文章编号：

黏着语形态分析的图状建模方法

徐春[1,2,3,4]，蒋同海[1,2]，阿布都热合曼·卡的尔[4]

（1.中国科学院 新疆理化技术研究所，新疆 乌鲁木齐 830011；
2.新疆民族语音语言信息处理重点实验室，新疆 乌鲁木齐 830011；
3.中国科学院大学， 北京 100049；
4.新疆财经大学 计算机科学与工程学院，新疆 乌鲁木齐 830012）

摘 要：为黏着语形态分析建立了一种图状结构的判别式模型，该模型将黏着语语句的形态分析结果建模为形态成分的图状结构，通过灵活丰富的特征设计描述了词语内部形态成分之间以及分属相邻词语的形态成分之间的关联约束。相比传统的线性模型，图状模型更好地考虑了各形态成分之间的语言学关联，从而有望取得更高的整句分析性能。在韩语和维吾尔语上的实验结果表明，图状模型相比线性模型取得了显著的性能提升，形态分析词级准确率分别提升了2.8和4.4个百分点。

关键词：形态分析；黏着语；图状模型；线性模型

中图分类号：TP391.1 **文献标志码：**A

Graphic Modeling Method for Morphological Analysis of the Agglutinative Language

XU Chun[1,2,3,4], JIANG Tonghai[1,2], Abdurahman Kadir[4]

(1.Xinjiang Technical Institute of Physics & Chemistry，Chinese Academy of Sciences,Xinjiang,Urumqi 830011，China；
2.Xinjiang Key Laboratory of Minority Speech and Language Information Processing,Xinjiang,Urumqi 830011,China;
3.University of Chinese Academy of Sciences，Beijing 100049，China；
4.College of Computer Science and Engineering, Xinjiang University of Finance and Economics, Xinjiang,Urumqi 830012，China)

Abstract: A discriminant model of the graphic structure is established for the morphological analysis of the agglutinative language. The model models the morphological analysis of the agglutinative language sentences into the graphic structure of morphological components, and describes the correlation between the morphological components of the words inside and the morphological components of the adjacent words through flexible and rich feature design. Compared with the traditional linear model, the pattern model is better to consider the linguistic association between the morphological components, and it is expected to achieve higher sentence analysis performance. The experimental results in Korean and Uyghur indicate that the graphic model achieves significant performance improvement comparing with the linear model, and the word level accuracy of morphological analysis respectively increases by 2.8 and 4.4 percentage points.

Key words: morphological analysis; agglutinative language; graphic model; linear model

1 引言

形态分析是自然语言处理任务的基础。对于形态复杂的黏着语如东亚的韩语和中国的维吾尔语，其词语形态和构词规律比较复杂，目前形态分析准确率仍有较大的提升空间[1,2]。对具有结构分析性质的自然语言处理任务而言，设计恰当的模型是取得较好精度的至关重要的前提。亚洲语言所特有的词语切分任务，意在将字符序列拆分为词语序列，这

收稿日期：2017-06-10； **定稿日期：**2017-07-25
基金项目：新疆自治区自然科学基金项目(2015211B034)；新疆自治区重点实验室开放课题(2015KL031)；新疆自治区重大科技专项课题(2016A03007-3)；国家自然科学基金项目(61363082)；国家自然科学基金项目(61662073)；
作者简介：徐春（1977—），女，博士研究生，副教授，主要研究方向：自然语言处理、无线光网络；蒋同海（1963—），男，博士，研究员，主要研究方向：自然语言处理、人工智能；阿布都热合曼·卡的尔（1975—），博士，副教授，主要研究方向：人工智能。

