民间文学文本命名实体识别方法

黄健钰[1,2],王笳辉[1,2],段 亮[1,2],冉 苒[3]

(1. 云南大学 信息学院;2. 云南省智能系统与计算重点实验室;3. 云南大学 文学院,云南 昆明 650500)

摘 要: 民间文学文本命名实体识别任务旨在从民间文学文本中判别实体并将其划分到预定义的语义类别,为民间文学的保存与传播奠定基础。民间文学区别于一般中文语料,其文本存在一词多义情况突出与领域名词众多的问题,导致常规命名实体识别方法难以准确充分地识别出民间文学文本中存在的实体及其类别。针对该问题,提出一种基于BERT的民间文学文本命名实体识别模型TBERT。该模型首先在通用中文BERT模型的基础上融合民间文学文本语料特征与实体类型特征;然后利用BiLSTM模型进一步提取序列依赖特征;最后结合CRF模型获取的标签约束信息输出全局最优结果。实验结果表明,该方法在民间文学文本数据集上具有良好表现。

关键词: 民间文学文本;命名实体识别;Fine-Tune;TBERT-BiLSTM-CRF;特征融合

DOI: 10.11907/rjdk.222235 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP391.1 **文献标识码:** A **文章编号:** 1672-7800(2023)010-0065-08

Named Entity Recognition Method for Folk Literature Texts

HUANG Jianyu[1,2], WANG Jiahui[1,2], DUAN Liang[1,2], RAN Ran[3]

(1. *School of Information Science and Engineering, Yunnan University*; 2. *Yunnan Key Laboratory of Intelligent Systems and Computing*; 3. *School of Chinese Language and Literature, Yunnan University, Kunming* 650500, *China*)

Abstract: The task of identifying named entities in folklore texts aims to identify entities from folklore texts and classify them into predefined semantic categories, laying the foundation for the preservation and dissemination of folklore. Folk literature is different from general Chinese corpus in that its text has prominent polysemy and numerous domain nouns, which makes it difficult for conventional named entity recognition methods to accurately and fully identify the entities and their categories present in folk literature texts. To address this issue, a folk literature text named entity recognition model TBERT based on BERT is proposed. This model first integrates the corpus features and entity type features of folk literature texts on the basis of the universal Chinese BERT model; Then, the BiLSTM model is used to further extract sequence dependent features; Finally, combine the label constraint information obtained from the CRF model to output the global optimal result. The experimental results show that this method performs well on the dataset of folk literature texts.

Key Words: folk literature texts; named entity recognition; Fine-Tune; TBERT-BiLSTM-CRF; feature fusion

0 引言

民间文学是由人民群众以口头方式创作并传播,且经过不断集体修改与加工的文学,常以民间传说、民间故事、神话诗歌等形式存在。保护民间文学有利于传承中华民族的传统文化,建立文化自信。命名实体识别(Named Entity Recognition, NER)任务旨在从非结构化文本中判别实体并将其分类为预定义的语义类别(如人名、组织和位置)[1-2]。NER技术可以快速识别民间文学文本中的关键词汇,在信息检索、自动文本摘要、问题回答等[3-4]各种自然语言处理任务中扮演着重要角色,为民间文学的保存与传播提供了技术支撑。

与通用语言文本不同,民间文学文本语言特点不一、

收稿日期: 2022-10-19

基金项目: 国家自然科学基金项目(62002311);云南省教育厅科研基金项目(2022Y010);云南大学研究生实践创新项目(2021Y174);云南大学研究生实践创新项目(2021Z44)

作者简介: 黄健钰(1998-),男,云南大学信息学院硕士研究生,研究方向为自然语言处理;王笳辉(1996-),男,云南大学信息学院博士研究生,研究方向为贝叶斯深度学习、自然语言处理;段亮(1986-),男,博士,云南大学信息学院副教授、硕士生导师,研究方向为机器学习、大数据分析;冉苒(1999-),女,云南大学文学院硕士研究生,研究方向为少数民族语言文学。本文通讯作者:王笳辉。

形式混杂，对其进行NER具有一定挑战。首先，民间文学文本中的一词多义问题突出，如语句“池塘生长着千瓣莲花”中的“千瓣莲花”表示一种物品，而语句“千瓣莲花姑娘”中的“千瓣莲花”表示角色“仙女”；语句“英勇的勐兰嘎”中的“勐兰嘎”表示一个角色，但在“勐兰嘎部落”中则表示一个组织；“赞颂”不仅为非实体动词，还在语句“他们给孩子取名叫做赞颂”中表示角色。由以上示例可以看出，如何准确识别民间文学文本中的实体及其具体类型十分困难，需要NER模型能够在给定语境中将该类多义词判定为其正确的实体类型，从而获得高质量的实体数据。此外，民间文学文本中存在较多领域专有名词，如“俄耶”在民间文学文本中表示“阿妈”；“粑粑”表示一种饼类食物；“国哈火塔”表示“凶猛的人”；“卡”表示“毒药”。这些领域名词未采用现代汉语中的常见释义，使得通用模型难以理解其语义，从而影响实体判定，导致识别结果无法达到预期。

