地理信息系统，进而更好地服务国家建设[1]。

目前，大部分中文地址仍然为人工手动输入。这种人工输入的地址依赖于输入者的习惯，容易产生地址语义不清晰、部分地名缺失等问题，从而导致无法解析定位等问题。解决这些问题的核心挑战是改进中文地址分词以及中文地址修正匹配方法。

中文地址分词属于中文分词任务中的子任务。传统的中文分词方法主要包括字典匹配法[2] [3]、机器学习方法[4] [5] [6] [7] [8]等。随着深度学习技术的发展，通过训练基于神经网络的深度学习模型进行中文分词成为了中文分词的主要研究方法。2006 年，张晓淼将反向传播(Back Propagate, BP)神经网络进行改进并将其运用于中文分词任务[9]。2015 年，Chen 等人将长短期记忆模型(Long Short-Term Memory, LSTM)用于中文分词以及词性标注任务[10]。Graves 提出了双向长短期记忆模型(Bidirectional-LSTM, Bi-LSTM)模型[11]，黄积杨将 Bi-LSTM 模型应用于中文分词任务[12]。张子睿等人则在 Bi-LSTM 模型的基础上引入了条件随机场(Conditional Random Field, CRF)层，构建了 BILSTM-CRF 模型[13]。张文静等人提出了格式 LSTM (Lattice-LSTM)模型，将词典信息输入到了字符序列中[14]。王玮提出了基于六词位标注的 Bi-LSTM 模型[15]，进一步提升了中文分词的效果。Jacob 等人提出了将 Transformer 中编码器改造为双向的结构，即基于 Transformer 的双向编码表征模型(Bidirectional Encoder Representations from Transformers, BERT)预训练模型[16]。目前，有学者继续基于 BERT 预训练提出了平均池化[17]、跨层参数共享[18]等改进方法。

当前，中文地址匹配的主流方法基于规则进行匹配[19] [20] [21]，很少考虑地名的多语义性，导致匹配过程中对于多语义性地址匹配准确度较低。为了解决上述问题，Bizer 等人提出了链接数据概念，将语义网络中不同类型的数据集链接起来，构成一个庞大的知识图谱[22]。2009 年，Akerkar 等人提出了知识存储数据库系统，为知识图谱的信息存储打下基础[23]。之后，Google 正式提出了基于语义网络的知识

图谱技术，并将该技术成功运用于其搜索系统，提高了搜索的丰富度以及准确率。但是，中文分词的边界不准确性以及中文地址的多样性极大地影响了模型在精确地址匹配任务上的准确度。

为了解决上述问题，本文提出 LEBERT-CRF 模型，对中文地址进行分词。并基于上述分词结果构建了具有多语义性的地址知识图谱库，对存在缺失、旧名、别名的地址进行修复补全。

为了验证我们所提出模型的有效性，我们在中文地址数据集上进行了实验验证。首先在中文分词任务中，我们所提出的 LEBERT-CRF 混合模型有效提升了分词效果。大量测试表明：在准确率、召回率以及 F 值上，新模型相较 BERT-BiLSTM-CRF 模型分别提升了 1.45%、1.89%和 1.67%；在地址匹配任务中，本文提出的地址修正补全模型在地址匹配任务上相比中文省份城市地区匹配器(Chinese Province City Area mapper, CPCA, https://github.com/DQinYuan/chinese_province_city_area_mapper)开源工具在省级、市级、区县级单位地址上分别提升了 2.12%、2.36%和 1.12%。

本文的结构如下：第二节介绍 LEBERT-CRF 模型；第三节介绍基于知识图谱的中文地址修正补全具体方法；第四节展示了基于本文提出方法的实验结果；最后一节总结本文的内容。

