融合多元特征表示的藏文命名实体识别方法

俄见才让[1,2]　**周毛克**[2,3]　**陈波**[1,2]　**赵小兵***[1,2]
[1]中央民族大学　信息工程学院　北京　100081
[2]国家语言资源监测与研究少数民族语言中心　100081
[3]中央民族大学　中国少数民族语言文学学院　北京　100081
*通讯作者：赵小兵
{nmzxb_cn@163.com}

摘要

本文针对基于音节嵌入方式的藏文命名实体识别(TNER)中词汇信息和音节部件信息忽略的问题，提出了基于交叉Transformer架构的MECT-TL模型，融合了藏文音节信息、词汇信息和音节部件信息的多元数据特征。MECT-TL通过平面网络结构将藏文音节与词汇信息结合，并整合音节部件信息，有效提升了藏文实体识别的准确性。实验结果显示，相较于主流的TNER基准模型BiLSTM-CRF，本文模型在F1值上提高了5.14个百分点，与基于Transformer架构的TENER模型相比提高了4.18个百分点。这表明，融合藏文词汇和音节部件信息的方法可以显著提高TNER任务的性能。

关键词： 词汇信息；音节部件信息；多元数据特征；交叉Transformer

Research on Tibetan Named Entity Recognition Using Multi-Feature Fusion Representation

Cairang Ejian[1,2]　**Maoke Zhou**[2,3]　**Bo Chen**[1,2]　**Xiaobing Zhao***[1,2]
[1]School of Information Engineering, Minzu University of China, Beijing,100081
[2]National Language Resources Monitoring and Research Center for Minority Languages,100081
[3]School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing,100081
*Corresponding author: Xiaobing Zhao
{nmzxb_cn@163.com}

Abstract

This paper addresses the issue of neglecting lexical and syllable component information in Tibetan Named Entity Recognition (TNER) based on syllable embedding. It proposes the MECT-TL model based on a cross-Transformer architecture, which integrates multiple data features including Tibetan syllable information, lexical information, and syllable component information. MECT-TL combines Tibetan syllables with lexical information through a flat network structure and integrates syllable component information, effectively improving Tibetan entity recognition accuracy. Experimental results show that compared to the mainstream TNER baseline model BiLSTM-CRF, the proposed model achieves a 5.14 percentage point improvement in F1 score, and a 4.18 percentage point improvement compared to the Transformer-based TENER model. This indicates that integrating Tibetan lexical and syllable component information significantly enhances the performance of TNER tasks.

Keywords: Lexical information , Syllable component information , Multiple data features , Cross-Transformer

1 引言

命名实体识别(NER)在信息理解会议(MUC-6)(Grishman and Sundheim, 1996)首次提出以来，迅速成为自然语言处理领域的核心研究课题。从文本中自动识别并分类具有特定意义的实体，对于深入理解和分析不同语言和文化背景下的文本至关重要。NER在信息检索 (Khalid et al., 2008)、知识图谱构建 (Riedel et al., 2013)、问答系统 (Diefenbach et al., 2018)、舆情分析 (Wang et al., 2016)、生物医学研究 (Settles, 2004)、推荐系统 (Karatay and Karagoz, 2015)和机器翻译 (Babych and Hartley, 2003)等多个下游任务中发挥关键作用。

藏文命名实体识别(TNER)由于其独特文化价值和语言特性，对于推动藏文智能处理和文化传承具有重要意义。在实现TNER任务时，通常采用基于藏文音节的方法。与基于词的方法相比，这种方法通过实验证明通常更优，因为它不依赖于分词，从而避免了分词错误带来的影响。然而，这种只关注单个字符的方法不仅无法充分捕捉词汇的语义信息，还忽略了藏文音节内部结构信息。

TNER模型进行训练或评估时，词汇在预测实体边界起到至关重要的作用，例如，输入数据通常是已经标注好的序列文本如图1(a)所示，输入藏文例句:“མཚོ་སྔོན་པོ་དྷ་ཏེན་རྩལ་གུང་སི་”被翻译为“青海布达科技公司”，这句话中包含了地理实体:“མཚོ་སྔོན་(青海)” 和组织实体“པོ་དྷ་ཏེན་རྩལ་གུང་སི་(布达科技公司)”。数据集文本序列使用BMES标注方案标注相应的实体类型标签，以帮助模型学习如何识别和分类文本中的不同实体。如图1(b)所示，基于藏文音节的藏文命名实体识别方法因处理每个单独的藏文字符，从而缺乏对整体词义和词边界的认识。导致模型在实体边界的预测上不够准确，错误地将图1(a)的输入文本序列中的实体识别为地理实体：“མཚོ་སྔོན་པོ་(青海湖)”，非实体：“དྷ་(达)”，组织实体：“ཏེན་རྩལ་གུང་སི་(科技公司)”。这种误判揭示了基于藏文音节方法在捕捉词义和辨认词边界上的不足，倾向于产生不准确的实体边界预测。直接影响到实体边界的准确预测，导致识别准确率下降，尤其开放领域的文本处理中，这一问题尤为突出，因为文本的多样性和复杂性更高，分词错误更难避免，其对实体识别的影响也更为显著。词汇信息在确定实体边界中扮演着至关重要的角色，如图1(c)所示，通过在藏文音节级别嵌入中融合藏文词汇信息，模型能够更准确地识别和划分实体。将输入序列文本正确识别为地理实体“མཚོ་སྔོན་(青海)”，组织实体:“པོ་དྷ་ཏེན་རྩལ་གུང་སི་(布达科技公司)”,与原来标注数据集标签完全对应。这种在藏文音节嵌入基础上结合词汇嵌入的方式可以融入更多文本特征信息以判定词汇边界不仅能够显著提高TNER的准确性和鲁棒性，还能增强模型对文本结构的理解能力。

