文章编号：1003-0077(2023)04-0052-11

融合多粒度特征的老挝语词性标注研究

唐 文[1]，周兰江[1]，张建安[2]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 战略支援部队 信息工程大学三院昆明大队，云南 昆明 650500)

摘 要：词性标注是自然语言处理领域的基础任务之一。语料稀缺、词形复杂、存在大量低频词和未登录词，句式较长，在数据传递过程中信息易丢失，这些都是导致老挝语词性标注不准确的主要原因。因此，该文提出一种融合多粒度特征的老挝语词性标注方法，构建了融合老挝词、字符和音节特征的 Transformer-CRF 模型。首先，在传统词向量的基础上融合老挝语字符和音节特征向量，使模型在三个粒度级别上充分利用语料信息；其次，使用 Transformer 对老挝语句子进行长远上下文信息提取，解决重要信息丢失问题；最后，使用 CRF 提取相邻词性约束关系，从而获取最优词性标签。实验结果表明，在语料有限的情况下，该模型与其他主流模型相比达到了更显著的效果，精确率、召回率和 F_1 值分别为 94.76%、93.93%、94.34%。

关键词：多粒度；老挝语；词性标注；Transformer

中图分类号：TP391 **文献标识码**：A

Multi-granularity Feature Fusion for Part-of-speech Tagging of Lao Language

TANG Wen[1], ZHOU Lanjiang[1], ZHANG Jian'an[2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Kunming Branch, No.3 College, PLA Information Engineering University, Kunming, Yunnan 650500, China)

Abstract: Part-of-speech tagging in Lao language is challenged by the lack of corpus, the complexity of word form, the large number of low-frequency words and out-of-vocabulary(OOV) words, and the long distance dependency, etc. This paper proposes a multi-granularity feature fusion approach via Transformer-CRF model, integrating the features of Lao words, characters and syllables. Firstly, it integrates the feature vectors of the Laos language characters and syllables based on the classical word vector. Secondly, it uses Transformer to extract the long-range context information of Lao sentences. Finally, it uses CRF to extract the constraints of adjacent parts-of-speech to determine the optimal tags. The experimental results show that, compared with other mainstream models, this model achieves much better accuracy rate, recall rate and F_1 value, reaching 94.64%, 93.82% and 94.23%, respectively.

Keywords: multi-granularity; Lao; part-of-speech tagging; Transformer

0 引言

词性标注(part-of-speech tagging)是指通过抽取上下文信息为句子中的每一个词标注一个正确词性，即确定每个词是名词、动词、形容词或者其他词性的过程，在句法分析、信息抽取、机器翻译等领域有重要的研究意义。

目前，老挝语词性标注研究存在以下挑战：

(1) 老挝语属于东南亚低资源语种，从巴利语、梵语、高棉语、泰语借入大量多音节词，还从汉语、法语、英语、越南语等借入少量词汇[1]。因此，老挝语组成成分复杂，且老挝词形态结构丰富，导致老挝语词性标注工作面临大量未登录词和低频词的问题。

(2) 老挝语中通常有多个定语修饰中心词，且

收稿日期：2020-08-02 **定稿日期**：2020-11-09

基金项目：国家自然科学基金(61662040)

语料中老挝语句式普遍过长，导致模型在进行词性标注任务时存在长远上下文信息丢失的问题。

传统的词性标注研究方法主要使用传统机器学习方法[2-3]，该方法往往需要大规模的语料数据和人工抽取大量的特征。近年来，基于神经网络模型的方法取得更好的效果[4-5]，但针对语料资源稀缺、形态复杂的语言，使用传统神经网络模型进行词性标注的效果不佳。

综上，本文在对老挝词和老挝句子研究的基础上，提出一种融合多粒度特征的老挝语词性标注方法，构建了融合老挝词、字符和音节特征的 Transformer-CRF 模型。首先，将字符向量和音节向量输入 CNN 中，自动获取含有丰富老挝词信息的字符词特征向量和音节词特征向量；其次，将字符词特征向量、音节词特征向量与预训练的词向量进行线性拼接，获得老挝词特征向量；然后，将老挝词特征向量输入 Transformer 层，得到老挝句式语义特征；最后，将 Transformer 层的输出作为 CRF 层的输入，获取相邻词性约束关系，从而获取最优的老挝词性标签。为验证本文所提方法对老挝语词性识别性能的提升，本文与其他 5 种主流模型进行实验对比。实验结果表明，在同一老挝语料集下，本文方法与主流方法相比，在老挝语词性标注任务中取得最显著的效果，评估指标均优于其他主流模型。其中，本文模型与当前 BiLSTM-Attention-CRF 模型（老挝语词性标注当前最优）对比，本文模型 P、R、F_1 值分别提升 1.9%、2.39%和 2.15%。另外，本文还利用不同粒度特征对模型的词性标注效果进行对比。实验结果表明，融合词、字符和音节特征能有效提升模型词性识别性能，尤其对未登录词和低频词的词性识别提升效果最为明显。

本文的主要贡献如下：

（1）提出一种多粒度老挝语词向量表示方法。首先，对老挝语的词、字符、音节进行分布式表示；然后，通过 CNN 神经网络获取老挝字符级词特征向量和音节级词特征向量；最后，将其与老挝语词向量线性拼接获取老挝语词特征向量。通过输入多粒度的老挝语词信息，丰富模型对老挝语词信息的提取，提升模型对老挝语词性标注的准确性。

（2）使用 Transformer-CRF 模型，解决老挝语句式过长导致上下文信息丢失的问题，有效提高老挝语词性标注模型的性能。

本文组织结构如下：第 1 节相关工作，综述词性标注相关文献；第 2 节给出老挝语字符和音节特征；第 3 节介绍模型结构；第 4 节介绍具体实验和详细分析；第 5 节进行总结。

1　相关工作

传统的词性标注研究方法主要使用机器学习方法。例如，CHANG 等人[6]利用隐马尔可夫模型（Hidden Markov Model，HMM）进行词性标注任务研究，虽然他们的研究取得一定效果，但是 HMM 模型[7]存在输出独立性假设问题，很难利用观测序列中的非独立特征（单词前后缀、周围单词、字符类型等特征）。Andrew 等人[8]将最大熵马尔可夫模型（Maximum Entropy Markov Model，MEMM）应用在词性标注任务中，虽能解决输出独立性假设的问题，但存在标注偏置的问题，容易陷入局部最优。Phuong 等人[9]利用条件随机场（Conditional Random Fields，CRF）进行词性标注研究，虽然缓解了标注偏置问题[10]，但需要语言学专家定制复杂特征模板。由于神经网络模型能自动提取自然语言特征，研究者使用基于神经网络的方法进行词性标注研究，旨在提取语言的更深层次信息，提高词性标注准确率。