2. 基于 LEBERT 的中文地址分词模型

本节主要介绍 LEBERT-CRF 模型的整体结构，LEBERT 层与 CRF 层的主要内容和作用。

2.1. LEBERT-CRF 模型

本文提出使用 LEBERT-CRF 来解决中文地址分词的任务，其主要的模型结构如图 1 所示。

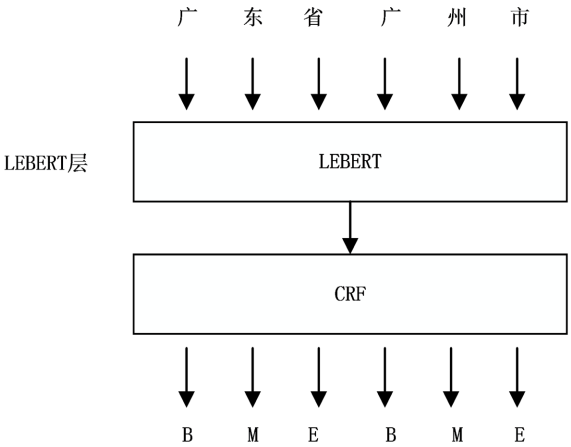

Figure 1. Structure of LEBERT-CRF model
图 1. LEBERT-CRF 模型结构

整个模型由 LEBERT 层和 CRF 层两部分组成，输入的地址字符串转换为词嵌入向量输入 LEBERT 层，经过 LEBERT 层进行编码后，最后再经过 CRF 层输出。

2.2. LEBERT 模型

BERT 模型在很多自然语言处理任务中取得了良好效果，但在中文分词领域中，由于中文句子中的词汇之间没有明显的分割符，输入均为单个字符，并不包含词汇层面特征等原因，其效果不佳，因此需要词典为模型提供额外的词汇信息。为能够充分利用字符的特征并将词典信息融入到字符中，Liu 等人提出了利用词汇增强的 BERT 模型——LEBERT (Lexicon Enhanced BERT) [24]。LEBERT 模型在 BERT 的

中间层中嵌入了一个词汇适配器(Lexicon Adapter, LA)，由此引入该词汇的词嵌入向量。LEBERT 模型结构图如图 2 所示。

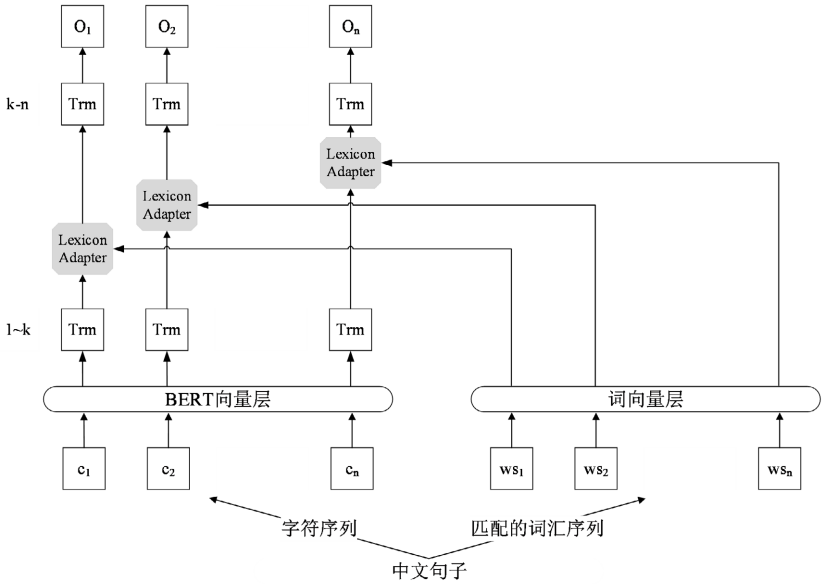

Figure 2. Structure diagram of LEBERT model
图 2. LEBERT 模型结构图

词典适配器的嵌入使得整个网络不仅能考虑字符层面的特征，还能考虑词汇层面的特征。如给定一个字符串序列 $s_c = \{c_1, c_2, \cdots, c_n\}$，并且字符串中的每个字符到定义好的词典 D 中进行匹配，找出可能包含该字符的词汇，然后组成字符－词汇对。词汇适配器的结构如图 3 所示。