是线性序列单向维度的结构分析，序列标注的经典模型如隐马尔科夫模型和条件随机场模型即可取得较好的精度。依存句法分析和成分句法分析任务，意在将线性的词语序列，分析为层次化的树状结构，这是更为复杂的两个维度上的结构分析，简单的序列标注模型在这里不再适用，更能反映语言结构化规律的模型如生成树模型可以取得更好的分析精度。黏着语的形态分析也是一类结构分析任务，它具有怎样的语言学和结构化特点，又该采用何种恰当的模型呢？与构词方式较为简单的英语相比，韩语和维吾尔语等形态丰富的黏着语，其构词规律更加复杂。这类语言的词语通常由词干和若干起修饰作用的词缀等两种形态成分构成，形态分析的任务就是解析出词语的词干和词缀结构，并且标定出它们的类别。在黏着语语句中，不仅词语内部的各形态成分之间存在关联约束，相邻词语的形态成分之间也存在语言学关联。因此，在英语和汉语上取得良好效果的序列标注模型[3, 4, 5]在这里变得不太适用。而研究者往往直接借用这些现成的线性序列模型，同时将任务限定为粗切分或标注，这使得系统的理论价值和实用性大打折扣。另一方面，传统的基于规则的形态分析模式[1]需要专门的语言学人才，往往耗费大量的精力调试搭建，而准确率和稳定性又不尽人意。因此，有必要构造更为恰当的统计模型，以更好的反应黏着语的语言学和结构化特点，尽可能准确地描述形态丰富语言的词法和句法规律，从而为黏着语搭建高性能的形态分析系统。

本文为黏着语形态分析建立了一种图状结构的判别式模型，该模型将黏着语语句的形态分析结果建模为以形态成分为节点的图状结构，并通过灵活丰富的特征设计，以图中的边描述词语内部形态成分之间以及分属相邻词语的形态成分之间的关联约束。具体地，在图状模型中，每个词语内部各词干和词缀之间都存在相应的边，对应相应的近距离特征；分属相邻词语的词干和词缀之间也存在相应的边，对应相应的远距离特征。这两类特征分别描述了词语内部和词语之间各形态成分之间的语言学关联约束关系。与传统的线性模型相比，图状模型更好地考虑了各形态成分之间的语言学关联，从而有望取得更高的整句分析性能。在为每个词枚举可能的形态分析候选时，借鉴了汉语分词中基于字符分类的序列标注模型。与传统的基于语言学规则的形态拆分相比，字符分类模式不依赖于语言学家的规则设计，从标注语料库中自动学习到的形态拆分规律具有更好的覆盖度和泛化性。黏着语普遍存在的音变现象，即由形态成分构成词时发生诸如元音脱落等词形变化，因此，需要在形态分析之前还原出各形态成分的原始形态即音变还原。同样地，为摆脱对语言学规则设计的依赖，将音变还原也建模为字符分类问题。通过动态规划的字符对齐算法，可以从标注语料中自动化地获取音变还原规律，并抽取相应的判别式分类实例，从而自动化地构建高效的音变还原模型。音变还原结果作为形态拆分的输入，形态拆分为上层的图状模型提供语句中各词语的形态分析候选。

在韩语和维吾尔语上进行实验。针对韩语，采用来自网络开放语料和手工标注的共 54358 句形态标注语料作为实验数据。针对维吾尔语，采用新疆大学多语种信息重点实验室开发的 72739 句形态标注语料作为实验数据。实验结果表明，在两种语言上，图状模型相比线性模型均取得了显著的性能提升。在韩语上，形态分析词级准确率提升了 2.8 个百分点，从 91.01%提升到 93.79%；在维吾尔语上提升了 4.4 个百分点，从 91.26%提升到 95.67%。

2 形态分析图状模型

描述图状模型的结构，就要与传统的线性建模方法进行对比。图状建模的优点在于，它可以刻画更大范围内形态元素之间的关联约束关系，比只考虑邻近元素的线性建模方法性能更好。在图状建模的基础上，如何进行有效地特征设计是关键。

2.1 图状建模

最直观的建模方式是将句中词语的形态元素表示为一个线性序列，如图 1 所示。对于一个黏着语词语，词干前面或后面可能缀接一个或多个词缀。这种表示方法使得词干之间的距离变长，同一词内不相邻形态元素之间的关联无法进行直接有效地刻画，更不必说不同词之间相隔更远的形态元素了，

