文章编号：1003-0077（2011）00-0000-00

融合多策略的维吾尔语词干提取方法

赛迪亚古丽·艾尼瓦尔[1]，向露[2]，宗成庆[2]，艾克白尔·帕塔尔[1]，艾斯卡尔·艾木都拉[1]

（1.新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830046；

2. 中国科学院自动化研究所 模式识别国家重点实验室，北京 100190）

摘要：维吾尔语是形态变化复杂的黏着性语言，维吾尔语词干词缀切分对维吾尔语信息处理具有非常重要的意义，但到目前为止，维吾尔语词干提取的性能仍存在较大的改进空间。本文以 N-gram 模型为基本框架，根据维吾尔语的构词约束条件，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。实验结果表明，词性特征和上下文词干信息可以显著提高维吾尔语词干提取的准确率，与基准系统比较，融入了词性特征和上下文词干信息的实验准确率分别达到了 95.19%和 96.60%。

关键词：维吾尔语；形态；词干提取；N-gram 模型；词性特征；上下文词干信息

中图分类号：TP391　　**文献标识码：**A

Approach to Uyghur Stemmer Using Combination of Multi-strategies

Sediyegvl Enwer[1], Lu Xiang[2], Chengqing Zong[2], Akbar Pattar[1], Askar Hamdulla[1]

(1. Institute of Information Science and Engineering，Xinjiang University, Urumqi　Xinjiang　830046;

2. National Laboratory of pattern Recognition, Institute of Automation, Beijing 100190)

Abstract: Uyghur is an agglutinative language with complex morphology, Uyghur words stem segmentation plays an important role in Uyghur language information processing. But so far, the performance of the Uyghur words stem segmentation still has much room for improvement .According to the constraints of Uyghur word formation, we proposed a stem segmentation model for Uyghur which fuses the part of speech feature and context information based on N-gram model. Experimental results show that, the part of speech feature and the context information of stem can increase the performance of Uyghur words stem segmentation significantly with the accuracy reaching 95.19% and 96.60% respectively compared to the baseline system.

Key words: Uyghur; Morphology; Stem Segmentation; N-gram model; Part of speech; Context information

1　引言

维吾尔语属于黏着性语言，黏着性语言的构词和构形都是以词根、词干缀接不同的词缀来实现语法功能。每一个词的构成和其语法意义的表示都是依赖于不同词缀的缀接，每个词缀都有独立的语法意义，词缀不仅改变词根的词义，也会决定一个词在句子中的作用。任意词根追加不同的词缀（所属性，时态，复数）都会生成不同的新词。所以正确切分维吾尔语词干和词缀才能正确揭示其词类词性和语法关系。另一方面维吾尔语中同形异义词数量较多，这使得维吾尔语词干提取歧义现象严重。所以设计一个高准确率的维吾尔语词干提取系统，对维吾尔语信息处理的研究具有重要的意义。

维吾尔语作为黏着语，它的语法形式都是通过在单词原形的后面或前面附加一定的附加成分来完成的。这就造成在真实维吾尔文本中，一个维吾尔语词对应多个字符串的形式。由于词典的规模是有限的，所以这些不同的形式不可能都录用在词典中。所以，有必要找出词干与相应的附加成分的关系。并且，维吾尔语词切分中，除了词干提取以外还要进行词缀的切分。这是因为构形附加成分与词干互相黏连，并且构形附加成分也互相黏连。因为，构形附加成分往往可以表示一定词汇意义或语法意义，所以，如果不将这些黏连在一起的构形附加成分完整的切分开，不能准确的领会整个单词的含义。并且，构形附加成分还能表示词与词之间的关系。所以，切分构形附加成分是很有必要的。同时，构形附加成分的切分对句法分析、语义分析、语用分析等更深层的自然语言处理的应用都有很重要的意义。