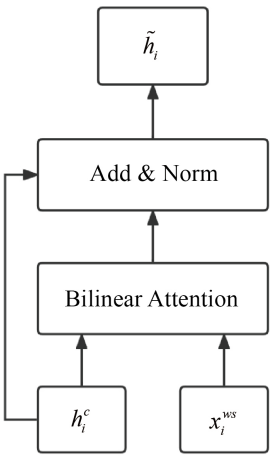

Figure 3. Lexicon adapter
图 3. 词典适配器

如公式(1)所示，词典适配器将匹配到的词典信息融入到当前字符中：

$$\tilde{h} = LA\left(h_i^c, z_i^w\right) \tag{1}$$

其中，z_i^w 表示第i个单词所匹配的词向量组，h_i^c 表示上一层的输出向量，LA 表示词典匹配器，其核心是将 z_i^w 和 h_i^c 通过双线注意力机制合并，其结果再和 h_i^c 相加并正则化，输入到下一层中。

2.3. CRF 层

LEBERT 模型对模型预测的标签序列进行输出时，不考虑预测标签序列的位置，而是输入各个预测标签的最大概率。CRF 层考虑标签序列整体的标签分布，对于标签序列的约束力更加充足。以“广东省广州市”为例。采用 BMES 编码方式，其中 B 表示命名实体的开始部分，E 表示命名实体的结束部分，中间部分则使用 M 表示，单字实体以 S 进行表示。则标签序列为{B-province, M-province, E-province, B-city, M-city, E-city}。LEBERT 模型可以很好地识别不同类别命名实体，比如“广东省”属于省级单位，所以标记为“X-province”，而“广州市”标记为“X-city”。但对于命名实体内部的关系，特别是命名实体的中间标志 M 很难准确识别。因此，可能会出现将“广东省”的标签序列识别为{M-province, E-province, M-province}这类错误。因此，我们在 LEBERT 模型层后引入了 CRF 层以对序列整体进行约束。CRF 层使用负对数似然函数(Negative Log Likelihood)作为损失函数，并在训练过程中最小化该损失函数，损失函数表达式如(2)所示：

$$Loss = -\sum \log\left(p\left(y|s\right)\right) \tag{2}$$

其中，$p\left(y|s\right)$ 表示在句子 s 的状态下，取得输出状态序列 $y=\{y_1,y_2,\cdots,y_n\}$ 的概率。$p\left(y|s\right)$ 计算方法如公式(3)所示：

$$p\left(y|s\right)=\frac{e^{\sum_i\left(O_{i,y_i}+T_{y_{i-1},y_i}\right)}}{\sum_{\tilde{y}}e^{\sum_i\left(O_{i,\tilde{y}_i}+T_{\tilde{y}_{i-1},\tilde{y}_i}\right)}} \tag{3}$$

其中，O_{i,y_i} 表示从第 i 个位置得到标签 y_i 的概率，T_{y_{i-1},y_i} 表示由 $i-1$ 的状态转移到i的状态的概率。$\tilde{y}$ 表示所有的位置的标签序列。

3. 基于知识图谱的中文地址修正补全方法

在本节中首先分析了目前中文地址中所存在的问题，然后通过构建地址知识图谱库，并基于知识图谱库的多语义性设计一套具有多重匹配功能的算法，最后对经过分词的中文地址进行修正补全，可以更好地解决此类问题。

3.1. 中文地址中存在的问题

中文地址经常由人为输入所产生，但由于个人习惯或者历史地名等原因，造成很多地址存在信息错误或者信息缺失等问题。中文地址存在的主要错误类型有地理信息缺失、地理信息有误(错误输入、旧名以及别名)等。

3.2. 地址知识图谱库的构建

从中国统计局公布的行政区划来看，中文地址主要存在四级分布，即：省(自治区)、市(州)、区(县)、街道(乡镇)。基于上述的四级行政区划，本文构建多语义地址图谱库用以解决以上提到的问题。标准地址通过抓取国家四级行政区划信息获取，地名的旧名由国家民政局公开的行政区划变更获取，而地区的别名或俗称是从网络查询获得。此外，本文获取的数据已经在预处理中实现了结构化。本文收集的各类实体数量如表 1 所示。