Huang 等人[11]使用 BiLSTM-CRF 模型，有效提取输入前后特征信息，提升模型序列标注性能（分词、词性标注、命名实体识别）。Ma 等人[12]在 BiLSTM-CRF模型的基础上，利用 CNN 神经网络自动获取字符级词特征向量，并将其与预训练的词向量联合输入 BiLSTM-CRF 网络中进行词性标注。该方法通过对字符特征的提取，提升模型对未登录词的词性识别效果。尽管如此，上述方法仍存在缺点，在反向传播过程中通常存在梯度消失和梯度爆炸的问题[13]，导致长远上下文信息丢失；且很难实现并行化。而注意力（Attention）机制可以考虑句子中的每个词对待标注词的词性影响，根据影响的大小分配不同权值。因此，Wu 等人[14]提出基于自注意力机制的 BiLSTM-CRF 模型，该模型虽然取得较好的成果，但是并行能力不佳，且长远信息提取能力有限。Vaswani 等人[15]提出 Transformer 模型，该模型只使用基于自注意力机制的全连接神经网络，解决了并行计算困难和长远上下文信息丢失的问题[15-17]。由于 Transformer 固有体系结构的优越性，在许多序列标注任务中表现出显著优势。Yan 等人[18]在命名实体识别任务中成功利用 Transformer-CRF 模型，并取得了最好的效果。Qiu 等人[19]利用 Transformer-CRF 模型在 8 个不同的数

据集上进行测试,实验结果表明,模型可以有效提升分词任务表现。

英语、汉语等高资源语言的词性标注研究较为成熟,尽管公布了很多有效的词性标注研究工作,但面向老挝语的词性标注研究屈指可数。Yang 等人[20]利用少量标注语料和未标注语料训练 HMM 模型,并使用维特比算法进行解码,实现老挝语词性标注研究;王兴金等人[21]在 HMM 模型的基础上结合老挝语规则的方法,对老挝语进行词性标注;王兴金等人[22]还利用深度学习的方法,在有限的老挝语语料的情况下有效识别稀疏词的词性。尽管他们取得一定效果,但是还存在并行能力欠缺,长远信息易丢失词特征提取不充分的问题。

2 老挝语字符和音节特征

2.1 老挝字符

老挝语存在相当数量的外来词,语言组成成分复杂,未登录词和低频词较多。老挝语的词拥有丰富的形态结构,且词缀信息能改变老挝语词性,如动词加前缀ການ或ຄວາມ转变为名词词性等,通过对老挝语字符特征的提取可以推断出单词的表示形式,获取老挝语复杂的词缀信息,可以有效解决难以识别未登录词的词性问题。

老挝语是一种拼音文字,它的字母由 65 个字符组成,字母表如表 1 所示。其中老挝语元音共有 28 个,包含 18 个单元音、6 个复合元音和 4 个特殊元音;辅音有 33 个,包含中辅音 8 个、高辅音 12 个、低辅音 13 个;声调符号 4 个[1]。

表 1 老挝字母表

老挝字母	Detail
28 元音	◌ະ ◌າ ◌ິ ◌ີ ◌ຶ ◌ື ◌ຸ ◌ູ ເ◌ະ ເ◌ ແ◌ະ ແ◌ ໂ◌ະ ໂ◌ ເ◌າະ ◌ໍ ເ◌ິ ເ◌ີ ເ◌ັຍ ເ◌ືອ ◌ົວະ ເ◌ຍ ເ◌ືອ ◌ົວ ໄ◌ ໃ◌ ເ◌ົາ ◌ຳ
33 辅音	ກ ຈ ດ ຕ ບ ປ ຢ ອ ຄ ງ ຊ ຍ ທ ນ ພ ຟ ມ ລ ວ ຮວ ຮ ຂ ສ ຖ ຫ ຜ ຝ ໜ ໝ ຫ ຫວ ຫງ ຫຍ
4 声调符号	◌່ ◌້ ◌໊ ◌໋

2.2 老挝音节

ISO/IEC10646 为老挝文共分配了 128 个码位,编码范围在 0x0E80～0x0EFF 之间。其中收录了 82 个老挝文字符,包括老挝语的辅音、元音、声调符号、数字及各类符号等[23]。老挝文字符、字形及编码如表 2 所示。

一个老挝语音节通常由主辅音、元音、尾辅音和声调符号构成,其中尾辅音和声调符号的个数可为 0。

表 2 国际标准的老挝文字符、字形及编码

	U+0E8x	U+0E9x	U+0EAx	U+0EBx	U+0ECx	U+0EDx	U+0EEx	U+0Ex
0	-			ະ	ເ	໐		
1	ກ		ມ	◌ັ	ແ	໑		
2	ຂ		ຢ	າ	ໂ	໒		
3	-		ຣ	◌ຳ	ໃ	໓		
4	ຄ	ດ	-	◌ິ	ໄ	໔		
5	-	ຕ	ລ	◌ີ	-	໕		
6		ຖ	-	◌ຶ	ໆ	໖		
7	ງ	ທ	ວ	◌ື		໗		
8	ຈ			◌ຸ	◌່	໘		
9		ນ		◌ູ	◌້	໙		
A	ຊ	ບ	ສ		◌໊	-		
B	-	ປ	ຫ	◌ົ	◌໋	-		
C		ຜ		◌ຼ	◌໌	ໜ		
D	ຍ	ຝ	ອ	ຽ	◌ໍ	ໝ		
E		ພ	ຮ	-	-	ໞ		
F		ຟ	ຯ	-	-	ໟ		

老挝语音节字符的编码序列按照 ISO/IEC10646 编码体系中的老挝文字符编码序列。其中，老挝文中无论元音出现在主辅音之前还是之后，ISO/IEC10646 编码中主辅音的编码顺序总是在元音之前，因此，尽管部分老挝语元音符号出现在主辅音符号之前，但是计算机输入时应先输入主辅音符号。如输入老挝语音节“ໂປ”时应先输入辅音符号“ປ”(U+0E9B)，然后再输入“ໂ◌”(U+0EC2)元音辅号。

