文章编号：1003-0077(2020)01-0045-06

乌兹别克语词干提取算法的比较研究

吾买尔江·买买提明[1,2]，古丽尼格尔·阿不都外力[1,2]，
买合木提·买买提[1,2]，卡哈尔江·阿比的热西提[1,2]，吐尔根·依布拉音[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重点实验室，新疆 乌鲁木齐 830046)

摘 要：黏着语的自然语言处理中，词干提取作为一项基础的预处理任务，对其他任务的性能影响较大。现有的乌兹别克语词干提取任务仍依赖基于规则的方法，且实验效果不太理想。该文将乌兹别克语词干提取任务视为序列标注问题进行处理，以字符为最小单位进行切分，分别构建了基于条件随机场(CRF)和门控循环单元网络(Bi-GRU)的乌兹别克语词干提取模型。实验结果表明，基于序列标注的乌兹别克语词干提取模型与基于规则的方法相比不仅降低了人工成本，而且在性能方面有较为显著的提升。

关键词：乌兹别克语；词干提取；序列标注

中图分类号：TP391 **文献标识码**：A

A Comparative Study of Uzbek Stemming Algorithms

WUMAIERJIANG Maimaitiming[1,2], GULINIGEER Abuduwaili[1,2], MAIHEMUTI Maimaiti[1,2],
KAHAERJIANG Abiderexiti[1,2], TUERGEN Yibulayin[1,2]

(1. School 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: As a basic task agglutinative languages processing, word stemming would directly influence the performance of other tasks. The existing Uzbek word stemming task still relies on rule-based approaches. This paper presents the application Conditional Random Field(CRF) and Bidirectional Gated Recurrent Unit(Bi-GRU) in this task, in which the minimum division unit is the character. The experimental results show that the proposed models, which are based on sequence labeling significantly improves the performance compared with the rule-based method.

Keywords: Uzbek language; stemming; sequence labeling

0 引言

乌兹别克语作为形态复杂的黏着语，单词由词干和词缀组成，其中词干表达词义，而词缀只能黏附在词干表达语法范畴。词干提取(stemming)是针对文本中的单词进行词干词缀切分，从而获得词干的过程[1]，是自然语言处理领域的基础性研究内容之一，其提取结果直接影响信息检索、机器翻译等下游任务的性能。

现阶段乌兹别克语的词干提取已具有初步的研究结果，基于规则和词典的方法[2]是目前比较主流的研究方法。虽然基于规则和词典的方法准确率较高，但是完全依赖于先前构建的词干库和语言学家制定的语言规则。因此，受到词典和语言规则的限制，无法覆盖所有的词法规则，而且成本较高。黏着语形态的复杂性是由不同语法范畴的词缀相互组合所导致的。随着规则的增多，逐渐会出现规则冲突问题。例如，在多义词或词干尾部包含与词缀相同的单词中，容易出现切分错误。因此，单独利用基于

收稿日期：2019-08-15 **定稿日期**：2019-09-16

基金项目：国家语委科研项目(ZDI135-54)；国家自然科学基金(61762084，61662077，61462083)；新疆维吾尔自治区重点实验室开放课题(2018D04019)

规则和词典的方法很难满足大数据背景下的乌兹别克语词干提取任务。在其他黏着语中,除了传统的方法之外,还使用基于统计和深度学习的方法进行词干提取。基于统计方法的词干提取方法,根据语料中单词特征的统计分布,能够有效地学习大部分语法规则,但是依然需要利用人工选择和提取特征;基于深度学习的方法,能够使用神经网络自动提取特征,通过不断地训练优化模型的权重参数,学习得到表现良好的词干提取模型。

为了解决基于规则和词典方法的局限性,本文分别采用基于条件随机场和门控循环单元网络的序列标注方法实现乌兹别克语词干提取。为验证这两种模型的有效性,本文首先在序列标注任务上引入两种不同的标注方法(BIO 和 BMES),验证 BMES 更能准确地反映词干词缀信息;其次,在借鉴 Imailov A[2] 的工作基础上,复现基于 Lovins 算法的乌兹别克语词干提取模型,并与基于序列标注模型比较词干提取性能;最后,在不同数据集上验证了当考虑上下文信息时对词干提取任务的质量影响。

