计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(6):169−174 [doi: 10.15888/j.cnki.csa.007269]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 BERT+BiLSTM+CRF 的中文景点命名实体识别①

赵　平[1], 孙连英[2], 万　莹[1], 葛　娜[1]

[1](北京联合大学 智慧城市学院, 北京 100101)
[2](北京联合大学 城市轨道交通与物流学院, 北京 100101)
通讯作者: 孙连英, E-mail: sunlychina@163.com

摘　要: 为解决旅游文本在特征表示时的一词多义问题, 针对旅游游记文本景点实体识别中景点别名的问题, 研究了一种融合语言模型的中文景点实体识别模型. 首先使用 BERT 语言模型进行文本特征提取获取字粒度向量矩阵, BiLSTM 用于上下文信息的提取, 同时结合 CRF 模型提取全局最优序列, 最终得到景点命名实体. 实验表明, 提出的模型性能提升显著, 在实际旅游领域内景点识别的测试中, 与以往研究者方法比较下准确率, 召回率分别提升了 8.33%, 1.71%.

关键词: BERT 语言模型; BiLSTM; 条件随机场; 景点实体识别

引用格式: 赵平,孙连英,万莹,葛娜.基于 BERT+BiLSTM+CRF 的中文景点命名实体识别.计算机系统应用,2020,29(6):169–174. http://www.c-s-a.org.cn/1003-3254/7269.html

Chinese Scenic Spot Named Entity Recognition Based on BERT+BiLSTM+CRF

ZHAO Ping[1], SUN Lian-Ying[2], WAN Ying[1], GE Na[1]

[1](Smart City College, Beijing Union University, Beijing 100101, China)
[2](College of Urban Rail Transit and Logistics, Beijing Union University, Beijing 100101, China)

Abstract: To solve the polysemy troublesome problem of tourism text in feature extraction, a Chinese scenic spot named entity recognition model based on fusion language model is studied for the problem of attraction alias in the visual recognition of tourist travel texts. Firstly, the BERT is used for tourism text feature extraction to obtain the word granularity vector matrix. BiLSTM is used to extract the context information. The CRF is used to obtain the global optimal sequence, and finally the tourist attraction entity is obtained. Experiments show that the performance of the proposed model is significantly improved. In the test of scenic spot identification in the actual tourism field, compared with the previous research, precision and recall rates are increased by 8.33% and 1.71%, respectively.

Key words: BERT model; BiLSTM; conditional random field; scenic spot entity recognition

1　引言

随着社会媒体的发展, 越来越多的旅游者喜欢通过游记分享旅游体验. 游记文本中景点的提取对旅游领域问答系统、个性化推荐等研究具有重要的意义.

1996 年, 命名实体识别 (Named Entity Recognition, NER) 一词在 MUC-6[1]上提出来的, 为自然语言处理的一项基础任务. 早期基于规则和词典[2]主要依赖语言学家根据上下文语义结构归纳的模板. 该方法对于难以归纳的总结无法解决, 识别效果不明显, 且归纳总结过程代价比较大, 所以学者们使用机器学习方法[3–5]来解

① 基金项目: 国家重点研发计划 (2018 YFC0807806); 中国物流学会项目 (2018CSLKT3-184)
Foundation item: National Key Research and Development Program of China (2018 YFC0807806); Project of China Society of Logistics (2018CSLKT3-184)
收稿时间: 2019-04-25; 修改时间: 2019-05-21; 采用时间: 2019-08-12; csa 在线出版时间: 2020-06-10

决这一问题, 机器学习的方法主要采用数学统计进行建模, 对 NER 问题分类 3 类小问题: 特征选择、机器学习策略、序列标注等. 在处理 NER 问题时, 使用大规模的标注语料让机器来训练模型, 通过训练好的模型对测试语料进行序列解码等, 得到命名实体. 但机器学习方法对文本特征提取要求较高. 目前, 基于深度学习的 NER 方法[6,7]比前两种方法得到了更广泛的应用, 目前流行的方法为 BiLSTM 方法. 由于 BiLSTM 是对序列中各个位置的分数值进行独立分类, 不能考虑相邻标签之间的信息. 而 CRF 能较好解决这个问题, 模型最后一层使用条件随机场模型作为句子级的序列标注, 如 Li 等[8]提出基于 LSTM-CRF 的命名实体识别方法.