维吾尔语属于阿尔泰语系突厥语族，是典型的黏着性语言，与汉语的字符顺次拼接的构词方法相比，日语、蒙古语、土耳其语和阿拉伯语等形态变

收稿日期：　　**定稿日期：**

基金项目：国家自然科学基金（61163032）；

化复杂的语言的构词规则更加复杂。词干提取在蒙古语，阿拉伯语，土耳其语等黏着性语言中与中文分词一样很重要。当前，阿拉伯语和其它黏着性语言的词法分析研究已经做到可用的水平，并取得了一定客观的成果：日语[1]、阿拉伯语[2]、蒙古语[3], 但对维吾尔语的词法分析研究起步比较晚，很多研究者提出了不同的方法。文献[4]提出了基于有限状态自动机和词典查询相结合的维吾尔语名词词干提取算法，此方法中由于维吾尔语的语音和谐，词缀与词干词尾相似导致过度切分的情况。文献[5]提出了最大熵模型和有限状态自动机相结合的维吾尔语词干提取方法。准确率已达到91.27%，这个方法对名词词干提取是有效的，但对其它词性的词语词干提取效果不理想。文献[6]提出了一个有向图模型来对维吾尔语进行词法分析，词干提取准确率达到94.7%，但是此模型会导致一个词有过多的非法候选，以致引入无谓的歧义。文献[7]提出了使用条件随机场的维吾尔语词干提取方法，这是一个纯统计的方法，准确率达到88.9%。

文献[8]提出了通过建立词干库、词缀库，规则和统计相结合的维吾尔语词干提取方法，词干词缀切分准确率可以达到95%。该工作提出了基于语素（包括词根和词缀）的n-gram语言模型的词干提取模型。此方法取得了较好的性能，但依赖于词干、词缀库，同时也存在切分过碎的问题。由于维吾尔语的构词和形态变化比较复杂，n-gram语言模型虽然可以取得一定的准确率，但是仍存在下述问题不能解决：

（1）对同一个词进行词干、词缀切分时，其词干出现歧义，如：

aldi（拿了）=al（拿）+di(第三人称单数，过去式词缀)；

aldi（前面）= aldi（前面）；

（2）词干的一部分被看成是词缀，出现错误切分,如：

dENiz（海）= dENiz（海）；

dENiz（海）=dE（说）+Niz（第一人称单数）；[错误切分]

（3）对同一个词进行词干、词缀切分时，其词缀出现错误，如：

ademler（人们的）=adem（人）+ler（第（二）三人称复数)；

ademler（人们的）=adem（人）+lar（第（二）三人称复数)；[错误切分]

（4）音变字母可以还原成不同的字母，而且都具有实际意义，如：

bErip（去了）=bar（去）+ip；

bErip（给了）=bEr（给）+ip；

为了解决上述问题，仅仅只考虑待切分的维吾尔语词本身和简单的词干、词缀统计信息是远远不够的，我们必须要考虑维吾尔语词本身的构词特点和语言特征。同时一个词语的意义往往受到特定上下文的影响，为了消除歧义切分，我们还必须考虑上下文信息。因此，在已有工作的基础上，我们提出了融合语言特征的词干提取模型：

1）在大规模文本语料库的基础上，对词干词性和词缀的连接形式进行统计，从而得到词干词性-词缀结构的初步表达模式。这样可以通过词干词性和词缀的连接模式解决过度切分和词干词缀连接形式不合法的问题。

2）利用大规模文本语料库来学习上下文词干之间的转移概率并利用此转移概率作为选择最优切分的依据，从而可以解决维吾尔语词切分歧义的问题。

本文第 2 节介绍维吾尔语构词特点。第 3 节介绍维吾尔语词词干提取方法。第 4 节是实验和结果分析。第 5 节是结论。

2 维吾尔语构词特点

维吾尔语是一种黏着语言，与汉语和英语有很大不同，词与词之间以空格隔开，具有比较复杂的形态变化。按附加词根的位置，附加成分有前接附加成分前缀和后接附加成分后缀，其中多数附加成分为后接附加成分，只有少数为前缀附加成分。在维吾尔语中，语音和语义结合的最小单位是语素，语素是由一个或一个以上的语素组成的，他们都有一定的意义或语法意义。维吾尔语的语素可以分为三类，即词根、构词附加成分和构形附加成分。维吾尔语单词的组成形式是“$\text{prefix} + \text{stem} + \text{suffix}_1 + \text{suffix}_2 + \cdots + \text{suffix}_n$”，结构如图 1 所示。

