基于 LEBERT 的中文命名实体识别方法

沈言玉[1] 王 芬[1] 赵宇航[1]
SHEN Yanyu WANG Fen ZHAO Yuhang

摘 要 预训练模型 BERT（bidirectional encoder representation from transformers）因其卓越的性能在中文命名实体识别任务中取得了显著成果，但 BERT 在处理中文文本时，未充分考虑词汇信息。为克服这一局限，文章提出了一种基于 LEBERT（lexicon enhanced BERT）的中文命名实体识别方法，结合 BiLSTM（bidirectional long short-term memory）和 CRFF（conditional random field）模型，进一步提升识别性能。LEBERT 在预训练阶段通过引入词汇嵌入，使得模型能够更好地捕捉词汇的语义信息；BiLSTM 用于捕捉序列数据中的双向依赖关系；CRF 层则用于解码最优的标签序列，不仅考虑到标签之间的转移概率，还避免了非法实体的出现。实验结果表明，该方法在 Weibo、Resume、OntoNotes 数据集分别取得 71.57%、96.53%、82.04% 的 F_1 值，优于其他主流方法。

关键词 中文命名实体识别；LEBERT；BiLSTM；词汇增强

doi：10.3969/j.issn.1672-9528.2025.06.030

0 引言

信息技术的飞速发展催生了海量文本数据，促使自动从文本中提取有用信息的需求激增。中文命名实体识别作为自然语言处理的关键任务，在信息检索、文本摘要、问答系统、知识图谱构建等多个领域发挥着重要作用[1]。中文命名实体识别旨在精准识别文本中的特定实体边界和类型，如人名、地名、机构名等。但中文的词边界模糊、形态学特征缺乏及新实体频出等特点，给中文命名实体识别带来了诸多挑战。中文命名实体识别技术历经基于规则、机器学习到深度学习三个阶段[2]。基于规则的方法虽准确，却因规则构建和维护的复杂性，难以应对新实体。机器学习方法虽简化了规则构建，却需大量特征工程，限制了性能提升。深度学习的兴起，尤其是 CNN 和 LSTM 模型，能自动学习高层次特征，无需复杂特征工程，显著提高了中文命名实体识别的准确性。

Ma 等人[3]提出了 SoftLexicon 方法将词汇信息融入字符表示，仅需对字符表示层进行简单调整即可引入词典信息。Zhang 等人[4]提出了基于 Lattice 结构的 LSTM 模型，同时编码字符序列与字典匹配的词汇信息。Chang 等人[5]提出基于 BERT（bidirectional encoder representation from transformers）的方法，利用 BERT 的双向编码能力和预训练语言表示，端到端完成命名实体识别。Li 等人[6]提出 FLAT 模型，通过 Transformer 的自注意力机制和优化的位置编码，利用字符 - 词晶格结构实现高效并行计算。Wu 等人[7]改进 FLAT（feature-rich lattice attention transformer）模型，提出 MECT（multimodal entity and context transformer）模型，通过融合部首、字符和词汇信息，利用双流 Transformer 提升性能。Liu 等人[8]提出 LEBERT（lexicon enhanced BERT）模型，在 BERT 中引入词典信息增强模块，将词汇信息与上下文表示相结合，提升了中文序列标注性能。Zhu 等人[9]提出边界平滑技术，通过将实体标注概率分配至周围 span，缓解模型过度自信问题。

1 模型介绍

模型主要分为三部分：基于 LEBERT 的字嵌入表示层、特征提取层 BiLSTM（bidirectional long short-term memory）层以及标签解码层 CRF（conditional random field）层，如图 1 所示。LEBERT 层对输入的文本序列进行编码，通过引入词汇信息来增强模型能力；BiLSTM 通过正向和反向两个方向进行训练，捕捉序列中的长期依赖关系；CRF 层对特征提取层提取的序列进行标签优化，得到每个位置最可能的标签序列。

1. 江苏开放大学 江苏南京 210036

[基金项目] 华东地区开放大学联盟联合科研攻关课题“融合知识图谱的开放教育：数字化转型的路径研究”（HDLMKT2217）；江苏开放大学 2024 年度信息化专项科研课题“生成式 AI 赋能校园管理服务的应用与实践”（2024XXHKY014）