在对于旅游领域内的景点识别研究, 现有的主要是基于机器学习的方法, 薛征山等[9]提出的基于隐马尔可夫模型的旅游景点识别方法, 该方法虽然在景点实体识别上有一定的效率, 但是其未能考虑到上下文之间的语义信息, 且在对文本提取特征的过程中未能解决文本特征表示的一词多义问题, 旅游领域景点词语一般会存在不同语境下不同含义, 比如“黄山”在不同语境下可以指安徽省黄山市, 属于地名, 也可以指旅游景区“黄山”等, 继而景点实体识别效率一般. 针对这个问题本文提出将深度学习方法应用到旅游领域景点识别中, 在现有研究基础上, 提出将 BiLSTM+CRF 方法应用旅游领域景点实体识别中. 郭剑毅等[10]提出的基于层叠条件随机场方法, 该方法过于依赖人工构建特征模板, 对于旅游领域, 景点实体数量过多, 无法一一列举, 且在人工构建特征模板的时候耗时耗力, 未能考虑到上下文语境和语义的信息. 针对该问题, 本文将 BERT 语言模型[11](Bidirectional Encoder Representation from Transformers, BERT) 融合到 BiLSTM-CRF 命名实体识别模型中. BERT 语言模型对自然语言处理任务效率有很大的提升, 利用该模型可以解决文本特征表示时的一词多义问题. BiLSTM 能够充分利用先验知识, 获取有效的上下文信息, CRF 可以考虑句子级相邻标签之间的信息, 并且获得全局最优序列. 在实际旅游领域内景点识别的测试中比以往学者的研究方法效率有显著提升. P 值, R 值, F 值分别为 8.33%, 1.71%, 6.81%.

2 BERT+BiLSTM+CRF 模型

2.1 模型框架

BERT+BiLSTM+CRF 模型由 BERT 模块、BiLSTM 和 CRF 3 个模块组成. 整体模型如图 1 所示. 首先使用 BERT 模型获取字向量, 提取文本重要特征; 然后通过 BiILSTM 深度学习上下文特征信息, 进行命名实体识别; 最后 CRF 层对 BiLSTM 的输出序列处理, 结合 CRF 中的状态转移矩阵, 根据相邻之间标签得到一个全局最优序列.

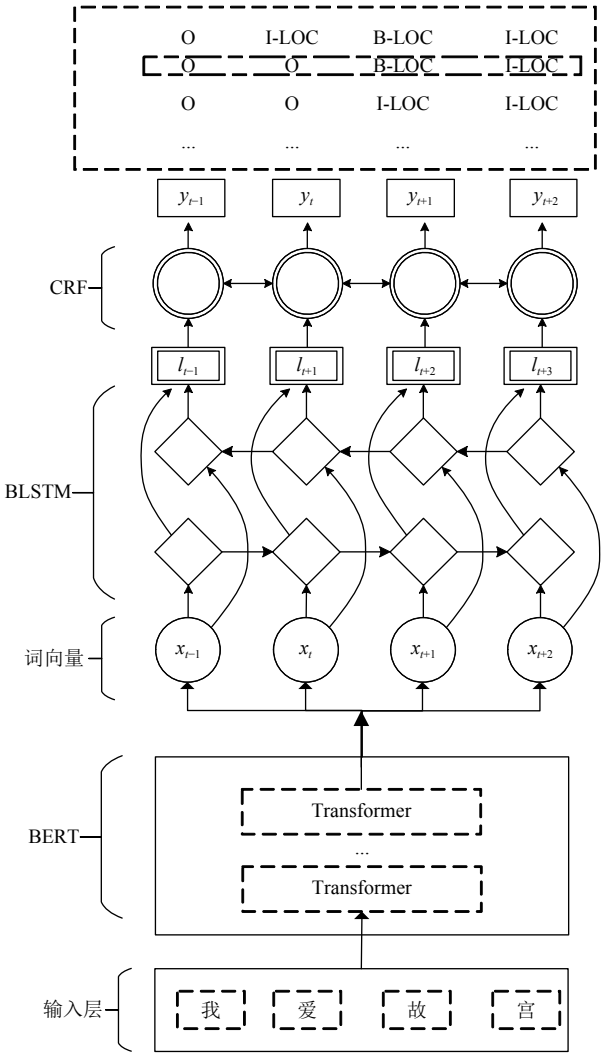

图 1 BERT+BiLSTM+CRF 模型图

模型第一层是利用预训练的 BERT 语言模型初始化获取输入文本信息中的字向量记为序列 $X=(x_1, x_2, x_3, \cdots, x_n)$, 所获取的字向量能够利用词与词之间的相互关系有效提取文本中的特征.