1 相关工作

黏着语中,单词由语素组成,语素是最小的语法单位[1]。根据语素在词中的位置不同,可分为词根与词缀。词缀根据功能不同,可分为派生词缀和屈折词缀,示例见表1。派生词缀又称为构词词缀,粘附在词根上构成新词;而屈折词缀又称为构形词缀,粘附在词根或词缀后,只能改变单词的形式,不能构成新词。通过屈折词缀来表示单词的语法范畴,而一个单词去除屈折词缀后的部分是词干(本文中只研究屈折词缀,并将其简称为词缀)。当词干与词缀相连接时,由于连接的不紧密,会发生一系列的音变现象(语音变化现象),增加词干提取的困难。

表1 乌兹别克语词缀类型例子

词缀类别	例子
派生词缀	bog'(名词,花园)+ bon(派生词缀)= bog'bon(名词,园丁)
屈折词缀	bog'(名词,花园)+ lar(屈折词缀)= bog'lar(名词,花园的复数)

由于乌兹别克语词干提取研究仍处于初期阶段,因此基于规则和词典的方法比较流行。Ismailov A[2]等首先介绍并分析了主流的英文词干提取方法的优缺点,其次对比了适用于乌兹别克语的算法(Lovins Stemmer 和 Paice/Husk Stemmer,其中 Lovins Stemmer 实验结果比较好),由于乌兹别克语中存在大量的前词缀,为了提高词干提取准确率,在 Lovins 算法中加入了处理前缀的步骤;艾孜海尔江等[3]利用基于规则和词典的方法构建了词干提取模型,并与最大熵模型相结合提出了融合乌兹别克语形态特征的最大熵名词标注模型。

除此之外,国内黏着语词干提取研究中,苏依拉等[4]将词干词缀库和逆向最大匹配方法相结合,缓解了蒙汉词对齐模型训练时出现的数据稀疏和长距离依赖的问题;那日松等[5]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;徐春等[6]利用外部信息模块和联合校验模块来优化模型,将词干提取之前的句子视为源语言、词干提取之后的句子为目标语言,提出了一种基于机器翻译的维吾尔语形态分析模型,其实验结果优于英文与中文的实验结果;赛迪亚古丽·艾尼瓦尔等[7]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;哈里旦木·阿布都克里木等[8]提出了基于双向门限递归单元神经网络的维吾尔语形态切分方法,将维吾尔语单词自动切分为语素序列,缓解了数据稀疏的问题;古丽尼格尔·阿不都外力等[9]提出了基于 BiLSTM-CRF 模型的维吾尔语词干提取模型,将字符作为最小切分单位,融入候选特征来缓解过度切分、不切分和歧义切分等问题带来的影响;李婧等[10]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%;Gulila 等[11]在分析哈萨克语词缀的基础上,首先提出了基于有限状态自动机的哈萨克语名词和动词词干提取方法,其次将全切分方法与词法分析相结合,实现了哈萨克语较精准的词干提取目标。

2 几种典型的词干提取方法

2.1 Lovins 算法

Lovins 算法[12]是最早提出的基于英文文本的

词干提取算法，由 294 种词尾、29 种构词条件和 35 种转化规则组成，能够有效处理英文单词中双写动词结尾和不规则单词的复数形式。该算法由两步实现：(1)识别和去除后缀；(2)转换剩余部分。由于乌兹别克语中存在较多的前缀，直接使用该算法会导致无法正确提取带有前缀的单词。因此，文献[2]中提出了基于改进的 Lovins 词干提取方法。改进算法中，首先增加了去除前缀的步骤；其次保留了去除后缀的步骤；最后，由于乌兹别克语中词干的形式是不会发生变化的，因此删除了转换剩余部分的步骤，算法流程图如图 1 所示。

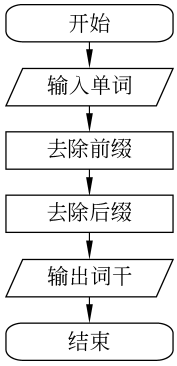

图 1　改进的 Lovins 算法流程图

2.2　CRF 模型

条件随机场(conditional random field，CRF)[13]是基于序列标注的算法，在序列标注问题上已取得了较高的效果。由于之前的序列标注模型存在一些局限性，比如最大熵马尔科夫模型(maximum entropy markov model，MEMM)很容易陷入局部最优，隐马尔科夫模型(Hidden Markov Model，HMM)不能考虑上下文特征等。因此，Lafferty 等提出了一种新的无向图模型——条件随机场，CRF 不仅结合了 MEMM 和 HMM 的优点，还克服了 MEMM 中的局部归一化问题、标记偏置问题和 HMM 受严格独立性假设限制的问题。

