

基于动态多尺度增强的藏文自动分词方法

李京阳, 彭展

(西藏民族大学信息工程学院, 陕西 咸阳 712000)

摘要: 高质量的藏文分词不仅是确保藏文文本处理质量与效率的关键, 还对推动藏文智能信息处理技术的发展具有重要意义。为此, 针对藏文特性所导致的紧缩词识别、歧义处理和未登录词识别等问题, 提出一种融合预训练模型、动态多尺度上下文增强与注意力机制的多策略藏文分词模型 TDMA。首先, 利用 TiBERT 预训练模型提取藏文深层语义特征; 其次, 通过动态多尺度上下文增强方法扩展上下文窗口, 以捕捉词缀与词根之间的复杂关系; 最后, 结合 BiLSTM-CRF 与注意力机制优化标签依赖关系。实验表明, 所提方法在 MLWS2021 语料和 tibetan-news 语料上的 F1 值分别提升了 3.78%、6.78%。

关键词: 藏文分词; 预训练模型; 动态多尺度上下文增强

DOI: 10.11907/rjdk.251637

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2025)012-0126-07

扫描二维码阅读全文:



A Tibetan Word Segmentation Method Based on Dynamic Multi-Scale Enhancement

LI Jingyang, PENG Zhan

(College of Information Engineering, Xizang Minzu University, Xianyang 712000, China)

Abstract: High quality Tibetan word segmentation is not only the key to ensuring the quality and efficiency of Tibetan text processing, but also of great significance in promoting the development of intelligent Tibetan information processing technology. To address the issues of tight word recognition, ambiguity handling, and unregistered word recognition caused by the characteristics of Tibetan language, a multi strategy Tibetan word segmentation model TDMA is proposed, which integrates pre-training models, dynamic multi-scale context enhancement, and attention mechanisms. Firstly, the TiBERT pre trained model is used to extract deep semantic features of Tibetan language; Secondly, the dynamic multi-scale context enhancement method is used to expand the context window and capture the complex relationship between affixes and roots; Finally, combining BiLSTM-CRF with attention mechanism to optimize label dependency relationships. The experiment showed that the proposed method improved the F1 score by 3.78% and 6.78% on the MLWS2021 corpus and tibetan news corpus, respectively.

Key Words: Tibetan word segmentation; pre-trained models; dynamic multi-scale context enhancement

0 引言

在全球化和信息化的大背景下, 语言技术的发展已成为连接不同文化和促进知识共享的重要桥梁。藏语作为中国五大少数民族语言之一, 其信息化建设对保护和传承藏族语言文化至关重要。藏文分词作为藏文信息处理的基础工作, 对实现语音合成、语音识别、信息检索等应用具

有关键作用^[1]。作为理解和处理藏语文本的核心环节, 不仅为机器翻译、搜索引擎、网络信息安全等领域研究奠定了基础, 更是后续句法分析、语义分析及篇章分析的重要前提。然而, 藏文结构的复杂性给分词带来了挑战, 尤其是紧缩词识别、歧义处理和未登录词识别等问题。

为了解决以上问题, 本文在引入预训练模型的基础上进一步结合了注意力机制和动态多尺度上下文增强方法, 以实现更精准的藏文分词效果。首先, 通过优化注意力分

收稿日期: 2025-09-11

软著编号: 基于 TDMA 模型的藏文分词软件 (2025SR0992415)

基金项目: 西藏自治区科技计划项目 (XZ202501ZY0094); 西藏民族大学研究生科研创新与实践项目 (Y2025128)

作者简介: 李京阳 (2000-), 女, 西藏民族大学信息工程学院硕士研究生, 研究方向为藏文信息处理、藏文自然语言处理; 彭展 (1986-), 男, 博士, 西藏民族大学信息工程学院讲师, 研究方向为文本处理、网络安全、网络舆情。

配,引导模型自主聚焦音节组合模式,减少对低信息量成分的依赖强度;其次,构建动态多尺度上下文增强模块,通过层次化特征融合提升文本序列的表示多样性,帮助模型在特征稀疏条件下更有效地捕获跨粒度语义特征。

1 相关研究

早期,藏文分词的研究从依赖规则和词典的方法,逐步过渡到统计学习方法,最终形成以深度学习技术为主导的研究趋势^[2]。扎西次仁^[3]设计了一个人机互助的藏文分词和词登录系统,标志着藏文分词研究的起步,这一时期的研究主要基于语言学规则和词典匹配。随后,陈玉忠等^[4]提出基于格助词和接续特征的书面藏文自动分词方案,这一阶段的研究开始引入统计模型。

尽管这些传统方法在一定程度上实现了藏文分词,但由于藏文的复杂性及紧缩词和未登录词的存在,难以适应大规模文本语料分词。才智杰^[5]提出还原法来提高藏文自动分词中紧缩词的识别准确率,解决了高频率紧缩词的识别问题,但仍无法处理未登录词+紧缩词的情况。孙媛等^[6]在使用格助词分块方法的基础上,采用双向匹配方式发现交集型歧义切分,利用词频信息处理歧义切分问题。史晓东等^[7]将基于HMM的汉语分词系统Segtag移植到藏文分词中,设计了央金藏文分词系统,实现了藏语的分词标注一体化。李亚超等^[8,9]开发的TIP-LAS系统是基于条件随机场模型实现的藏文分词系统,采用最大熵模型并融合音节特征来实现藏文词性标注系统。

