文章编号：1003-0077(2019)08-0060-07

基于 Bi-LSTM-CRF 模型的维吾尔语词干提取的研究

古丽尼格尔·阿不都外力[1,2]，吐尔根·依布拉音[1,2]，
卡哈尔江·阿比的热西提[1,2]，王路路[1,2]

（1. 新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830046；
2. 新疆大学 新疆多语种信息技术实验室，新疆 乌鲁木齐 830046）

摘　要：词干提取是维吾尔语自然语言处理中的基础性研究，其提取质量直接影响其他任务的性能。但目前维吾尔语词干提取研究存在过度切分、不切分和歧义切分等问题，这些问题导致词干提取质量不高，对后续任务的性能影响较大。因此该文提出了基于 Bi-LSTM-CRF 的维吾尔语词干提取模型，将字符作为最小切分单位，选取维吾尔语字符特征、音类特征以及语音特征为候选特征，结合模型进行实验。实验表明，该文提出的 Bi-LSTM-CRF 模型在维吾尔语词干提取任务上，$F1$ 值达到了 88%，在融入手工提取的候选特征之后，$F1$ 值提高了 1.8 个点，有效提高了词干提取的准确性，缓解了上述问题带来的影响。

关键词：维吾尔语；词干提取；Bi-LSTM-CRF
中图分类号：TP391　　**文献标识码**：A

Research on Uyghur Stemming Based on Bi-LSTM-CRF Model

GULINIGEER Abudouwaili[1,2], TUERGEN Yibulayin[1,2], KAHAERJIANG Abiderexiti[1,2], WANG Lulu [1,2]

(1. College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China; 2. Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University, Urumqi, Xinjiang 830046, China)

Abstract: Stemming is a basic research in Uyghur Natural-language Processing (NLP), which is still challenged by issues of over-segmentation, non-segmentation and ambiguity segmentation in Uyghur stemming. This paper propose a neural network model of Bi-LSTM-CRF, which is based on bidirectional (Bi) long short-term memories (LSTMs) and conditional random fields (CRFs). It uses Uyghur character as minimum language unit to extract Uyghur character features, phonological features and phonetic features, and use them as the candidate features. The stemming result shows that an F-score of 88% for the Bi-LSTM-CRF model of Uyghur stemming, with further 1.8% increase after incorporating the manual features.

Keywords: Uyghur language; stemming; Bi-LSTM-CRF

0　引言

维吾尔语是典型的形态丰富的黏着语。黏着语种的单词由词干和词缀组成，词干主要表达词的意义，而词缀提供语法信息（所属性，形态，复数）。作为维吾尔语自然语言处理中的基础性研究，词干提取的质量会直接影响维吾尔语言处理的其他任务，如词性标注、命名实体识别等[1]。除此之外，维吾尔语中词干与词缀相连接时，连接处由于结合的不规则性，会发生一系列的音系现象[2]，这种音系现象对词干提取带来了一定的困难。

收稿日期：2019-02-01　　**定稿日期**：2019-02-13
基金项目：国家自然科学基金（61762084，61662077，61462083）；国家语委科研项目（ZDI 135-54）；国家重点研发计划（2017YFB1002103）

维吾尔语自然语言处理技术还处于发展初期[3]，目前维吾尔语中的词干提取大致可以分成基于词典/规则的方法[4]、基于统计的方法[5]和基于神经网络的方法[6]。基于词典/规则的方法工作量较大，需要语言学家制定语言学规则并构造限制条件。这种方法虽然结果更加准确，但需要大量的语言学知识，受词干提取词典大小的限制，而且语言学规则只适用于常规词形变换，缺乏全面性。基于统计的方法是通过词的分布统计规律进行词干提取，能较好地处理 OOV 现象和一般构词规律构成的词形。基于统计学习的维吾尔语词干提取研究虽然有了初步的成果，但需要人工选择和提取特征，而且还存在着过度切分、不切分和歧义切分等问题。基于神经网络的方法是一种特征学习的过程，通过后向传播算法学习出最适合维吾尔语词干提取模型的参数。此方法通过自动学习数据中的特征表示来缓解人工选择和提取特征的过程中成本较大的问题，但仍然存在过度切分、不切分和歧义切分的问题。

为了解决以上问题，本文提出了基于 Bi-LSTM-CRF 神经网络的维吾尔语词干提取方法。该方法将采用 BIO2 标记，引入字符特征、音类特征以及语音特征作为候选特征。为了进一步证明模型的有效性，本文将分两组做实验对比：