老挝语音节的具体拼写规则有辅音与元音组合添加死闭音节、活闭音节以及活闭音节和声调符号，其中需注意的有辅音与特殊元音组合时不能添加尾辅音，辅音与单元音、复合元音组合添加死闭音节时不能添加声调符号等规则。

音节是语音结构的基本单位，老挝语的单词由单音节的单纯词和多音节的连绵词组成[1]，如单词ລັດຖະບານ由[ລັດ ຖະ ບານ]三个音节构成，以音节为标注基本单位符合老挝语的语言特点，可以通过音节在词中的位置特性来获取词的内部结构信息，可以有效提升识别的准确性。此外，字符构成音节，音节构成单词，音节介于字符和单词之间，数量远远超过字符数量，使用音节粒度对于低频词的识别效果有较大地提升。如表 3 所示的例句为将句子按照单词、音节和字符三个不同粒度进行切分。

表 3　句子不同粒度切分示例表

句子	ລັດຖະບານໂປໂລຍໄດ້ເພີ່ມໃນເຮືອນແລະເພີ່ມລາ ຄາ
单词	ລັດຖະບານ ໂປໂລຍ ໄດ້ ເພີ່ມ ໃນ ເຮືອນ ແລະ ເພີ່ ມ ລາຄາ
音节	ລັດ ຖະ ບານ ໂປ ໂລຍ ໄດ້ ເພີ່ມ ໃນ ເຮືອນ
老挝字符	◌ັ ລ ດ ຖ ◌ະ ບ ◌າ ນ ໂ◌ ປ ລ ຍ ໄ◌ ດ ◌້ ◌່ ເ◌ີ ພ ມ ໃ◌ ເ◌ືອ ຮ ນ ແ◌ະ ເ◌ີ ຄ

3　基于 Transformer-CRF 的词性标注模型

3.1　模型结构

本文构建一个基于老挝语的单词、字符和音节多粒度语义信息的网络模型，该模型由多特征嵌入层 Enbedding、Transformer 层和 CRF 层组成，共同进行词性标注。模型结构如图 1 所示。

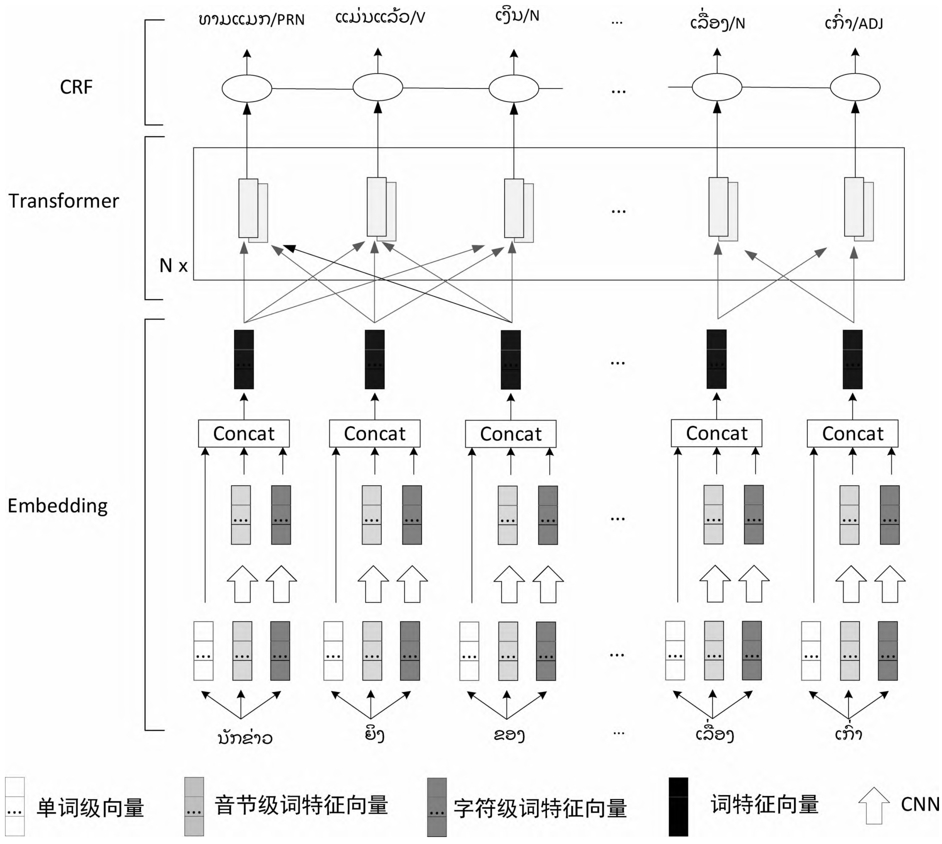

图 1　融合多粒度特征的老挝语词性标注模型结构图

3.2 多粒度特征嵌入层

对已分词的老挝语句子进行词性标注前，需要将老挝词进行分布式表示后再输入模型，且分布式表示的质量对词性标注效果有很大影响。本文采用Glove模型[24]引入共现矩阵计算语词向量，生成具有全局统计信息的词向量矩阵。由于词向量难以获取单词之间形态结构相似性和词内部的结构特征，为了获取更加丰富的老挝语单词语义信息，本文在词向量的基础上融合字符特征向量和音节特征向量，每个词向量由三个向量拼接组成：词向量、字符向量和音节向量，词向量捕获句法和语义信息，字符向量捕获词形态信息，音节向量捕获词内部结构信息，使模型在三个粒度级别上充分利用语料信息。

为更好地解决语料中低频词和未登陆词的词性标注问题，本文需要提取字符和音节特征向量。Labeau和Wang等人[25-26]在词性标注任务中成功使用CNN和Bilstm提取字符特征，实验表明，CNN和BiLSTM的性能相差不大，且CNN还减少了计算成本。为此，本文选择使用CNN提取字符特征，音节特征的提取采用相同的方法，具体过程如图2所示。