随着深度学习技术在自然语言处理领域的突破,学者们开始尝试将深度学习技术应用于藏文分词。李博涵等^[10]实验验证了多个网络模型在藏文分词任务中的有效性,推动了深度学习技术在藏文分词中的应用。王莉莉等^[11]基于BiLSTM-CRF的藏文分词方法,通过学习藏文音节的序列特征,结合上下文信息更准确地识别词边界,尤其是在处理藏文中常见的紧缩词和未登录词时表现出色。王康^[12]在双向长短期记忆网络模型中引入卷积神经网络和膨胀卷积迭代算法,通过IDCNN作为特征提取器来提升模型分词效果。于永斌等^[13]引入注意力机制和音节扩展方法,提升模型对上下文信息和当前音节之间联系的关注能力。然而,上述方法无法表示多义词,提取的只是一个缺少上下文信息的静态单词向量。一方面,没有考虑词汇在不同语境下的信息;另一方面,会遗漏一些文本的语法和语义信息,因此分词效果还有待提高。

随着Bert等预训练语言模型的兴起。索朗次仁^[14]基于藏文子词词典构建和遮蔽预训练方法,提出适用于藏文的预训练语言模型,并在词性标注任务中达到了较高的准确率。尹宗鹤等^[15,16]提出基于ALBERT预训练语言模型和音节特征融合的ASBC藏文分词模型。色差甲等^[17]利用UniLM架构探索预训练模型在藏文分词中的应用,通过

预训练模型捕捉语法和语义层面的信息,辅助模型更好地锁定目标音节的标签,进一步提升了分词性能。杨杰等^[18]提出一种重构预训练模型标记器的方法,通过增强标记器切分结果与标注信息的平行性,进一步增强了模型对单音节特征信息的获取能力。格桑加措等^[19]提出一种结合Bi-LSTM与CRF的藏文分词方法,引入<unk>标记处理未登录音节字,显著提升了模型在机器翻译任务中的表现。王怡如^[20]探索了基于图神经网络(Graph Neural Network, GNN)的藏文分词与词性标注方法,引入虚拟节点和多头注意力机制,有效缓解了藏文数据稀疏和结构复杂的问题。常芳玉等^[21]提出基于八词位标签的BiLSTM-CRF模型,通过更细粒度的标签设计,提升了分词的准确性和语义表示能力。

2 TDMA 模型

2.1 模型架构

图1为本文提出的TDMA (Tibert-based Dynamic Multi-Scale Context and Attention)模型架构。

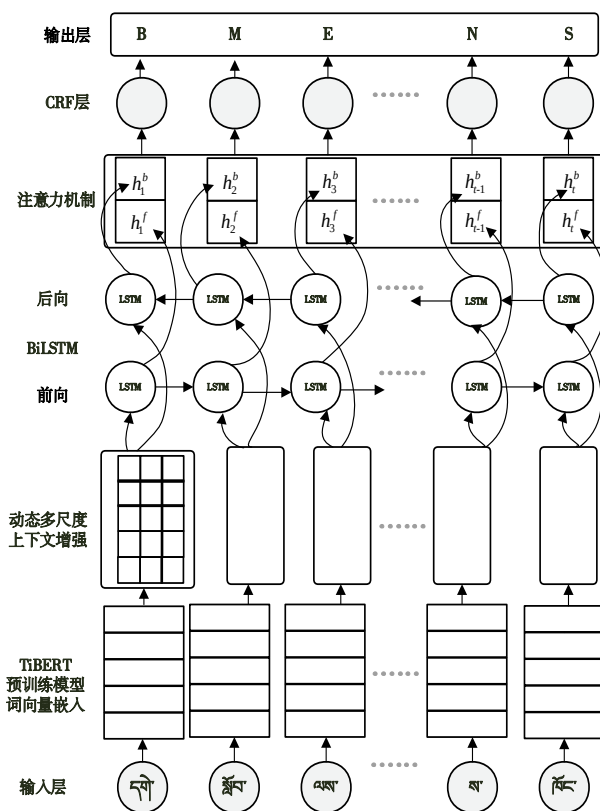


Fig. 1 TDMA model architecture

图1 TDMA 模型架构

首先,采用藏文预训练模型TiBERT提取藏文文本的深层语义和句法特征,生成词向量^[22];其次,通过动态多尺度卷积门控机制增强上下文表征,捕获特有的前后缀组合规律,使模型更准确地捕捉藏文的音节结构和语言特性;再次,将增强后的特征输入双向长短期记忆网络(BiL-

STM)^[23],利用双向序列捕捉长距离依赖关系;接下来,在解码阶段引入注意力机制,动态聚焦关键上下文信息辅助分词决策;最后,通过CRF层建模标签依赖关系,完成分词任务^[24]。

2.2 预训练模型 TiBERT

目前,word2vec、GloVe等常用的单词表示方法,只能从数据中学习静态、非语境化的词嵌入,无法区分一个词的多重含义,因此需要引入预训练的语言模型来生成上下文词向量表示。TiBERT是一个藏语单语预训练语言模型,源于青海师范大学省部共建藏语智能信息处理及应用国家重点实验室等提供的开源模型^[22]。该模型采用与BERT相同的架构,但在词汇表构建和子词分割方面对藏语进行了优化。