分界点　　分界点

（前缀）　∧　词根　　∧　复合词缀

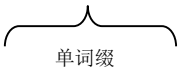


图1　维吾尔语单词的结构

其中，prefix 是前缀，stem 是词干，$\text{suffix}_1 + \text{suffix}_2 + \cdots + \text{suffix}_n$ 是复合词缀，$\text{suffix}_i\ (i = 1,2,\cdots,n)$是单词缀，复合词缀是由多个单词缀连接构成的。附加成分的追加成分是多层次的，表现出不同的形态和不同的语法意义。

如：

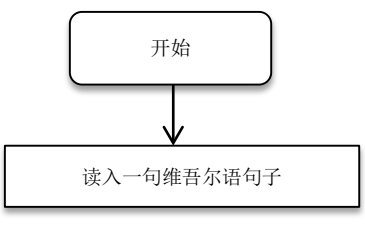
yaz	写（词干）
yazdim	我写了
yazalidim	我能写了
yazganliktin	由于（她）写了
yazguzganidim	我让别人写了
…	…

由于维吾尔语词是字符序列，词干和词缀之间没有明确的间隔标记。维吾尔语词缀种类多，连接形式各式各样，而且有限的词干附加各种词缀可以生成无限的新词并表示不同的语法意义，这些构词特点大大加剧了维吾尔语信息处理的难度。为了提高维吾尔语信息处理的性能，进行维吾尔语词干提取工作具有重要的现实意义。

3 维吾尔语词干提取方法

3.1 维吾尔语词干提取过程

本文对维吾尔语进行词干提取的流程如图 2 所示。



第一步，读入一个维吾尔语句子。

第二步，使用正向匹配和逆向匹配算法对句子中的每个词进行切分得到词的切分候选集合。具体步骤为[9]：

（1）如果待切分词有前缀，使用正向匹配，切分出前缀。

（2）接上一步，使用正向匹配对剩余部分进行切分，将剩余部分切分成词干和复合词缀的形式。

（3）使用逆向匹配法，将复合词缀切分成单词缀的形式。

（4）将待切分词写成“前缀+词干+单词缀”的形式。

第三步，利用词性特征检查切分候选集合中词干、词缀连接的合法性，并计算权重。

第四步，使用 n-gram 统计语言模型算出每个切分候选的概率[9]。

用变量W代表文本中一个词的任意语素序列，它

由顺序排列的 n 个语素组成，即$\mathrm{W}=w_1w_2\ldots w_n$，$w_i$是词的任意一个语素，则该语素序列 W 的概率可以展开为：$\mathrm{P(W)}=\mathrm{p}(w_1)\mathrm{P}(w_2|w_1)\mathrm{P}(w_3|w_1w_2)\ldots\mathrm{P}(w_n|w_1w_2\ldots w_{n-1})$

一般，用二元语言模型来计算每个切分候选的概率，这里，词干-词缀的边界是切分的重要部分：

$$\mathrm{P(stem-suffix\ boundary)}=\frac{c(stem,firstSuffix)}{c(stem)}$$

第五步，通过上下文词干信息计算出相邻词干的转移概率。

最后，结合第四步和第五步的概率值选出最优切分结果。

3.2 维吾尔语词性特征

维吾尔语词语分为 12 类，12 类里面 7 个是实词（主要分类），5 个是虚词（助词类）。实词分为动词和静词，静词包括名词、形容词、数词、量词、代词、副词和拟声词等词类。虚词包括后置词、连词、感叹词和语气词 [10]。

本文我们收集训练语料库中的词干，对它们进行人工词性标注，如：名词、数词、动词和形容词等。标注的时候如果一个词干同时属于两个或两个以上的词性时，检查词干在训练语料库中哪个词性类的出现次数高来判断词干的词类。维吾尔语中有大量的词缀成分，以上词类都可以附加各自的词缀和有自己特定的词缀，而且不同词性的词干能够连接的词缀不同 [11]。由于维吾尔语附加成分的追加形式多变复杂，种类繁多，连接形式各式各样 [12]，所以我们将从训练语料中集中学习每个词类词干各自的单词缀信息，有了词干词性信息和词缀信息，我们就可以为词干词缀的连接形式设计出初步的知识表达模式了。

通过收集 9025 条各个领域的句子（包括小说，新闻稿，科学读物等）作为训练语料库，统计出 11,114 个词干，313 个单词缀，并对所有词干进行词性标注，学习出了 12 类词干能连接地单词缀分布情况。见表 1.