由于在抓取行政区划地名数据时存在明显的层级结构，因此需将每相邻的两级行政区划抽取出从属

关系，即[本级实体－从属于→上一级实体]。另外，由于引入了别名、旧名等概念，因此抽取出了[“原名”实体－别名→“别名”实体]以及[“现名”实体－旧名→“旧名”实体]这两个关系三元组。最终结果如图 4 所示。另外，根据地名具备的属性，又将实体划分为六种本体：“省”、“市”、“区”、“街道”、“旧名”以及“别名”。抽离本体可以更好地区分实体与实体之间的区别，目的是为之后的数据存储与后续算法匹配工作进行准备。

Table 1. Statistics table of entities and its amount
表 1. 实体及其数量统计表

实体名	值
省级	34
市级	342
区县级	3029
街道	43067
别名	34
旧名	62

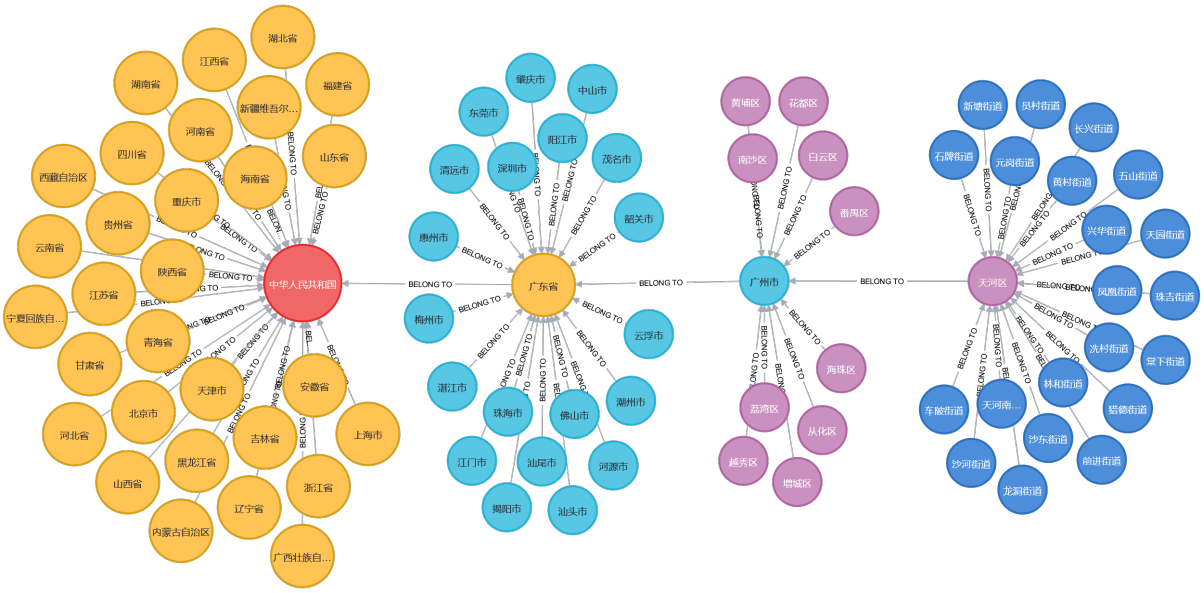

Figure 4. Knowledge graph based on addresses information
图 4. 基于地址信息建立的知识图谱

3.3. 修正补全算法设计

在地址知识图谱构建完成后，我们开始对中文地址文本序列进行分词，然后再对分词后的地址序列进行补全和修正。根据地名中存在特点(如重名、缺失信息的位置、地名是否是旧名别名的位置等)，进行分类，采用不同的匹配重构算法，将地址中四级地址匹配出来。根据中文地址文本中存在的几种问题，我们提出针对以下四种问题的基于知识图谱的四种修正方案：重名情况下前文全缺失重构、重名情况下前文只有相邻缺失重构、前文缺失重构以及旧名别名重构。其基本流程图如图 5 所示。