在分词方面,将BERT使用的WordPiece算法替换为Sentencepiece算法,采用Unigram语言模型生成多个候选的子词序列,使模型对噪声和子词分割错误更具鲁棒性,最终构建的词汇表覆盖了数据集中99.95%的字符。表1为TiBERT预训练模型的参数,均源原始论文。

Table 1 TiBERT training parameters

表1 TiBERT训练参数

参数名称	值
hidden size	300
embedding size	768
batch size	64
dropout	0.3
learning rate	0.1
num epochs	20
vocab size	30 005

2.3 动态多尺度上下文增强

藏文与汉语和英语的不同之处在于,其属于拼音文字系统,字母体系由30个辅音字母和4个元音字母构成。藏文的音节结构以一个基字为核心,其他字母则作为上加字、下加字、前加字、后加字、再后加字及元音部分,围绕基字进行前后或上下的叠加,共同构成完整的藏文音节字结构,如图2所示。

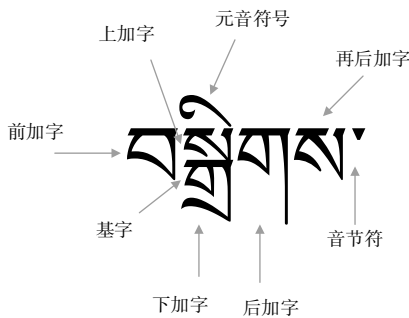


Fig. 2 Tibetan syllable structure

图2 藏文音节字结构

一个藏文音节字最少情况下包含一个基字,最多情况下包含6个辅音字母和1个元音字母。本文鉴于此结构特点,提出了一种动态多尺度上下文增强方法,如图3

所示。

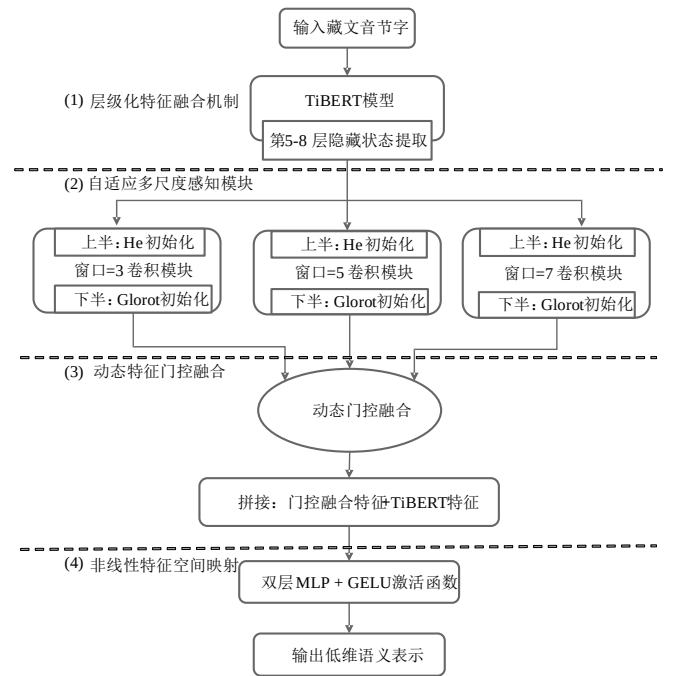


Fig. 3 Dynamic multi-scale context enhancement method

图3 动态多尺度上下文增强方法

通过分层特征融合与自适应特征提取机制,有效提升了藏语形态特征的代表能力。同时,在保持TiBERT原有架构的基础上,通过4个关键环节实现上下文增强:

(1)层级化特征融合机制。针对藏文字符中基字与加字的结构特性,对TiBERT中间层(第5-8层)的隐藏状态进行特征融合。该设计基于两个重要考量:①中间层特征通常包含丰富的形态和句法信息,而高层更偏向语义表征,选择5-8层可有效捕获藏文字符的构形特征;②通过分层特征融合,能建立基字与加字之间的非线性映射关系,增强模型对藏文字素组合模式的感知能力。

(2)自适应多尺度感知模块。为了适应藏文上下加字的纵向组合特性,设计了 $k=3, 5, 7$ 等3种窗口尺度的多分支一维卷积模块。首先,模块接收层级融合特征 $F_{mid} \in \mathbb{R}^{n \times d}$;其次,对每个尺度执行卷积运算,如式(1)所示。

$$F_{conv}^{(k)}[t, :] = \text{ReLU} \left(\sum_{i=0}^{k-1} W_k[:, :, i] \cdot F_{mid} \left[t + i - \left\lfloor \frac{k}{2} \right\rfloor, : \right] + b_k \right) \quad (1)$$

式中: $W_k \in \mathbb{R}^{d \times d \times k}$ 为卷积核; $b_k \in \mathbb{R}^d$ 为偏置项。