模型第二层为双向 LSTM 层, 第一层获取的 n 维字向量作为双向长短时记忆神经网各个时间步的输入, 得到双向 LSTM 层的隐状态序列 $\overrightarrow{h_t}$(表示前向) 和 $\overleftarrow{h_t}$(表

示后向), 待前向与后向全部处理完, 对各个隐状态序列进行按照位置拼接得到完整的隐状态序列记为 $h_t=(h_1,h_2,\cdots,h_n)\in R^{n\times m}$,接着线性输出层将完整的隐状态序列映射到 s 维 (s 维为标注集的标签类别数目), 记提取的句子特征为全部映射之后的序列为矩阵 $L=(l_1,l_2,\cdots,l_n)\in R^{n\times s}, l_i\in R^s$ 的每一维 $l_{i,j}$ 分别对应其字 x_i 对应每个类别标签 y_i 的分数值. 如果此时直接对每个位置的分数值进行独立分类, 选取每个分值最高的直接得到输出结果, 则不能考虑相邻句子之间的信息, 不能得到全局最优, 分类结果不理想. 所以引入模型最后一层.

2.1.1 BERT 模型

BERT[11]是一种自然语言处理预训练语言表征模型. BERT 能够计算词语之间的相互关系, 并利用所计算的关系调节权重提取文本中的重要特征, 利用自注意力机制的结构来进行预训练, 基于所有层融合左右两侧语境来预训练深度双向表征, 比起以往的预训练模型, 它捕捉到的是真正意义上的上下文信息, 并能够学习到连续文本片段之间的关系. 模型预训练结构图如图 2 所示.

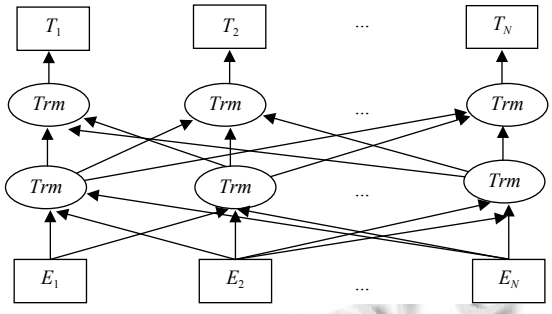

图 2 BERT 模型预训练结构图

图 2 中, *Trm* 表示[11]自注意力机制 (Transformer) 编码转换器, E_1,E_2,$\cdots$, E_N 表示模型的输入, 为词向量, 而 T_1, T_2, $\cdots$, T_N 表示模型的输出. 由于一般的语言模型不能很好理解句子之间的关系, 而在命名实体识别中句子之间的语义关系是非常重要的, 所以 BERT 模型拼接句子 L 和 M, 并预测 M 是否位于原始文本中 L 之后. 语言模型的预训练在文本特征提取时, 能解决一词多义问题继而能够改进命名实体识别的任务, 所以本文将 BERT 语言模型结合到命名实体识别的任务中, 取得了显著的效果.

2.1.2 BiLSTM

长短时记忆神经网[12]是 1997 年提出来的, 是目前最流行的递归神经网络, 其不仅对短期的输入比较敏感, 更能保存长期的状态. LSTM 的主要由 3 个开关来控制单元的输入输出.

(1) 遗忘门: 单元状态 c_{t-1} 保留到当前时刻 c_t 的决策, 计算公式如式 (1):

$$f_t=\sigma(W_{fh}\cdot h_{t-1}+W_{fx}\cdot X_t+b_f) \tag{1}$$

式中, W_{fh} 对应输入项 h_{t-1}; W_{fx} 对应输入项 X_t; W_{fh} 和 W_{fx} 组成遗忘门的权重矩阵 W_f, b_f 为偏置项, σ 为激活函数.

(2) 输入门: 当前输入 X_t 保存到 c_t 的决定, 计算公式如式 (2):

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{2}$$

式中, W_i 为权重矩阵, b_i 是偏置项.

用 $\widetilde{c}_t$ 表示当前输入的单元状态, 由上一次的输出和当前的输入确定, 如式 (3):

$$\widetilde{c}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c) \tag{3}$$

当前时刻单元状态 c_t, 如式 (4):

$$\tilde{c}_t=f_t^{\circ}c_{t-1}+i_t^{\circ}\tilde{c}_t \tag{4}$$

式中, c_{t-1} 表示前一个的单元状态, f_t 为遗忘门. 符号 $^{\circ}$ 表示按元素乘.