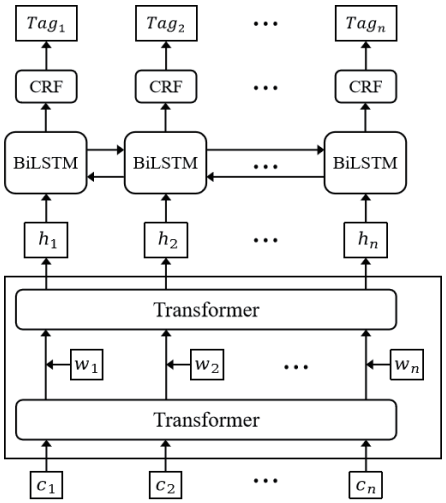

图 1 模型结构图

1.1 嵌入表示层 LEBERT

LEBERT 模型通过在 BERT 模型的 Transformer 层之间引入词典适配器（Lexicon Adapter）使词典信息可以有效地融合在 Bert 模型中，从而增强特征信息。这种设计使得模型能够在 BERT 的较低层进行词典知识的深度融合。在中文分词领域，由于中文句子中的词汇之间没有明显的分割符，LEBERT 通过在 BERT 的中间层嵌入一个词典适配器，引入词汇的词嵌入向量，使得模型能够同时考虑字符层面和词汇层面的特征。LEBERT 模型引入了字符 - 词汇对作为模型的输入特征，这有助于模型同时捕捉字符层面和词汇层面的特征。

1.1.1 字符 - 词汇对序列

对于输入文本中的每个字符，LEBERT 找出在输入文本中匹配到的所有词汇，形成字符 - 词汇对，这种结构使得模型能够捕捉到每个字符对应的词汇信息。为匹配输入序列中的词汇，LEBERT 会提前构建一棵字典树。这棵树用于找出输入文本中每个字符所匹配到的所有词汇，从而得到每个字符对应的词序列。每个字符和包含该字符的词汇组成词汇对，表示为 (c_i, w_i)，其中 c_i 是句子中的第 i 个字符，w_i 是包含 c_i 词汇组成的集合。基于此，整个句子就被转换成了一个字符 - 词汇对序列，即 s_{cw}={ (c_1, w_1), (c_2, w_2),…, (c_n, w_n) }。

1.1.2 Lexicon Adapter

在构建了字符 - 词汇对序列之后，句子中的每个位置都包含了字符特征和词汇特征。为了将这些特征融合，LEBERT 设计了 Lexicon Adapter 结构，如图 2 所示。

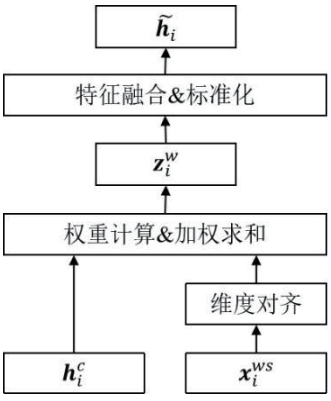

图 2 词典适配器

该结构在 BERT 的 Transformer 层之间注入词汇特征，通过在第 k 层和第（k+1）层 Transformer 之间注入词汇信息，Lexicon Adapter 接收两个输入：字符向量 $\boldsymbol{h}_i^c$ 和成对的词向量 $\boldsymbol{x}_{ij}^w$，由于字符向量 $\boldsymbol{h}_i^c$ 和词向量 $\boldsymbol{x}_{ij}^w$ 的维度不一致，首先将词向量进行非线性映射，将其与字符向量进行维度对齐，其公式为：

$$\boldsymbol{v}_{ij}^w = \boldsymbol{W}_2\left(\tanh\left(\boldsymbol{W}_1\boldsymbol{x}_{ij}^w + b_1\right)\right) + b_2 \tag{1}$$

式中：$\boldsymbol{W}_1 \in \mathbb{R}^{d_c \times d_w}$；$\boldsymbol{W}_2 \in \mathbb{R}^{d_c \times d_c}$；$b_1$ 和 b_2 是偏置项；d_c 和 d_w 分别表示 BERT 的隐藏层大小和词嵌入的维度。