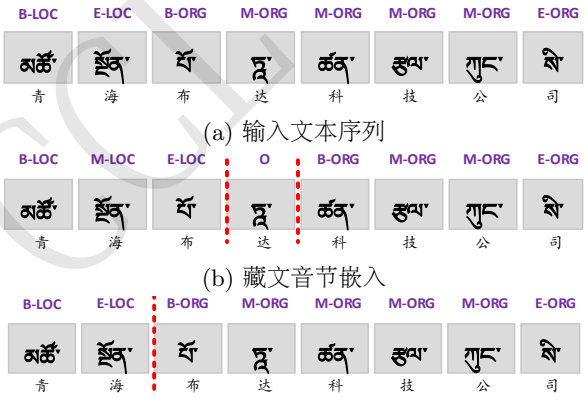


图 1. 藏文词汇信息对TNER的影响

每一个藏文音节都是由精细的音节部件构成，每一部件都遵循藏文的语法规则，共同形成具有特定含义的藏文音节。每个部件都承载着特定的语义和功能。一个完整的藏文音节由一个基字或者基字基础上1至6个音节部件构成，这些部件包括前加字、上加字、下加字、元音、后加字以及重后加字。例如，图2所示“བཀྲིགས(排列)”这一藏文音节，就由上述六种部件组成。

藏文音节部件不仅在视觉上具有区分度，而且在语言处理中扮演着独特的角色，它们对于区分词汇的含义至关重要。因此，基于藏文音节嵌入的方法忽视这些音节内部的特征，等同于忽略了藏文本身的一些关键信息，这可能会导致模型在识别过程中

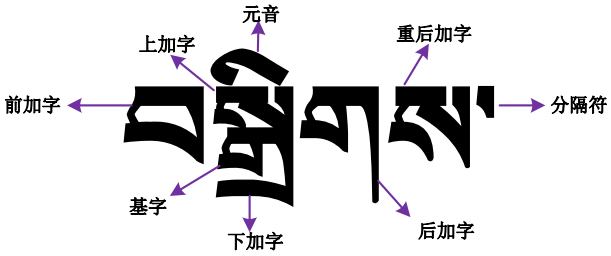


图 2. 藏文音节构成示例

的信息损失，从而影响最终的识别精度。此外，藏文句子中，部分功能性虚词 (多拉, 2011)如:“འི”、“འིས”、“ར”、“འོ”、“འམ”等通常不独立出现，而是附着在前面的藏文音节之后，形成了黏着现象。这种现象会改变原有藏文音节的结构，增加了TNER复杂度。例如，在句子“གླུ་བ་ནམ་མཁའི་གླུ་གཞས་ལ་ཉན་བཞིན་ལྷ་སར་འགྲོ།(意为“听着歌手南木夸的歌声去拉萨)”中，虚词“འི”黏着在人名实体“ནམ་མཁའ(南木夸)”之后形成黏着现象“ནམ་མཁའི”，而地名实体“ལྷ་ས(拉萨)”后面紧接的虚词“ར”形成黏着现象“ལྷ་སར”，都改变了原藏文音节的结构。另外，藏文人名如:“ཨོབྡན་བཀྲིས(俄见扎西)”是“ཨོ་རྒྱན་བཀྲ་ཤིས(俄见扎西)”的缩写，其中“ཨོ་རྒྱན་(俄见)”缩写为“ཨོབྡན(俄见)”，“བཀྲ་ཤིས(扎西)”缩写为“བཀྲིས(扎西)”，这样的缩写方式同样导致原藏文音节结构发生变化给TNER任务带来挑战。

针对目前藏文NER任务中藏文词汇信息和藏文音节部件信息被忽略的问题，本文提出一种有效结合藏文词汇信息和部件信息的藏文NER方法，采用基于Transformer编码器的平面网格结构，在藏文音节的基础上融合了藏文词汇信息。同时，再用一个Transformer编码器来提取藏文音节部件特征。通过交换两个编码器的查询值(Query)的方式实现了藏文多元数据特征的融合，从而有效地解决TNER任务中的挑战。主要贡献包括:

- 在藏文音节基础上融合了藏文词汇和藏文音节部件特征，采用多元藏文数据特征融合方法实现了藏文命名实体识别任务。
- 构建了藏文词汇词典和藏文音节部件词典，为后续相关研究提供了基础数据支持，利用交叉Transformer架构有效实现了对藏文多元数据的融合。
- 在数据集TibNER[1]上，本文所提出的方法相比基线模型表现出显著的提升，验证了该方法的有效性和实用性。