(3) 输出门: 计算如式 (5):

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_c) \tag{5}$$

输入门和单元状态确定了长短时记忆神经网络的输出, 如式 (6):

$$h_t=o_t^{\circ}\tanh(c_t) \tag{6}$$

神经网络可以根据文本中词的分布式表示自动提取特征, 字向量的 BiLSTM-CRF 模型, 在 BiLSTM 输出预测曾后, 由 CRF 层利用上下文已经预测的标签, 找到全局最优的标注序列, 实验对比分析见文第四部分.

2.1.3 CRF

CRF[13]用来分割和标记序列数据, 根据输入的观察序列来预测对应的状态序列, 同时考虑输入的当前状态特征和各个标签类别转移特征, 被广泛应用于 NER 的问题中. CRF 应用到 NER 的问题中主要是根据 BiLSTM 模型的预测输出序列求出使得目标函数最优化的序列.

两个随机变量 X 和 Y, 在给定 X 的条件下, 如果每个 Y_V 满足未来状态的条件概率与过去状态条件独立[13], 如式 (7):

$$P(Y_v|X,Y_u,u\neq v)=P(Y_v|X,Y_u,u\sim v) \tag{7}$$

则 (X,Y) 为一个 CRF. 常用的一阶链式结构 CRF[13] 如图 3 所示.

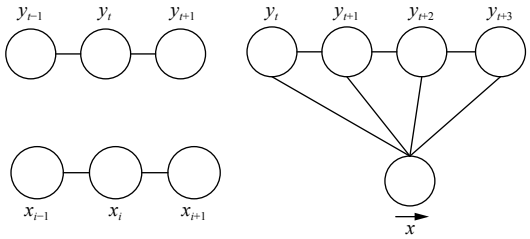

图 3 条件随机场一阶链式结构

CRF 应用到 NER 中是在给定需要预测的文本序列 $X=\{x_1,x_2,\cdots,x_n\}$, 根据 BERT-BiLSTM 模型的输出预测序列 $Y=\{y_1,y_2,\cdots,y_n\}$, 通过条件概率 $P(y|x)$ 进行建模, 则有式 (8):

$$P(y|x)=\frac{1}{z(x)}\exp(\sum_{i,m}\lambda_m\mu_m(y_{i-1},x,i)\cdot\sum_{i,n}\beta_n t_n(y_i,x,i)) \quad (8)$$

其中, i 表示当前节点在 x 中的索引, m, n 表示在当前节点 i 上的特征函数总个数. t_n表示节点特征函数, 只和当前位置有关. μ_m 表示局部特征函数, 只与当前位置和前一个节点位置有关. $\beta_n\lambda_m$ 分别表示特征函数 t_n和μ_m对应的权重系数, 用于衡量特征函数的信任度. $z(x)$归一化因子, 如式 (9):

$$Z(x)=\sum_y\exp(\sum_{i,m}\lambda_m\mu_m(y_{i-1},x,i)\cdot\sum_{i,n}\beta_n t_n(y_i,x,i)) \quad (9)$$

2.2 算法描述

算法 1. 景点实体提取算法

输入: 旅游游记文章
输出: 景点实体集

1. get_train_example(data_dir),get_test_example(data_dir),get_labels() /*获取训练数据 examples、测试数据 predict_examples、标签集 labels;
2. convert_single_example() /* 分析样本, 将字、标签全部转化为 id,次数对文本进行按照序列截断, 在句子开头结尾加上标识符, 结构化存储到 InputFeature 对象中, 存为一个类*/
3. TFRecordWriter(output_file) /*将步骤 2 中的数据转化为 TF_Record 格式*/
4. for (ex_index, example) in enumerate (examples) /*遍历所有训练样本重复步骤 2 和步骤 3*/
5. create_model(), model_fn()/*构建模型, 初始化参数,使用 BERT 加载获取每个字对应的 embedding, 训练基于 BERT-BiLSTM-CRF 的实体识别模型*/
6. filed_based_convert_examples_to_features()/*使用步骤 2 中的 predict_examples 作为模型的输入, 得到实体识别结果 result*/
7. end for
8. return result

3 数据集

3.1 构建数据库

本文从马蜂窝等互联网旅游网站上通过爬虫技术获取 1 万余篇旅游游记文章, 将数据解析成 TXT 文件, 进行数据清洗, 通过正则表达式去除无用的网址、特殊的标点符号以及一些符号化的字等信息, 按照优先级处理特殊符号, 但是保留逗号, 句号等重要的标点符号. 数据预处理流程如图 4 所示.