再次,采用垂直方向差异化初始化策略对卷积核 W_k 按高度分区初始化。其中,上半部分($j < \lceil k/2 \rceil$)采用He进行初始化,以适应高频特征(如音节“ ལྟོ ”中的基字“ ལྟ ”),如式(2)所示;下半部分($j \geq \lceil k/2 \rceil$)采用Glorot进行初始化,以匹配低频特征(如音节“ ལྟོ ”中的加字“ ལྟོ ”),如式(3)所示。

$$W_k[:, :, j] \sim N\left(0, \sqrt{\frac{2}{k \cdot d}}\right) \quad (2)$$

$$W_k[:, :, j] \sim N\left(0, \frac{1}{d}\right) \quad (3)$$

最后, 这种基于文字空间结构的初始化方法使模型在训练初期即具备对纵向组合特征的感知能力, 以便生成多尺度特征集合, 如式(4)所示。

$$\mathcal{F}_m = \{F_{\text{conv}}^{(3)}, F_{\text{conv}}^{(5)}, F_{\text{conv}}^{(7)}\} \subset \mathbf{R}^{n \times d} \quad (4)$$

此外, 各分支通过独立卷积运算提取不同感受野的局部特征(如 $k=3$ 窗口捕获紧密组合“ཨ་ཨ་ཨ་”, $k=7$ 窗口捕获完整音节“ཨ་ཨ་ཨ་”), 以形成互补的多粒度表征。

(3) 动态特征门控融合。为解决多尺度特征融合的权重分配问题, 引入可学习的门控注意力机制。通过计算各尺度特征的相对重要性权重, 实现特征的自适应融合。具体为, 针对每个卷积分支的输出特征 $F_i \in \mathbf{R}^{d \times n}$, 计算门控系数 α_i , 再进行特征融合。通过与原始 TiBERT 特征进行纵向拼接, 既保留了全局语义的连贯性, 又增强了局部形态特征的区分度。

(4) 非线性特征空间映射。为了降低增强特征的维度冗余, 构建了包含 GELU 激活函数的双层 MLP 投影网络。该设计利用 GELU 激活函数的连续可微特性 ($\text{GELU}(x) = x\Phi(x)$, Φ 为标准正态分布累积函数), 相较于传统 ReLU 函数, 在特征压缩过程中能更好地保持信息的拓扑结构。第一层 MLP 将维度扩展至 4D 建立高维交互空间, 第二层压缩至目标维度 D, 通过这种“扩展—压缩”策略实现特征蒸馏, 有效提升了模型对复杂字形组合的泛化能力。

2.4 注意力机制

注意力机制的核心是计算每个输入音节对于当前预测的重要性。对于每个时间步 t , 令 $h_t = [h_t^f, h_t^b]$ 表示 BiLSTM 在时间步 t 的隐藏状态, h_t^f 为前向 LSTM 的隐藏状态, h_t^b 为后向 LSTM 的隐藏状态, $h_t \in \mathbf{R}^{2m}$, m 为 BiLSTM 单侧的隐藏层维度。注意力机制具体的实现步骤为:

步骤 1: 计算注意力分数。对于每个时间步 t , 将 BiLSTM 的输出与可学习的权重向量 w_b 进行点积, 再通过一个激活函数 f 计算原始注意力分数 α'_t 。

$$\alpha'_t = f(W_a h_t + b_a) \quad (5)$$

式中: $W_a \in \mathbf{R}^{k \times 2m}$ 为权重矩阵; $b_a \in \mathbf{R}^k$ (k 为注意力隐空间维度) 为可学习的偏置项。

步骤 2: 归一化权重。使用 softmax 函数对所有时间步的原始注意力分数进行归一化, 得到注意力权重 α_t 。

$$\alpha_t = \frac{\exp(\alpha'_t)}{\sum_{i=1}^T \exp(\alpha'_i)} \quad (6)$$

式中: T 为序列的总时间步数。

步骤 3: 计算上下文向量。使用归一化的注意力权重对 BiLSTM 的输出进行加权求和, 得到上下文向量 c 。

$$c = \sum_{i=1}^T \alpha_i h_i \quad (7)$$

3 实验结果与分析

3.1 实验环境

本文实验硬件为 NVIDIA GeForce RTX 3090 GPU, 操作系统为 Windows 10, Python 版本为 3.8.20, 开发工具为 PyCharm, 深度学习框架采用 TensorFlow 2.10.0。

3.2 实验数据

实验数据集分别为 MLWS2021 藏文分词语料和 tibetan-news 语料^[25,26]。前者是第二届少数民族语言分词技术评测的官方语料, 包含新闻、法律、经济、小说和语言文字等领域综合语料; 后者是由西藏大学自然语言处理团队 (UTibetNLP) 建立的藏语新闻数据集, 包含政务、教育、文化、旅游多方面, 如表 2 所示。为了使模型在训练、验证和测试过程中充分利用数据, 按照 8:1:1 的比例划分训练集、验证集和测试集。

Table 2 Experimental dataset

数据集	训练集	验证集	测试集
MLWS 2021 语料 ^[25]	21 983	2 748	2 748
tibetan-news 语料 ^[26]	174 705	21 838	21 838

3.3 实验设置

为了全面评估所提多策略融合分词框架的性能, 选取传统机器学习、深度学习和大模型 3 类具代表性的模型进行比较。传统机器学习模型选用了基于条件随机场的 TIP-LAS 模型^[9]。深度学习模型采用了 BiLSTM-CRF 和 ASBC 模型, BiLSTM-CRF 结合了双向 LSTM 的上下文特征提取能力和 CRF 的标注优化能力, 是目前常用的分词模型^[15]; ASBC 模型基于 ALBERT 预训练语言模型和音节特征融合, 具有较强的特征表示能力。大模型选择基于 LLaMA2 的 T-LLaMA 模型, 参数量达 70 亿, 属于大规模语言模型, 具备更强的通用语言理解与生成能力, 能更好地捕捉藏语复杂的语法结构和语义信息^[27]。

