SL-BiLSTM-CRF 藏文命名实体识别方法[1]

洛桑嘎登[1]，群诺[1]，索南尖措[1]，仁增多杰[1]*

（1. 西藏大学信息科学技术学院，西藏自治区拉萨市，850000）

摘要：藏文命名实体识别是藏语自然语言处理的基础任务，命名实体识别的结果将影响下游的任务。随着藏文信息处理研究的深入，藏文信息处理研究已从计算机工具类研究慢慢发展为知识挖掘的深入研究。藏文文本作为知识的重要载体，分析藏文文本信息，挖掘文本中的知识，对完成机器翻译，网络舆情检测，知识图谱构建等任务具有重要意义。藏文命名实体识别面临两个难题，一是和汉语一样表示实体词的边界难以区分；二是藏文存在黏着词的特性，黏着词特性加大了词语边界区分的难度。为此，在本文中，我们提出了一种融合藏文音节部件信息的藏文命名实体识别神经网络模型 SL-BiLSTM-CRF。该方法有效利用了藏文音节的部件信息和藏文音节表征信息，使得模型充分考虑藏文命名实体识别任务中词语边界问题。实验证明，我们的方法在藏文命名实体识别任务中具有更好的表现。

关键词：藏文；命名实体识别；深度学习

中图分类号：TP391　　**文献标志码：** A

1. 前言

命名实体识别是自然语言处理的基础任务[1]。命名实体识别被应用在很多下游的自然语言处理任务中，比如，机器翻译，构建知识图谱，智能问答系统，信息抽取，信息检索等。通常，命名实体识别是指识别文本中人名、地名、组织机构名等 [2]。传统的命名实体识别方法主要是基于规则和词典匹配的实体识别方法以及基于统计机器学习的命名实体识别方法。随着神经网络算法的不断发展，深度学习技术在自然语言处理领域的很多任务中取得了很好的成绩，包括命名实体识别任务。基于规则和词典匹配的命名实体识别方法实现简单，但是存在两个问题，一是需要依赖语言学家的知识总结规则。二是，需要有一个庞大的词典作为支撑。显然，这类方法有着一定的局限性，因为很难构建出一个符合所有范式的规则和一个包含所有实体的词典库。基于统计机器学习的方法，早期这类方法是命名实体识别研究的热点，主要方法包括，隐马尔科夫、最大熵、条件随机场[3-5]等。这类方法是在基于大规模的标注数据上通过人工构建特征实现。同样，这类方法也存在一定的局限性，它需要依赖复杂的特征工程，需要依赖专家的知识，人工构建特征。这将耗费大量的人力物力。

随着计算机算法、算力、数据的不断发展，深度学习方法在自然语言处理领域大放异彩。命名实体识别作为自然语言处理的基础任务，深度学习方法在命名实体识别任务[6-9]中取得了很好的成绩。

在面向英语的实体识别任务中，基于双向长短记忆网络结合条件随机场模型（BiLSTM-CRF）融入字符信息取得了英文命名实体识别任务最好的结果[10]。面向汉语的命名实体识别任务中，主要解决汉语词边界难以区分的问题，和英语一样采用的基准模型是 BiLSTM-CRF，只在基准模型的输入端，结合词典信息或改变网络的结构以适应汉语的命名实体识别

[1]**基金项目：**国家重点研发计划重点专项（项目编号：2017YFB1402200）;西藏自治区自然基金项目（XZ2017ZRG-08）；2021 年武汉理工大学-西藏大学“西藏经济社会发展与高原科学研究共建创新基金”专项项目（项目批准号：lzt2021008）；中央引导地方科技发展资金(ZX202102YD0018C);西藏大学成长计划（08000013）