下面逐一对每种方案进行介绍。

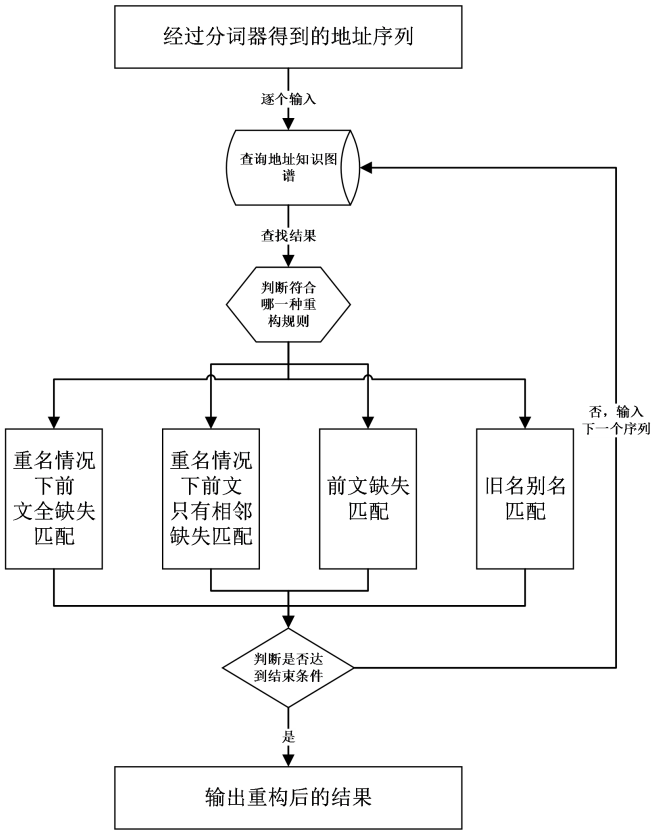

Figure 5. Multiple matching schemes based on the knowledge graph
图 5. 基于知识图谱的多种匹配模型

3.3.1. 重名情况下前文全缺失匹配方案

假设使用某个地名到知识图谱库中进行查询，结果发现并不唯一，也就是存在相同地名的情况，并且符合下一级地名实体存在以及该级地名的前级地名为缺省的状态下，可以利用本级和下一级的地名进行组合构建实体关系对，然后可以利用该关系对进行匹配，将缺失的前级地名匹配出来。

3.3.2. 重名情况下前文只有相邻缺失匹配方案

假设使用某个地名到知识图谱库中进行查询，结果发现并不唯一，但缺失前一级的单位，且更前一级单位是存在的，这种情况可以用当前地名以及前前级的地名构建起关联关系对，利用这个关联关系在地址知识图谱库中将缺失的前级地名实体匹配出来。

3.3.3. 前文缺失匹配方案

假设某个地址实体的上一级地址实体缺失，且本级地址实体在地址知识图谱库中没有重名地址实体，则根据该地址实体的名字去数据库中进行匹配，将前面所缺失的地名实体匹配出来进行补全。

3.3.4. 旧名别名匹配方案

如果对地名进行实体匹配后，得到的是“旧名”或者“别名”的实体，那么需要将匹配出的实体的关联“原名”实体匹配出来，然后再利用这个“原名”重复以上操作。

4. 实验结果分析

本节对我们提出的 LEBERT-CRF 模型进行实验，并通过实验和已有的方案进行对比分析。

4.1. 数据处理

模型训练采用的数据是中文地址数据，实验共采用五万余条中文地址数据进行训练，并将数据集中的 80%作为训练集、10%作为验证集以及 10%作为测试集。由于地址信息涉及用户隐私，我们不公开此数据集。在地址数据中，基本的中文地址标注要素如表 2 所示。