表 1 词性-词缀分布统计

序号	词干类别	占所有词干的比例/%	能连接的单词缀的比例/%
1	名词	79.14	78.59
2	形容词	8.62	64.85
3	动词	8.87	45.36
4	数词	1.63	15.65
5	量词	0.2	3.8
6	副词	0.6	18.12
7	代词	0.4	24.28
8	后置词	0.1	9.58
9	连词	0.6	16.61
10	感叹词	0.2	1.59
11	语气词	0.2	11.5
12	模拟词	0.03	0

从表 1 可以看出，维吾尔语中每个词类的词干只能连接部分词缀，以上词类可以附加各自的词缀以及都有各自特定的词缀，不同词性的词干能够连接的词缀集合不同。以名词类词干为例，在整个 11,114 个词干中有 79.14%的词干是名词性词干，名词性词干的数量很多，而在 313 个单词缀中能缀接在名词性词干后面的词缀占 78.59%，剩下的 21.41%的词缀不能连接在名词性词干后面。此外，模拟词不需要进行词干提取。这个约束条件有利于检查一个单词缀是否能够合法地连接在某一词性类的词干后面，从而可以降低词干-词缀连接错误的问题。

根据上述的语言约束条件，我们可以初步设计词性词干-词缀的连接模式。本文提出的切分规则定义如下：假设一个维吾尔语词语“W（$S_1S_2\ldots S_n$）”，W，提出，$S_1\in$ Td，其中 W 是词干，Td 是单词缀库，$S_1S_2\ldots S_n$是复合词缀，Wd是词性词干表，S_1为词干 W 连接的第一个词缀。如果词干 W 为某个词性词干表中的词干，且S_1是单词缀库词缀且满足该词性词干对词缀的要求，那么使用公式（3-1）计算维吾尔语词每种切分的概率值。

$$P_{word}=\lambda P_{\mathrm{W}}+\ （1-\lambda）P_{S_1S_2\ldots S_n} \qquad （3\text{-}1）$$

其中，

$$\lambda=\mathrm{P}\left(\mathrm{Pos}_i,\ \mathrm{suffix}_i\right)=\frac{\mathrm{count}(\mathrm{Pos_i},\mathrm{suffix_i})}{\sum\mathrm{count}\left(\mathrm{suffix_{pos}}\right)}$$

其中，Pos_i是当前词干的词性，suffix_i是缀接在当前词干后面的单词缀，$\sum\mathrm{count}\left(\mathrm{suffix_{pos}}\right)$是当前词性能连接的所有词缀在训练语料中出现的次数，$\mathrm{count}(\mathrm{Pos}_i,\mathrm{suffix_i})$是当前词性词干连接当前词缀在

训练语料中出现的次数。如果当前词干不在词性词干表中时，为了避免数据稀疏问题，我们将给λ赋一个极小值。P_W和$P_{S_1S_2\cdots S_n}$分别是用 n-gram 语言模型计算的词干和词缀的概率值。

例如，维吾尔语词 birlexme(联合)本身是一个名词词干，通过前向逆向匹配法可以得到该词的 5 个候选切分，我们可以对这 5 个候选切分进行词干词缀连接合法性的检查，如表 2 所示。

表 2 词性特征

序号	切分候选	词干	词干词性	第一个词缀	是否合法
1	birlexme	birlexme（联合）	名词	-	是
2	birlex + me	birlex（结合）	动词	me	是
3	bir + lex + me	bir（一）	数词	lex	否
4	ber + lex + me	ber（给）	动词	lex	否
5	bar + lex + me	bar（有）	动词	lex	否

从表 2 可以看出，当 birlexme 切分成 birlex+me 时，由于 birlex 是动词性词干，词缀 me 可以连接在动词性词干后面，因此 birlex+me 这种切分是合法的；而当 birlexme 切分成 bir+lex+me 时，bir 是数词性词干，由于词缀 lex 不能连接在数词性词干后面，因此这种切分是不合法的。因此，词性词干-词缀连接形式可以有效的减少非法候选导致的歧义性问题。