在评估指标方面, 采用精确率 (P)、召回率 (R) 和 F1 值作为主要的评价标准, 以全面、客观地反映模型的分词性能。

$$P = \frac{TP}{TP + FP} \quad (8)$$

$$R = \frac{TP}{TP + FN} \quad (9)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (10)$$

3.4 实验结果

3.4.1 比较试验

本文在两种不同语料上进行比较实验, 以验证所提模型在不同数据集上的有效性, 实验结果如表 3、表 4 所示。

Table 3 Experimental results on MLWS 2021 corpus

表3 MLWS 2021 语料上的实验结果 (%)

方法	模型	P	R	F1
传统机器学习模型	TIP-LAS	90.25	91.34	90.79
	BiLSTM-CRF	92.04	92.02	92.03
深度学习模型	ASBC	93.18	94.13	93.65
	本文模型-TDMA	95.84	95.79	95.81
大模型	T-LLaMA	96.52	96.38	96.45

Table 4 Experimental results on tibetan-news corpus

表4 tibetan-news 语料上的实验结果 (%)

方法	模型	P	R	F1
传统机器学习模型	TIP-LAS	87.61	88.44	88.02
	BiLSTM-CRF	89.23	86.12	87.65
深度学习模型	ASBC	92.38	93.57	92.97
	本文模型-TDMA	94.25	94.62	94.43
大模型	T-LLaMA	94.71	94.89	94.80

为了更直观地展示不同模型在各个数据集上的性能差异,绘制了柱状图,如图4、图5所示。其中,柱状图横轴表示P、R和F1值,纵轴表示每个模型性能得分情况。

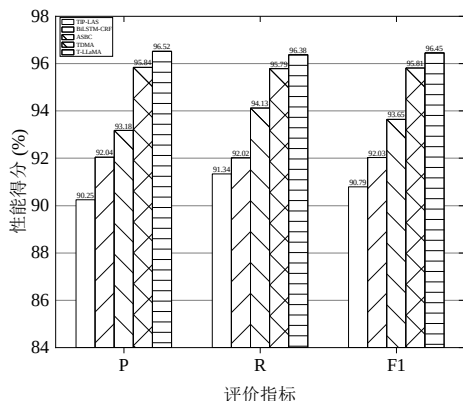


Fig. 4 Model performance on MLWS 2021 corpus

图4 MLWS 2021 语料上的模型性能

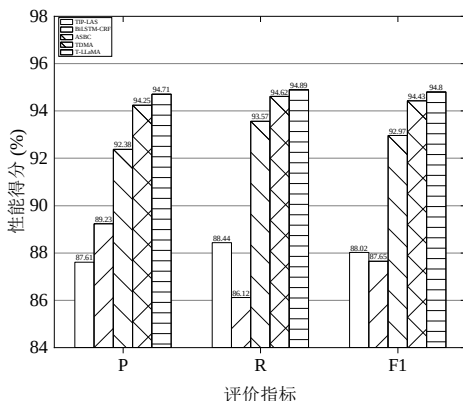


Fig. 5 Model performance on tibetan-news corpus

图5 tibetan-news 语料上的模型性能

实验表明,基于TiBERT预训练语言模型与动态多尺度上下文增强的藏文分词模型TDMA,在跨数据集测试中的表现优越。由表3、表4的数值分析和柱状图的可视化展示,TDMA在MLWS 2021语料、tibetan-news语料上的F1值分别达到95.81%、94.43%,相较于最优基线模型ASBC

分别提升2.16%、1.46%。相较于T-LLaMA大模型而言,TDMA在参数量差距极大的情况下,性能差距仍能控制在1%以内,体现了优异的质量效率平衡。

模型通过TiBERT的预训练特征提取模块,可有效捕获藏文音节粘连和复杂构词规律,尤其在低资源场景(MLWS 2021语料)中仍能保持高度的泛化能力,TDMA的F1值相较于传统机器学习模型TIP-LAS提升5.02%,召回率相较于BiLSTM-CRF提升3.77%。

同时,动态多尺度上下文增强模块通过自适应扩展窗口,显著提升了音节级特征区分度,有效解决了藏文复合词分割中的长距离依赖问题。在MLWS 2021语料中,TDMA的F1值相较于ASBC提升2.16%,精确率、召回率均显著高于ASBC,说明自适应扩展窗口不仅避免了固定窗口导致的局部特征过拟合,还通过动态调整上下文感知范围,同步提升了复合词分割的精确率和召回率。在tibetan-news语料中,TDMA的召回率相较于BiLSTM-CRF提升8.5%,验证了该模块对长距离音节依赖关系的解析能力优秀。