对于每个字符，计算其所匹配到的每个词向量的权重，使用双线性注意力机制得到 $\boldsymbol{a}_i$，$\boldsymbol{a}_i$ 计算公式为：

$$\boldsymbol{a}_i = \mathrm{softmax}\left(\boldsymbol{h}_{c_i}\boldsymbol{W}_{\mathrm{attn}}\boldsymbol{V}_i^{\mathrm{T}}\right) \tag{2}$$

式中：$\boldsymbol{W}_{\mathrm{attn}} \in \mathbb{R}^{d_c \times d_c}$；$\boldsymbol{V}_i = \{\boldsymbol{v}_{i1}^w, \boldsymbol{v}_{i2}^w, \ldots, \boldsymbol{v}_{im}^w\}$是是分配给第 i 个字符的所有词向量的集合。

对于每个字符，根据其匹配到的每个词汇的权重，对词向量进行加权求和，得到该字符的加权词汇向量$\boldsymbol{z}_i^w$：

$$\boldsymbol{z}_i^w = \sum_{j=1}^{m} \boldsymbol{a}_{ij}\,\boldsymbol{v}_{ij}^w \tag{3}$$

将字符向量$\boldsymbol{h}_{c_i}$与加权词汇向量$\boldsymbol{z}_i^w$相加，得到特征融合向量$\tilde{\boldsymbol{h}}_i$。最后将$\tilde{\boldsymbol{h}}_i$进行 dropout、层标准化和残差连接等操作，得到 Lexicon Adapter 的最终输出。

1.2 特征提取层 BiLSTM

在中文命名实体识别中，文本数据具有时间序列或事件序列特性，使用 BiLSTM 可以有效地捕捉序列中的上下文信息及其长期依赖关系。BiLSTM 模型由两个独立的 LSTM 单元组成：一个用于前向处理输入序列，另一个用于后向处理输入序列。

给定一个输入序列 $\boldsymbol{X}$，BiLSTM 层会分别用前向 LSTM（LSTM_f）和后向 LSTM（LSTM_b）按正序和逆序处理该序列。因此，对于每个时刻 t，用公式表示为：

$$\overrightarrow{\boldsymbol{h}_t} = \mathrm{LSTM}_f\left(\boldsymbol{x}_t, \overrightarrow{\boldsymbol{h}_{t-1}}\right) \tag{4}$$

$$\overleftarrow{\boldsymbol{h}_t} = \mathrm{LSTM}_b\left(\boldsymbol{x}_t, \overleftarrow{\boldsymbol{h}_{t+1}}\right) \tag{5}$$

式中：$\overrightarrow{\boldsymbol{h}_t}$代表前向 LSTM 在时刻 t 的输出；而$\overleftarrow{\boldsymbol{h}_t}$代表后向 LSTM 在时刻 t 的输出。

最终，BiLSTM 层的输出是对这两个输出的合并，可以计算 BiLSTM 的输出，即：

$$\boldsymbol{y}_t = \mathrm{BiLSTM}(\boldsymbol{X}) = \left[\overrightarrow{\boldsymbol{h}_t}; \overleftarrow{\boldsymbol{h}_t}\right] \tag{6}$$

通过 BiLSTM 层，可以计算每个词的上下文表示，进而把每一时刻的输出合并，如图 3 所示。该层的输出可以被用作特征提取器，提取输入序列中每个词的上下文信息。

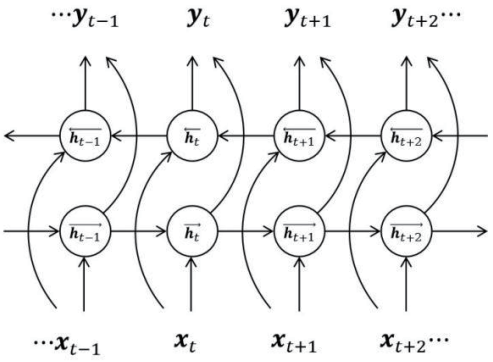

图 3 BiLSTM 结构

1.3 标签解码层 CRF

CRF 层的作用是将序列标注任务中的标签依赖关系纳入考虑，以提高预测的准确性。CRF 通过特征函数捕捉输入序列与标签之间的复杂关系，并利用这些特征函数的权重来优化模型。在中文命名实体识别中，CRF 层将 BiLSTM 层输出的每个字符对应各个标签的发射概率，结合标签之间的转移概率，进一步优化标签序列的预测。