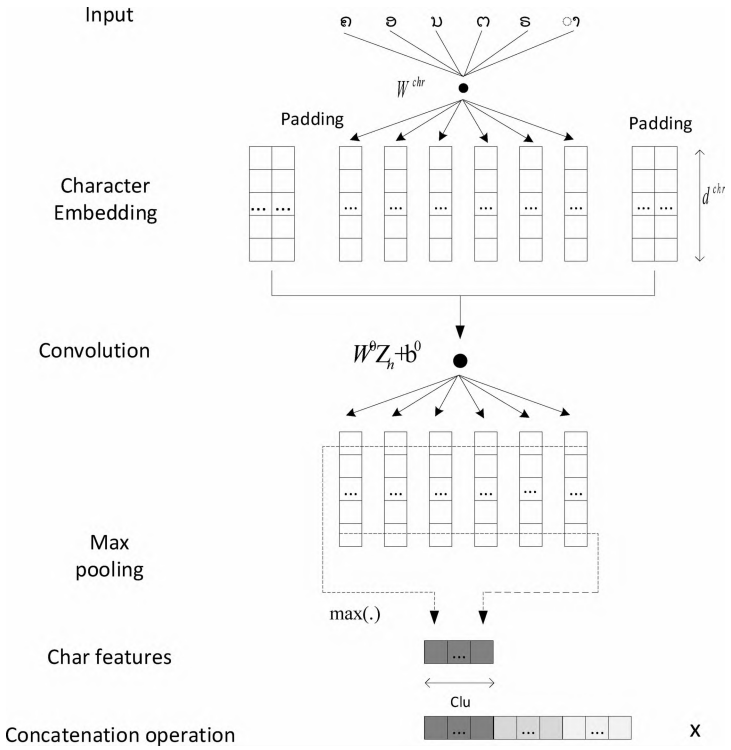

图2 卷积神经网络提取字符特征图

给定一个由 N 个字符$\{c_1,c_2,\cdots,c_N\}$组成的单词 W，使用Word2vec[27]为矩阵向量转换工具，将每个老挝字符转换为字符向量$\{\boldsymbol{r}_1^{\text{chr}},\boldsymbol{r}_2^{\text{chr}},\cdots,\boldsymbol{r}_n^{\text{chr}}\}$，每个老挝字符 c 的向量 $\boldsymbol{r}^{\text{chr}}$ 由矩阵向量相乘得到，如式(1)所示。

$$\boldsymbol{r}^{\text{chr}}=\boldsymbol{W}^{\text{chr}}v^{c} \tag{1}$$

其中，$\boldsymbol{W}^{\text{chr}}$为字符向量矩阵，$v^{c}$ 在索引为 c 的位置为1，其他位置为0。在得到字符向量$\{\boldsymbol{r}_1^{\text{chr}},\boldsymbol{r}_2^{\text{chr}},\cdots,\boldsymbol{r}_n^{\text{chr}}\}$，后，将其作为卷积层的输入，卷积层以 $\boldsymbol{Z}_n\in\mathbb{R}^{d^{\text{chr}}k^{\text{chr}}}$为卷积核在输入序列中，每次选择大小为 k^{chr} 的窗口提取字符局部特征，如式(2)、式(3)所示。

$$\boldsymbol{Z}_n=(\boldsymbol{r}_{n-(k^{chr}-1)/2}^{\text{chr}},\cdots,\boldsymbol{r}_{n+(k^{chr}-1)/2}^{\text{chr}})^{T} \tag{2}$$

$$\boldsymbol{r}_x=\boldsymbol{W}^{0}\boldsymbol{Z}_n+\boldsymbol{b}^{0} \tag{3}$$

其中，$\boldsymbol{W}^{0}$ 是卷积层的权重矩阵，$\boldsymbol{b}^{0}$ 是偏置项。为了降低计算复杂度，提高特征提取的准确性，采用“MAX POOLING”操作，即取每个字符局部特征中最大值，舍弃其他值，提取一个固定大小的由字符特征向量组成的词特征向量。Max pooling操作如式(4)所示。

$$\tilde{\boldsymbol{r}}_x=\max\{\boldsymbol{r}_x\} \tag{4}$$

最后，将字符词特征向量、音节词特征向量与词向量拼接组成词特征向量。如式(5)所示。

$$\boldsymbol{x}=\tilde{\boldsymbol{r}}_1\oplus\cdots\oplus\tilde{\boldsymbol{r}}_{\text{clu}}\oplus\tilde{\boldsymbol{v}}_1\oplus\cdots\tilde{\boldsymbol{v}}_{\text{clu}}\oplus\tilde{\boldsymbol{w}} \tag{5}$$

其中，$\oplus$表示拼接操作，clu代表卷积核个数，$\tilde{\boldsymbol{v}}_x$ 代表音节特征向量。经过上述操作，老挝语句

子中的每个单词$[w_1,w_2,\cdots,w_n]$都得到对应的词特征向量$[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}n]$，第 i 个词特征向量 $\boldsymbol{x}_i$ 包含老挝词的形态结构、内部结构和语义特征。

3.3　Transformer 模型

为解决老挝句式过长导致长远上下文信息丢失问题，本文使用完全基于自注意力机制的 Transformer 模型。该模型本质是编码器/解码器(Encoder/Decoder)结构。Transformer 结构如图 3 所示。

对于给定的待标注老挝语句子 $s=[w_1,w_2,\cdots,w_n]$，其中，w_i 表示句子中第 i 个单词。首先，将老挝词特征向量与位置编码相加，得到添加位置信息后的老挝词特征，向量记为 $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$。其次，编码器将输入序列 $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ 经过 Self-attention 层、前馈神经网络层、残差连接、归一化处理后映射为连续序列 $\boldsymbol{z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n\}$。最后，解码器对于任何一个给定的 z，在经过 n 个时间步之后，生成输出序列 $\boldsymbol{y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$。其中，在每一个时间步中，模型都会将之前的输出作为额外的输入，共同生成下一步序列的输出。