2 相关工作

近年来，命名实体识别领域取得了显著进展。 (Huang et al., 2015)首次将Bi-LSTM-CRF模型引入序列标注，取得了前所未有的成果。 (Zhang and Yang, 2018)提出了针对中文命名实体识别的Lattice-LSTM模型。该模型基于Bi-LSTM-CRF模型，通过综合编码中文词汇和字符信息，显著提高了对中文实体的识别准确性。 (Gui et al., 2019a)提出基于卷积神经网络(CNN)的中文字符级命名实体识别方法LR-CNN，采用重新思考机制整合词典，实现了更快速的识别速度和更好的性能。 (Gui et al., 2019b)提出了一种名为LGN的新型模型，采用图神经网络(GNN)作为基础，利用词典构建图结构，以捕获局部组合和全局语义。 (Yan et al., 2019)提出采用改进的Transformer编码器的NER模型TENER，证明了Transformer在NER任务中的有效性。 (Ma et al., 2019)提出的Soft_Lexicon方法简单有效地将词典融入字符表示中，使其易于迁移到任何中文命名实体识别模型上，实现了词汇词典的简化使用。 (Li et al., 2020)提出了基于Transformer的中文命名实体识别模型FLAT，将字符和词汇格子结构转换为跨度组成的平面结构，以实现对格子信息的充分利用，从而实现了出色的性能和效率。 (Wu et al., 2022)提出的NFLAT模型针对FLAT模型的冗余注意力计算进行了优化，显著降低了计算和内存成本，是用于中文命名实体识别的一种高效模型。 (Wu et al., 2021)提出了一种基于多元数据嵌入的交叉Transformer方法MECT，用于中文命名实体识别，通过融合汉字的结构信息来提高性能。此外，DAMO-NLP团队 (Wang et al., 2022)在MultiCoNER (Malmasi et al., 2022)共

[1]本文所用数据集已公开，地址：https://github.com/EJ-Cairang/TibNER_up

享任务中利用基于维基百科的多语言知识库，为命名实体识别模型提供相关的上下文信息，取得了10项胜利。接着，在此基础上，他们团队 (Tan et al., 2023)提出了U-RaNER检索增强系统，通过引入Wikidata知识库和注入方法来增强多语言细粒度NER任务的检索上下文，在MultiCoNER2 (Fetahu et al., 2023)共享任务的挑战中赢得了13项比赛中的9项冠军。

随着深度学习的发展，藏文命名实体识别(TNER)也取得了显著进展。 (朱亚军and 拥措, 2022)基于BiLSTM-CRF模型对藏文人名地名进行识别任务。 (环科尤et al., 2022)基于TS-BiLSTM-CRF网络识别格萨尔史诗文本中的六种格萨尔命名实体。 (洛桑嘎登et al., 2022)提出了一种融合藏文音节部件和藏文音节特征的模型SL-BiLSTM-CRF，验证了该方法对TNER任务的有效性。 (朱亚军et al., 2023)在基于音节的藏文BERT模型构建了BERT-BiLSTM-CRF模型，用于藏医药医学实体识别。 (徐泽辉et al., 2023)提出利用级联技术将任务分解为两个子任务的Cascade-BiLSTM-CRF模型，结合藏文预训练模型，以简化模型结构和缩短训练时间。 (格勒尼玛et al., 2023)将藏文音节特征和分词特征融合到了BiLSTM-CRF模型中，以预测藏文实体识别结果，验证了该方法的有效性。

当前用于TNER的基座模型主要采用LSTM网络结构，然而，LSTM网络在建立长距离依赖关系和并行处理能力方面存在一定限制。受到FLAT (Li et al., 2020)和MECT (Wu et al., 2021)方法在中文命名实体识别任务中的启发，本文提出了MECT-TL模型，采用基于Transformer的架构，在藏文音节基础上结合藏文词汇和音节部件信息，实现了藏文多元数据的融合。相较于现有TNER方法，本文方法具有更大的改进和创新。

3 MECT-TL:融合藏文词汇和音节部件特征的方法

3.1 整体框架方法

本文的总体模型框架**MECT-TL**(**MECT** for **T**ibetan **L**anguage)如图3所示，是对 (Wu et al., 2021)为中文命名实体识别所设计的MECT模型针对藏文特性进行改进，以适用于藏文命名实体识别任务。具体改进包括对藏文词汇信息融合和藏文音节部件信息融合。 MECT-TL模型

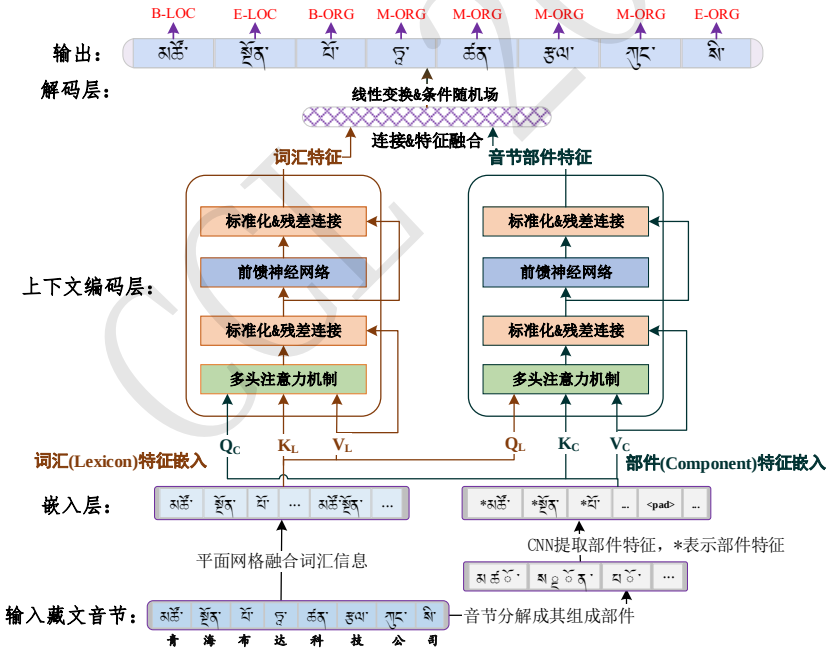


图 3. MECT-TL模型总体框架

包括两个Transformer编码器：融合词汇信息和提取音节部件信息。在这两个编码器中，输入文本序列分别经过词汇特征嵌入的Transformer编码器和音节部件特征嵌入的Transformer编码器，从而分别提取藏文词汇信息和音节部件信息。词汇特征嵌入的Transformer编码器在藏文音节的基础上融合相关的词汇信息，得到词汇特征；而音节部件特征嵌入的Transformer编码器则首先将藏文音节序列分解成对应的构成音节部件，随后通过CNN网络提取这些部件的特征，并将其送入专门用于整合藏文音节部件特征的编码器中，得到音节部件特征。通过交替整合词汇