相当程度上弱化了统计模型的能力。

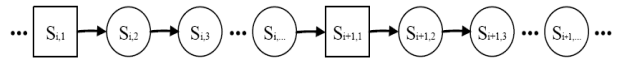

图 1 形态切分线性建模

还有一种直观的方式是将之建模为树状结构，如图 2 所示。该建模方式的优点是拉近了相邻词词干的距离，使得词干之间的语言学关联，可以通过对词干元组设置参数或设计特征的方式进行有效地建模。相比线性建模，树状建模更好地考虑了词干之间的联系。但事实上，对于大多数的黏着语，邻近词的词缀与词干之间以及词缀与词缀之间同样存在语言学关联。

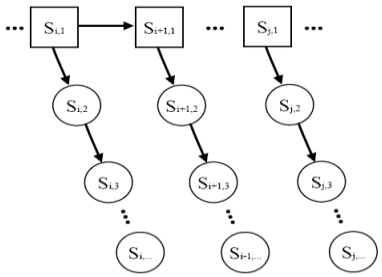

图 2 形态切分树状建模

本文提出了一种可以有效刻画邻近词中任意形态元素语言学关联的图状建模方式，如图 3 所示。在图状模型中，节点代表形态元素包括词干和词缀，节点之间的边表示形态元素之间的关联关系。对于同一词内和相邻词间的任意一对形态元素，都存在直接的边进行关联。因此，图状建模可以视为前面提到的树状建模扩展，而树状建模又可以视为线性建模的扩展。借助图状建模，可以针对可能的形态元素间的关联关系，自由灵活地设计相应的参数（生成式方法）或特征模板（判别式方法）。相比于生成式方法，更便于判别式方法通过丰富灵活的特征设计，充分挖掘黏着语的构词和构句规律；即以判别式方法进行形态分析图状模型的建模、训练和解码。

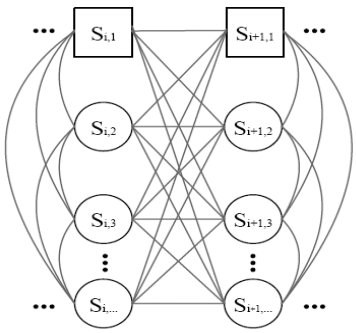

图 3 形态切分图状建模

给定待分析语句x，判别式图状模型的解码算法输出最可能的分析结果$\tilde{y}$：

$$\tilde{\boldsymbol{y}} = \underset{y \in \mathrm{GEN}(x)}{argmax} S(y|\vec{a}, \Phi, x)$$

$$= \underset{y \in \mathrm{GEN}(x)}{argmax} f(x,y) \cdot \vec{a} \quad (1)$$

其中，$\vec{\boldsymbol{\alpha}}$是模型的训练算法输出的权重向量，$f(x,y)\cdot\vec{\boldsymbol{\alpha}}$是特征向量$f(x,y)$和权重向量$\vec{\boldsymbol{\alpha}}$的内积。权重向量通过一定的训练算法习得，采用感知机算法，如表 1 所示:

表 1 感知机训练算法

Algorithm 1	
Input:训练实例$(\boldsymbol{X}, \boldsymbol{X})$	
$\vec{\boldsymbol{a}} \leftarrow 0$	
for $t \leftarrow 1..T$ do	▷T 轮迭代
for $(x,y) \in (X,Y)$ do	
$\tilde{\boldsymbol{y}} \leftarrow argmax_{z \in GEN(x)} \Phi(x,z).\vec{\boldsymbol{a}}$	
if $\tilde{\boldsymbol{y}} \neq y$ then	
$\vec{\boldsymbol{a}} \leftarrow \vec{\boldsymbol{a}} + \Phi(x,y) - \Phi(x,\tilde{\boldsymbol{y}})$	▷更新权重
end if	
end for	
end for	
Output:权重向量$\vec{\boldsymbol{a}}$	