对于给定的可观察序列 $X=x_1x_2\cdots x_n$ 和相应的序列标签为 $Y=y_1y_2\cdots y_n$，则可以用条件概率公式计算 Y 的概率值，如式(1)所示。

$$p_W(Y\mid X)=\frac{1}{Z(X)}\exp\left(W\cdot\sum_{i=1}^{n}\Phi(y_{i-1},y_i,x)\right) \tag{1}$$

其中，$\Phi(y_{i-1},y_i,x)$为特征函数，W 为参数，$Z(X)$为规范化因子。

2.3　GRU 模型

循环神经网络(recurrent neural network，RNN)，是一种通过隐含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 在长序列数据的训练中可能会有梯度爆炸或消失的问题。因此，为了解决长距离依赖的问题，Hochreiter S 等[14]提出了改进之后的循环神经网络，即长短时记忆网络(long short-term memory neural network，LSTM)，LSTM 是在 RNN 的基础上增加了控制门和一个细胞状态，能够决定状态是否遗忘。控制门包括遗忘门、输入门和输出门。

LSTM 有很多变体，其中门控循环单元网络[15](gated recurrent unit neural network，GRU)是效果比较好而且网络结构更简单的网络变体。GRU 由两个控制门组成，分别是更新门和重置门，GRU 网络结构如图 2 所示。

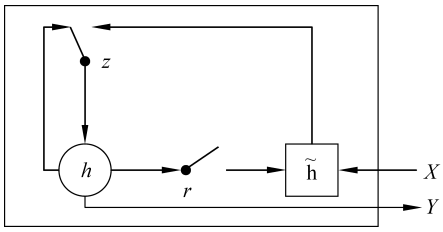

图 2　GRU 网络结构

更新门控制前一时刻的状态信息被带入到当前状态中的程度，更新值越大说明前一时刻的状态信息带入得越多，如式(2)所示。

$$r_t=\sigma(W_r\cdot[h_{t-1},x_t]) \tag{2}$$

重置门控制前一时刻状态信息的忽略程度，重置值越小说明忽略得越多，如式(3)所示。

$$z_t=\sigma(W_z\cdot[h_{t-1},x_t]) \tag{3}$$

其中，W 为权重矩阵，h_{t-1}为前一时刻隐含节点的输出，x_t 为输入，sigmoid 根据 h_{t-1}与 x_t 产生一个 0～1 之间的数值。

当前隐含节点的候选状态，如式(4)所示。

$$\tilde{h}_t=\tanh(W\cdot[r_t*h_{t-1},x_t]) \tag{4}$$

当前隐含节点的状态，如式(5)所示。

$$h_t=(1-z_t)*h_{t-1}+z_t*\tilde{h}_t) \tag{5}$$

LSTM[14]和 GRU[15]都能通过各控制门保留重要的信息，并且保证信息不丢失。但是，在相同的实验效果下，GRU 的网络结构更为简单，也更节省时间。

在实际问题中，当前的输出不仅跟上一时刻的输出有关，还跟下一时刻的输出有关。因此，为了充分考虑上下文信息，会采用正逆向的神经网络来解决此类问题。如果从左向右生成的正向隐含层的输出序列为$\{\vec{h}_1,\vec{h}_2,\cdots,\vec{h}_t,\cdots,\vec{h}_T\}$，从右向左生成的逆向隐含层的输出序列为$\{\overleftarrow{h}_1,\overleftarrow{h}_2,\cdots,\overleftarrow{h}_t,\cdots,\overleftarrow{h}_T\}$，那么在第$t$个时刻网络的输出为如式(6)所示。

$$h_t=(\vec{h}_t,\overleftarrow{h}_t) \tag{6}$$

3 标记集

本文将词干提取视为序列标注问题，因此在标注字符(乌兹别克语字母)过程中，为了更有效地表示和验证上下文语义关系，采用了BIO和BMES方法来标记词干与词缀，标记集合分别定义为{B-S,I-S,O}和{B-S,M-S,E-S,S,B-SS,M-SS,E-SS,S-S}。在{B-S,I-S,O}中，将每个字符分为三类：B-S(词干首字符)、I-S(词干中部)、O(非词干)；在{B-S,M-S,E-S,S,B-SS,M-SS,E-SS,S-S}中，将每个字符分为八类：B-S(词干首字符)、M-S(词干中间字符)、E-S(词干结束字符)、S(单字符成词干)、B-SS(词缀首字符)、M-SS(词缀中间字符)、E-SS(词缀结束字符)、S-S(单字符成词缀)，下面将举例说明：