(1) 将 Bi-LSTM-CRF 模型应用到维吾尔语词干提取上，并与 CRF、LSTM、Bi-LSTM、LSTM-CRF 模型做实验对比，验证 Bi-LSTM-CRF 模型能有效地解决词干提取时出现的过度切分、不切分和歧义切分等情况；

(2) 引入不同的候选特征，验证当逐步加入字符特征、音类特征以及部分语音特征组时，特征集对维吾尔语词干提取质量的影响。

1　相关工作

1.1　词干提取

维吾尔语属于典型的黏着语，在黏着语中词是最重要的语法单位，是由语素构成(最小的语法单位)。根据语素在词中的不同作用将其分成词根和词缀[2](构形词缀和构词词缀)，词干由词根和构词词缀组成，是词的核心部分，词义由词干体现，而词缀(本文只考虑构形词缀)只能黏附在词根或词干的语素上，它本身不能单独构成词，其主要表达语法含义，如“مەكتەپ(学校)”与“ىمىز(表示第二人称复数的词缀)”连接成“مەكتىپىمىز(我们的学校)”，再与“دە(在)”连接成“مەكتىپىمىزدە(在我们学校)”。而词干提取是根据语言形态中的规律来去除词缀，从而获得词干的过程。

除了维吾尔语，国内少数民族语言中属于黏着语的还有蒙古语、哈萨克语等。由于国内少数民族语言的词干提取技术发展得比较晚，因此基于词典/规则相结合的方法比较多。史建国等[7]利用词典和规则的方法对蒙古文进行词切分，得到了性能较好的斯拉夫蒙古文词切分系统；李婧等[8]采用基于规则、字典查找和最大匹配相结合的方法对哈萨克语进行词干提取，并提出了结合哈萨克语元音和谐规律、词干词性和词尾缀接顺序切分词尾的方法，使得词干提取正确率达 95.26%；早克热·卡德尔等[9]首先构造了名词的有限状态自动机，并用最大熵模型给有限状态自动机加入了歧义词缀识别能力，建立了基于规则和信道噪声模型的元音和谐处理方法。随着统计学习模型在自然语言处理领域中的广泛应用，词干提取也从传统的方法逐步过渡到了统计的方法。赛迪亚古丽·艾尼瓦尔等[5]以 N-gram 为基准模型，根据维吾尔语构词规律，提出了融合词性特征和上下文词干信息的维吾尔语词干提取模型，由于语料库规模较小，模型依赖于上下文特征和词性特征，而且可能存在一些重复单词等原因，当语料库规模逐渐增大时，模型准确率提升较缓慢；那日松等[10]设计了两组对比实验，将蒙古文的分词问题转化为序列标注问题，使用了四词位标注集，利用 CRF 模型，以上下文词形和蒙古文连写的构形附加成分作为特征，实验结果表明，上下文作为特征的实验组比附加成分作为特征的实验组效果更好；李文等[11]将维吾尔语和蒙古语作为研究对象，介绍了基于最大后验概率模型非监督式形态切分方法，在非监督式切分的基础上，通过加入调参的方式，使模型更适用于特定的语言。实验结果表明，虽然切分的准确性提高了，但此方法只适合用于特定的语言，而且也有过渡切分的问题；姜文斌等[12]将维吾尔词语的层次结构引入到词法分析研究中，提出了维吾尔词法分析的有向图模型，对于音系现象又提出了基于词内字母对齐算法的自动还原模型，其词干提取的正确率达到了 94.70%，但由于只根据从训练集中自动抽取的词干表和词缀作为当前切分词的递归

穷举可能的候选结构，因此导致过多的候选，而且只限制于词干库表和词缀库表；哈里旦木·阿布都克里木等[6]提出了基于语素序列的维吾尔语形态切分方法，将单词切分成若干个语素（词根和词缀），从而缓解了数据稀疏问题。

1.2 CRF模型

条件随机场（Conditional Random Field，CRF）[13]是一种无向图模型，近年来已经广泛应用到其他自然语言处理任务中，如分词、词性标注、命名实体识别等。其结合了最大熵（MEM）和隐马尔可夫（HMM）的特点，通过考虑上下文中标签之间的相关性来防止HMM和MEM中的有限特征选择。除此之外，CRF可以通过全局特征归一化的过程获得全局最优，CRF链式结果如图1所示。