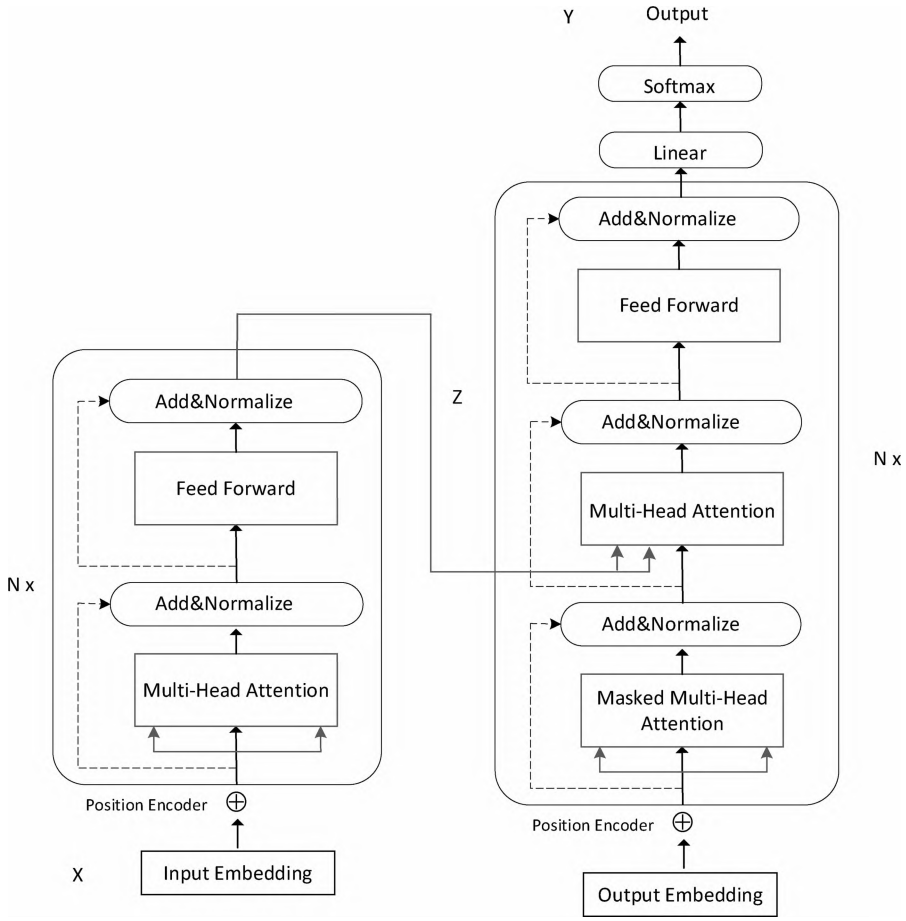

图 3　Transformer 结构图

3.3.1　位置编码

由于自注意力机制不能意识到不同序列的位置，使其无法捕捉句式的顺序特征。为了解决这个问题，使用由频率变化的正弦波产生的位置编码，如式(6)、式(7)所示。

$$\mathrm{PE}_{t,2i}=\sin(t/10000^{2i/d}) \tag{6}$$

$$PE_{\mathrm{t},2\mathrm{i}+1}=cos(\mathrm{t}/10000^{2\mathrm{i}/\mathrm{d}}) \tag{7}$$

其中，i 表示位置嵌入的维数指数，d 表示嵌入向量的维数。然后将词特征向量与位置编码相加，得到添加位置信息后的词特征，向量记为 $\boldsymbol{x}$，如式(8)所示。

$$\boldsymbol{x}=[\boldsymbol{x}_1+\mathrm{PE}_1;\boldsymbol{x}_2+\mathrm{PE}_2;\cdots;\boldsymbol{x}_n+\mathrm{PE}_n] \tag{8}$$

3.3.2　自注意力机制

自注意力机制可以对老挝句子中的不同词之间的关系进行注意力计算，从而学习句子内部词之间的依赖关系，捕获句式内部的结构信息，且无须考虑词之间的位置关系，因此可以更好地处理长远信息丢失问题。

自注意力机制的输入是添加位置编码后的词特征向量 $\boldsymbol{x}$。首先，$\boldsymbol{x}$ 经过一个线性变换后得到查询

$\boldsymbol{Q}$，$\boldsymbol{x}$ 经过第二个线性变换得到键 $\boldsymbol{K}$，经过第三次线性变换得到值 $\boldsymbol{V}$；其次，将 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 这三个向量进行缩放点积注意力计算；最后，得到注意力得分，如式(9)所示。

$$\boldsymbol{Z}=\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{9}$$

为了增强自注意力的表现效果，我们使用多头自注意力来增加训练参数，防止过拟合；同时允许模型从不同的子空间学习相关信息，加强注意力机制的表现能力。其中，多头注意力的本质是多个独立的注意力计算。使用不同的线性变换得到多个 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 向量之后并行执行注意力操作，生成多个输出值 head_i，且这些值之间相互独立；将这些值进行拼接和线性变换，生成输出值 $\boldsymbol{Z}_i$，如式(10)、式(11)所示。

$$\begin{aligned}\boldsymbol{Z}_i&=\text{Multi}\boldsymbol{H}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})\\&=\text{Concat}(\text{head}_1,\cdots,\text{head}_h)\boldsymbol{W}^0\end{aligned} \tag{10}$$

$$\text{head}_1=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{11}$$

其中，h 为多头数，$\boldsymbol{W}_i^Q$，$\boldsymbol{W}_i^K$，$\boldsymbol{W}_i^V$，$\boldsymbol{W}^0$ 为参数矩阵。

3.3.3 前馈神经网络

多头注意力的输出输入前馈神经网络进行进一步处理，如式(12)所示。

$$\text{FFN}(\boldsymbol{x})=\max(0,\boldsymbol{x}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{12}$$

其中，$\boldsymbol{W}_1$，$\boldsymbol{W}_2$，$\boldsymbol{b}_1$，$\boldsymbol{b}_2$ 均为可学习参数。

3.4 CRF层

在老挝语中，相邻词性之间存在一定的约束关系，如在“ພວມ、ໄດ້、ຈະ、ສາມາດ”后接动词、量词不能重复出现等规则。使用CRF可以有效利用不同标签之间的依赖关系。本模型中CRF的输入为Transformer的输出序列 $y=\{y_1,y_2,\cdots,y_n\}$，其预测标签序列为 $Y=\{Y_1,Y_2,\cdots,Y_n\}$，则状态序列线性链CRF的条件概率可定义如式(13)所示。

$$P(Y_1,\cdots Y_n\mid y,\lambda)=1/Z(y)\times\exp\left(\sum_j\lambda_j t_j(Y_{i-1},Y_i,X,i)+\sum_k\mu_k s_k(Y_i,X,i)\right) \tag{13}$$

其中，$Z(y)$ 表示归一化因子(Normalization Factor)，可使所有可能状态序列概率之和为1，可由式(14)计算得到。$t_j(Y_{i-1},Y_i,X,i)$ 表示整个观察序列的转移特征函数，Y_i 表示标签的标注序列；$s_k(Y_i,X,i)$ 是观察序列 i，X 的状态特征函数；λ_i 和 μ_k 分别是 t_j 和 s_k 对应的权重参数。

$$Z(y)=\sum_Y exp\left(\sum_j\lambda_j t_j(Y_{i-1},Y_i,X,i)+\sum_k\mu_k s_k(Y_i,X,i)\right) \tag{14}$$