传统的NER方法通常采用Word2Vec技术[5]计算词之间的语义相似度，将文本字符转化为词向量，通过BiLSTM-CRF（Bidirectional Long Short-Term Memory- Conditional Random Fields）模型进行序列建模与特征提取并输出预测标签，难以针对一词多义问题准确划分实体类型，也难以识别具有领域特色的实体。BERT预训练模型能够抽取文本特征，产生蕴含丰富句法与语义信息的词嵌入[6]，但一般中文BERT（Bidirectional Encoder Representation from Transformers）预训练模型基于维基百科与大型书籍语料训练获得，在民间文学文本NER中存在一定局限性，仍有改进空间。

1 相关研究

NER技术主要分为基于规则的识别方法和基于语言模型的识别方法两大类[7]。基于规则的识别方法要求研究者对于领域知识具备一定了解，能够根据研究领域的知识特点总结出相关规则并应用于问题的解决方法中；基于语言模型的识别方法则不要求研究者具备专业领域知识，其将NER作为一种序列标注和预测任务，通过对现有机器学习模型迁移学习后再进行识别。

在通用领域，郑玉艳等[8]利用元路径探测种子实体间的潜在特征以扩展实体集合，尝试解决最优种子的选择问题；Ju等[9]在BILSTM+CRF模型上叠加平面NER层以提取嵌套实体特征，该方法对于深层次实体的识别效果较为明显；琚生根等[10]利用关联记忆网络结合实体标签信息特征以提高模型的整体分类能力，但对部分少样本实体分类效果不明显；Xu等[11]在字符嵌入中添加汉字部首特征，获得了良好的模型表现，证实了在不同粒度中同时利用多个嵌入的有效性；武惠等[12]利用迁移学习算法缓解了模型对于少量实验数据学习能力不足的问题，以自动捕获特征的方式有效解决了领域知识的需求问题；Wang等[13]利用已训练完成的NER模型提取旧类数据特征以合成新数据，通过实体数据增量方法提升了模型训练效果；Nie等[14]提出一种对语义进行扩充的方法，提升了模型对于稀疏实体的识别效果。以上方法考虑了中文通用领域知识的特点，通过提取汉字特征、实体结构特征等方式提升模型性能，而民间文学中存在着大量领域专有名词，以上方法难以识别。

在垂直领域，余俊康[15]利用交叉共享结构学习多个相关任务的特征，克服了通用模型需要大量领域标注数据的问题；杨锦锋等[16]分阶段规范标注法则，借助领域知识特点抽取中文电子病历实体关系，但该方法对实体的一致性要求较高；Li等[17]建立临床命名实体识别（CNER）模型，分别使用LSTM和CRF提取文本特征和解码预测标签，同时在模型中添加医学字典特征，可有效识别和分类电子病历中的临床术语；Wang等[18]提出一个建立在BiLSTM-CRF模型基础上的多任务学习方法，通过共享不同医学NER模型的特征提升性能；李丽双等[19]利用大量未标记的生物医学语料与医学词典进行半监督学习，获得了更深层次的语义特征信息，提高了模型性能；王得贤等[20]利用自注意力机制获取法律文书的内部特征表示，有效确定了证据名、证实内容和卷宗号等实体边界。以上方法将部分领域知识特征应用于NER任务中，而民间文学文本一词多义问题更加突出，要求模型具备更强的分类能力，常规模型难以满足需求。

为此，针对民间文学文本中存在的一词多义与实体分类问题，本文提出TBERT-BiLSTM-CRF模型，修改传统BERT模型的嵌入层结构，增加实体类别标签表征，从而使词向量包含实体类别信息，增强了字符对应向量的表达能力，亦加强了模型对于实体类别的划分能力。针对民间文学文本中存在较多领域专业名词的问题，利用未标记的民间文学专有领域语料增量预训练BERT模型，在一般中文BERT模型的基础上添加了民间文学文本语义特征，使得模型输出更符合民间文学文本的语境。该模型的创新之处在于通过添加类型嵌入层使传统BERT模型具备表征实体标签的能力，通过民间文学语料增量预训练进一步优化了TBERT模型的输出，结合BiLSTM-CRF模型根据序列依赖特征与标签约束规则输出全局最优结果，改善了传统NER方法对于民间文学文本NER任务的局限性。

2 TBERT-BiLSTM-CRF模型构建

民间文学文本的NER问题可被视作一项序列标注任务。例如，给定一段民间文学文本序列$S=\{w_1, w_2, \cdots, w_n\}$，其中$w_i$为序列中的第$i$个字符（$i\geq1$），民间文学文本NER任务旨在准确充分地预测出该字符序列对应的标签序列$L=\{l_1, l_2, \cdots, l_n\}$，以最终识别出其中所有实体的位置和类别。本文提出的TBERT-BiLSTM-CRF

模型总体框架如图 1 所示,主要包括 TBERT、序列依赖学习与实体识别 3 个部分:①TBERT。TBERT 模型学习实体类别特征,同时利用民间文学文本语料进行增量预训练进一步优化 TBERT 模型的输出,从而将输入文本转化为含有字符类型信息与文本语义信息的字符表示;②序列依赖学习。BiLSTM 模型学习序列上下文依赖特征并对序列进行建模;③实体识别。CRF 模型对序列进行解码,根据标签依赖规则输出全局最优结果。