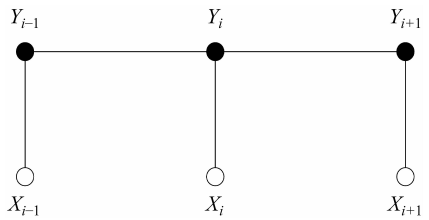

图1 CRF链式结构

现给定可观察序列 $W=w_1w_2\cdots w_n$，与之相应的标记序列为 $Y=y_1y_2\cdots y_n$，则条件概率定义如式(1)所示。

$$p_\lambda(Y\mid W)=\frac{1}{Z(W)}\exp\Big(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,W,t)\Big) \tag{1}$$

其中，f_k 为特征函数，λ_k 为参数，$Z(W)$ 为规一化因子，使给定所有可能状态序列的概率之和为1。而观察序列需要搜索概率最大的 $Y^*=\arg\max p(Y|W)$。

1.3 LSTM模型

循环神经网络（Recurrent Neural Network，RNN），是一种通过隐藏层节点周期性的连接来获得序列化数据中动态信息的神经网络，可以对序列化的数据进行分类。但是，RNN对长跨度时间可能会有梯度消失或爆炸的问题。为了解决长距离依赖的问题，Hochreiter S等[14]提出了一种改进的循环神经网络——长短时记忆网络（Long Short Term Memory Network，LSTM），LSTM可以选择性忘记历史信息以及更新存储的信息，这将有效地解决RNN的梯度消失或爆炸问题，LSTM网络结构如图2所示。

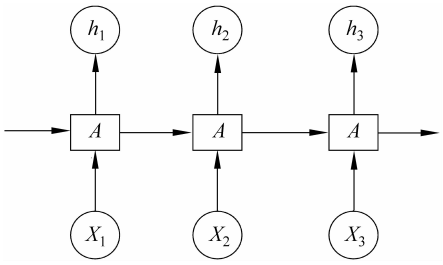

图2 LSTM网络结构

LSTM单元由三个门（遗忘门、输入门、输出门）和一个细胞状态组成，其结构如图3所示。

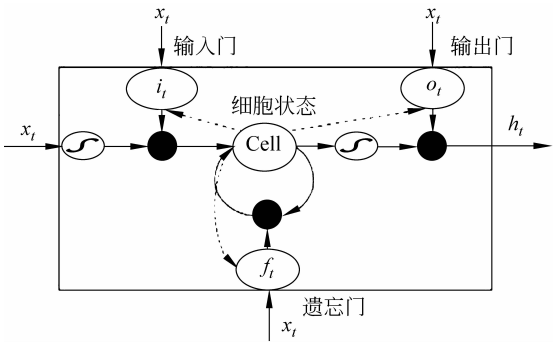

图3 LSTM单元模型结构

遗忘门决定历史细胞状态的保留信息，这由sigmoid函数来控制，它会根据上一时刻的输出和当前的输入来产生一个0～1的 f_t 值，来决定上一时刻学到的信息是否通过以及通过多少，计算如式(2)所示。

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{2}$$

输入门控制将新的信息中哪些部分保存到细胞状态中，首先用sigmoid函数来决定哪些值用来更新，而用tanh函数来生成新的后选值，并将这两部分生成的值进行结合并更新，计算如式(3)～式(5)所示。

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{3}$$

$$\widetilde{C}_t=\tanh(W_C\cdot[h_{t-1},x_t]+b_C) \tag{4}$$

$$C_t=f_t*C_{t-1}+i_t*\widetilde{C}_t \tag{5}$$

决定输出门控制全部更新后的细胞状态中哪些部分被输出，首先通过sigmoid函数得到初始的输出，之后用tanh函数将 C_t 值映射到－1到1的区间，再通过初始输出值逐对相乘，最终得到输出，计算如式(6)、式(7)所示。

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{6}$$

$$h_t=o_t*\tanh(C_t) \tag{7}$$

2　基于 Bi-LSTM-CRF 的维吾尔文词干提取

2.1　Bi-LSTM-CRF 模型

Bi-LSTM-CRF 模型[15]是由 Bi-LSTM 和 CRF 模型结合的模型，从 Bi-LSTM 输出的向量作为 CRF 模型的输入值，Bi-LSTM-CRF 模型不仅能保留 Bi-LSTM 上下文信息，而且能通过 CRF 层考虑前后的标签信息。Bi-LSTM-CRF 网络结构如图 4 所示。