Table 2. Tagging elements in Chinese addresses
表 2. 中文地址标注要素

大类	标志词	例子
一级行政区划	省	广东省
二级行政区划	市、地区、自治州、盟	广州市
三级行政区划	区、县、县级市、自治县、旗	海珠区
四级行政区划	街道	赤岗街道
五级行政区划	社区	赤岗办事处
路	路、街、巷、道	阅江西路
门(楼)址	号、栋、单元	222 号
兴趣点、标志物	公司、大楼	广州塔
定位词	方向	东、东面、向东
	距离	100 米
	助词	附近、路口
其他	符号	“()”

基于上表中的标注要素规则，将中文地址按要素进行切分，如“广东省广州市海珠区赤岗街道阅江西路北 222 号观光区 108 层”将切分为:“广东省/广州市/海珠区/赤岗街道/阅江西路/北/222 号/观光区/108 层”。本模型对于地址序列的标注方法为 BMES 编码，在对上文提及的例子进行标注后的标注序列为{B, M, E, B, M, E, B, M, E, B, M, M, E, B, M, M, E, S, B, M, M, E, B, M, M, E, B, M, E}。

4.2. 模型训练

我们使用向量化 Skip-Gram [25]作为词嵌入向量集，词向量维度为 200。本文提出的 LEBERT-CRF 模型参数如表 3 所示。

Table 3. Model’s parameters
表 3. 模型参数

参数名	值
Transformer 层数	12
注意力头数量	12
注意力层 Dropout 率	0.1
最大位置嵌入向量维度	512

此外，在模型训练中的其他参数如表 4 所示。

Table 4. Model's training parameters
表 4. 模型训练参数

参数名	值
学习速率	0.0001
每个 GPU 训练批量(Batch)	16
训练轮数(Epoch)	12

本文 LEBERT-CRF 模型的实验环境是 torch 1.6.0，Transformer 3.4.0，TensorFlow 2.3.1，编程语言为 Python 3.7.0，开发环境为 PyCharm，其他依赖包有 NumPy 1.18.5、skicit-learn 0.23.2 等。使用的硬件环境为：56 核 CPU 的 Intel(R) Xeon(R) Gold 5117 CPU 2.00 GHz；内存 64 G，GPU 为 GeForce RTX 2080Ti。

4.3. 结果对比

在本节，本文对 LEBRET-CRF 模型分别进行了中文地址数据分词实验以及基于知识图谱的地址修正补全实验。

4.3.1. 中文地址数据分词

在中文分词或者命名实体识别任务中，有三个最常用的评价指标：准确率 P (Precision)、F-值(F-score)及召回率 R (Recall)。

我们对常见的几种中文分词模型进行了实验对比，结果如表 5 所示。

Table 5. Results comparison table of Chinese segmentation experiments
表 5. 中文分词模型实验结果对比表

模型	准确率 P(%)	召回率 R(%)	F 值(%)
HMM	84.68	82.88	83.77
BiLSTM	93.47	93.47	93.47
CRF	95.59	95.87	95.63
BiLSTM-CRF	96.70	96.70	96.70
BERT-BiLSTM-CRF	95.72	95.34	95.53
LEBERT-CRF	**97.17**	**97.23**	**97.20**

从表 5 可以看出，相较于其他模型，本文提出的 LEBERT-CRF 模型表现更优。在准确率、召回率、F 值上新模型相比 BiLSTM-CRF 模型分别提高了 0.47%、0.53%以及 0.50%。特别是相对于基于 BERT 预训练模型的 BERT-BiLSTM-CRF 模型，如图 6 所示，我们提出的 LEBERT-CRF 混合模型，通过词典的引入进一步提升了分词的效果，从而在各个轮次的表现均优于 BERT-BiLSTM-CRF 模型。最终在准确率、召回率以及 F 值上，新模型相较 BERT-BiLSTM-CRF 分别提升了 1.45%、1.89%和 1.67%。