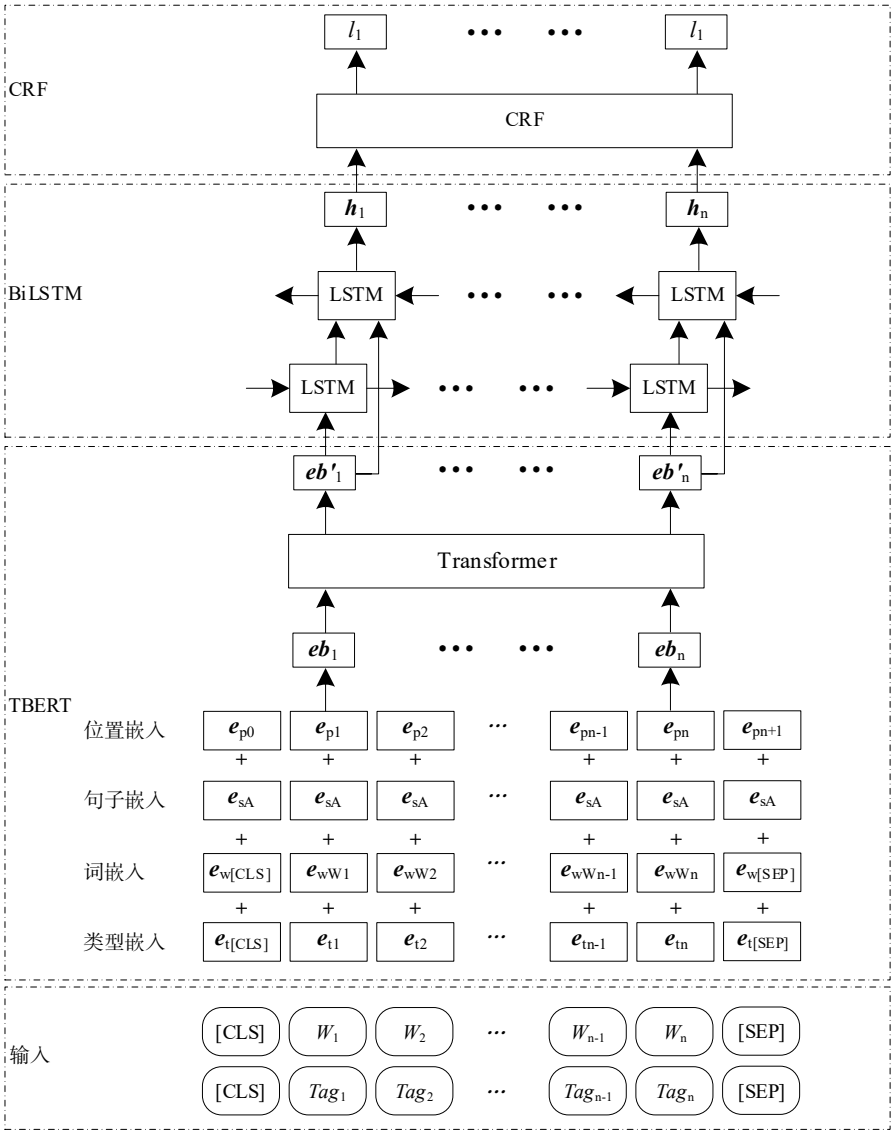

Fig. 1　Main framework of TBERT-BiLSTM-CRF

图 1　TBERT-BiLSTM-CRF 模型总体框架

2.1　语料预训练

预训练模型能够挖掘文本中的深层语义知识并通过语言模型进行表达,针对民间文学领域的预训练模型,若采用通用的 BERT 模型则难以恰当地表达出存在着较多领域专属名词的民间文学的语境。因此本文首先利用未经标记民间文学的文本语料对 BERT 进行预训练,使最终模型中的字符表示包含民间文学领域相关深层特征知识。BERT 模型采用遮蔽策略(Masked Language Modeling, MLM)以[MASK]标记对输入的字符随机遮蔽,并根据其上下文语义预测被遮蔽的词。此外,模型还针对训练语句进行预测下一句任务(Next Sentence Prediction, NSP),若输入的两个句子为前后句关系,则使用[isNext]标记,反之则以[notNext]标记,通过这种方式能够捕获句子级别的上下文关系。

设 Encoder 中的参数为 θ,被遮蔽的单词集合为 M,输出层中 MLM 任务使用的参数分别为 θ_1,词典为 V,则模型采用负对数似然函数计算其损失。表示为:

$$\mathcal{L}_1(\theta,\theta_1)=-\sum_{i=1}^{M}\log P(m=m_i|\theta,\theta_1),m_i\in[1,2,\cdots,|V|] \quad (1)$$

若 NSP 任务的输出层参数为 θ_2,预测标签集合为 N,则模型计算 NSP 任务的损失函数表示为:

$$\mathcal{L}_2(\theta,\theta_2)=-\sum_{i=1}^{N}\log P(n=n_i|\theta,\theta_2),n_i\in[\text{isNext},\text{notNext}] \quad (2)$$

重新训练模型将花费巨大开销，因此本文采用增量训练方式，使用BERT模型的初始权重，在保留通用领域知识的基础上对模型进行民间文学领域知识扩展，从而使其融合民间文学文本的语义特征。