* **通信作者：gaden168@163.com**

任务[11-13]。最近被提出的 FLAT[14]方法，在汉语命名实体识别中取得了最好的效果。

面向藏文的命名实体识别方法研究相对滞后，大多数研究方法还停留在基于统计机器学习的方法。导致藏语命名实体识别方法滞后的原因有几个，首先是目前还没有一个公开可用的藏文分词技术。其次，目前没有一个公开的训练好的藏文词向量，同时也没有形成公开可用的基于藏文的预训练模型比如 Bert 模型等，这给解决基于深度学习的藏文自然语言处理带来了难度[15-16]。藏文与英语相比，藏文词语之间没有明显的分隔符[17]，词语边界不清晰，这给命名实体识别任务带来了挑战。与汉语相比，藏文中又存在黏着词的特性[18]，这是藏文命名实体识别任务的又一难题。针对以上问题，本文提出了一种基于藏文音节特征表征形式的神经网络框架 SL-BiLSTM-CRF（Syllable Level BiLSTM-CRF），可以充分提取藏文音节特征和构成藏文音节部件的特征，可以提高藏文命名实体识别任务中词边界的识别和黏着词的处理效果。实验结果表明，在我们所收集的数据上，相比于其他方法，该方法取得了更好的效果。

2 方法实现

2.1 问题定义

藏文是一种拼音文字，但是不像英文，藏文词之间没有明显的分隔符。藏文命名实体识别需要解决两个难题。一是，如何确定表示实体词语的边界。二是，针对带有黏着词的藏文音节，如何解决黏着词的问题。

命名实体识别可以看成是一个多任务学习。首先在句子中识别出实体，其次根据预先定义的类别对识别的实体进行的分类。在具体实现过程中可以根据 CRF 层的标记方式的不同，也可以实现在识别实体边界的同时对实体进行分类。本文采用的标记方式是识别实体边界的同时对实体进行分类。