感知机训练过程旨在学习一个从输入$x \in X$到输出$\mathrm{y} \in Y$的判别式模型，其中X是训练语料的句子集合，Y是对应的标注结果集合。函数$\mathrm{GEN}(x)$用以枚举当前输入语句的候选分析结果，如公式 2：

$$\mathrm{GEN}(x) = \{(y1, y2, \ldots, y_{|x|}) | y_i \in \mathrm{GEN}(x_i), 1 \le i \le |x|\} \quad (2)$$

整个语句的候选结果集合由句中各词的候选结果集通过笛卡尔乘积得来。函数f将训练实例$(x,y) \in (X,Y)$映射到对应的特征向量$f(x,y) \in R^d$，其中d是特征向量的维度。为获得更好的训练效果，在上述算法的基础上采用平均参数技术[6]。

2.2 特征设计

相比于参数训练算法，特征设计通常是影响系统最终性能的更关键因素。能够正确反映句中形态元素之间的语言学关联约束，具有良好的描述能力和泛化能力的特征模板，正是特征设计所要解决的问题。基于之前描述的形态分析图状模型，图中的任意一个节点对应为一项一元特征，一条边及其连接的两个节点对应为一项二元特征。一元特征描述的是一个形态元素作为特定词句法角色的事件，二元特征则描述两个形态元素作为相应词句法角色的事件之间的关联约束。采用的特征模板如表 2 所示：

表 2 形态分析图状模型特征模板

类型	特征模板	
1 阶	$S_{0,i} \circ P(i, S_0)$	$1 \le i \le \lvert S_0 \rvert$
	$S_{0,i} \circ S_{0,j} \circ D(i, j, S_0)$	$1 \le i < j \le \lvert S_0 \rvert$
2 阶	$S_{-1,i} \circ P(i, S_{-1}) \circ S_{0,j} \circ P(j, S_0)$	$1 \le i \le \lvert S_{-1} \rvert,\ 1 \le j \le \lvert S_0 \rvert$
3 元组	$S_{*,i-2} \circ S_{*,i-1} \circ S_{*,i}$	$S_* = [S_{-2}, S_{-1}, S_0],\ \ \lvert S_* \rvert - \lvert S_0 \rvert \le i \le \lvert S_* \rvert$

其中，$\boldsymbol{S_0}$、$\boldsymbol{S_{-1}}$和$\boldsymbol{S_{-2}}$为由形态元素构成的向量，

分别表示当前词及其前面两个词的候选分析结果，S_*表示这三个向量顺序拼接而成的向量。函数P表示形态元素在分析结果向量中的相对位置，对于向量中的首尾元素，函数分别返回 L（left）和 R（right），对于其他元素则返回 M（middle）。函数D返回形态切分结果向量中两个形态元素的距离，对于相邻的形态元素返回 N（near），对于不相邻的情形则返回 F（far）。

3 词语形态候选生成

为待分析语句中的每个词语枚举出可能的形态分析候选，是形态分析图状模型的前序阶段。给定可能的词干和词缀集合，一个词的形态分析候选可以通过简单的递归枚举求得。但这带来两方面的问题：对于形态切分歧义严重的词，可能会生成过多的不合理形态分析候选；而对于未登录词，则无法生成正确的形态分析候选。本节描述一种高效且具有泛化能力的形态拆分算法，该方法借鉴基于字符分类的汉语分词原理，将词语形态切分建模为词内字母的序列标注问题。

对于许多黏着语而言，音变现象广泛存在，即由形态成分构成词时发生诸如元音脱落等词形变化。因此，需要还原出各形态成分的原始形态，这一过程即音变还原。同样地，为摆脱对语言学规则设计的依赖，将音变还原也建模为字符分类问题。通过从标注语料中自动化地获取音变还原规律并抽取相应的判别式分类实例，可以自动化地构建高效的音变还原模型。

形态拆分为上层的图状模型提供语句中各词语的形态分析候选。对于存在音变现象的语言，则需在形态拆分之前先进行音变还原。接下来，介绍基于字符分类原理的形态拆分和音变还原。