CRF 通过定义转移特征和状态特征来捕捉序列中标记之间的依赖关系以及标记与输入序列之间的关系。转移特征关注相邻标记间的联系，而状态特征则关联当前标记与输入序列的特征。这些特征的权重通过训练数据学习得到，使模型能够更准确地预测每个字符的标签。

2 实验结果

2.1 实验数据集

本实验在公开的 3 个数据集上进行实验，分别是 Weibo、Resume、OntoNotes 4.0。

Weibo 数据集是一个基于微博平台的中文命名实体识别数据集。不仅标注了 4 种不同的实体，同时还标注了具体和泛指。这个数据集涵盖了社交媒体上用户生成内容的丰富性和多样性，对于研究社交媒体文本中的实体识别具有重要价值。

Resume 数据集是面向简历的中文命名实体识别数据集，标注了 8 种不同类型的实体，包括教育背景、工作经验等，对于简历信息提取和自动化处理具有重要意义。

OntoNotes 4.0 数据集来源于多种文本类型，包括新闻报道、访谈记录、网络评论和论坛帖子，涵盖了丰富的语言场景和多样的文本风格，OntoNotes 4.0 数据集的命名实体标注包含人名、地名、组织名以及地缘政治实体。

2.2 评价指标

实验中采用 3 个经典评价指标精确率（Precision）、召回率（Recall）、F_1 值来衡量命名实体识别模型的性能，计算方式分别为：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

式中：TP 为正确预测为正例的数量；FP 为错误预测为正例的数量；FN 为实际为正例但被预测为负例的数量。

2.3 实验环境

本研究的实验环境包括使用 NVIDIA GeForce RTX 3080 GPU 进行模型训练，使用的 Python 版本为 3.8，PyTorch 版本为 1.10.0，本实验中使用的主要参数如表 1 所示。

表 1 实验参数设置

超参数	设置值
训练轮次	30
批处理大小	32
学习率	5e-5
CRF 学习率	1e-3
输入序列的最大长度	150
优化器	AdamW

2.4 实验结果

本文采用 BERT-Softmax、BERT-CRF、LEBERT-Softmax、LEBERT-CRF、BERT-BiLSTM-CRF 模型作为对比算法进行实验，这些模型分别运用了 LSTM、CRF、Transformer、BERT 等技术，涵盖了多种常用的命名实体识别方法。

表 2 在 Weibo 数据集上各个对比模型的评测指标

模型	精确率 P/%	召回率 R/%	F_1 值 /%
BERT-Softmax	69.30	69.81	69.56
BERT-CRF	71.00	68.60	69.78
BERT-BiLSTM-CRF	69.25	71.26	70.24
LEBERT-Softmax	70.86	69.32	70.09
LEBERT-CRF	69.91	71.26	70.57
LEBERT-BiLSTM-CRF	72.64	70.53	71.57

表 3 在 Resume 数据集上各个对比模型的评测指标

模型	精确率 P/%	召回率 R/%	F_1 值 /%
BERT-Softmax	95.55	96.38	95.96
BERT-CRF	95.91	96.50	96.20
BERT-BiLSTM-CRF	96.04	96.93	96.48
LEBERT-Softmax	95.91	96.62	96.26
LEBERT-CRF	96.15	96.74	96.44
LEBERT-BiLSTM-CRF	96.33	96.20	96.53

表 4 在 OntoNotes 4.0 数据集上各个对比模型的评测指标

模型	精确率 P/%	召回率 R/%	F_1 值 /%
BERT-Softmax	81.07	80.77	80.91
BERT-CRF	78.89	83.72	81.23
BERT-BiLSTM-CRF	81.47	81.47	81.47
LEBERT-Softmax	82.07	81.63	81.31
LEBERT-CRF	81.46	81.63	81.55
LEBERT-BiLSTM-CRF	81.91	82.17	82.04