3.3 上下文词干信息

维吾尔语词汇中同形异义词较多，出现频率较高，而且同一个词在不同上下文中切分结果是不同的。如：

uniN ismi turdi（他的名字叫 吐尔地 ）

u ornidin turdi（他站起来了）

其中，单词 turdi 在两个句子中形式是一样的，但是在第一句中 turdi 是一个人名，词干就是其本身。而在第二句中 turdi 是由词干 tur 加词缀 di 构成的，并且词干词缀的连接形式是合法的。如果不考虑上下文信息，仅仅简单地使用统计方法对 turdi 进行词干词缀切分会得到 tur+di 的切分结果，而这种切分结果在第一句的上下文环境中是不正确的。对于这类问题，我们可以利用上下文词干信息来帮助找出正确的切分结果，从而解决维吾尔语词切分歧义的问题。

在训练语料库中，我们利用词干转移概率来捕捉上下文信息。由于维吾尔语词汇量的庞大，就算是再大的语料库，也很难学习上下文词之间关系，所以本文利用上下文词干转移概率来捕捉上下文信息，而不是上下文词的转移概率。相邻两个词干之间的转移概率由公式（3-2）来计算：

$$\mathrm{P}(W_2|W_1)=\frac{P(W_1,W_2)}{P(W_1)}=\frac{\mathrm{Count}(W_1,W_2)}{\mathrm{Count}(W_1)} \quad (3\text{-}2)$$

其中，W_1和W_2是句子中相邻两个词的词干，如图 3 所示，$\mathrm{Count}(W_1,W_2)$是W_1和W_2的共现次数，$\mathrm{Count}(W_1)$是W_1出现的次数。通过转移概率，我们就能获知在词干W_1出现的情况下W_2出现的概率，从而帮助我们找到在特定上下文中一个维吾尔语单词最有可能的切分。

如图 3 所示，$word_1,word_2,\cdots,word_n$是一个维吾尔语句子，$W_1W_2\cdots W_n$是该维吾尔语句子中各单词的词干，$W_i$表示词干，$S_{ij}$表示连接在$W_i$后的第j个词缀。那么，一个维吾尔语句子最优切分的概率由公式（3-3）计算：

$$\mathrm{P}(word_1,word_2,\ldots,word_n)=argmax\{\sum_{i=1,2,\cdots,n}P_{word_i}+\sum_{i=1,2,\cdots,n-1}P(W_{i+1}|W_i)\} \quad (3\text{-}3)$$

其中，P_{word_i}是利用公式（3-1）计算的每个单词的切分概率，$P(W_{i+1}|W_i)$是相邻两个单词词干的转移概率。

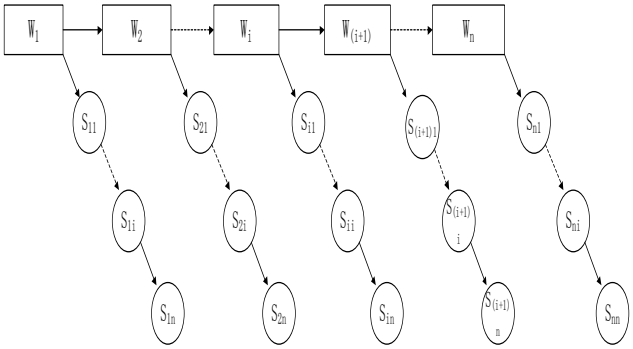


图 3 维吾尔语句子结构分析图

例如，对于维吾尔语句子 uniN ismi turdi（他的名字叫 吐尔地 ），一共有三个单词，其中 turdi 在不同上下文中有不同的意思。首先用前向、后向匹配算法得到每个单词的切分候选，然后检查词干词缀连接的合法性并对每个候选中的词干部分进行排列组合得到 12 种切分候组合，如图 4 所示。之后我们

就利用公式（3-3）找出这 12 种切分组合中的最优切分作为最终结果。

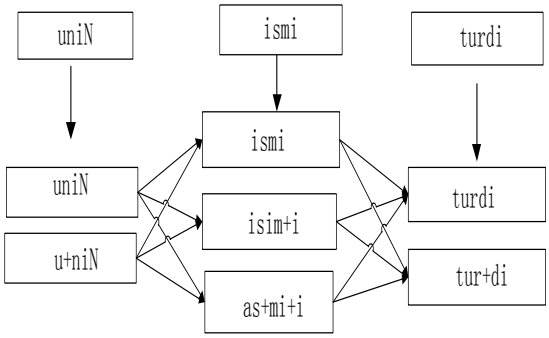


图 4 切分候选组合图

在不考虑上下文信息的情况下，我们得到的切分结果是 u+niN isim+i tur+di，其中最后一个单词出现切分错误，而当我们引入上下文词干信息后就得到了正确的切分结果 u+niN isim+i turdi。因此，考虑上下文词干信息可以有效的解决维吾尔语词切分歧义的问题。