2.2 TBERT模型

民间文学文本中的同一个字符可能表示不同类型的实体。目前通用BERT模型采用词嵌入$\boldsymbol{e}_{\mathrm{w}}$、句子嵌入$\boldsymbol{e}_{\mathrm{s}}$与位置嵌入$\boldsymbol{e}_{\mathrm{p}}$相加的方式生成文本的向量表示，其中词嵌入生成字符本身的向量，反映了其语义；句子嵌入表示当前句子的归属，使模型具备一定的文本分类能力；位置嵌入记录字符的位置信息以保证文本输入的时序性。对于民间文学文本NER任务而言，实体类别信息作为最终的预测目标直接影响NER结果的好坏，能否高效准确地表达字符的类别信息是NER任务的关键所在，而通用BERT模型无法表征实体类别信息。因此，本文提出Type based Bidirectional Encoder Representation from Transformers（TBERT）模型，利用嵌入层生成实体类别标签向量，并与文本字符向量相结合对实体类别特征进行捕获，使模型能够到学习实体类别信息，以更好地完成序列标注任务。

将文本字符与其对应的实体类别标签作为TBERT模型的输入，利用模型原始的3层嵌入对文本字符进行表征以生成字符向量$(\boldsymbol{e}_{\mathrm{w}}, \boldsymbol{e}_{\mathrm{s}}, \boldsymbol{e}_{\mathrm{p}})$。该模型额外增加了一层类型嵌入，由于实体标签间不存在明显的上下文语义联系，采用One-Hot技术对各实体及非实体类型进行统一编码并对齐BERT模型向量，然后将实体类别向量与字符向量相加得到最终向量表示$\boldsymbol{eb}$。公式为：

$$\boldsymbol{e}_{\mathrm{t}} = \mathrm{OneHot}(Tag) \tag{3}$$

$$\boldsymbol{eb} = \boldsymbol{e}_{\mathrm{t}} + \boldsymbol{e}_{\mathrm{w}} + \boldsymbol{e}_{\mathrm{s}} + \boldsymbol{e}_{\mathrm{p}} \tag{4}$$

TBERT模型堆叠使用全连接Transformer编码器（Encoder）结构，具体如图2所示，主要包括多头注意力机制（Multi-head Attention）、前馈神经网络（Feedforward Network）与归一化操作（Add and Norm）。

Attention机制的通用表达式如式（5）所示，其中$\boldsymbol{V}$表示输入，$\boldsymbol{Q}$与$\boldsymbol{K}$表示计算注意力的权重，三者由$\boldsymbol{eb}$经过线性变换得到；d_k表示$\boldsymbol{Q}$与$\boldsymbol{V}$的维度。通过Softmax函数对$\boldsymbol{Q}$与$\boldsymbol{K}$的点积运算结果作归一化处理并乘以$\boldsymbol{V}$获得输出向量。

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}})\boldsymbol{V} \tag{5}$$

多头注意力机制利用多个Attention层计算文本语句权重以获取字符关系信息，将各Attention层的结果整合输出。为了避免Attention机制对于上述操作的拟合程度不够，Encoder结构使用前馈神经网络对结果修饰，并再次归一化处理获得最终输出$\boldsymbol{eb}'$，如公式（6）所示，其中n表示Attention的头数，$\boldsymbol{W}$表示权重矩阵，$\boldsymbol{b}$表示偏置。

$$\boldsymbol{eb}' = \mathrm{Softmax}(\tanh(\boldsymbol{W} \cdot \sum_{i=1}^{n} Attention_i(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + \boldsymbol{b})) \tag{6}$$

在BERT预训练的基础上，利用TBERT再次对标记字符类型数据进行微调更新，使得模型增加字符的类型信息。TBERT结合了文本字符信息与对应实体类别标签信息的词向量，可更轻易地区分一词多义类实体。例如："千瓣莲花"在语句1中表示角色，在语句2中表示物品。原始BERT模型能够在考虑当前语境的情况下将语句1中的"千瓣莲花"以向量$\boldsymbol{v} = (\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{\mathrm{n}})$表示，语句2中的"千瓣莲花"以向量$\boldsymbol{v}' = (\boldsymbol{v}'_0, \boldsymbol{v}'_1, \cdots, \boldsymbol{v}'_{\mathrm{n}})$表示，但由于其未采用实体类别信息，导致两者在数值上近似而令模型难以区分。而TBERT模型能够将实体类别标签转化为向量并叠加至原有字符向量中，扩大了$\boldsymbol{v}$与$\boldsymbol{v}'$的数值差距，从而有效增强了模型对于实体的分类能力。

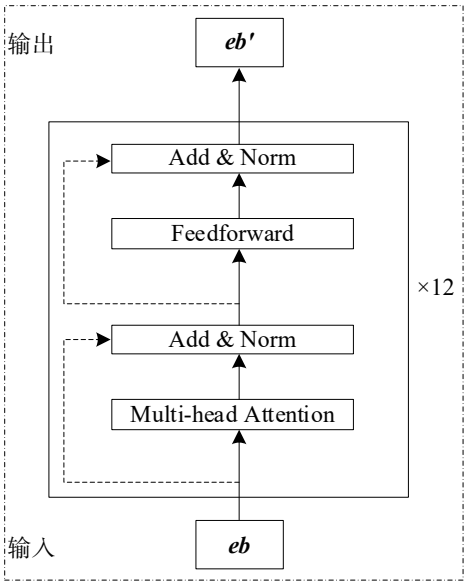

Fig. 2 Framework of Transformer encoder

图2 Transformer编码器结构

2.3 序列依赖学习

民间文学文本NER作为一项序列标注任务旨在输出文本序列对应的标签序列，因此利用BiLSTM模型对TBERT产生的词嵌入进行编码以学习序列上下文依赖特征。BiLSTM模型由前向LSTM层与后向LSTM层组成，解决了循环神经网络的梯度消失问题[21]，从而更适用于民间文学中长文本的编码工作。