通常，命名实体识别当成是一种序列标注的过程。命名实体序列的标注有基于词的标注和基于音节（字）的标注，例如：

对于藏文句子：

ཨ་དབང་ལེ་ཧོང་ནས་ཡོང་། （译：次旺来自北京）

（1）基于词的标注：

ཨ་དབང་/PER ལེ་ཧོང་/LOC ནས་/O ཡོང་།/O

（2）基于音节的标注：

ཨ་/B-PERདབང་/I-PERལེ་/B-LOCཧོང་/I-LOCནས་/Oཡོང་།/O

基于词的标注需要先进行分词，这就存在 OOV 和分词错误传播的问题[12]。本文采用第二种方法，基于音节（字）的标注。

我们的模型 SL-BiLSTM-CRF 如图 1 所示：

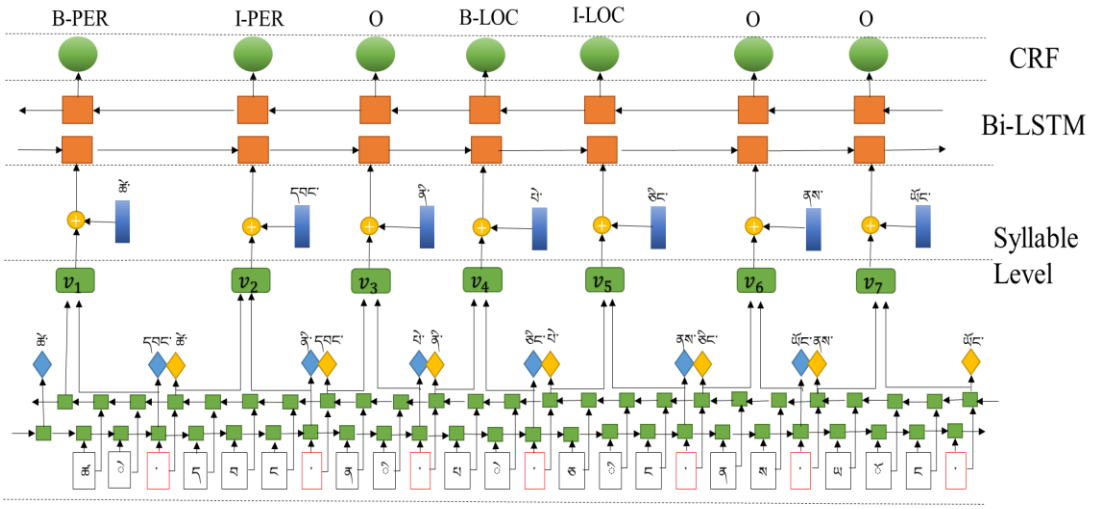

图 1 SL-BiLSTM-CRF 网络架构

Fig. 1 SL-BiLSTM-CRF Neural Architecture

模型分为三个模块。第一模块是基于藏文音节的特征表征形式（Syllable Level），它融合了藏文音节的特征和组成音节部件的特征；第二模块是一个双向 LSTM 层；第三个模块是 CRF 层。

2.2 藏文音节特征表征层

在英文和汉语的命名实体识别中，采用基于字符的编码作为模型的输入，所谓字符对应到藏文中我们称其为音节。藏文的每个音节由前加字、基字、上加字、下加字、后加字、再后加字和元音七部分按照一定的规则组成。藏文同汉语一样词和词之间没有明显的分隔标记，在汉语命名实体识别中已经证明字向量作为神经网络的输入比词向量作为输入的效果更好[11]。因此，本文采用基于字（音节）向量的编码方式，但是藏文与汉语不同的是藏文是一种拼音文字，藏文中存在黏着词的特性，有些实体词以黏着词的形式出现在文本中，例如

བཀྲ་ཤིས་ཟླ་བའི་ /PERག་ཡུལ་/Oའི་/Oལྷ་སར་/LOCཡིན། （译：扎西达娃的故乡在拉萨）

表 1 实体包含黏着词例子

Tab.1 Examples of Entity Contains Adhesive Words

原文	分解之后
བཀྲ་ཤིས་ཟླ་བའི་	བཀྲ་ཤིས་ཟླ་བ + འི་
ལྷ་སར་	ལྷ་ས + ར་

上句中，人名 བཀྲ་ཤིས་ཟླ་བ་(扎西达娃)，地名 ལྷ་ས་(拉萨)这两个词都以黏着词结尾。‘བཀྲ་ཤིས་ཟླ་བའི་’是‘བཀྲ་ཤིས་ཟླ་བ་’和黏着词‘འི་’的缩写，而’ལྷ་སར་‘是’ལྷ་ས་‘和黏着词’ར་‘的缩写。可见，藏文中黏着词在文本中出现频率高而且给词边界的确定带来了困难。为了充分考虑藏文黏着词的特性，本文设计的方法是在模型的输入端，完成字（音节）向量编码之前，先进行构成藏文音节部件的编码，即把音节拆分成每个部件，通过两个 LSTM 模型对音节部件进行编码，如图 1 所示，我们把基于音节部件的向量表征和基于整个完整音节（字）的表征向量进行拼接，我们称之为 Syllable Level 层。之后，将这部分作为 BiLSTM-CRF 层的输入，训练网络，得到最终的结果。

2.3 BiLSTM 层

BiLSTM 模型即，双向长短记忆网络。它是前向 LSTM 模型和后向 LSTM 模型的结合。

BiLSTM 模型克服了单向 LSTM 模型只能考虑之前出现的信息，而无法考虑上下文信息的弊端。每个 LSTM 模块都有遗忘门、输入门、输出门三个'门'控机制实现，具体的计算过程如下：

$$f_t = \sigma\left(w_f \cdot [h_{t-1}, x_t] + b_f\right)$$

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

$$\widetilde{C_t} = \tanh(w_c \cdot [h_{t-1}, x_t] + b_c)$$

$$C_t = f_t \cdot c_{t-1} + i_t \cdot \widetilde{C_t}$$

$$O_t = \sigma(w_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = O_t \cdot \tanh(c_t) \qquad (1)$$