4 实验和结果分析

4.1 实验设置

本文采用的数据是我们人工标注的 10,025 条维吾尔语句子，使用其中的 9025 句作为训练语料，剩下的 1000 作为测试语料，语料统计情况如表 3 所示。

表 3 语料统计情况

	句子个数	单词个数
训练语料	9025（句）	123788（词）
测试语料	1000（句）	6737（词）
OOV	-	2489（词）

维吾尔语的词法分析比较复杂，考察的方面较多，为了能够更好地体现系统性能，我们使用了以下指标来考察系统的性能:

1.词干级正确率Pstem

以词干为单位，仅考察词干是否被正确提取，而不考虑词缀的情况；

2. 词级准确率Pword

以词为单位，仅当词内词干正确，且各词缀切分正确时，才认为分析正确；

本文采用切分准确率的定义如（4-1）式所示：

$$P = \frac{\text{正确切分词数}}{\text{总词数}} \times 100\% \qquad (4\text{-}1)$$

4.2 实验及结果分析

实验 1: 针对词干级和词级的实验

实验设置：在同一个测试集上，我们分别使用了 N-gram 模型，N-gram 模型+词性特征以及 N-gram 模型+词性特征+上下文词干信息来进行词干词缀切分的实验。实验结果如表 4 所示。

表 4 实验结果

	N-gram 模型	N-gram 模型+词性特征	N-gram 模型+词性特征+上下文词干信息
Pstem	95.04%	95.35%（+0.31）	97.09%(+2.05)
Pword	95.02%	95.19%(+0.17)	96.60%(+1.58)

实验结果分析：

从表4可以看出，不同的系统对维吾尔语的词干级分析能力和词级分析能力有所不同。在N-gram模型的基础上，加入了词性特征之后，不论在词干级别还是在词级别，系统的性能均有一定的提升（分别达到了95.35%和95.19%）。而这个提升是有限的，通过错误分析，我们发现这主要是由于维吾尔语中存在很多同形异义词，从而导致标注歧义。例如:对于kokrek一词，当这个词表示“蓝一点”时是形容词，形容词词干“kok”连接词缀“rek”是合法的，而该词还可以作为名词，表示“胸部”。当在词性特征的基础上进一步加入上下文词干信息后，我们发现与基线系统相比较，我们系统的性能有了显著的提升，在词干级别和词级别分别提升了2.05% 和 1.58%。这表明了本文提出的词性特征和上下文信息能够显著提升维吾尔语词干提取的性能。通过对数据的分析，我们发现加入上下文词干信息后，能够更有效的解决对同一个词进行词干、词缀切分时，其词干出现歧义，词干的一部分被当作词缀等问题。

实验 2：针对语料库规模对系统性能影响的实验

实验安排:固定测试集不变，而从训练集中每次提取不同规模的子集训练三个不同的系统，并考察各个系统在测试集上的表现。整个训练集含 9025 条句子，我们分别取训练集的 5%，10%，30%，50%及80%等不同规模的子集来分别训练三个切分系统，

并对测试集进行切分。实验的评价标准是准确率。图 4 为不同系统准确率随训练规模增加的变化曲线。

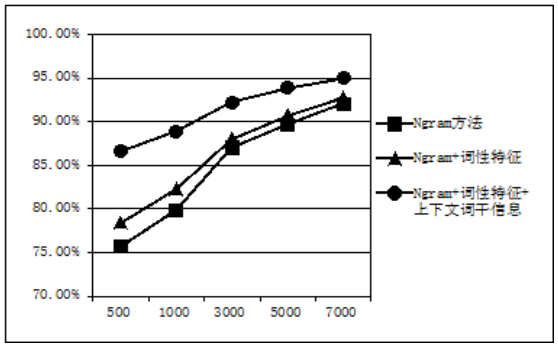


图 4 训练语料规模与系统准确率曲线

实验结果分析：

分析曲线可以发现，训练数据的规模会影响系统的性能。训练集由总规模的 5%提高到 30%时系统准确率明显提高，同时随着语料库规模增大系统准确率提高的幅度趋于缓慢；而在不同规模的训练集上，融入词性特征后，系统的性能均有一定的提高，说明词性特征对于维吾尔语词干的提取是有用的；而进一步加入上下文信息后，系统的性能有了更加明显的提升，这更进一步说明了上下文信息能够有效提高维吾尔语词干提取的性能。