从实验结果来看，所提出的 LEBERT-BiLSTM-CRF 模型在所有数据集上均取得了最佳的 F_1 值，验证了模型的优越性能。如表 2 所示，在 Weibo 数据集上，LEBERT-BiLSTM-CRF 的 F_1 值为 71.57%，相比其他模型具有明显的性能优势，尤其在精确率上达到了 72.64%，优于所有对比模型；如表 3 所示，在 Resume 数据集上，LEBERT-BiLSTM-CRF 的 F_1 值为 96.53%，相较于 BERT-BiLSTM-CRF 和 LEBERT-CRF 均有小幅提升，且精确率和召回率表现均衡；如表 4 所示，在 OntoNotes 4.0 数据集上，LEBERT-BiLSTM-CRF 的 F_1 值为 82.04%，同样优于其他模型，进一步验证了其强大的序列标注能力。综合各项指标，LEBERT-BiLSTM-CRF 在精确率、召回率和 F_1 值上均表现出色，超越基于 BERT 的其他模型以及部分基于 LEBERT 的模型。

实验结果表明，LEBERT-BiLSTM-CRF 模型能够有效结合 LEBERT 提供的全局上下文语义和词汇级信息，通过 BiLSTM 加强对局部依赖关系的建模，并利用 CRF 层进行全局优化，显著提升了模型的序列标注能力。在 Weibo 数据集上，LEBERT 的词汇增强能力在处理中文社交媒体的噪声文本时表现尤为突出，而 BiLSTM 和 CRF 的结合有效缓解了标签预测中的错误传播问题。在 Resume 和 OntoNotes 4.0 等较为规范的数据集上，LEBERT-BiLSTM-CRF 的性能稳定且略有提升，表明其对不同领域任务具有较强的泛化能力。总体而言，该模型充分结合了预训练语言模型的语义建模优势和序列标注任务的特定需求，为中文命名实体识别提供了一个高效且鲁棒的解决方案。

3 总结

针对中文命名实体识别中未能考虑中文文本中词汇信息的问题，本文提出了一种基于 LEBERT-BiLSTM-CRF 的中文命名实体识别模型。该模型通过在字符表示层面引入 LEBERT 模型，该模型能够有效融合外部词汇信息，增强对语义的理解，BiLSTM 层使模型可以捕捉字符序列间的双向依赖关系，CRF 则通过对标签之间的转移概率进行建模进一步提高了模型的准确率。最后实验部分在 3 个公开的数据集的实验结果表明本文提出的方法优于其他主流方法，证明了该方法的有效性。

参考文献：

[1] 赵继贵，钱育蓉，王魁，等 . 中文命名实体识别研究综述 [J]. 计算机工程与应用 ,2024,60(1):15-27.

[2] LIU P, GUO Y M, WANG F L, et al. Chinese named entity recognition: the state of the art[J]. Neurocomputing, 2022, 473: 37-53.

[3] MA R T, PENG M L, 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. Brussels: ACL, 2020:5951-5960.

[4] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.Brussels:ACL,2018:1554-1564.

[5] CHANG Y, KONG L, JIA K J, et al. Chinese named entity recognition method based on BERT[C]//2021 IEEE international conference on data science and computer application (ICDSCA). Piscataway: IEEE, 2021: 294-299.

[6] LI X N, YAN H, QIU X P, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Brussels: ACL, 2020: 6836-6842.

[7] WU S, SONG X N, FENG Z H. MECT: multi-metadata embedding based cross-transformer for Chinese named entity recognition[EB/OL].(2021-07-12)[2024-05-19]. https://doi.org/10.48550/arXiv.2107.05418.

[8] LIU W, FU X Y, ZHANG Y, et al. Lexicon enhanced Chinese sequence labeling 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. Brussels: ACL, 2021:5847-5858.

[9] ZHU E W, LI J P. Boundary smoothing for named entity recognition[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.Brussels: ACL, 2022: 7096-7108.

【作者简介】

沈言玉（1996—），女，安徽六安人，硕士，研究方向：自然语言处理、人工智能安全。

王芬（1995—），女，安徽池州人，硕士，研究方向：网络安全、数据分析。

赵宇航（1996—），男，江苏宿迁人，硕士，研究方向：知识图谱、推荐系统。

（收稿日期：2025-01-14 修回日期：2025-06-03）