如式（7）、式（8）、式（9）、式（10）所示，LSTM网络使用细胞状态$\tilde{c}_t$记录当前最重要的信息，同时利用遗忘门$\boldsymbol{f}_{\mathrm{t}}$与输入门$\boldsymbol{i}_t$控制$\tilde{c}_t$中信息的更新，通过Sigmoid函数$\sigma$将输出值控制在0~1之间，其中0表示完全舍弃，1表示完全保留。

$$\boldsymbol{f}_{\mathrm{t}} = \sigma\left(\boldsymbol{W}_{\mathrm{f}} \cdot \left[\boldsymbol{h}_{\mathrm{t-1}}, \boldsymbol{x}_{\mathrm{t}}\right] + \boldsymbol{b}_{\mathrm{f}}\right) \tag{7}$$

$$\boldsymbol{i}_{\mathrm{t}} = \sigma(\boldsymbol{W}_{\mathrm{i}} \cdot [\boldsymbol{h}_{\mathrm{t-1}}, \boldsymbol{x}_{\mathrm{t}}] + \boldsymbol{b}_{\mathrm{i}}) \tag{8}$$

$$\tilde{c}_{\mathrm{t}} = tanh(\boldsymbol{W}_{\mathrm{c}} \cdot [\boldsymbol{h}_{\mathrm{t-1}}, \boldsymbol{x}_{\mathrm{t}}] + \boldsymbol{b}_{\mathrm{c}}) \tag{9}$$

$$\boldsymbol{c}_{\mathrm{t}} = \boldsymbol{f}_{\mathrm{t}} * \boldsymbol{c}_{\mathrm{t-1}} + \boldsymbol{i}_{\mathrm{t}} * \tilde{c}_{\mathrm{t}} \tag{10}$$

式中：$\boldsymbol{W}$表示权重矩阵，$\boldsymbol{b}$表示偏置量，t表示时刻，$\boldsymbol{h}_{\mathrm{t-1}}$表示$t-1$时刻的隐藏状态，$\boldsymbol{x}_{\mathrm{t}}$表示$t$时刻的输入，$\boldsymbol{o}_{\mathrm{t}}$表示输出门。

根据细胞状态，利用tanh函数确定最终的输出值$\boldsymbol{h}_{\mathrm{t}}$。

计算公式为:

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) \tag{11}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * tanh(\boldsymbol{c}_t) \tag{12}$$

最后将双向LSTM层的结果进行拼接作为CRF层的输入进行解码操作。

2.4　实体识别

BIO(Begin-Inside-Outside)标注规则中,"I-X"标签只可能存在于实体的中间位置,不可能出现在实体的开头或单独出现。若仅使用一个线性层选取BiLSTM输出中概率最高的标签作为最终结果,则很可能产生不合理的序列,如"B-CHA O O I-CHA"。因此,本文利用CRF模型对BiLSTM层的输出进行修正并计算出全局最优序列[22]。

对于给定的输入$\boldsymbol{h} = (\boldsymbol{h}_0, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_n)$,其对应的预测输出标签$L = \{l_0, l_1, \cdots, l_n\}$的得分计算公式为:

$$score(\boldsymbol{h}, L) = \sum_{i=0}^{n} \boldsymbol{W}_{l_i, l_{i+1}} + \sum_{i=1}^{n} PR_{i, l_i} \tag{13}$$

式中:$\boldsymbol{W}_{l_i, l_{i+1}}$表示标签从$l_i$转移至$l_{i+1}$的概率,$PR_{i, l_i}$表示第$i$个词被标记为$l_i$的概率。

$PR(L|\boldsymbol{h})$为$\boldsymbol{h}$的预测结果为L的概率,计算公式为:

$$PR(L\Big|\boldsymbol{h}) = \frac{e^{score(\boldsymbol{h}, L)}}{\sum_{L' \in L_h} e^{score(\boldsymbol{h}, L')}} \tag{14}$$

式中:L'为真实标签,L_h为所有可能存在标签组合。

在最终预测阶段,根据式(15)输出最优结果:

$$L^* = \arg\max_{L' \in L_h} score(\boldsymbol{h}, L') \tag{15}$$

2.5　算法描述

本文提出的TBERT-BiLSTM-CRF模型在BERT模型利用未经标记的民间文学文本语料进行增量预训练的基础上,通过字符类型嵌入并再次优化产生含有字符类型信息与文本语义信息的字符表示,然后由BiLSTM模型进行序列依赖学习,经CRF模型预测输出最优结果。该模型算法具体步骤为:

输入:原始未标记民间文学文本语料,BERT模型,带类型标记数据

$S = \{w_1, w_2, \cdots, w_{n'}\}$:句子

输出:L^*:句子对应标签序列

1. BERT增量预训练

2. TBERT基于BERT,利用标记类型数据进行微调

3. For Each w_i In S Do

4. $\boldsymbol{eb'}_i \leftarrow$ TBERT(w_i) //使用TBERT获取句子中字符的基础向量表示

5. End For

6. For i = 1 To n' Do

7. $\boldsymbol{h}_i \leftarrow$ BiLSTM($\boldsymbol{eb'}_i$) //通过BiLSTM模型获取句子中字符的深层特征表示