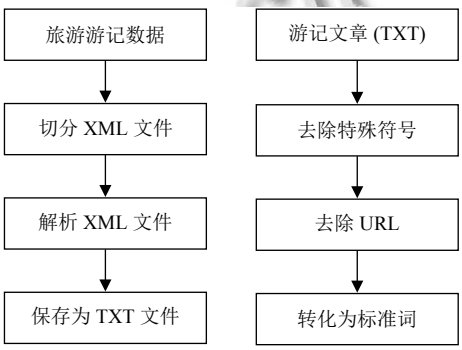

图 4 数据清洗预处理

词边界特征能很好地表示边界字符的位置信息, 有助于确定实体边界, 所以本文按照 BIO 标注格式 (B 表示景点开始标志, I 表示词的中间部分, O 表示其他非景点的词) 进行自动化标注, 并建立自己的旅游游记数据库 (TDB). 标注实例如表 1 所示, 数据分布情况如表 2 所示.

表 1 标注实例

字	我	喜	欢	去	鸟	巢
词性	r	v	v	v	n	n
标注	O	O	O	O	B-SE	I-SE

表 2 数据分布情况

数据	实体个数	非实体个数	总量
训练集	37 861	213 733	251 594
测试集	36 569	243 307	279 876

4 实验过程

4.1 评价指标

本文采用 MUC 评测会议上所提出的命名实体识别的评价指标, MUC-2 上[1]提出的 NER 的最初评价指标: 精确率 (Precision, P), 召回率[1](Recall, R). 本文中主要采用 P、R 和 F 值 (F 值为召回率和精确率的加权调

和平均值) 作为评价指标计算式 (10)~式 (12). 如表 3 所示.

表 3 评价指标相关解释

	实际值为景点	实际值为非景点
测试值为景点	N_{correct}	$N_{\text{incorrect}}$
测试值为非景点	N_c	N_d

$$R = \frac{N_{\text{correct}}}{N_c + N_d} \tag{10}$$

$$P = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{incorrect}}} \tag{11}$$

$$F1 = \frac{2PR}{P+R} \tag{12}$$

当$\alpha = 1$时, 式 (12) 为最常见的 $F1$ 值, 计算公式如式 (12), 当 $F1$ 值较高说明实验方法比较有效.

4.2 模型分层测试实验结果

为了验证本文所提出模型的有效性, 从 TDB 数据库随机抽取 19 965 条句子作为训练集和 19 690 条句子作为测试集进行实验. 本文设置了 4 组对比试验, 分别与 CRF 模型, BiLSTM 模型, BiLSTM+CRF 模型进行分层测试对比, 来验证每个模块的重要性. 以下实验训练数据和测试数据均为同一数据集. 实验对比分析如表 4 所示.

表 4 模型验证实验分析 (%)

模型	P	R	F 值
CRF	86.67	87.84	87.25
BiLSTM	93.25	87.98	90.53
BiLSTM+CRF	94.97	92.10	93.52
本文	**95.83**	**97.41**	**96.61**

由表 4 可知, 本文所提出的方法 P 值, R 值, F 值在相对于其他 3 组对比实验中效果最好的, P 值, R 值, $F1$ 值上分别提升了 0.86%, 5.31%, 3.09%.

(1) 单层 BiLSTM 模型

观察实验测试数据可知, 由于 CRF 能够充分考虑标注序列的顺序性, 得到全局最优标注序列, 所以缺少 CRF 层会将一个完整实体拆分 (如“故宫博物院”) 拆分成“故宫”“博物院”两个实体, 而 BiLSTM 虽说能够考虑上下文信息, 但是其输出序列只根据当前词输出得分最大值, 容易将完整实体细分. 所以对于 BiLSTM、BiLSTM+CRF 两种方法而言, 后者识别效果较好.

(2) 单层 CRF 模型

由于 CRF 只是传统的机器学习方法, 过于依赖人工构建的特征模板, 缺乏深度学习方法中上下文信息的特征, 而景点实体的识别对上下文语义理解依赖较大, 所以 BiLSTM+CRF 方法相比较而言, 在 P 值上比 CRF 提高了 8.3%, R 值提高了 9.57%.