此外,TDMA在两个语料集上均表现出良好的稳定性,F1值差异仅为1.38%,而BiLSTM-CRF的差异达到4.38%,表明TDMA具有更好的领域适应性。

3.4.2 消融实验

为了进一步分析各模块对模型性能的影响,还在MLWS 2021语料上进行消融实验,逐步移除关键模块以定量评估各组件对分词性能的影响,如表5所示。由此可知,移除动态多尺度增强后,F1骤降3.66%,紧缩词错误率由8.3%飙升至15.7%,证明了多尺度卷积纵向捕获“基字—加字”组合关系,在解析紧缩结构时的决定性作用。将TiBERT替换为静态嵌入后,F1跌至90.47%,未登录词错误率更从5.2%激增至14.3%,表明预训练语言模型提供的上下文感知表示对消歧多义词与识别稀有词至关重要。去掉注意力机制后,F1小幅下降1.58%,但长句错误率增加2.1%,短句几乎不受影响,印证了该模块主要用于提升模型对长距离依赖的建模与关键信息聚焦能力。

Table 5 Ablation experiment results

表5 消融实验结果 (%)

模型	F1	$\Delta F1$	紧缩词错误率	未登录词错误率
TDMA(完整)	95.81	—	8.3	5.2
无动态增强	92.15	-3.66	15.7	6.1
无注意力机制	94.23	-1.58	9.8	5.9
静态词嵌入	90.47	-5.34	11.2	14.3
BiLSTM-CRF基线	92.03	-3.78	13.5	8.7

为进一步验证TiBERT的有效性,比较了不同预训练模型在相同架构下的表现,如表6所示。由此可见,TiBERT在藏文分词任务上的性能显著优于其他预训练模型。

Table 6 Comparison results of pre-trained models				
表 6 预训练模型比较结果 (%)				
预训练模型	F1	ΔF1	紧缩词错 误率	未登录词 错误率
TiBERT	95.81	—	8.3	5.2
Tibetan-BERT-wwm ^[28]	93.27	-2.54	11.5	7.8
多语言 BERT ^[29]	91.34	-4.47	13.2	12.1

3.4.3 模型效率与性能

为了全面评估 TDMA 模型的实用性,对模型参数量、训练时间及推理速度进行了详细分析,如表 7 所示。由此可知,TDMA 模型总参数量为 4.16 M(4 157 843 个参数),词嵌入层占比 92.4%,动态多尺度增强模块占比 3.6%,BiLSTM 层占比 4.0%。相较于基线模型 BiLSTM-CRF(3.91 百万参数),仅增加了 6.4% 的参数量,但性能提升显著。

Table 7 Performance indicators of TDMA model	
表 7 TDMA 模型的性能指标	
评估指标	数值
总参数量/M	4.16
内存占用/MB	15.86
每轮时间/min	4.17
推理速度/句/s	312

从训练效率方面而言,TDMA 模型在 MLWS 2021 语料中上完成训练的平均耗时为 4.17 min/轮。推理性能表明,模型在批大小为 32 时,处理速度达到 312 句/s,单句平均延迟为 3.21 ms,可满足藏文信息处理系统的实时性要求。

从综合参数量、训练成本和推理效率而言,TDMA 在保持模型性能竞争力的同时,具有较低的部署门槛和优异的计算效率,在当前硬件条件下无需额外地模型轻量化或知识蒸馏处理,即可直接应用于实际生产环境。

3.4.4 案例分析

为了更直观地展示所提模型在实际应用中的效果,选取几个具有代表性的案例进行分析。在紧缩词处理方面,对于案例“ལྷ་པོ་”,本文模型正确切分为“ལྷ་/པོ་”,而 BiLSTM-CRF 错误合并为单字“ལྷ་པོ་”,表明本文模型能更好地处理藏文的紧缩词结构,通过结合上下文信息和注意力机制准确识别紧缩词边界,解决了紧缩词切分歧义的问题。在测试集中选取一些未登录词进行测试时,例如“རྩེ་པོ་”。本文模型能正确识别并切分,而 BiLSTM-CRF 无法正确处理,这得益于预训练模型 TiBERT 强大的语义表示能力和动态多尺度上下文增强模块,可有效捕捉上下文信息,使模型能更好地识别未登录词,提升了模型的泛化能力和适应性。在复杂句子切分方面,对于案例“མི་ལོ་མཆོག་རྩེ་པོ་”,本文模型正确切分为“མི་ལོ་མཆོག་/རྩེ་པོ་”,而 BiLSTM-CRF 存在切分错误,表明本文模型通过结合上下文信息和注意力机制,能准确识别句子中的紧缩词和未登录词,解决了复杂句子的切分问题。在长句处理方面,对于案例“མི་ལོ་མཆོག་རྩེ་པོ་ལྷ་པོ་”,本文模型正确切分为“མི་ལོ་མཆོག་/རྩེ་པོ་/ལྷ་པོ་”,而 BiLSTM-CRF 在部分切分上存在错误,表明本文模型通过注

意力机制优化了长句的处理,能更好地聚焦关键信息,减少切分错误。

综上,上述案例表明本文模型在处理藏文分词任务时优势显著,能解决紧缩词切分、未登录词识别、复杂句子切分和长句处理等问题,提升了模型性能和泛化能力。然而,TDMA 模型在处理专业领域术语时仍存在一定的局限性,例如在古典医学中出现的“རྩེ་པོ་ལྷ་པོ་”(镜面玛瑙花鬘草药)这一复杂复合词,TDMA 错误地切分为“རྩེ་པོ་/ལྷ་པོ་”,未能识别这是描述一种特定藏药材的完整专有名词。同样地,在佛教著作中出现的“འོ་ལྷ་པོ་ལྷ་པོ་”(立论所依境)这一高度专业术语,模型将其过度切分为“འོ་/ལྷ་པོ་/ལྷ་པོ་”,破坏了其整体性。由此可知,TDMA 因缺乏领域知识引导而存在过度切分现象,这些错误主要源于训练语料中专业文本比例不足及模型对领域特定语义结构理解有限。