信息和音节部件信息的Query以及连接操作，实现了对藏文音节、藏文词汇和音节部件的多元特征融合。最终，利用条件随机场(CRF)(?)进行解码，输出预测的标记序列，从而完成藏文命名实体识别任务。

3.2 词汇特征抽取

3.2.1 构建词汇词典

本文旨在探究融合不同粒度词汇对TNER的影响，本文采用Word2Vec处理了1.76GB的藏文文本，并构建了三种不同粒度的50维词向量词典，以捕捉词语之间的上下文关系：

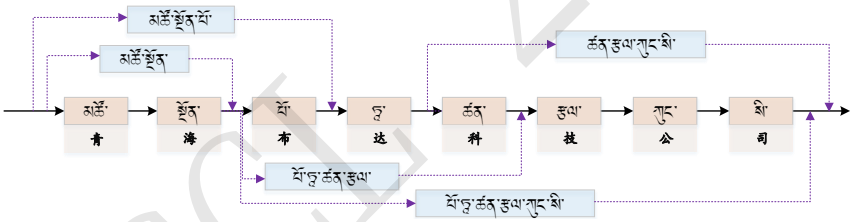
音节词典：包含22,973个藏文音节。有助于藏文的理解能力，改善语言模型。

双音节词典：由113,670个双音节藏文词汇构成。这增强了对双音节词汇语义的理解。

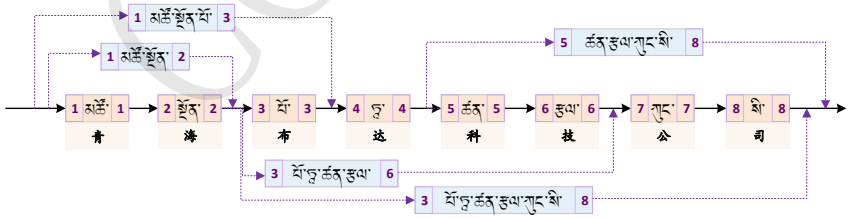
综合词典：包含单、双和多音节藏文词汇，共包含909,401个词汇。为复杂文本分析提供了强大支持。

3.2.2 融合词汇信息

基于藏文音节方法在捕捉词义和辨认词边界上的不足，缺乏关键信息导致产生不准确的实体边界预测。对输入藏文音节序列与藏文词典进行匹配，以便在词典中识别出所有潜在的词汇，从而识别出所有可能的词汇，并构建出词汇信息的晶体网格结构。如图4(a) 所示在这个结构中，每个节点不再是简单的藏文音节，而是扩展到了所有匹配到的潜在词汇。所形成的晶体结构是一个复杂的有向无环图，包括了整个藏文音节及其潜在的词汇。每个词汇在晶体中的位置是由它的首尾字符位置来明确界定的。为后续的实体辨识提供了清晰的边界指示。在这个多维结构中，识别特定词汇如人名、地名或机构名称至关重要，这是NER任务的核心。尽管晶体网格结构能够有效地融合藏文词汇信息，但其复杂性和动态性使得利用GPU并行计算能力变得困难，导致推理速度下降。图4(b)是带有位置索引的晶体结构。通过将晶体结构转化为平面结构图4(c)，平面结构中每个跨度对应于藏文音节或潜在藏文词汇及其在原始晶体网格中的位置，平面结构简化了表示方式。这种转换使模型能够更充分地利用网格信息，显著提高并行化能力，从而提升模型的性能和效率。



(a) 融合词汇信息的晶体网格结构



(b) 带有位置索引的晶体网格结构

(c) 融合词汇信息的平面网格结构

图 4. 融合词汇信息的晶体网络结构转平面网络结构

平面网格结构中，开始位置索引和结束位置索引表示每个跨度的第一个和最后一个藏文音节或藏文词汇的位置索引，它们表示文本单元在网格中的位置。对于藏文音节来说，它的开始位置编号和结束位置编号是一样的。有一种简单的算法可以将平面网格结构恢复到原始的晶体