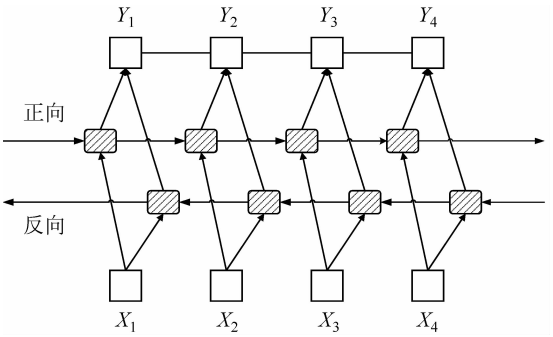

图 4　Bi-LSTM-CRF 网络结构图

在 Bi-LSTM-CRF 模型中，通过 Bi-LSTM 层提取特征并输入到 CRF 层，利用 CRF 层对序列建模的能力对特征解码。因为模型的特征是由 RNN 网络结构学习得到，所以特征会分为标签间的转移特征 $h_p(s_{t-1},s_t)$ 和标间特征 $h_q(s_t,l_0^{t+d})$。故 CRF 的目标函数将定义为式(8)。

$$H(s_{t-1},s,l_{t-d}^{t+d})=\sum_{m=1}^{M}\lambda_m h_m(s_{t-1},s,l_0^{t+d})=\sum_{p}^{P}\lambda_p h_p(s_{t-1},s_t)+\sum_{q}^{Q}\lambda_q h_q(s_t,l_0^{t+d}) \tag{8}$$

维吾尔语中词干和词缀拼接时，一般在词干或词缀中会出现音系现象(弱化、增音、脱落等)，这将严重影响切分准确度，也成为了维吾尔语词干提取过程中的难点。由图 5 我们可以发现，Bi-LSTM-CRF 模型克服了 LSTM 模型只记录上文信息、不考虑下文信息的缺点，将通过 Bi-LSTM 得到的两个隐藏层单元输出结果进行拼接，作为整体网络隐藏层输出，并将其输出结果输入到 CRF 层里，将维吾尔语词干提取转变成序列标注的过程。

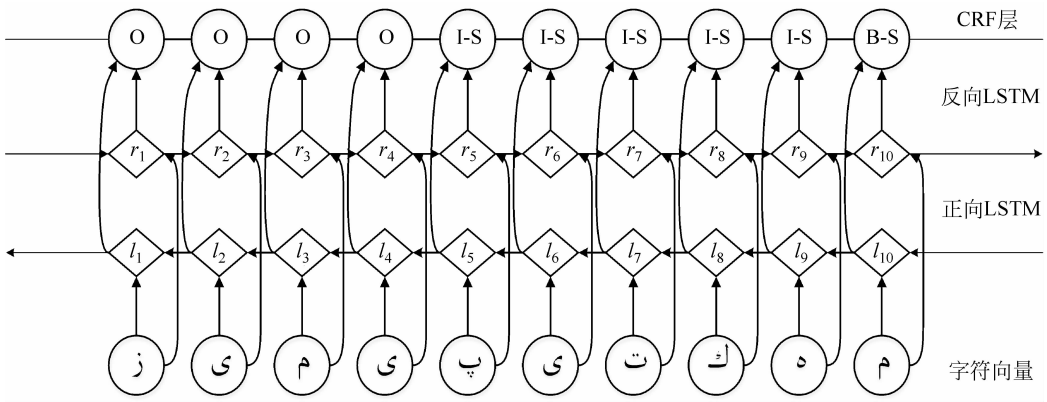

图 5　基于 Bi-LSTM-CRF 模型的维吾尔语词干提取结构

2.2　特征选择与标记集

本文中我们考虑几种候选特征作为特征集合，确定哪一个特征对词干提取有较为显著的影响，选取候选特征时，我们参考了文献[16]提出的特征，分别为当前字符的字符特征 C(字符本身)、音类特征 S(当前字符为元音，则特征为 V；当前字符为辅音，则特征为 C)和语音特征 P1、P2、P3(当前音类为元音时，则根据元音发音时横向舌位、纵向舌位和展圆情况进行分类；当前音类为辅音时，则根据发音时声带的振动情况、发音部位和发音方式进行分类)。