4 实验及分析

4.1 实验语料与参数配置

4.1.1 实验数据与模型设置

本文实验采用的老挝语语料从维基百科老挝语版上爬取获得，对其进行降噪处理和过滤，最终得到2 495条句子。实验数据如表4所示，其中，将已分好词的训练集和测试集语料与老挝语词典(共31 719词)进行匹配，若在词典中匹配到对应的老挝词，便匹配成功，若未在词典中匹配到对应的词则为未登录词，根据未出现的词在训练集和测试集语料中所占比例获得未登录词比例。

表4 实验数据

数据	句子数	单词数	未登录词比例/%	低频词比例/%
训练集	2 120	53 015	15.29	—
测试集	375	8 250	10.6	9.3

根据老挝语语言特点将老挝语词性标注集共分为21大类，将大类再细分共为31种词性，由老挝语语言专家进行人工分词与标注，老挝语词性标注集如表5所示，标注样例如图4所示。

表5 老挝语词性标注集

类别	标记	说明	类别	标记	说明
名词	n	普通名词等	感叹词	y	感叹词
动词	v	判断、存在动词等	缩略词	j	缩略词
形容词	a	状态、性质形容词	方位词	f	方位词
副词	d	程度、时间副词等	终结词	e	终结词

续表

类别	标记	说明	类别	标记	说明
代词	r	人称、指示代词等	时间词	t	时间词
介词	p	排除、对象介词等	成语词	i	成语词
数词	m	基数词、序列数词等	习惯词	I	习惯语
量词	q	名量词、动量词等	语素	g	语素
助词	u	语气、疑问助词等	非语素	k	非语素
连词	c	递进、转折连词等	标点符	w	标点符号
拟声词	o	拟声词			

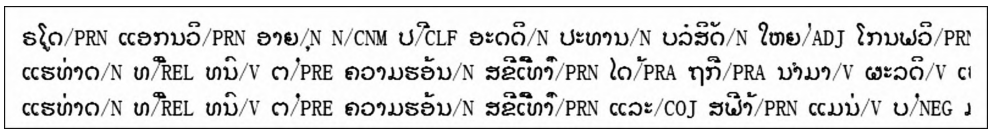

图 4　人工标注老挝语词性样例

本实验采用 Python 语言及 Tensorflow 框架。由于数据集较少，本文只使用一层 Transformer，经多次调整模型参数如表 6 所示。

表 6　模型参数配置

Parameters	Value	说明
num _identical	6	Attention 层数
num _heads	8	多头数目
dmodel _dimension	512	嵌入编码维度
batch _size	32	最小批处理尺寸
hidden _units	512	隐藏层维数
Dropout rate	0.3	丢弃率，防过拟合
Learning rate	0.000 1	学习率

4.1.2　评价指标

本实验使用 P 精确率(Precision)，R 召回率(Recall)、F_1 值作为指标来综合评估模型词性标注的效果。P、R、F_1 值的具体计算如式(15)～式(17)所示。

$$P=\frac{\text{正确识别老挝语词性个数}}{\text{识别出来的老挝语词性个数}}\times 100\% \quad (15)$$

$$R=\frac{\text{正确识别老挝语词性个数}}{\text{语料中老挝语词性个数}}\times 100\% \quad (16)$$

$$F_1=\frac{(2\times P\times R)}{P+R}\times 100\% \quad (17)$$

4.2　模型对比实验

本文使用 Transformer-CRF 模型，在此基础上融合老挝字符和音节特征来丰富词的形态结构和内部结构信息。为验证 Transformer-CRF 模型对老挝语词性标注的有效性，在同一语料集下，与其他 5 种主流的词性标注模型进行比较分析，结果如表 7 所示。

表 7　本文模型与主流模型实验结果对比

（单位：%）

序号	Model	P	R	F_1
1	CRF	90.34	89.62	89.97
2	CNN-CRF	91.54	90.18	90.85
3	Bilstm-CRF	91.97	91.23	91.59
4	Bilstm-Attention-CRF	92.86	91.54	92.19
5	Transformer-CRF	93.73	92.68	93.20
6	Our	94.76	93.93	94.34

由表 7 可知，本文模型的 P、R、F_1 值均超过所有主流模型，F_1 值最大提升为 4.37%，充分证明主流模型在老挝语词性预测效果略有不足，本文模型对老挝语词性预测性能实现了有效的改进。实验 1 和实验 2 相比较，说明 CNN 神经网络在提取特征上的有效性。实验 3 和实验 4 相比较，说明 attention 机制在词性标注任务上的有效性。实验 3 和实验 5 相比较，说明了 Transformer-CRF 模型比 BiLSTM-CRF 模型在词性标注任务上更具有优势。本文模型与当前 BiLSTM-Attention-CRF（老挝语词性标注当前最优）对比，本文模型 P、R、F_1 值分别提升 1.9%、2.39%和 2.15%，充分验证了我们模型的鲁棒性和有效性，也证明本文模型在完全避免人工制定特征的情况下，通过融合多粒度老挝语特征可以有效提升老挝语未登录词和低频词的词性标注准确率。

4.3　不同粒度特征方法对比实验

为验证本文融合不同粒度特征对模型结果产生的影响，在同一语料集下，进行不同粒度的对比实验，实验结果如表 8 所示。

表 8 不同粒度特征对模型性能的影响 （单位：%）

序号	Model	Distributed Presentation	P	R	F_1
1	BiLSTM-Attention-CRF	Word Level	92.86	91.54	92.19
2	BiLSTM-Attention-CRF	Word + Char Level	92.90	92.37	92.63
3	BiLSTM-Attention-CRF	Word + Char + Syllable Level	93.31	93.15	93.22
4	Transformer-CRF	Word Level	93.73	92.68	93.20
5	Transformer-CRF	Word + Char Level	94.20	93.62	93.90
6	Transformer-CRF	Word + Char + Syllable Level	94.76	93.93	94.34