网格结构。这样的处理既保持了数据的完整性，也便于充分利用Transformer的优势，利用位置编码将词汇信息在实体边界的预判中发挥作用，如图5所示。

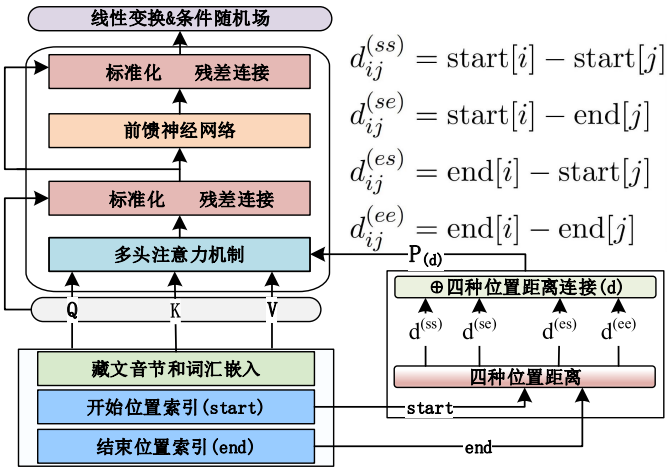


图 5. 平面网格结构的Transformer编码器

start[i]和end[i]表示跨度X_i的开始位置和结束位置。针对跨度X_i和X_j之间的关系，可以定义图中的$d_{ij}^{(ss)}$, $d_{ij}^{(se)}$, $d_{ij}^{(es)}$和$d_{ij}^{(ee)}$四种相对距离来描述它们之间的相互位置关系(相交、包含和相离)。$d_{ij}^{(ss)}$表示X_i和X_j的起始位置的距离，$d_{ij}^{(se)}$表示X_i的起始位置和X_j的结束位置的距离、$d_{ij}^{(es)}$表示X_i的结束位置和X_j的起始位置的距离、$d_{ij}^{(ee)}$表示X_i和X_j的结束位置的距离。相对位置编码的表达式为：

$$R_{ij} = \text{ReLU}\left(W_r\left(P_{d_{ij}^{(ss)}} \oplus P_{d_{ij}^{(se)}} \oplus P_{d_{ij}^{(es)}} \oplus P_{d_{ij}^{(ee)}}\right)\right) \tag{1}$$

其中P_d计算公式为：

$$P_d^{(2k)} = \sin\left(\frac{d}{10000^{2k/d_{\text{model}}}}\right), \quad P_d^{(2k+1)} = \cos\left(\frac{d}{10000^{2k/d_{\text{model}}}}\right) \tag{2}$$

这里的P_d指的是相对位置的嵌入表示。然后使用Transformer-XL (Dai et al., 2019)中的相对位置编码将R_{ij}加到Attention机制里面然后再使用自注意力的变形来计算跨度间相对位置编码的注意力分数，公式为：

$$A_{i,j}^* = W_q^T E_{x_i}^T E_{x_j} W_{k,E} + W_q^T E_{x_i}^T W_{k,R} R_{i-j} + u^T E_{x_j} W_{k,E} + v^T W_{k,R} R_{i-j} \tag{3}$$

最后将$A_{(i,j)}^*$代入到自注意力的计算公式中，可以得到整个序列的注意力值分布：

$$\text{Attention}(A, V) = \text{softmax}(A) \cdot V \tag{4}$$

$$A_{ij} = \left((Q_i + u)^T K_j\right) + \left((Q_i + v)^T R_{ij}^*\right) \tag{5}$$

$$[\text{Q}, \text{K}, \text{V}] = E_x[W_q, W_k, W_v] \tag{6}$$

其中$R_{ij}^* = R_{ij} \cdot W_R$，$u$和$v$是可学习的参数。这实现了在藏文音节基础上融合藏文词汇信息。

3.3 音节特征抽取

3.3.1 音节部件词典构建

本文自行构建了一个包含18776个藏文音节及其构成部件对应的词典，几乎覆盖了藏文文本中可能出现的所有音节。每个音节都在基字(表1中粗体表示)的基础上被详细地分解为前加字、上加字、下加字、元音、后加字以及重后加字等组成部分。该词典精细地记录了每个音节的构成部件，为深入理解和分析藏文音节提供了有力的支持。表1展示了一些藏文音节及其对应构成部件的例子，这些示例反映了词典大致内容。表中符号“-”用以表示特定位置没有相应的音节部件。

藏文音节	前加字	上加字	基字	下加字	元音	后加字	重后加字
བསྒྲིགས	བ	ས	ག	ྒྲ	ི	ག	ས
དབྱིབས	ད	–	བ	ྱ	ི	ས	ས
ཚུལ་གས	–	ས	ག	ུ	་	ས	ས
མངའ་གས	མ	–	ང	–	ཱི	ས	ས
འཕྲུལ	འ	–	ཕ	ྲུ	–	ལ	–
རྟེད	–	–	ར	–	ེ	ད	–
གསན	ག	–	ས	–	–	ན	–
བབྱུ	ད	–	བ	ྱུ	–	–	–
ཨ, ཨཿ	–	ཨ	ཨ	–	ཿ	–	–
ཨཾ	–	–	ཨ	–	–	–	–

表 1. 藏文音节及其对应构成部件示例

3.3.2 融合音节部件信息

藏文音节由1至7个音节部件构成，这些部件按照严格的藏文语法组合，赋予每个音节特定的语义。因此，音节部件序列在向量空间中能够有效地映射出音节的内在特征，对于理解音节的语义层面至关重要。为了更好地捕捉藏文音节序列的局部上下文特征，本文用一个基于卷积神经网络的模型。该模型能够通过局部特征提取强化音节嵌入的表征能力。如图6所示，展现了藏文音节部件特征提取的过程。图中的示例输入的藏文音节:“མཚོ་(湖)”，可分解为三个部件：前加字:“མ”、基字:“ཚ”、上加字:“ོ”。这些部件随后通过卷积神经网络进行卷积运算、最大池化等操作，获得了带有特征的藏文音节部件特征。 然后将特征输送到Transformer编码器来获取藏