8. End For

9. For Each L Do

10. $score(\boldsymbol{h}, L) \leftarrow \sum_{i=0}^{n} \boldsymbol{W}_{l_i, l_{i+1}} + \sum_{i=1}^{n} PR_{i, l_i}$ //计算路径得分

11. End For

12. $L^* \leftarrow \arg\max_{L' \in L_h} score(\boldsymbol{h}, L')$ //最大得分路径L^*作为输出预测标签序列

13. RETURN L^*

3　实验方法与结果分析

3.1　数据集

本文使用的民间文学文本语料包括《千瓣莲花》、《傣族文本》、《娥并与桑洛》与《云南少数民族古典史诗全集》,字数信息如表1所示。

Table 1　Word count for the corpus of folk literature texts

表1　民间文学文本语料字数信息

文本名称	字数(万)
千瓣莲花	3.4
傣族文本	1.0
娥并与桑洛	2.0
云南少数民族古典史诗全集	181.9

由于民间文学文本语料规模庞大且需要人工标注,且以上4则文本在内容与形式上具有相似性,挑选其中1 824句能够反映民间文学文本一词多义等特点的语句,利用BIO标注的方式产生数据集,共计5 921个标签。人名(PER)、地点(LOC)与组织(ORG)是3种广泛应用于NER任务的标签,其同样适用于民间文学的NER工作。考虑到民间文学中不仅会出现人名,还会有许多拟人化的动植物角色,因此将PER替换为角色(CHA)。此外,民间文学中描述了一些对于剧情发展具有推动作用的"宝物",本文对该类实体也进行了标注,并用物品标签(OBJ)表示。表2为序列标签集。数据集以8:2的比例随机划分为训练集与测试集,其中语句和各类实体的分布情况如表3所示。

Table 2　The sequence labels

表2　序列标签集

标签名	表示含义
B-CHA	角色类实体的开头
I-CHA	角色类实体的中间或结尾
B-LOC	地点类实体的开头
I-LOC	地点类实体的中间或结尾
B-ORG	组织类实体的开头
I-ORG	组织类实体的中间或结尾
B-OBJ	物品类实体的开头
I-OBJ	物品类实体的中间或结尾
O	非实体

Table 3　Statistics of the datasets

表3　数据集统计信息

数据集	语句	CHA	LOC	ORG	OBJ
训练集	1 444	2 289	791	203	1 370
测试集	380	671	200	66	331

3.2　评价指标

NER任务旨在识别出文本中的预定义语义类别,能否准确全面地进行识别在NER模型性能评价中占据重要地位,因此本文使用准确率P、召回率R与$F1$值评价实验结

果。计算公式分别为：

$$P = \frac{N_p}{N_F} \tag{16}$$

$$R = \frac{N_p}{N_A} \tag{17}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{18}$$

式中：N_p表示模型识别的正确实体数量，N_A表示测试集中的实体数量，N_F表示模型识别的实体数量。P、R和$F1$取值范围均为0~1，其值越大越好。

3.3 实验平台与参数设置

实验平台为Intel（R）Xeon（R）CPU E5-2650 v3 @ 2.30GHz处理器，RTX 2080Ti GPU，256 GB内存，Ubuntu 20.04.1操作系统，Python 3.6语言，Tensorflow-gpu 1.11.0框架。

预训练在chinese_L-12_H-768_A-12模型的基础上进行，民间文学文本数据经处理后生成tf.record文件。同时设置最大句子长度为128，batch_size为32，学习率为2e-5进行训练。模型参数设置见表4。

Table 4 Model parameter settings

表4 模型参数设置

参数	数值
batch_size	128.000
dropout	0.500
learning_rate	0.001
max_seq_length	128.000
max_epoch	100.000

3.4 实验结果与分析

3.4.1 不同模型比较

在民间文学数据集上对本文模型（TBERT-BiLSTM-CRF）与目前广泛应用于NER任务的BERT-BiLSTM-CRF[17]、BiLSTM-CRF[18]、BiLSTM[21]、CRF[22]模型的表现进行比较，结果见表5。可以看出，将CRF与BiLSTM结合后，3项评价指标相较单独结构均有明显提高；在此基础上添加一般中文语料BERT预训练模型后，3项指标比BiLSTM-CRF模型分别提高了1.15%、2.24%、1.75%；将一般中文语料BERT预训练模型更换为本文方法生成的TBERT模型后，相比BERT-BiLSTM-CRF 3项指标分别提高了3.61%、2.14%、2.89%。说明同时利用民间文学的语义特征与实体类别特征可使模型理解民间文学的领域知识并加强对实体的划分，从而在识别出更多实体的同时确保分类的准确率。

Table 5 Experimental result comparison of each model

表5 各模型实验结果比较

模型	P(%)	R(%)	$F1$(%)
CRF	84.20	84.35	84.27
BiLSTM	84.08	83.85	83.96
BiLSTM-CRF	84.45	86.33	85.38
BERT-BiLSTM-CRF	85.42	88.38	86.87
TBERT-BiLSTM-CRF	**88.51**	**90.27**	**89.38**