由表 8 可知，本文模型未使用多粒度特征时，P、R、F_1 值均有下降，由此可以证明多粒度特征计算可以有效提升模型老挝语词性标注的性能。当模型一致时，融合词、字符、音节的多粒度特征的模型性能最高，在 BiLSTM-Attention-CRF 模型中，F_1 值最大提升为 0.45%；在 Transformer-CRF 模型中，F_1 值最大提升为 1.14%。当模型不一致时，无论是在单粒度特征、双粒度特征还是多粒度特征，本文模型比 BiLSTM-Attention-CRF 模型都更有优势，在基于老挝的词向量的基础上，本文模型 F_1 值提升 1.01%；在融合词和字符特征向量粒度上，本文模型 F_1 值提升 1.27%；在融合词、字符、音节的多粒度上，本文模型 F_1 值提升 1.12%；以上充分证明通过融合多粒度特征能有效提升模型对未登录词和低频词的词性标注效果。

4.4 交叉验证实验结果

由于本实验的数据规模较小，本文采用增加交叉验证的实验结果。在表 9 中，本文采用 k-折交叉验证，其中 $k=\{5,6,7,8,9,10\}$。交叉验证实验具体方法为：将语语料均分为 k 组，选取其中 k-1 组作为训练集，其余 1 组作为测试集，分别重复进行 k 次实验，且每次实验选取的训练集不同，实验结果取 k 次实验结果的平均值。

表 9 交叉验证的实验结果 （单位：%）

交叉验证	P_BiLSTM-CRF	P_BiLSTM-Att-CRF	P_Trans-CRF	P_Our
1:4	90.76	91.23	92.08	92.43
1:5	90.92	91.23	92.52	92.81
1:6	90.89	91.09	92.49	92.89
1:7	91.86	92.80	93.83	94.64
1:8	91.97	92.86	93.73	94.76
1:9	91.64	92.96	93.45	94.50

4.5 长远上下文信息丢失测试

对于老挝语句式过长导致上下文信息丢失的问题，本文使用两组不同的模型进行比较说明。图 5 明确显示了所提方法在处理不同长度的词性标注方面的性能，其中 x 轴对应按长度排序的老挝语句子。由图 5 可，知 BiLSTM-CRF 模型随着句子长度的不断增长，词性标注性能明显逐渐退化，原因在于，BiLSTM 虽然能在一定程度上缓解长期依赖的问题，但是对于特别长的长期依赖有限，在长句表现上会受到限制[15]。而本文模型对特别长的句子准确率只有轻微下降，可以更好地解决长远上下文信息丢失长问题。

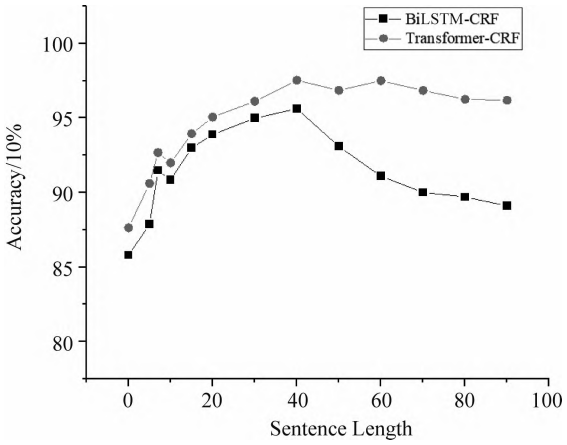

图 5 不同模型在不同句子长度上的表现

4.6 未登录词和低频词识别测试

对于未登陆词和低频词的词性识别效果，本文使用三组不同的模型进行比较说明。图 6 显示了本文模型在未登陆词和低频词的词性识别上的表现，其中 x 轴是按平均词频等级排序的老挝词。从图 6 可见，所提方法在识别未登录词和低频词词性时的准确率比其他两个模型有明显提升。虽然模型没有未登录词的信息，但是在融入字符特征后，模型可以

通过对字符特征的学习获取词的形态结构信息，推断出未登录词的表示，提高模型对未登录词的词性识别准确率；模型通过对音节特征的学习获取词的内部结构信息，提升模型对低频词的词性识别准确率。

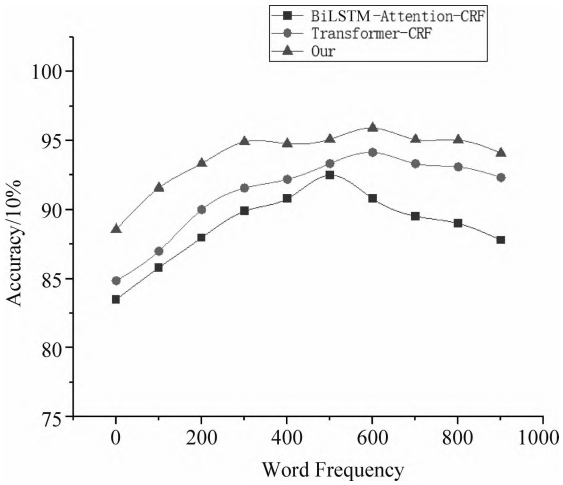

图 6　不同模型识别未登录词和低频词的词性表现

4.7　标注结果分析

图 7 显示，融合词、字符和音节特征的 Transformer-CRF 模型对主要词性标注效果的影响。首先，对副词(ADV)、专有名词(PRN)、普通名词(N)和形容词(ADJ)的标注提升效果最为显著，特别是副词和专有名词，原因是语料中的专有名词很多来自外来词，存在大量未登录词和低频词，融合多粒度特征可以学习形态结构特征，推测词的表达形式，使专有名词的识别效果得到很大的提升；在老挝语中，根据不同的语境有些副词可以表示指示代词、泛指代词、关系代词，Transformer-CRF 模型能够充分提取语义和长距离特征来唯一地确定词性。其次，对代词(CLT)、助动词(PRA)和动词(V)识别效果的提升基本达到 0.8%及以上。介词(PRE)和连词(COJ)的提升效果没那么明显，对量词(CLF)标注效果提升较差。最后，对数词标注提升效果为 0，原因在于日常生活中，老挝人也都通用阿拉伯数字[1]，本文使用正则表达式的方法融合老挝数词特征。

经过对数据的分析，发现主要错误是量词被标记为名词的情况较多，占 26.8%，因此量词的准确率对模型的影响较大，原因是老挝语中量词的数量和表达均丰富，在句子中用法多变，用法特殊，量词的基本用法是名量词，即用名词充当量词，模型在标注时易混淆，因此提升对量词标注的准确率对模型的性能影响比其他词性大。