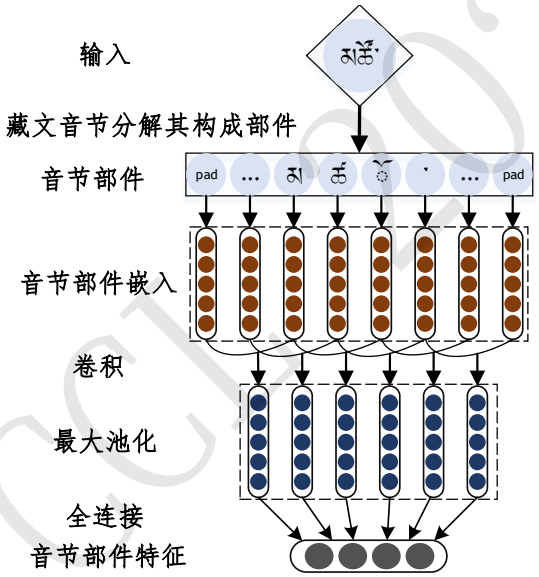


图 6. 基于卷积神经网络的音节部件特征抽取

文音节内部结构的补充语义信息。同时，利用了藏文音节的上下文和藏文词汇信息，丰富了藏文的语义表达，实现了多元特征的融合。输入$Q_L(Q_C)$、$K_L(K_C)$、$V_L(V_C)$是通过对藏文词汇融合和音节部件特征嵌入进行线性变换获得的:

$$\begin{bmatrix} Q_{L(C),i} \\ K_{L(C),i} \\ V_{L((C),i} \end{bmatrix}^T = E_{L(C),i} \begin{bmatrix} W_{L(C),Q} \\ I \\ W_{L(C),V} \end{bmatrix}^T \tag{7}$$

其中E_L和E_C分别是藏文词汇信息嵌入和音节部件信息嵌入，I是单位矩阵，每个W是一个可学习的参数。然后使用3.2.2中表述的相对位置编码来表示词的边界信息，并在模型中计算注意力得分:

$$\text{Attention\_}L\left(A_C, V_L\right) = \text{softmax}\left(A_C\right) V_L \tag{8}$$

$$\text{Attention\_}C\left(A_L, V_C\right) = \text{softmax}\left(A_L\right) V_C \tag{9}$$

$$A_{L(C),ij} = \left(Q_{L(C),i} + u_{L(C)}\right)^T K_{C(L),j} + \left(Q_{L(C),i} + v_{L(C)}\right)^T R^*_{L(C),ij} \tag{10}$$

其中u和v是方程(10)中注意力偏置的可学习参数，A_L 是藏文词汇的注意力得分，A_C 表示音节部件注意力得分。$R^*_{ij} = R_{ij} \cdot W_C$，其中$W_C$是可学习的参数。相对位置编码$R_{ij}$ 的计算如下：

$$R_{ij} = \text{ReLU}\left(W_r\left(p_{s_i - s_j} \oplus p_{e_i - e_j}\right)\right) \tag{11}$$

其中，s:start, e:end表示跨度的开始位置索引和结束位置索引。为了实现藏文词汇嵌入和音节部件特征嵌入所需的注意力偏置，该模型采用了随机注意力以提升性能。随机注意力能更好地适应两个子空间的分布。随机注意力是一个随机初始化的参数矩阵$B \in \mathbb{R}^{max_len \times max_len}$，它被添加到之前的注意力得分上，从而得到总的注意力得分：

$$\begin{aligned} V^*_L &= \text{Softmax}\left(A_C + B\right) V_L \\ V^*_C &= \text{Softmax}\left(A_L + B\right) V_C \end{aligned} \tag{12}$$

为了减少信息损失，直接连接藏文词汇特征和音节部件特征，并将它们输入到一个全连接层进行信息融合：

$$\text{Fusion}\left(V^*_L, V^*_C\right) = \left(V^*_C \oplus V^*_L\right) W^* + b \tag{13}$$

其中，符号$\oplus$的连接操作负责将来自音节部件和词汇信息的特征向量合并，形成一个全面的特征表示，这对于增强模型对藏文文本的解析能力至关重要。W^*和b是作为模型中的可学习参数，在训练过程中持续优化，以促进有效的特征融合。将融合后的特征被输入到条件随机场(CRF) (Lafferty et al., 2001)层进行识别任务。

4 实验与分析

4.1 实验数据与评估标准

4.1.1 实验数据

本文利用藏文实体词典半自动地构建了藏文命名实体识别数据集——TibNER。为保证数据集的质量，我们对自动标注结果进行了人工校审。TibNER规模为55227句，标注的实体包括人名(PER)、地名(LOC)和组织机构名(ORG)，共计43678个实体。按照8:1:1的比例将TibNER数据集随机划分为训练集、验证集和测试集，数据详情如表2所示。

数据集	句子数量	实体类型及数量			实体总数
		人名	地名	组织机构名	
训练集	44182	7790	15700	8764	32254
验证集	4997	933	1787	917	3637
测试集	6048	1068	2065	1155	4288
总计	**55227**	**9791**	**19552**	**10836**	**40179**

表 2. TibNER数据集统计信息

4.1.2 评估标准

命名实体识别的评价标准对于评估模型性能至关重要。1996年 (Grishman and Sundheim, 1996)等提出了一套综合性评价标准，正确识别的实例需要系统同时正确识别其边界和类型。更具体地说，假阳性(F_P)、假阴性(F_N)和真阳性(T_P)的数量用于计算精确度($P = \frac{T_P}{T_P+F_P} \times 100\%$)、召回率($R = \frac{T_P}{T_P+F_N} \times 100\%$)和F1值($F1 = \frac{2PR}{P+R} \times 100\%$)。