本文采用 BIO2 的组块(chunk)方法来标记词干，标记集合定义为{B,I,O}，即将每个字符分三类：B-S(词干首字符)、I-S(词干中部)、O(非词干)，如“مەكتېپىمىز(我们的学校)”，标记为：

“O/م O/ى I-S/پ I-S/ى I-S/ت I-S/ك I-S/ه B-S/م”

。“O/ز O/ى”

通过这种表示方法，将单词根据标注语料映射成由独立标记组成的功能块，即可将词干提取任务转换成序列标注问题。

3 实验数据与结果分析

3.1 实验数据

目前为止，由于维吾尔语词干提取公开的标注数据集或语料库还未见公开，因此本文将从天山网爬取新闻数据，并进行人工校对和人工提取词干(数据大小：15 万)，按词长进行由长到短的排序，并选出其中最长的 1 万个单词进行预处理，采用交叉验证法对标记语料进行分割产生训练集、测试集和验证集(分割比为 0.75∶0.15∶0.1)，语料具体统计如表 1 所示。

表 1 语料统计表

数据类型	Token	Char
训练集	7 500	100 116
测试集	1 500	20 015
验证集	1 000	13 326

标记集在数据集中的分布统计如图 6 所示。

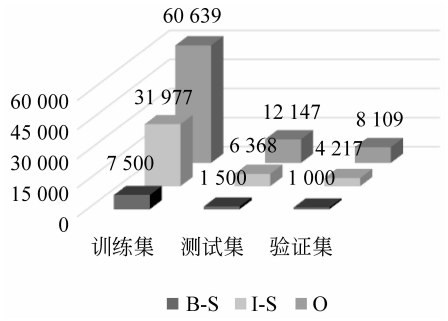

图 6 标记集在数据集中的分布

数据集中最长的单词长度、词干长度、词缀长度和最短的单词长度、词干长度、词缀长度(由于数据是基于词的，因此只考虑了字符特征)如表 2 所示。

表 2 单词、词干、词缀长度

类型	单词长度	词干长度	词缀长度
最长	32	20	22
最短	8	2	4

数据集有以下特点：

① 包含的单词、词干和词缀长度比较长；

② 包含较多的外来词、不规则词；

③ 以字符(维吾尔文字母)作为最小的分割单位；

④ 由无重复的维吾尔语单词构成，没有上下文语言环境。

3.2 实验设计与结果分析

为了进一步验证模型和特征对词干提取的影响，在本节中分别设计不同模型、特征的对比实验，寻找最适合词干提取的模型和特征，确定最佳的提取效果。在实验过程中，将使用 F 值(F_1)作为评测指标，衡量词干提取效果。

本文利用 CRF^{++} 开源工具①和 Lample 等[17]提出的 Bi-LSTM-CRF 模型作为基准模型，构建基于维吾尔语的词干提取模型，Bi-LSTM-CRF 网络结构超参数参考了 Lample 提出的网络，超参数如表 3 所示。

表 3 神经网络超参数

实验参数	值
Dropout rate	0.5
迭代次数	50
输入向量维度	50
隐含层	100

3.2.1 不同模型的对比实验

本组实验中，将对 CRF、LSTM、Bi-LSTM、LSTM-CRF 和 Bi-LSTM-CRF 等模型分别做实验对比，其实验结果如表 4 所示。

表 4 实验结果(%)

模型	验证集	测试集
CRF	78.70	77.95
LSTM	38.87	37.95
Bi-LSTM	64.82	63.74
LSTM-CRF	72.20	70.73
Bi-LSTM-CRF	90.00	88.00

(1) 从表中可见，Bi-LSTM-CRF 模型的词干提取明显高于 CRF、LSTM、Bi-LSTM 和 LSTM-CRF 模型，F 值分别提升了 10.05、50.05、24.26、17.27 个点。实验结果说明，Bi-LSTM-CRF 模型比其他模型更加准确地识别了词干和词缀，而且也正确地切分了词干和词缀。

(2) LSTM-CRF 模型和 Bi-LSTM-CRF 模型的识别效果都高于 LSTM 和 Bi-LSTM，而且 CRF 模

① https://taku910.github.io/crfpp/

型也高于 LSTM 模型和 Bi-LSTM，其实验结果说明，采用序列标注方法对维吾尔语进行词干提取时，对提取结果是有一定的帮助的。

（3）LSTM 模型和 LSTM-CRF 模型分别低于 Bi-LSTM 模型和 Bi-LSTM-CRF 模型，其原因可能是通过双向的 LSTM 模型有效地考虑了上下文信息，并且对于单向的 LSTM 模型，双向的具有一定的互补性，因此对形态复杂的维吾尔语进行词干提取时，双向的神经网络明显优越于单向的神经网络。