比较TBERT-BiLSTM-CRF、BiLSTM-CRF、BERT-BiLSTM-CRF 3种模型对民间文学数据集中4种不同类型实体的准确率、召回率与F1值，结果见表6、表7、表8。

由表6可知，TBERT-BiLSTM-CRF模型对各类实体识别的准确率均优于其他模型约2%，说明利用标签信息能使模型更好地区分一词多义类实体，使识别更加准确。

Table 6 Precision comparison of each model for different entity categories

表6 各模型对不同类型实体识别准确率比较 (%)

实体类别	BiLSTM-CRF	BERT-BiLSTM-CRF	TBERT-BiLSTM-CRF
CHA	92.03	92.83	**94.52**
LOC	76.92	75.74	**77.13**
OBJ	77.41	77.62	**85.64**
ORG	77.85	80.23	**82.89**

由表7可知，TBERT-BiLSTM-CRF模型对3类实体的召回率超出其他模型0.7%~3%，表现优异。

Table 7 Recall comparison of each model for different entity categories

表7 各模型对不同类型实体召回率比较 (%)

实体类别	BiLSTM-CRF	BERT-BiLSTM-CRF	TBERT-BiLSTM-CRF
CHA	93.45	94.36	**95.12**
LOC	78.21	78.99	**81.42**
OBJ	78.29	85.28	**88.36**
ORG	**85.26**	79.49	80.93

由表8可知，TBERT-BiLSTM-CRF模型对各类实体识别的F1值均超过其他模型1%~5%。由于F1值的计算综合考虑了模型识别准确率与召回率，说明TBERT-BiLSTM-CRF模型较目前广泛使用的NER模型能够更准确完整地识别出民间文学文本中存在的实体。

Table 8 F1 value comparison of each model for different entity categories

表8 各模型对不同类型实体F1值比较 (%)

实体类别	BiLSTM-CRF	BERT-BiLSTM-CRF	TBERT-BiLSTM-CRF
CHA	92.73	93.59	**94.82**
LOC	77.56	77.33	**79.22**
OBJ	77.85	81.27	**86.98**
ORG	81.39	79.86	**81.90**

3.4.2 案例分析

以下列举了BiLSTM-CRF、BERT-BiLSTM-CRF模型与TBERT-BiLSTM-CRF模型对于民间文学文本具体句子案例的识别结果，其中加粗部分表示模型识别的实体，括号内记录其对应的类型。

（1）BiLSTM-CRF模型。**国王（CHA）**的第七个姑娘她说的每一句话会变成一朵**千瓣莲花（OBJ）**漂在天上，喷发出馥郁的清香，世上千万个美丽的姑娘我一个也不看在心上，我单单爱上**千瓣莲花（OBJ）**姑娘，我真有福气能来到与**莫板森林（LOC）**遇见美丽的**莲花（OBJ）**姑娘。

(2)BERT-BiLSTM-CRF模型。**国王(CHA)**的第七个姑娘她说的每一句话会变成一朵**千瓣莲花(OBJ)**漂在天上,喷发出馥郁的清香,世上千万个美丽的姑娘我一个也不看在心上,我单单爱上**千瓣莲花(OBJ)**姑娘,我真有福气能来到与**莫板森林(LOC)**遇见美丽的**莲花(CHA)**姑娘。

(3)TBERT-BiLSTM-CRF模型。**国王(CHA)**的第七个姑娘她说的每一句话会变成一朵**千瓣莲花(OBJ)**漂在天上,喷发出馥郁的清香,世上千万个美丽的姑娘我一个也不看在心上,我单单爱上**千瓣莲花(CHA)**姑娘,我真有福气能来到**与莫板森林(LOC)**遇见美丽的**莲花(CHA)**姑娘。

可以看出,BiLSTM-CRF模型利用Word2Vec技术生成的词向量较为单一,导致涉及"莲花"的实体皆判断为物品,且其因缺少民间文学领域知识而未将实体"莫板森林"完整地识别出来。这个问题同样出现在BERT-BiLSTM-CRF模型的识别结果中,该模型虽然能够根据上下文将"莲花姑娘"中的"莲花"正确判断为角色,但对于文字表述上完全相同的两个"千瓣莲花"并没有进行区分而均判断为物品。TBERT-BiLSTM-CRF模型因融合了民间文学的语义特征与实体类别特征,在实体识别的准确度方面表现良好。

嵌套类实体会对TBERT-BiLSTM-CRF模型造成干扰,如表9中的案例1与案例2所示,其分别为地名嵌套角色名实体与组织名嵌套角色名实体。对于前后文关系紧密的民间文学文本,若前后文割裂输入进模型,则模型难以根据前后文判断出实体的正确类别,如表9中的案例3所示。

Table 9　Error analysis

表9　错误分析

结果	案例1 (回到召捧勒的家中)	案例2 (他来自英趴一族)	案例3 (共麻拉来到千瓣莲花姑娘身边,触摸了千瓣莲花)
正确结果	召捧勒的家(LOC)	英趴一族(ORG)	千瓣莲花(CHA)
预测结果	召捧勒(CHA)	英趴(CHA)	千瓣莲花(OBJ)