4.2 实验设置与对比系统

实验均在Linux 操作系统环境下进行，使用Tesla V100 GPU，CUDA 版本为12.2，PyTorch 版本为1.5.1。为了确定TibNER数据集上的最佳模型超参数配置，采用了SMAC算法来搜索最佳超参数。这种策略允许我们在预定义的参数空间内随机选择不同的超参数组合，随机搜索的超参数范围表3所示，并通过重复实验来评估它们对模型性能的影响。

超参数	范围
output_dropout	[0.1, 0.2, 0.3]
lattice_dropout	[0.1, 0.2, 0.3]
radical_dropout	[0.1, 0.2, 0.3, 0.4]
warm_up	[0.1, 0.2, 0.3]
d_{head}	[16, 20]
d_{model}	[128, 160]
lr	[1e-3, 25e-4]
radical_lr	[6e-4, 25e-4]
momentum	[0.85, 0.97]

表 3: 超参数范围

MECT-TL与主流的TNER模型CRF、BiLSTM、BiLSTM、TENER进行了评估:
CRF:考虑了标签序列之间的依赖关系，通过最大化条件概率来确定最佳标签序列。
BiLSTM:通过双向LSTM网络捕捉序列数据中的长依赖关系和上下文信息。
BiLSTM-CRF:结合了双向长短期记忆网络(BiLSTM)和条件随机场(CRF)的模型，既能捕捉序列数据中的长依赖关系和上下文信息，又考虑了标签序列之间的依赖关系，通过最大化条件概率来确定最佳标签序列。目前TNER任务中常用的基准模型。
TENER:专门为命名实体识别任务开发的中文NER模型，在Transformer编码器中巧妙融合位置信息的自注意力机制。这种机制有效地捕捉了文本序列中的长距离上下文依赖关系，使得TENER能够识别和处理文本中的复杂地长依赖关系。

4.2.1 主实验

在TibNER数据集上与主流TNER模型进行了对比实验，结果见表4所示。

TNER模型	P(%)	R(%)	F1(%)
CRF	81.49	84.13	82.79
BiLSTM	74.02	83.99	78.69
BiLSTM-CRF	83.84	84.47	84.15
TENER	84.96	85.25	85.11
MECT-TL(Our)	**88.49**	**90.11**	**89.29**

表 4. TNER模型性能评估实验

本文模型MECT-TL在F1值上达到了89.29%，表现最佳。相比当前主流的TNER基准模型BiLSTM-CRF，本文模型提高了5.14% 的F1值；而相较于基于Transformer架构的TENER模型，则提高了4.18% 的F1值。MECT-TL模型在TNER任务中表现出卓越的性能。该模型在藏文音节的基础上融合了藏文词汇和音节部件信息，为其提供了丰富的多元数据特征，使其能够全面获取知识。可见，藏文词汇增强和音节部件增强方法对提升TNER任务性能具有显著的增益效果。

4.2.2 消融实验

通过逐步融合藏文的单音节词典、双音节词典以及包含单音节、双音节及更多音节的综合词典，以评估不同粒度的词汇信息对模型性能的影响。接着，比较仅融合音节部件的效果与在最佳增益效果的综合词典上再融合音节部件的模型性能。实验结果表5所示。表中“✓”表示有对应的词典融合，“✗”表示没有对应的词典融合。

实验结果表明，融合不同粒度的词汇信息对提升模型的藏文命名实体识别性能具有显著影响。其中，融合双音节词典比融合单音节词典具有更大的增益效果，这表明词汇有助于正确判定实体边界，并捕捉更复杂的语义信息。综合词汇信息融合后，性能达到最优，反映了融合多音节及更丰富词汇信息对模型理解和处理藏文文本的重要性。特别是对于藏文这种结构复杂的语言，有效的词汇信息融合显著提升了命名实体识别的准确度。再通过比较仅融合音节部件的效果与在最佳增益效果的综合词典上再融合音节部件的模型实验，可以发现融合词汇的效果比

藏文词汇词典			音节部件词典	P(%)	R(%)	F1(%)
单音节	双音节	综合				
✓	✗	✗	✗	84.14	85.61	84.67
✗	✓	✗	✗	85.06	86.45	85.75
✗	✗	✓	✗	87.15	89.98	88.54
✗	✗	✗	✓	84.83	86.04	85.43
		✓	✓	**88.49**	**90.11**	**89.29**

表 5. 消融实验结果

融合音节部件的增益效果更为显著。同时融合两者将能取得最佳的结果，说明融合多元特征的方法对TNER的有效性，这为未来藏文处理模型的优化和其他低资源语言的命名实体任务提供了启示。

4.2.3 推理效率对比实验

为了验证MECT-TL的计算效率，在TibNER数据集上进行了推理速度的对比实验，包括非批量处理(batch=1)和批量处理(batch=16)条件。结果如图7所示，其中我们将BiLSTM-CRF设为基准速度(值为1)以便于对比。