上述公式中w_f 、w_i 、w_o 的分别是输入门、遗忘门和输出门的权重矩阵，b_f、b_i、b_o 代表偏置。当前时刻的单元状态 C_t是由上一次的单元状态 c_{t-1} 按元素乘以遗忘门f_t，再用当前输入的单元状态 C_t按元素乘以输入门 i_t，再将两个积加和。最后，在计算输出门的基础上，得到当前时刻的隐藏层输出h_t。

2.4 CRF 层

BiLSTM 模型层，尽管能算出当前词的标签概率，但是没有考虑相邻标记之间的转移概率。为了让模型充分考虑相邻标签之间的关系，我们采用 CRF 模型。如果记一个长度等于句子长度的标签序列$y = y_1, y_2, y_3, \ldots, y_n$ 那么模型对于句子 x 的标签等于 y 的打分为：

$$\text{score}(x, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} \qquad (2)$$

可以看出整个序列的打分等于各个位置的打分之和，而每个位置的打分由两部分得到，一部分是由 LSTM 输出的 p_i决定，另一部分则由 CRF 的转移矩阵 A 决定。进而利用 Softmax 得到归一化后的概率。

$$P(y|x) = \frac{\exp(\text{score}(x,y))}{\sum_{y'} \exp(\text{score}(x,y'))} \qquad (3)$$

最终得到输出标签的概率。

3. 实验对比

3.1 数据和评价指标

数据来源。本文研究所用数据均通过网络爬虫技术，经过爬虫技术抓取数据，清洗数据，人工矫正等步骤，总共收集了近 50M 左右的文本语料，包含藏文字符数约 400W。语料均来自公开的藏文网站、博客等。实验语料中的标签设置情况如表 1 所示。

表 2 实验语料中标签设置情况

Tab.2 Label settings in the experimental corpus

实体类型	标签
人名（PER）	B-PER，I-PER
地名（LOC）	B-LOC，I-LOC
组织机构名（ORG）	B-ORG，I-ORG
非实体	O

为了调节模型选择中超参数和获得更好的近似估计模型的泛化能力，我们把语料按照 6:2:2 的比例进行了训练集、验证集和测试集的划分。

表 4 训练和验证数据划分情况

Tab.4 Training and validation data

	字符个数（万）	语料大小（M）
训练集	288	30
验证集	96	10
测试集	96	10

每类实体在实验语料中的分布个数统计如下表 2 所示。

表 3 实验语料中各类实体信息统计

Tab.3 Statistics of entity information in experimental corpus

实体类型	实体个数	训练集	验证集	测试集
PER	40109	28981	7234	7428
LOC	65987	49573	9842	11414
ORG	66135	51517	9623	12618

模型评价指标。模型的评价指标主要包含三个：精确率（P）、召回率（R）和 F1 值。假设，我们用 TP（true positive）表示真正例，即把正例正确预测为正例；FP（false positive）表示假正例，即把负例错误预测为正例；FN（false negative）表示假负例，即把正例错误预测为负例；TN（true negative）表示真负例，即把负例正确预测为负例。那么，

$$\text{精确率(Precision)} = \frac{TP}{TP + FP}$$

$$\text{召回率(Recall)} = \frac{TP}{TP + FN}$$

$$\text{F1} = \frac{2 * P * T}{P + T}$$

3.2 模型训练

模型使用学习率为 0.001 的 Adam 优化算法，训练集次数 epoch 设置为 50 次。音节部件的表征编码维度设置为 8，完整音节的编码维度为 100。我们在 BiLSTM 的输入和输出部分使用了采样概率为 0.5 的 dropout 层，以避免训练时过拟合。操作系统环境为 Windows 10。深度学习框架使用 Python 开源库 Pytorch1.8.1 版本。