4 结语

本文针对藏文特性所导致的紧缩词识别、歧义处理和未登录词识别等问题,提出一种融合预训练模型、动态多尺度上下文增强与注意力机制的多策略藏文分词模型 TDMA。通过融合预训练模型、动态多尺度上下文增强与注意力机制,以精准捕捉藏文的深层语义特征、词缀与词根关系及上下文信息。实验表明,该模型在 P、R 和 F1 值上均优于其他基线模型,提升了藏文分词准确率与鲁棒性,为藏文信息处理技术的发展提供了有力支持。

未来,将进一步优化模型结构,探索更多融合策略以进一步提升模型性能:①引入领域自适应词典增强、多任务学习等策略,以进一步提升模型在专业场景下的分词准确性;②将结合藏文词性标注对藏文分词展开深入研究,探索分词与词性标注的联合模型,实现两者协同优化,进一步提升藏文分词的准确率和鲁棒性。

参考文献:

[1] CHEN Y Z, YU S W. Tibetan information processing: past, present, and future[C]// Proceedings of the Second China-Japan Natural Language Processing Joint Research Promotion Conference, 2002: 336-346.
陈玉忠,俞士汶. 藏文信息处理技术的研究现状与展望[C]// 第二届中日自然语言处理专家研讨会, 2002: 336-346.

[2] LIU X T, ZHAO X B. Research of Tibetan automatic word segmentation[J]. Journal of Minzu University of China (Natural Sciences Edition), 2022, 31(2): 63-66.
刘晓彤,赵小兵. 藏文自动分词技术研究[J]. 中央民族大学学报(自然科学版), 2022, 31(2): 63-66.

[3] ZHAXI C R. The design of a human computer assisted Tibetan word segmentation and word login system[C]// Collection of Modern Chinese Minority Languages, 1999: 322-327.
扎西次仁. 一个人机互助的藏文分词和词登录系统的设计[C]// 中国少数民族语言文字现代化文集, 1999: 322-327.