根据过度切分、不切分和歧义切分三类现象，对比了非 Bi-LSTM-CRF 模型（CRF、LSTM、Bi-LSTM、LSTM-CRF）和 Bi-LSTM-CRF 模型在维吾尔语词干提取时的切分结果。在实例“ئۆردەك(鸭子)”中，非 Bi-LSTM-CRF 模型将“دەك”误认为词缀；实例“مېنىڭچە(以我看来)”中，非 Bi-LSTM-CRF 模型没有切分词缀“چە”；实例“مەكتېپىمىز(我们的学校)”中，模型少切分了“مەكتېپىمىز。因此，将 Bi-LSTM-CRF模型应用到维吾尔语词干提取时，可以较正确地切分词干和词缀，如表 5 所示。

表 5　维吾尔词干提取实例分析

现象	实例	非 Bi-LSTM-RF 模型切分结果	Bi-LSTM-CRF 模型切分结果
过度切分	ئۆردەك	ئۆر+دەك	ئۆردەك
不切分	مېنىڭچە	مېنىڭچە	مېنىڭ+چە
歧义切分	مەكتېپىمىز	مەكتېپى+مىز	مەكتېپ+ىمىز

3.2.2　不同特征的对比实验

在对比实验（1）的基础上将对 CRF 模型和 Bi-LSTM-CRF模型引入手工提取的特征，如字符特征（C）、音类特征（S）、语音特征（P1，P2，P3）等（候选特征的输入维度为 30），实验结果如表 6 所示。

（1）当 Bi-LSTM-CRF 模型不加候选特征的 F 值比 CRF 模型加特征的 F 值提高了 8.2 个点，说明不加特征的 Bi-LSTM-CRF 模型词干提取的效果比加候选特征的 CRF 模型更好。

（2）当输入所有候选特征、模型不同时，Bi-LSTM-CRF模型与 CRF 模型相比 F 值提升了 9.33 个点。

（3）当模型相同、输入候选特征不同时，与不加特征的 Bi-LSTM-CRF 模型相比，F 值分别提升了 1.47、0.93、0.6 和 1.8 个点，实验结果说明，通过神经网络模型进一步提高词干提取性能时，可以考虑加入候选特征。

表 6　实验结果（%）

基本模型	特征	验证集	测试集
CRF	None	78.70	77.95
	$C+S+P_1+P_2+P_3$	80.60	79.80
Bi-LSTM-CRF	None	90.00	88.00
	C	90.10	89.47
	$C+S$	90.40	88.93
	$C+S+P_1$	91.30	88.60
	$C+S+P_1+P_2$	**91.30**	**89.80**
	$C+S+P_1+P_2+P_3$	90.90	89.13

（4）有些候选特征对词干提取影响不同，例如，特征 $C+S+P_1+P_2$ 组合时，其 F 值最高，提升了 1.8 个点，但当所有特征组合在一起时，其 F 值没有比特征组 $C+S+P_1+P_2$ 提升的高。（网络模型参数参考表 3）。

除此之外，通过分析实验结果发现以下两种情况对实验结果的准确率有较大的影响：

① 当前词为动词且与词干相连接的词缀种类较多时，会出现词缀的歧义切分。例如，“ئىشلىتەلمەيۋاتىمىز(我们还不能使用)”中，Bi-LSTM-CRF 模型将其切分成“ئىشلىتەلمەي+ۋاتىمىز”。（正确切分为“ئىشلى+تەلمەيۋاتىمىز”）

② 词干、词缀切分时，会出现词干歧义。例如，“ئالما(苹果，不要拿)”中，根据词性的不同，存在不同的切分方式。当“ئالما”为名词苹果时，词干为“ئالما”，不切分；当“ئالما”为动词不要拿时，词干为“ئال+ما”，需切分。

以上情况可能是由于在构建语料库中没有考虑词性特征或上下文语言环境所造成的。

4　结论

本文将维吾尔语词干提取看成序列标注问题，以字符为切分粒度来表征维吾尔语的构成机制，采用 CRF、LSTM、Bi-LSTM、LSTM-CRF 及 Bi-LSTM-CRF 模型对比维吾尔语词干提取效果和处理过度切分、不切分和歧义切分的能力，并在此基础上分析维吾尔语字符特点，引入字符特征、音类特征以及语音特征，对比几个特征组对维吾尔语词干提取影响。本文采用的基于 Bi-LSTM-CRF 模型在维吾尔语词干提取上的取得了较好的效果。实验结果表明：①Bi-LSTM-CRF模型能比较准确地识别维吾尔语