3.3 实验结果与分析

为了验证音节部件编码信息对提升模型效果的有效性，本文在相同的实验语料上，分别使用隐马尔可夫模型（HMM），条件随机场模型（CRFs），双向 LSTM 模型，BiLSTM-CRF 模型和本文提出的 SL-BiLSTM-CRF 模型我们的模型进行了对比实验。实验中的前四个模型是将音节(字)向量作为深度神经网络的输入的，而最后一个即本文提出的模型是在音节(字)向量上联合了音节部件信息的表征向量。

表 5 给出了各类模型在当前实验数据上的表现。

表 5 各类模型的效果

Tab.5 Results of various models

Model	P (%)	R(%)	F1(%)
HMM	76.46	70.10	73.14
CRFs	85.65	75.43	80.22
BiLSTM	79.24	80.63	79.93
BiLSTM-CRF	87.01	82.02	84.44
SL-BiLSTM-CRF	**89.65**	**82.68**	**86.02**

从结果可以看出，HMM 和 CRFs 两个传统的统计机器学习方法相比，CRFs 模型的精确率、召回率、F1 值分别比 HMM 模型高 9.19%，5.33%，7.08%。那是因为 HMM 模型状态的转移仅考虑之前的那一个状态，而 CRFs 模型可以考虑上下文的状态转移概率。BiLSTM 模型和 BiLSTM-CRF 模型相比，虽然都使用了双向 LSTM 网络架构，但是最后接了 CRF 层的模型比仅使用 BiLSTM 模型相比，精确率、召回率、F1 值分别高出 7.77%, 1.39%, 4.51%。那是因为 CRF 层的加入在模型计算出当前标签结果的概率基础上引入了相邻状态的转移概率，可以帮助输出结果更加合理。相比于以上四类模型，本文提出 SL-BiLSTM-CRF 模型，在 BiLSTM-CRF 基础模型架构不变的前提下，改变输入端的向量表征形式，在原有的音节（字）向量上融合了音节部件的向量表征，输入端的表征向量学到更小感受野的特征，所以不管是精确率、召回率、F1 值均比上面四个模型中表现最好的 BiLSTM-CRF 模型还要高出 2.64%, 0.66%, 1.58%。这个是因为模型的输入融合音节和音节部件的信息，感受野变小了，学到了更具体的表征，而藏文中的黏着词恰好在文本中作为音节部件的一部分出现，所以模型的结果有所提高。

表 6 分析了不同模型在不同实体上的精确率、召回率和 F1 值。

表 6 不同模型在三种实体识别上的 F1 值对比

Tab.1 F1 value of different models on three kinds of entity recognition

Model	PER			LOC			ORG		
	P	R	F	P	R	F	P	R	F
HMM	83.43	80.87	82.13	70.23	68.65	69.43	68.80	67.11	67.87
CRFs	90.24	85.45	87.78	82.34	71.33	76.44	80.03	82.32	76.43
BiLSTM	86.68	86.20	86.44	74.46	81.34	77.75	76.66	70.61	75.60
BiLSTM-CRF	92.65	85.35	88.85	83.67	83.01	83.34	82.22	79.33	81.13
SL-BiLSTM-CRF	94.24	86.16	**90.02**	85.22	84.86	**85.04**	84.02	81.24	**83.01**

从上表我们可以看出，在三个实体中，人名（PER）的精确率、召回率、F1 相对较高，这个可能和大多数新闻语料中提及的人名为政治人物或公众人物，这类人物相对来说具有一定的规范性，而且写法相对同一，所以效果更好些。地名（LOC）和组织机构名（ORG）的精确率、召回率和 F1 值偏低。分析结果发现，地名和组织机构名之间存在多种嵌套，如图 2 所示，例子中“青海”一词即是地名也是组织机构名“青海师范大学”的一部分，这类错误导