[4] CHEN Y Z, LI B L, YU S W. The design and implementation of a Tibetan

- word segmentation system[J]. Journal of Chinese Information Processing, 2003, 17(3): 16-21, 66.
- 陈玉忠,李保利,俞士汶. 藏文自动分词系统的设计与实现[J]. 中文信息学报, 2003, 17(3): 16-21, 66.
- [5] CAI Z J. Identification of abbreviated word in Tibetan word segmentation[J]. Journal of Chinese Information Processing, 2009, 23(1): 35-37, 43.
- 才智杰. 藏文自动分词系统中紧缩词的识别[J]. 中文信息学报, 2009, 23(1): 35-37, 43.
- [6] SUN Y, LUOSANG Q B, YANG R, et al. Research on segmentation method of Tibetan intersectional ambiguity field[C]// Lhasa: 12th National Conference on Minority Languages Information Processing, 2009.
- 孙媛,罗桑强巴,杨锐,等. 藏语交集型歧义字段切分方法研究[C]// 拉萨:第十二届全国少数民族语言文字信息处理学术研讨会, 2009.
- [7] SHI X D, LU Y J. A Tibetan segmentation system—Yangjin[J]. Journal of Chinese Information Processing, 2011, 25(4): 54-56.
- 史晓东,卢亚军. 央金藏文分词系统[J]. 中文信息学报, 2011, 25(4): 54-56.
- [8] LI Y C, JIA Y J, ZONG C Q, et al. Research and implementation of Tibetan automatic word segmentation based on conditional random field[J]. Journal of Chinese Information Processing, 2013, 27(4): 52-58.
- 李亚超,加羊吉,宗成庆,等. 基于条件随机场的藏语自动分词方法研究与实现[J]. 中文信息学报, 2013, 27(4): 52-58.
- [9] LI Y C, JIANG J, JIA Y J, et al. TIP-LAS: an open source toolkit for Tibetan word segmentation and POS tagging[J]. Journal of Chinese Information Processing, 2015, 29(6): 203-207.
- 李亚超,江静,加羊吉,等. TIP-LAS: 一个开源的藏文分词词性标注系统[J]. 中文信息学报, 2015, 29(6): 203-207.
- [10] LI B H, LIU H D, LONG C J, et al. Tibetan word segmentation based on deep learning[J]. Computer Engineering and Design, 2018, 39(1): 194-198.
- 李博涵,刘汇丹,龙从军,等. 基于深度学习的藏文分词方法[J]. 计算机工程与设计, 2018, 39(1): 194-198.
- [11] WANG L L, WANG H Y, BAIMA Q Z, et al. Tibetan word segmentation method based on BiLSTM-CRF model[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2020, 32(4): 648-654.
- 王莉莉,王宏渊,白玛曲珍,等. 基于BiLSTM-CRF模型的藏文分词方法[J]. 重庆邮电大学学报(自然科学版), 2020, 32(4): 648-654.
- [12] WANG K. Research on word segmentation and part-of-speech of Tibetan on neural network[D]. Lanzhou: Lanzhou University, 2020.
- 王康. 基于神经网络的藏语分词与词性标注研究[D]. 兰州:兰州大学, 2020.
- [13] YU Y B, LU R J, NIMA Z X, et al. Tibetan word segmentation model based on LSTM and CRF[J]. Technology Intelligence Engineering, 2023, 9(2): 108-116.
- 于永斌,陆瑞军,尼玛扎西,等. 基于长短时记忆和条件随机场藏文分词模型[J]. 情报工程, 2023, 9(2): 108-116.
- [14] SUOLANG C R. Research on Tibetan word segmentation and part-of-speech tagging based on pre-trained language models[D]. Lhasa: Tibet University, 2023.
- 索朗次仁. 基于预训练语言模型的藏文分词与词性标注研究[D]. 拉萨:西藏大学, 2023.
- [15] YIN Z H, NIMA C R, YU T, et al. Research on Tibetan automatic word segmentation method based on ASBC model[J]. Computer & Digital Engineering, 2023, 51(6): 1227-1230, 1237.
- 尹宗鹤,尼玛次仁,于韬,等. 基于ASBC模型的藏文自动分词方法研究[J]. 计算机与数字工程, 2023, 51(6): 1227-1230, 1237.
- [16] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite bert for self-supervised learning of language representations[DB/OL]. <https://arxiv.org/abs/1909.11942>.
- [17] SECHIA J, SANGJIE D Z, CAI R J, et al. A Tibetan word segmentation method based on pre-training model[J]. Journal of Chinese Information Processing, 2023, 37(12): 70-75.
- 色差甲,桑杰瑞珠,才让加,等. 一种基于预训练模型的藏文分词方法[J]. 中文信息学报, 2023, 37(12): 70-75.
- [18] YANG J, NIMA Z X, RENQING D Z, et al. Tibetan word segmentation system based on pre-trained model tokenization reconstruction[J]. Journal of Computer Applications, 2025, 45(4): 1199-1204.
- 杨杰,尼玛扎西,仁青东主,等. 基于预训练模型标记器重构的藏文分词系统[J]. 计算机应用, 2025, 45(4): 1199-1204.
- [19] GESANG J C, ABUDUREXITI R H M, NIMA Z X, et al. Research on Tibetan word segmentation method combining Bi-LSTM and CRF[J]. Journal of Minzu University of China (Natural Sciences Edition), 2024, 33(3): 40-46.
- 格桑加措,阿卜杜热西提·热合曼,尼玛扎西,等. Bi-LSTM和CRF结合的藏文分词方法研究[J]. 中央民族大学学报(自然科学版), 2024, 33(3): 40-46.
- [20] WANG Y R. Research on Tibetan word segmentation and part-of-speech tagging based on GNN[D]. Lanzhou: Lanzhou University, 2023.
- 王怡如. 基于GNN的藏文分词和词性标注研究[D]. 兰州:兰州大学, 2023.
- [21] CHANG F Y, CAI Z J. An eight-word-position tag for Tibetan word segmentation via BiLSTM-CRF[J]. Journal of Chinese Information Processing, 2024, 38(10): 64-70, 79.
- 常芳玉,才智杰. 一种基于八词位标签的BiLSTM-CRF藏文分词方法[J]. 中文信息学报, 2024, 38(10): 64-70, 79.
- [22] LIU S, DENG J, SUN Y, et al. TiBERT: Tibetan pretrained language model[C]// IEEE International Conference on Systems, Man, and Cybernetics, 2022: 2956-2961.
- [23] SIAMI N S, TAVAKOLI N, NAMIN A S. The performance of LSTM and BiLSTM in forecasting time series[C]// IEEE International Conference on Big Data, 2019: 3285-3292.
- [24] JOHN D L, ANDREW M, FERNANDO C N. Conditional random fields: probabilistic mode-LS for segmenting and labeling sequence data[C]// 18th International Conference on Machine Learning, 2001: 282-289.
- [25] ZHAO X B, GAO L, GAO D G, et al. A dataset of word segmentation technology evaluation for minority languages (MLWS2021)[J]. China Scientific Data, 2022, 7(2): 2-10.
- 赵小兵,高璐,高定国,等. 少数民族语言分词技术评测数据集MLWS2021[J]. 中国科学数据(中英文网络版), 2022, 7(2): 2-10.
- [26] ZHANG J Y, KAZHUO D J, GADENG L S, et al. Research and application of Tibetan pretraining language model based on BERT[C]// 2nd International Conference on Control and Intelligent Robotics, 2022: 519-524.
- [27] LYU H, PU C, DUO L, et al. T-LLaMA: a Tibetan large language model based on LLaMA2[J]. Complex & Intelligent Systems, 2025, 11(1): 72.
- [28] LIANG Y, LYU H, LI Y, et al. Tibetan-BERT-wm: a Tibetan pre-trained model with whole word masking for text classification[J]. IEEE Transactions on Computational Social Systems, 2024, 11(5): 6268-6277.
- [29] DEVLIN J, CHANG M W, LEE K T, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.