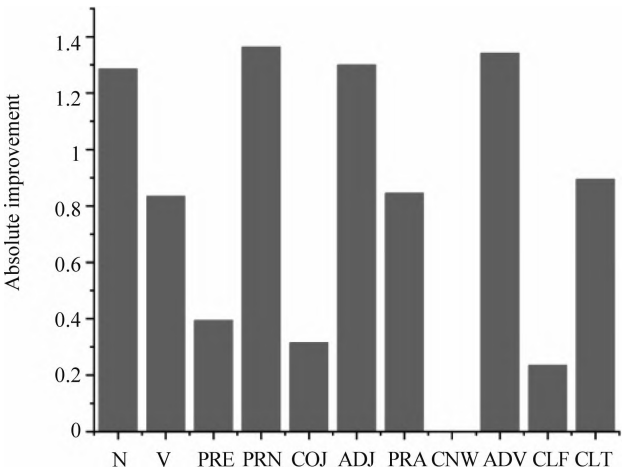

图 7　主要词性标注的绝对提升率

5　结论

本文根据老挝语的特点，提出融合词、字符和音节特征的老挝语词性标注方法，通过在 Transformer-CRF 模型中融入词、字符和音节特征，有效提升了模型对老挝语句子词性标注的效果。实验结果表明，与现有方法相比，本文提出的方法在老挝语语料稀少的情况下得到了更优的结果，F_1 值达到 94.34%。

虽然本文取得了一定的效果，但是老挝语语料稀少，包含的领域较为单一，严重制约了老挝语的研究。在接下来的工作中，将进一步扩大老挝语的语料库，并以此来提升并验证老挝语词性标注研究模型的健壮性；还可以进一步考虑使用标注后的语料对老挝语进行机器翻译等相关工作。

参考文献

[1] 张良民. 老挝语言用语法[M]. 北京：外语教学与研究出版社，2009.

[2] RATNAPARKHI A. A Maximum entropy model for part-of-speech tagging[C]//Proceedings of the Conference on Empirical Method for Natural Language Processings, 1996: 133-143.

[3] AYOGU I I, ADETUNMBI A O, OJOKOH B A, et al. A comparative study of hidden Markov model and conditional random fields on a Yorùba part-of-speech tagging task [C]//Proceedings of the International Conference on Computing Networking and Informatics, 2017.

[4] WANG P, QIAN Y, SOONG F K, et al. A unified tagging solution: Bidirectional LSTM recurrent neural network with word embedding[J]. arXiv preprint arXiv:7511.00215, 2015.

[5] VASWANI A, BISK Y, SAGAE K, et al. Supertag-

ging with LSTMs[C]//Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistic, 2016: 232-237.
[6] CHANG C H. HMM-based part-of-speech tagging for chinese corpora[C]//Proceedings of the Workshop on Very Large Corpora, 1993: 40-47.
[7] BAUM L E., PETRIE T, Statistical inference for probabilistic functions of finite state Markov chains[J]. Ann. Math. Stat, 1966: 1554-1563.
[8] ANDREW M, DAYNE F, FERNANDO C, et al. Maximum entropy markov models for information extraction and segmentation[C]//Proceedings of the International Conference on Machine Learning. Morgan Kaufmann, 2000: 591-596.
[9] PHUONG L H, PHAN X H, The trung tran. on the effect of the label bias problem in part-of-speech tagging[C]//Proceedings of the IEEE Rivf International Conference on Computing & Communication Technologies. IEEE, 2013: 103-108.
[10] LAFFERTY J, MCCALLUM A, Pereira F. 2001 Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the International Conference on Machine Learning, Williams, MA. 2001: 282-289.
[11] HUANG Z, XU W, YU K, et al. Bidirectional LSTM-CRF models for sequence tagging.[J]. arXiv: preprint arXiv:1508.01991, 2015.
[12] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1064-1074.
[13] JUERGEN S. Gradient Flow in recurrent nets: The difficulty of learning long-term dependencies[M]. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long Term Dependencies. Wiley-IEEE Press, 2001.
[14] WU G, TANG G, WANG Z, et al. An attention-basedBiLSTM-CRF model for Chinese clinic named entity recognition[J]. IEEE Access, 2019, 7:113942-113949.
[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5998-6008.
[16] LI H, ADELE Y C, LIU Y, et al. An augmented transformer architecture for natural language generation tasks[C]//Proceedings of the International Conference on Data Mining, 2019: 1131-1137.
[17] RAGANATO A, TIEDEMANN J. An analysis of encoder representations in transformer-based machine translation[C]//Proceedings of the EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP, 2018: 287-297.
[18] YAN H, DENG B, LI X, et al. TENER: Adapting transformer encoder for named entity recognition[J]. arXiv preprint arXiv: 1911.04474,2019.
[19] QIU X, PEI H, YAN H, et al. Multi-criteria Chinese word segmentation with transformer[J]. arXiv preprint arXiv:1906.12035,2019.
[20] 杨蓓. 老挝语分词和词性标注方法研究[D]. 昆明: 昆明理工大学硕士学位论文,2016.
[21] 王兴金,周兰江,张金鹏,等. 融合词预测的半监督老挝语词性标注研究[J]. 小型微型计算机系统,2019, 40(12): 2500-2505.
[22] 王兴金,周兰江,张建安,等. 融合词结构特征的多任务老挝语词性标注方法[J]. 中文信息学报,2019,33 (11): 39-45.
[23] International Organization for Standardization, Information Technology-Universal Coded Character Set (UCS), ISO/IEC 10646[S/OL]. http://standards.iso.org/ittf/PubliclyAvailableStandards/2017.
[24] RASHADUL H R, AMINUL I, EVANGELOS M. Improving text relatedness by incorporating phrase relatedness with word relatedness[J]. Computational Intelligence, 2018, 34(3):939-966.
[25] LABEAU M, LOSER K, ALLAUZEN A, et al. Non-lexical neural architecture for fine-grained POS tagging[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 232-237.
[26] WANG L, CHRIS D, ALAN W B, et al. Finding function in form: Compositional character models for open vocabulary word representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 1520-1530.
[27] MIKOLOV T. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013, 26: 3111-3119.

唐文(1996—),硕士,主要研究领域为自然语言处理。
E-mail: 862304414@qq.com

张建安(1972—),硕士,副教授,主要研究领域为信息安全和机器学习。
E-mail: zjaemail@163.com

周兰江(1964—),通信作者,硕士,副教授,主要研究领域为信息检索,机器学习和自然语言处理。
E-mail: 915090822@qq.com