致了地名和组织机构名分类不准确的问题。

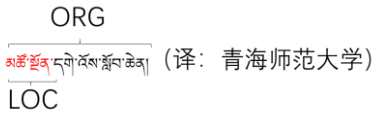

图 2 实体嵌套的例子

Fig. 1 Examples of entity nested

4. 总结

藏文命名实体识别是机器翻译，研究知识图谱的基础，本文提出一种融合藏文音节（字）向量和构成音节部件的向量作为 BiLSTM-CRF 深度神经网络架构的输入。实验结果发现，相比于之前在藏文命名实体识别任务上使用过的 HMM、CRFs、BiLSTM、BiLSTM-CRF 相比，在加入了藏文音节部件信息的表征向量之后，模型的精确率、召回率和 F1 值均有提高。

相比于词向量作为网络的输入，基于音节（字）向量的输入可以避免在训练词向量时依赖分词结果的准确性，避免了分词任务的错误传播到命名实体识别任务中。但是，单独基于音节（字）向量的输入难以学习一个音节内部的构成元素。因此，本文提出了一种融合音节部件信息特征和音节（字）向量信息的表征作为网络的输入。实验结果发现，该方法在藏文命名实体识别任务中取得了一定的效果。这给藏文其他基础任务的处理提供了新的思路。当然，本文实验部分仅考虑了单个命名实体的标注结果，没有考虑在日常文档中存在的实体名称嵌套的问题，实体名称嵌套是一种很常见的文本现象，下一步本文需要进一步去研究探讨，如何解决实体名称嵌套的问题。此外，本文提出的方法仅在命名实体识别任务上做了实验研究，下一步在其他任务上是否也有类似的表现值得研究。

5. 参考文献

[1] 宗成庆.统计自然语言处理[M].清华大学出版社,2008:150-178
[2] 江荻,康才晙.书面藏语排序的数学模型及算法[J].计算机学报,2004(04):524-529.
[3] 李亚超. 基于条件随机场的藏文分词与命名实体识别研究[D].西北民族大学,2013.
[4] 头旦才让,仁青东主,尼玛扎西.基于 CRF 的藏文地名识别技术研究[J].计算机工程与应用,2019,55(18):111-115.
[5] 加羊吉,李亚超,于洪志.CRF 与规则相结合的藏文人名识别方法[J].西北民族大学学报(自然科学版),2016,37(03):41-45.
[6] 贡保才让. 深层神经网络的藏文命名实体识别研究[D].青海师范大学,2018.
[7] 王志娟,刘飞飞,赵小兵,宋伟.基于置信度的藏文人名识别的主动学习模型研究[J].中文信息学报,2019,33(08):53-59.
[8] 珠杰,李天瑞.深度学习模型的藏文人名识别方法[J].高原科学研究,2017,1(01):112-124.
[9] 华却才让,姜文斌,赵海兴,刘群.基于感知机模型藏文命名实体识别[J].计算机工程与应用,2014,50(15):172-176.
[10] Liu L,Shang J,Xu F,et al. Empower Sequence Labeling with Task-Aware Neural Language Model. 2017.
[11] J He, H Wang. Chinese Named Entity Recognition and Word Segmentation Based on Character.
[12] Yue Z , Jie Y. Chinese NER Using Lattice LSTM[C]// The 56th Annual Meeting of the Association for Computational Linguistics (ACL). 2018.
[13] Yuan M,Li Y. Bi-Lattice LSTM Model with Self-Attention for Chinese NER[C]// 2020 IEEE 20th International Conference on Communication Technology (ICCT). IEEE, 2020.
[14] 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. 2020.
[15] 才智杰,才让卓玛,孙茂松.一种多基元联合训练的藏文词向量表示方法[J].中文信息学报,2020,34(05):44-49.
[16] 李亮. 基于 ALBERT 的藏文预训练模型及其应用[D].兰州大学,2020.
[17] 色差甲,慈祯嘉措,才让加,华果才让.基于神经网络的藏文正字检错法[J].中文信息学报,2020,34(12):48-53+64.

[18] 洛桑嘎登,杨媛媛,赵小兵.基于知识融合的 CRFs 藏文分词系统[J].中文信息学报,2015,29(06):213-219.