5 结束语

本文提出了基于多策略融合的维吾尔语词干提取方法，我们以 N-gram 模型为基本框架，根据维吾尔语的构词约束条件，提出了融合词性特征和上下文词干信息的词干提取模型。实验结果表明，词性特征和上下文词干信息可以显著提高维吾尔语词干提取的准确率，与基准系统比较，融入了词性特征和上下文词干信息的实验准确率分别达到了 95.19%和 96.60%。

由于本文实验所使用的语料库规模较小，能够使用的词性特征和上下文信息有限，维吾尔语词干提取的整体效果有待进一步提高。

下一步工作中，我们要更加深入地了解维吾尔语词汇的内部构件特征，学习更多的词干-词缀和词干信息，通过词干提取结果的错误分析，进一步修正系统，最终提高维吾尔语词干提取结果并将系统运用到各种领域和网络语言中。

参考文献

[1] Nagata, Masaaki, A stochastic Japanese morphological analyzer using a forward- DP backward-A* N-best search algorithm, Proceedings of the 15th conference on Computational linguistics-Volume 1, pp. 201–207, 1994, Association for Computational Linguistics.

[2] Buckwalter, Tim, Buckwalter fArabicg Morphological Analyzer Version 1.0, 2002.

[3] 姜文斌,吴金星,乌日力嘎等. 蒙古语有向图形态分析器的判别式词干词缀切分[J]. 中文信息学报,2011,25（04）:30-34.

[4] 早克热·卡德尔，艾山等. 维吾尔语名词构形词缀有限状态自动机的构造[J].中文信息学报，2009, 23(6): 116-121.

[5] 古丽拉·阿东别克，米吉提·阿布力米提.维吾尔语词切分方法初探[J].中文信息学报，2004，18（6）:61-65.

[6] 麦热哈巴·艾力，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）:3115-3129

[7] Aisha B. A Letter Tagging Approach to Uyghur Tokenization[C]// 2010 International Conference on Asian Language Processing：IEEE Computer Society, 2010:11-14.

[8] Ablimit M, Eli M, Kawahara T. Partly supervised Uyghur morpheme segmentation. In: Proc. of the Oriental-COCOSDA Workshop.2008. 71−76.

[9] 米吉提·阿布力米提,库尔班·吾布力. 在多文种环境下的维吾尔语文字校对系统的开发研究[J]. 系统工程理论与实践,2003,05:117-124.

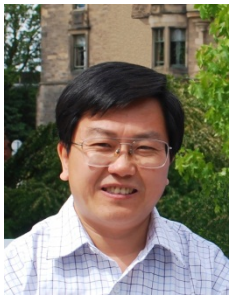
[10]哈力克·尼亚孜.基础维吾尔语[M].乌鲁木齐:新疆大学出版社. 1997: 73.

[11]哈米提·铁木尔著.现代维吾尔语语法[M].北京:民族出版社. 1987: 47 -48.

[12]米热古丽·艾力，米吉提·阿不力米提，艾斯卡尔·艾木都拉.基于词法分析的维吾尔语元音弱化算法研究.中文信息学报[J]. 2008，04:43-47.

赛迪亚古丽·艾尼瓦尔（1988-），女，硕士研究生，主要研究领域为自然语言处理。Email：852567316@qq.com

向露（1988-），女，助理工程师，主要研究领域为自然语言处理。Email：lu.xiang@nlpr.ia.ac.cn



宗成庆（1963-），男，研究员，主要研究领域为自然语言处理、机器翻译和情感分类。Email：cqzong@nlpr.ia.ac.cn

艾克白尔·帕塔尔（1958-），男，副教授，主要研究领域为维吾尔语词法分析。Email：akbarpattar@gmail.com

《通信作者》艾斯卡尔·艾木都拉（1972-），男，博士，教授，主要研究领域为语言文字信息处理。Email：askar@xju.edu.cn