中词干和词缀，有效缓解过度切分、不切分和歧义切分等现象；②本文引入的候选特征对维吾尔语的词干提取是有效的，其特征集中特征组字符特征(C)、音类特征(S)以及部分语音特征(P1 和 P2)的提取效果最佳。

本文还有一些局限性，比如没有研究词干与词缀连接时所出现的音系现象或词干提取时还原原词干(由于音系现象，词干中的一些字母会发生变化)等问题。故在以后的研究中，考虑更多特征因素，通过改进模型来提高维吾尔语词干提取的效果。

参考文献

[1] 艾孜尔古丽，阿力木·木拉提，玉素甫·艾白都拉. 基于形态分析的现代维吾尔语名词词干识别研究[J]. 中文信息学报，2015，29(6)：208-212.

[2] 叶蜚声，徐通锵. 语言学纲要[M]. 北京：北京大学出版社，2006.

[3] 吐尔根·依布拉音，袁保社. 新疆少数民族语言文字信息处理研究与应用[J]. 中文信息学报，2011，25(6)：149-157.

[4] 热娜·艾尔肯，李晓，艾尼宛尔·托乎提. 基于混合方法的维吾尔语词干提取方法研究[J]. 计算机应用研究，2015，32(1)：112-114.

[5] 赛迪亚古丽·艾尼瓦尔，向露，宗成庆，等. 融合多策略的维吾尔语词干提取方法[J]. 中文信息学报，2015，29(5)：204-210.

[6] 哈里旦木·阿布都克里木，程勇，刘洋，等. 基于双向门限递归单元神经网络的维吾尔语形态切分[J]. 清华大学学报(自然科学版)，2017(1)：1-6.

[7] 史建国，侯宏旭，飞龙. 基于词典、规则的斯拉夫蒙古文词切分系统的研究[J]. 中文信息学报，2015，29(1)：197-202.

[8] 李婧，刘海峰. 现代哈萨克语词干提取研究[J]. 信息通信，2015(7)：103-104.

[9] 早克热·卡德尔，艾山·吾买尔，吐尔根·依布拉音，等. 混合策略的维吾尔语名词词干提取系统[J]. 计算机工程与应用，2013，49(1)：171-175.

[10] 那日松，淑琴，齐力格尔. 基于 CRF 模型的蒙古文分词及词性标注的研究[J]. 内蒙古大学学报(哲学社会科学版)，2016(2)：23-28.

[11] 李文，李淼，等. 一种带权值参数的非监督式形态切分方法[C]//少数民族青年自然语言处理技术研究院与进展——第三届全国少数民族青年自然语言信息处理、第二届全国多语言知识库建设联合学术研讨会，2010.

[12] 姜文斌，王志洋，等. 维吾尔语词法分析的有向图模型[J]. 软件学报，2012，23(12)：94-100.

[13] Lafferty J D，Mccallum A，Pereira F C N. Conditional random fields：Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of 18th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2001：282-289.

[14] Hochreiter S，Schmidhuber J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735-1780.

[15] Huang Z，Xu W，Yu K. Bidirectional LSTM-CRF models for sequence tagging [J]. arXiv：1508.01991. 2015.

[16] 力提甫·托乎提. 现代维吾尔语参考语法[M]. 北京：中国社会科学出版社，2012.

[17] Lample G，Ballesteros M，Subramanian S，et al. Neural architectures for named entity recognition[C]//Proceedings of the Conference of the North American Chapter of the Associantion for Computtational Linguistics：Human Language Technologies (NAACL-HLT)，2016：260-270.

古丽尼格尔·阿不都外力(1993—)，博士研究生，主要研究领域为自然语言处理。
E-mail：1506254371@qq.com

卡哈尔江·阿比的热西提(1984—)，博士研究生，讲师，主要研究领域为自然语言处理，信息抽取。
E-mail：kaharjan@xju.edu.cn

吐尔根·依布拉音(1958—)，通信作者，博士生导师，教授，主要研究领域为自然语言处理。
E-mail：turgun@xju.edu.cn