“daraxtlarning(树的复数)”中词干为“daraxt(树)”，词缀为“larning”；使用BIO方法标记为“d/B-S a/I-S r/I-S a/I-S x/I-S t/I-S l/O a/O r/O n/O i/O n/O g/O”；使用BMES方法标记为“d/B-S a/M-S r/M-S a/M-S x/M-S t/E-S l/B-SS a/M-SS r/M-SS n/M-SS i/M-SS n/M-SS g/E-SS”。

4 实验

4.1 实验数据

由于到目前为止，未发现公开的乌兹别克语词干提取的数据集，因此本文构建了小规模的词干提取数据集。数据集源于乌兹别克斯坦宪法，共包括7 435个单词、568个句子，其中非重复单词1 986个。为了验证上下文语境对词干提取效果的影响，分别构建了单词级数据集DATA-SET1和句子级数据集DATA-SET2；数据集DATA-SET1由去重后的1 986个单词组成，不存在句子，因此没有语境；而对数据集DATA-SET2没有做任何处理，该数据集中包含7 435个单词和568个句子，具有一定范围内的上下文语境(滑动窗口大小为3，即考虑上一个词、当前词和下一个词)。实验数据采用交叉验证的方法获取训练集、测试集和验证集(切分比例为：0.8∶0.1∶0.1)，实验数据统计如表2所示。

表2 实验数据统计 (单词)

数据集	DATA-SET1	DATA-SET2
是否重复	否	是
是否带上下文	否	是
训练集	1 590	5 947
测试集	198	744
验证集	198	744
总计	1 986	7 435

通过进一步对实验数据分析，分别统计了数据集的相关特征，单词相关信息统计如表3所示。

表3 单词相关信息统计

类型	数量
最长单词	20
最短单词	1
平均词长	9
不同的词缀数	58
最长词缀	11
单词平均重复	3.7
最短词缀	1

4.2 实验设计和结果

为了对比不同模型和数据标注方法对词干提取的影响，首先复现了Lovins算法[2]，然后分别采用CRF模型和BiGRU模型进行实验。基于无监督学习的Morfessor算法和单向的GRU模型在相近的黏着语上的性能表现均弱于CRF模型[8]，本文未采用这两种模型进行对比实验。

4.2.1 不同标注方法的对比实验

本组实验中，以CRF为基本模型，分别使用BIO标注方法和BMES标注方法标注数据，实验结果如表4所示。BMES标记集的实验结果明显优于BIO标记集的实验结果。因此在后面的对比实验中都将采用BMES标记集。

表 4　不同标注集的 CRF 实验对比

数据格式	准确率	召回率	F_1 值
BIO	89.76	92.28	91.00
BMES	92.45	90.72	91.49

4.2.2　不同模型和不同数据集的对比实验

在标注集对比实验的基础上，分别在不同的数据集（DATA-SET1 和 DATA-SET2）上进行了 Lovins 算法、CRF 模型和 GRU 网络的对比分析实验，结果如表 5 所示。

表 5　实验结果

模型	数据集	准确率	召回率	F_1 值
Lovins	DATA-SET1	74.43	91.34	81.57
	DATA-SET2	77.35	90.55	82.89
CRF	DATA-SET1	92.45	90.72	91.49
	DATA-SET2	97.20	96.02	96.59
BiGRU	DATA-SET1	95.06	93.52	94.25
	DATA-SET2	99.57	98.53	99.05

1）模型对比实验中的发现

（1）基于 Lovins 算法的词干提取模型提取结果明显不如其他两种方法。实验结果说明，基于规则的方法使用规则库进行向前和向后匹配从而切分单词，由于语言规则无法覆盖所有的单词，规则之外的单词通常切分出错，因此采用基于规则的方法缺点较多，主要受限于规则库的规模以及规则之间的冲突。

在表 6 中，列出了常见的歧义现象，当词干尾部包含与词缀相同的字符时，例如，“Ertaga（明天）”单词中（此单词中没有词缀），尾部出现了与向格词缀一样的字符串“ga”，如果使用基于规则的方法切分词干与词缀，会出现过度切分情况；当出现多义词时，例如“turdi（起来【动词】，吐尔迪【人名】）”，如果不根据上下文语境来确定单词含义，会出现错误切分情况。由于语言歧义现象分布较离散，因此无法使用规则来进行映射。