另外，从训练时间上对比来看，BERT-BiLSTM-CRF 模型在 GeForce RTX 2080Ti 单核心上单个轮次的平均训练时间为 617 秒，而本文模型 LEBERT-CRF 由于相比 BERT-BiLSTM-CRF 模型去除了双向 LSTM 层，因此训练时长更短，在相同实验环境下单个轮次的平均训练时间为 462 秒，训练效率约为

BERT-BiLSTM-CRF 模型的 1.33 倍。

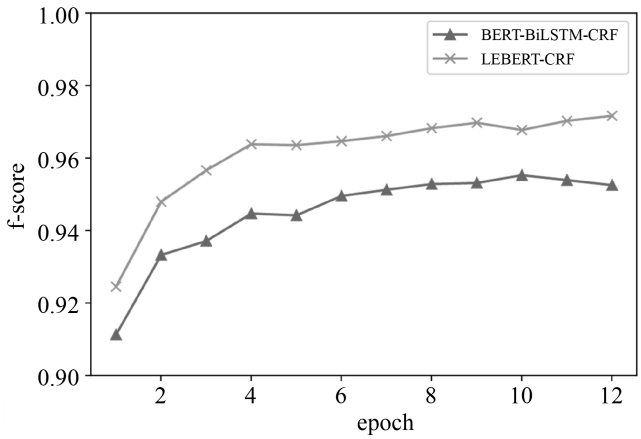

Figure 6. Comparison of F-score between BERT-BiLSTM-CRF model and LEBERT-CRF model
图 6. BERT-BiLSTM-CRF 模型与 LEBERT-CRF 模型的 F 值对比

4.3.2. 地址补全

在基于知识图谱对中文地址的修正补全实验中，我们使用的数据集是从经过预处理的数据集中随机抽取的 5000 条地址文本数据，我们通过搜索引擎，地图软件等方式将地址文本的信息补全完整并将正确完整的地址作为数据集的标签。

由于本文任务为地址中层级地址信息的修正补全，在计算匹配准确率 P 时，我们按照各级地址的匹配情况而非地址整体的匹配情况进行计算。P 值计算方法如公式 4 所示：

$$p = \frac{TP}{TP + FP} \tag{4}$$

上式表示的意义是：某一级正确匹配的地址总量除以数据总量。

我们对本文提出的地址知识图谱以及开源地址匹配工具 CPCA 对于各级地址的匹配结果进行了对比试验，实验结果如表 6 所示。

Table 6. Results comparison table of address matching experiments (p-value)
表 6. 地址匹配实验结果对比表(p 值)

地址级别	LEBERT-CRF	CPCA
省级地址单位	**99.04**	96.92
市级地址单位	**97.88**	95.52
区县级地址单位	**89.94**	88.82
乡镇街道地址单位	**13.70**	-

5. 总结

本文首先针对中文地址匹配补全任务的特点，提出了 LEBERT-CRF 模型用于中文地址分词并取得了良好效果。然后本文还结合多语义地址知识图谱，实现了四级中文地址修正补全。其中，中文地址分词任务相比目前常用的实体识别模型，在准确率、召回率以及 F 值均有较大的提升，验证了 LEBERT-CRF

模型在中文地址分词问题上的优越性。本文所提出的基于知识图谱的修正匹配方案，能够适应更加复杂的地址文本问题，且可以在保持较好性能的前提下，更加准确地匹配修正地址文本。

参考文献

[1] 国务院办公厅关于印发“十四五”城乡社区服务体系建设规划的通知(国办发[2021] 56 号) [J]. 中华人民共和国国务院公报, 2022(5): 69-77.

[2] 王思力. 面向大规模信息检索的中文分词技术研究[D]: [硕士学位论文]. 北京: 中国科学院计算技术研究所, 2006: 9-27.

[3] 张科. 多次 Hash 快速分词算法[J]. 计算机工程与设计, 2007, 28(7): 1716-1718.

[4] 李家福, 张亚非. 一种基于概率模型的分词系统[J]. 系统仿真学报, 2002, 14(5): 544-546.