4　结语

本文针对民间文学文本领域名词众多和一词多义的特点提出TBERT-BiLSTM-CRF模型,将民间文学的语义特征与实体类别特征融入一般中文BERT模型,使其具备识别具有领域特色实体与多重词义实体的能力;同时结合BiLSTM模型与CRF模型,根据上下文信息与序列间存在的强依赖关系使模型获得全局最优结果。与经典模型CRF、BiLSTM、BiLSTM-CRF与BERT-BiLSTM-CRF相比,本文模型在民间文学文本数据集上获得了最高的准确率、召回率与F1值。然而,本文仍存在一定不足:一方面是并未构建较为完备的民间文学文本数据集,另一方面是模型效果未在其他领域数据集中得到验证。未来将进一步探索完备数据集上的领域知识NER工作。

参考文献:

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

[2] PENG J Y, FANG Y, HUANG C, et al. Cyber security named entity recognition based on deep active learning[J]. Journal of Sichuan University (Natural Science Edition), 2019, 56(3): 457-462.
彭嘉毅, 方勇, 黄诚, 等. 基于深度主动学习的信息安全领域命名实体识别研究[J]. 四川大学学报(自然科学版), 2019, 56(3): 457-462.

[3] GUAN C Y, CHENG Y H, ZHAO H. Semantic role labeling with associated memory network[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 3361-3371.

[4] BAHDANAU D, CHO K H, YOSHUA B. Neural machine translation by jointly learning to align and translate[C]// Proceedings of International Conference on Learning Representation, 2015: 1-4.

[5] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the Advances in Neural Information Processing Systems, 2013: 3111-3119.

[6] JACOB D, CHANG M W, KENTON L, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

[7] ZHAO S, LUO R, CAI Z P. Survey of Chinese named entity recognition [J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(2): 296-304.
赵山, 罗睿, 蔡志平. 中文命名实体识别综述[J]. 计算机科学与探索, 2022, 16(2): 296-304.

[8] ZHENG Y Y, TIAN Y, SHI C. Method of entity set expansion based on frequent pattern under meta path[J]. Journal of Software, 2018, 29(10): 2915-2930.
郑玉艳, 田莹, 石川. 一种元路径下基于频繁模式的实体集扩展方法[J]. 软件学报, 2018, 29(10): 2915-2930.

[9] JU M, MIWA M, ANANIADOU S. A neural layered model for nested named entity recognition[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 1446-1459.

[10] JU S G, LI T N, SUN J P. Chinese fine-grained name entity recognition based on associated memory networks[J]. Journal of Software, 2021, 32(8): 2545-2556.
琚生根, 李天宁, 孙界平. 基于关联记忆网络的中文细粒度命名实体识别[J]. 软件学报, 2021, 32(8): 2545-2556.

[11] XU C W, WANG F Y, HAN J L, et al. Exploiting multiple embeddings for Chinese named entity recognition[C]// Proceedings of the 28th Association for Computing Machinery International Conference on Information and Knowledge Management, 2019: 2269-2272.

[12] WU H, LYU L, YU B H. Chinese named entity recognition based on transfer learning and BiLSTM-CRF[J]. Journal of Chinese Computer Systems, 2019, 40(6): 1142-1147.
武惠, 吕立, 于碧辉. 基于迁移学习和BiLSTM-CRF的中文命名实体识别[J]. 小型微型计算机系统, 2019, 40(6): 1142-1147.

[13] WANG R, YU T, ZHAO H D, et al. Few-shot class-incremental learning for named entity recognition[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 571-582.

[14] NIE Y Y, TIAN Y H, WAN X, et al. Named entity recognition for social media texts with semantic augmentation[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 1383-1391.

[15] YU J K. Chinese electronic medical record naming entity recognition based on multi-task learning [J]. Software Guide, 2021, 20(11): 42-46.
余俊康. 多任务学习的中文电子病历实体识别研究[J]. 软件导刊, 2021, 20(11): 42-46.

[16] YANG J F, GUAN Y, HE B ,et al. Corpus construction for named entities and entity relations on Chinese electronic medical records[J]. Journal of Software, 2016, 27(11): 2725-2746.
杨锦锋, 关毅, 何彬, 等. 中文电子病历命名实体和实体关系语料库构建[J]. 软件学报, 2016, 27(11): 2725-2746.

[17] LI X Y, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics, 2020, 107: 103422.

[18] WANG X, ZHANG Y, REN X, et al. Cross-type biomedical named entity recognition with deep multi-task learning[J]. Bioinformatics, 2019, 35(10): 1745-1752.

[19] LI L S, HE H L, LIU S S, et al. Research of word representations on biomedical named entity recognition[J]. Journal of Chinese Computer Systems, 2016, 37(2): 302-307.
李丽双, 何红磊, 刘珊珊, 等. 基于词表示方法的生物医学命名实体识别[J]. 小型微型计算机系统, 2016, 37(2): 302-307.

[20] WANG D X, WANG S G, PEI W S, et al. Named entity recognition based on JCWA-DLSTM for legal instruments[J]. Journal of Chinese Information Processing, 2020, 34(10): 51-58.
王得贤, 王素格, 裴文生, 等. 基于JCWA-DLSTM的法律文书命名实体识别方法[J]. 中文信息学报, 2020, 34(10): 51-58.

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

[22] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the 18th International Conference on Machine Learning, 2001: 282-289.

(责任编辑:尹晨茹)