表 6　常见歧义现象

例子	模型	切分结果	判断
Ertaga（明天）yakshanba. 明天是星期天。	Lovins	Ertaga＝Erta＋ga	错误
	BiGRU	Ertaga＝Ertaga	正确
Uo'rnidan *turdi*（起来）. 他站起来了。	Lovins	turdi＝tur＋di	正确
	BiGRU	turdi＝tur＋di	正确
Uning ismi *turdi*（吐尔迪）. 他的名字叫吐尔迪。	Lovins	turdi＝tur＋di	错误
	BiGRU	turdi＝turdi	正确

（2）基于序列标注的词干提取模型中，使用 BiGRU 神经网络模型获得了最优的提取效果。因此，相比于以往基于规则的方法，将词干提取视为序列标注任务进行处理结果较好。

2）数据集对比实验中的发现

不同的数据集，对基于规则的方法没有明显的变化，但是对序列标注模型而言，效果有明显提升。可能的原因是 DATA-SET2 是句子级别的数据集，模型在训练过程中学习当前单词的上下文信息，有助于提高模型性能。

此外，分析实验数据发现，主要的单词词缀源于名词，且词缀类型比较相似。由于数据采集的领域单一，导致基于序列标注方法的实验结果表现较好。但随着数据集的增多以及领域的扩充，歧义单词会出现更多、规则库在数据集上的覆盖率将更低，基于规则的方法性能可能会进一步降低。

5　结论

本文在不同数据集上采用 Lovins 算法、CRF 模型和 BiGRU 网络，对比了乌兹别克语的词干提取效果。实验结果表明：（1）在基于序列任务的词干提取中，采用 BMES 标记方法获得了更好的实验结果；（2）比起基于规则的词干提取方法，基于序列标注的方法对词干提取更为有效；（3）包含上下文信息的数据集有利于序列标注模型学习更多的信息。综上所述，本文提出的基于 CRF 模型和 BiGRU 网络的词干提取方法在乌兹别克语的词干提取任务上是有效的。由于语料领域的单一性，没有更深入地研究语言的歧义现象和音变现象等问题。在以后的研究中，我们将进一步考虑使用乌兹别克语的语言特征来提高词干提取的效果。

参考文献

[1] 吴思竹，钱庆，胡铁军，等. 词干提取方法及工具的对比分析研究[J]. 图书情报工作，2012，56(15)：109-115.

[2] Ismailov A，Jalil M M A，Abdullah Z，et al. A comparative study of stemming algorithms for use with the Uzbek language[C]//Proceedings of the International Conference on Computer & Information Sciences.

IEEE，2016：7-12.
[3] 艾孜海尔江，祖力克尔江，艾孜尔古丽. 基于多策略的乌孜别克语名词词干识别研究[J]. 中文信息学报，2018，32(09)：35-40.
[4] 苏依拉，赵亚平，牛向华. 基于统计的蒙汉机器翻译中词对齐方法研究[J]. 中文信息学报，2018，32(06)：44-51.
[5] 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.
[6] 徐春，杨勇，蒋同海.基于机器翻译的维吾尔语形态分析研究[J].计算机工程与应用，2017，53(14)：138-142.
[7] 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.
[8] 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.
[9] 古丽尼格尔·阿不都外力，吐尔根·依布拉音，卡哈尔江·阿比的热西提，等. 基于 Bi-LSTM-CRF 模型的维吾尔语词干提取的研究[J]. 中文信息学报，2019，33(8)：60-66.
[10] 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.
[11] Gulila ALTENBEK，Xiaolong WANG. Kazakh segmentation system of inflectional affixes[C]//Proceedings of the CIPS-SIGHAN Joint Conference on Chinese Language Processing. Red Hook，NY，USA，2010：183-190.
[12] Lovins J B. Development of a stemming algorithm. [J]. Mechanical Translation & Computational Linguistics，1968，11：22-31.
[13] Lafferty J D，Mccallum A，Pereira F C N. Conditional random fields：Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 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] Cho K，Van Merrienboer B，Gulcehre C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ，2014：1724-1734.

吾买尔江·买买提明(1994—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：107551700948@stu.xju.edu.cn

买合木提·买买提(1980—)，通信作者，博士，主要研究领域为自然语言处理及机器翻译。
E-mail：mahmutjan@xju.edu.cn

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：107556518131@stu.xju.edu.cn