[5] Mccallum, A., Freitag, D. and Pereira, F.C.N. (2000) Maximum Entropy Markov Models for Information Extraction and Segmentation. *Proceedings of the* 17*th International Conference on Machine Learning*, Stanford, 29 June-2 July 2000, 591-598.

[6] Low, J.K., Ng, H.T. and Guo, W. (2005) A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the* 4*th SIGHAN Workshop on Chinese Language Processing*, Jeju Island, 14-15 October 2005, 161-164.

[7] Lafferty, J., Mccallum, A. and Pereira, F. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the* 18*th International Conference on Machine Learning*, Williamstown, 28 June-1 July 2001, 282-289.

[8] 陈晴. 基于条件随机场的自动分词技术的研究[D]: [硕士学位论文]. 沈阳: 东北大学, 2005: 7-14.

[9] 张晓淼. 基于神经网络的中文分词算法的研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2006: 25-38.

[10] Chen, X., Qiu, X., Zhu, C., *et al*. (2015) Long Short-Term Memory Neural Networks for Chinese Word Segmentation. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, 17-21 September 2015, 1197-1206. https://doi.org/10.18653/v1/D15-1141

[11] Graves, A., Fernández, S. and Schmidhuber, J. (2005) Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. In: *International Conference on Artificial Neural Networks*, Springer, Berlin, 799-804. https://doi.org/10.1007/11550907_126

[12] 黄积杨. 基于双向 LSTMN 神经网络的中文分词研究分析[D]: [硕士学位论文]. 南京: 南京大学, 2016: 38-48.

[13] 张子睿, 刘云清. 基于 BI-LSTM-CRF 模型的中文分词法[J]. 长春理工大学学报(自然科学版), 2017, 40(4): 87-92.

[14] 张文静, 张惠蒙, 杨麟儿, 等. 基于 Lattice-LSTM 的多粒度中文分词[J]. 中文信息学报, 2019, 33(1): 18-24.

[15] 王玮. 基于 Bi-LSTM-6tags 的智能中文分词方法[J]. 计算机应用, 2018, 38(z2): 107-110.

[16] Jacob, D., Ming, W.C., Kenton, L., *et al*. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, Minneapolis, 2-7 June 2019, 4171-4186.

[17] Zhao, S., Zhang, T., Hu, M., *et al*. (2022) AP-BERT: Enhanced Pre-Trained Model through Average Pooling. *Applied Intelligence*, **52**, 15929-15937. https://doi.org/10.1007/s10489-022-03190-3

[18] Lan, Z.Z., Chen, M.D., Sebastian, G., *et al*. (2019) ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. ArXiv: 1909.11942.

[19] 黄丹丹, 郭玉翠. 融合 Attention 机制的 BI-LSTM-CRF 中文分词模型[J]. 软件, 2018, 39(10): 260-266.

[20] 张琛, 陈张建, 刘江涛, 等. Lucene 自适应分词的地址匹配方法改进与实现[J]. 测绘科学, 2021, 46(10): 185-193.

[21] 姚心宇. 中文地址识别系统中的地址表达与匹配[D]: [硕士学位论文]. 上海: 华东师范大学, 2012.

[22] Bizer, C., Heath, T. and Berners-Lee, T. (2011) Linked Data: The Story So Far. In: *Semantic Services, Interoperability and Web Applications*: *Emerging Concepts*, IGI Global, Hershey, 205-227. https://doi.org/10.4018/978-1-60960-593-3.ch008

[23] Akerkar, R. and Sajja, P. (2009) Knowledge-Based Systems. Jones & Bartlett Publishers, Sudbury.

[24] Liu, W., Fu, X., Zhang, Y. and Xiao, W. (2021) Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing*, Vol. 1, 5847-5858. https://doi.org/10.18653/v1/2021.acl-long.454

[25] Song, Y., Shi, S., Li, J., *et al*. (2018) Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Vol. 2, 175-180. https://doi.org/10.18653/v1/N18-2028