(3) 双层 BiLSTM+CRF 层

去除 BERT 模型时, 由于在文本特征提取的时候不能解决同一个单词不同语境下的特征表示问题, 针对一词多义问题不能得到很好的解决, 比如“北京海洋馆”中的“海洋”在不同语境下可以指人名也可以指景点名, 对于此类问题不能得到解决, 导致准确率, 召回率等下降.

(4) BERT+BiLSTM+CRF 模型

结合三层的模型, 可以通过 BiLSTM 获取上下文有效信息特征, 可以解决特征表示的一词多义问题, 结合 BiLSTM+CRF 的优势, 识别效率相对较高.

4.3 相关工作对比分析

经调研发现, 目前对旅游领域内景点识别的方法最好的方法为薛征山[9]和郭剑毅[10]两人所提出的, 为验证本文所提出方法的应用性, 从所构建的 TDB 数据库中随机抽取 19 965 个句子作为训练集, 和 19 690 个句子作为测试集进行实验设计了 3 组实验对比分析, 对旅游领域内的游记文章进行景点实体识别, 并与以往研究者薛征山[9]提出的基于 HMM 的中文旅游景点识别方法与郭剑毅[10]所提出的基于层叠条件随机场方法进行对比分析; 使用 3.3 节中的评价指标得到实验结果如表 5.

表 5 景点识别验证结果 (%)

方法	P	R	F 值
文献[10]	83.4	95.7	89.1
文献[9]	87.5	92.3	89.8
本文	95.83	97.41	96.61

观察实验结果可知, 本文所提出的基于深度学习方法比机器学习方法在识别效率上有大幅度的提升, 主要原因为深度学习能够学习文本上下文语义信息, 而本文在此基础上解决了文本特征表示时的一词多义问题, 所以该模型在旅游领域内景点识别相对以往研究者效率有一定提升, 其中 P 值和 F 值相对于薛征山[9]分别提高了 8.33% 和 6.81%, R 值相对于郭剑毅[10]提高了 1.71%.

5 结论

本文研究设计了一种融合新的语言模型 BERT 的 BiLSTM+CRF 景点实体识别方法. 利用 BERT 语言模型能够解决在文本特征表示的一词多义问题, 结合 BiLSTM 深度学习方法充分学习上下文信息的特点以及 CRF 机器学习方法提取全局最优标注序列, 得到景点实体. 在实验中进行了验证, *P* 值, *R* 值和 *F* 值均高达 95% 以上, 且 *P*, *R*, *F* 值相比以往研究者所提出的方法分别提高了 8.33%, 1.71%, 6.81%. 解决了旅游景点实体识别效率一般的问题, 将为解决从旅游游记文本中自动提取旅游线路的问题提供了技术支撑.

参考文献

1 Grishman R, Sundheim B. Message understanding conference-6: A brief history. Proceedings of the 16th Conference on Computational Linguistics. Stroudsburg, PA, USA. 1996. 466–471.

2 包敏娜, 斯·劳格劳. 基于词典匹配的蒙古文命名实体识别研究. 中央民族大学学报 (哲学社会科学版), 2017, 44(3): 165–169.

3 McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the 7th Conference on Natural Language Learning. 2003. 151–156.

4 Chieu HL, Ng HT. Named entity recognition: A maximum entropy approach using global information. Proceedings of the 19th International Conference on Computational Linguistics. Taipei, China. 2002. 1–7.

5 Bender O, Och FJ, Ney H. Maximum entropy models for named entity recognition. Proceedings of 2003 Conference on Computational Natural Language Learning. Edmonton, Canada. 2003. 121–126.

6 Wang X, Zhang Y, Ren X, *et al*. Cross-type biomedical named entity recognition with deep multi-task learning. arXiv: 1801.09851, 2018.

7 Cai XL, Dong SB, Hu JL. A deep learning model incorporating part of speech and self-matching attention for named entity recognition of Chinese electronic medical records. BMC Medical Informatics and Decision Making, 2019, 19(S2): 65. [doi: 10.1186/s12911-019-0762-7]

8 Li LS, Jiang YX. Integrating language model and reading control gate in BLSTM-CRF for biomedical named entity recognition. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018. [doi: 10.1109/TCBB.2018.2868346]

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

10 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. 中文信息学报, 2009, 23(5): 47–52. [doi: 10.3969/j.issn.1003-0077.2009.05.007]

11 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805, 2019.

12 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

13 Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, San Francisco, CA, USA. 2001. 282–289.