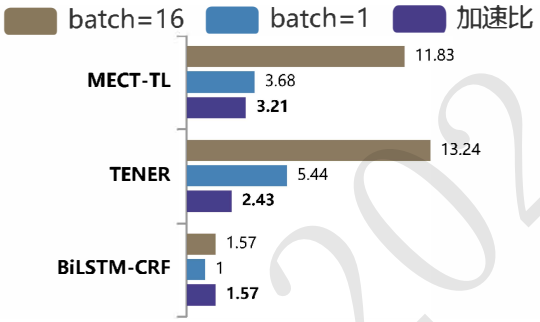


图 7. 不同模型计算效率对比实验

在非批量条件下，本文模型MECT-TL尽管在藏文音节特征基础上多融合了两种不同特征数据，但相较于BiLSTM-CRF的基准模型，其推理速度快了3.68倍，相比TENER模型慢了1.76倍。而在批量条件下，BiLSTM- CRF的加速比约为1.75，TENER的加速比约为3.06，表明模型在较大批处理下有一定的性能提升。而本文模型MECT-TL加速比达到了3.76，表现出了最佳的模型效率。虽然TENER推理速度比MECT-TL快，但MECT-TL模型在较大批处理下展现出了更为明显的性能提升。这表明本文模型在推理速度方面展现了高效的并行处理潜力，并未因多种特征融合而导致显著的推理速度下降。进一步突显了其在结构简洁性和计算效率方面的设计成功。MECT-TL模型表现出了强大的数据并行处理能力和快速有效的推理长距离依赖能力。

5 总结

本文针对传统藏文命名实体识别(TNER)方法的只依赖藏文音节层面嵌入的问题，通过采用交叉Transformer的先进方法，在藏文音节上融合了藏文词汇信息和藏文音节部件信息，实现了多元藏文数据特征融合表示。藏文词汇的融合使得模型更深入理解藏文词汇的语义信息。模型通过计算字词之间包含、相交、相离的距离关系得到的词汇注意力分数在判断文本中的实体边界起到关键的辅助作用，提高了实体识别的准确率。藏文音节部件信息的融合让模型利用这些承载语义信息的部件，提取到了藏文音节的内部结构，从而提高了TNER识别任务的性能。通过一系列细致的实验和分析，展示了所提方法在提升TNER任务的准确率和效率方面的优势。

致谢

感谢所有匿名审稿人的宝贵意见。本项研究成果受国家社会科学基金重大项目

（22&ZD035）、中国语言资源保护工程（课题编号：YB2404A003）、中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)资助。

参考文献

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information systems*, 55:529–569.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. Semeval-2023 task 2: Fine-grained multilingual named entity recognition (multiconer 2). *arXiv preprint arXiv:2305.06586*.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 volume 1: The 16th international conference on computational linguistics*.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019a. Cnn-based chinese ner with lexicon rethinking. In *ijcai*, volume 2019.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuan-Jing Huang. 2019b. A lexicon-based graph neural network for chinese ner. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1040–1050.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Deniz Karatay and Pinar Karagoz. 2015. User interest modeling in twitter with named entity recognition. In *5th Workshop on Making Sense of Microposts*.

Mahboob Alam Khalid, Valentin Jijkoun, and Maarten De Rijke. 2008. The impact of named entity normalization on information retrieval for question answering. In *European Conference on Information Retrieval*, pages 705–710. Springer.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. *arXiv preprint arXiv:2004.11795*.

Ruotian Ma, Minlong Peng, Qi Zhang, and Xuanjing Huang. 2019. Simplify the usage of lexicon in chinese ner. *arXiv preprint arXiv:1908.05969*.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Multiconer: A large-scale multilingual dataset for complex named entity recognition. *arXiv preprint arXiv:2208.14536*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 74–84.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pages 107–110.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, et al. 2023. Damo-nlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. *arXiv preprint arXiv:2305.03688*.

Zhibo Wang, Xiaohui Cui, Lu Gao, Qi Yin, Lei Ke, and Shurong Zhang. 2016. A hybrid model of sentimental entity recognition on mobile social media. *EURASIP Journal on Wireless Communications and Networking*, 2016:1–12.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, et al. 2022. Damo-nlp at semeval-2022 task 11: A knowledge-based system for multilingual named entity recognition. *arXiv preprint arXiv:2203.00545*.

Shuang Wu, Xiaoning Song, and Zhenhua Feng. 2021. Mect: Multi-metadata embedding based cross-transformer for chinese named entity recognition. *arXiv preprint arXiv:2107.05418*.

Shuang Wu, Xiaoning Song, Zhenhua Feng, and Xiao-Jun Wu. 2022. Nflat: Non-flat-lattice transformer for chinese named entity recognition. *arXiv preprint arXiv:2205.05832*.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.

多拉. 2011. 藏语语义理解中功能性虚词研究. 西藏大学学报*(*社会科学版*)*, 26:106–112.

徐泽辉, 珠杰, 许泽洲, 汪超, 严松思, and 刘亚姗. 2023. 结合级联技术的藏文预训练命名实体识别模型. 中文信息学报, 37:23–28.

朱亚军and 拥措. 2022. 基于深度学习的藏文人名地名识别. 信息与电脑*(*理论版*)*, 34:66–68.

朱亚军, 拥措, and 尼玛扎西. 2023. 基于藏文bert的藏医药医学实体识别. 计算机与现代化, pages 43–48.

格勒尼玛, 群诺, 项秀才让, 洛桑嘎登, and 尼玛扎西. 2023. 结合分词特征的藏文命名实体识别方法. 高原科学研究, 7:106–114.

洛桑嘎登, 群诺, 索南尖措, and 仁增多杰. 2022. 融合音节部件特征的藏文命名实体识别方法. 厦门大学学报*(*自然科学版*)*, 61:624–629.

环科尤, 华却才让, 才让当知, and 多杰才让. 2022. 基于深度学习的格萨尔史诗命名实体识别研究. 中文信息学报, 36:46–53.