3.1 基于字符分类的形态拆分

对于给定的待切分词语W：

$$W = C_{1:n} = C_1 C_2 \dots C_n$$

其中C_i（$1 \le i \le n$）是W中的第i个字母，n为字母序列的长度。词干词缀切分即为字母序列的划分问题：

$$C_1 C_2 \dots C_n \rightarrow C_{1:e_1} C_{e_1+1:e_2} \dots C_{e_{m-1}+1:e_m}$$

其中，$e_m = n$，字母序列$C_{1:n}$划分为m个子序列，每个子序列对应为一个形态元素。一般而言，第一个子序列$C_{1:e_1}$是词干，剩余的字母序列是词缀。

这是典型的序列划分问题，可以用序列标注的方式进行建模。将其与基于判别式字符分类的汉语分词进行类比，将每个字母C_i分类为如下四种类别之一：

b：词干或词缀的开始字母
m：词干或词缀的中间字母
e：词干或词缀的结束字母
s：单字母作为词干或词缀

当对整个词的字母序列完成标注之后，标注为bm*e或者s的字母子序列即为词干或词缀，相应地得到一个候选的词干词缀切分结果。对字符分类所采用的特征，是以该字符为中心的特定长度窗口中的字符元组。所用的特征模板列在表 3 中。其中，C_0表示当前考察的字母，C_{-i}/C_i表示C_0左边/右边的第i个字母。借助这些特征模板，从训练语料中抽取字母分类实例，以此训练字符分类器。

表 3 形态拆分和音变还原特征模板

类型	特征模板	
1 元组	C_i	$-4 \le i \le 4$
2 元组	$C_{i-1} \circ C_i$	$-3 \le i \le 4$
3 元组	$C_{i-2} \circ C_{i-1} \circ C_i$	$-2 \le I \le 4$
4 元组	$C_{i-3} \circ C_{i-2} \circ C_{i-1} \circ C_i$	$-1 \le I \le 4$

考虑到词干词缀切分的歧义性，为待分析语句中的每个词及其变形形态都生成 N 个最佳的切分方案。通过为 N 选择合适的值，可以在保证分析速度的同时取得较高的分析精度。

3.2 基于字符分类的音变还原

黏着语普遍存在音变现象，词干之上缀接词缀时，有些元音和辅音会弱化为另外一个音，或者出现丢失（脱落）和增加（增音）等情况。这些音变现象给黏着语的形态分析带来了较大困难，需要在形态拆分之前先进行音变还原。通过语言学规则的方式进行音变还原耗时耗力，而且需要相关语言专家知识的支撑。为实现更好的语言无关性，提出一种自动化的基于字符分类的音变还原方法。

基于字符分类的音变还原模型，将音变还原建模为对词中每个字母进行判别式分类的问题。黏着语词语可以看成它所包含的字母的线性序列，假设音变现象会发生在每个字母上，每个字母既可以保持不变（即变为其自身），又可以变为其他形式（包括变为空的情形）。一个字母在音变过程中变为何种形式，是可以由其上下文信息预测得知的。由此，音变还原问题建模为字符的判别式分类问题。

以字符分类模式求解音变还原，需要知道词中的每个字母在音变之后转换为何种字母串（它自身，空串或其他字符串）。首先要确定音变还原前后词内字母的对应关系。借鉴最短编辑距离算法的思想提

出一种词内字母对齐算法。音变前的一个字母可以对应到音变后的 0 个或多个字母，如果对应到一个字母且恰好为该字母本身，则得分加 1，否则得分为 0。通过类似于求解最短编辑距离的动态规划算法，即可快速的找到得分最高的匹配模式。在动态规划搜索过程中，不仅记录当前最高的匹配得分，还要记录此得分对应的匹配模式，以便搜索结束后回溯出最优路径对应的整个词的匹配模式。具体算法不再详述。

在经过字母对齐的词语集合上，可以提取出用以判别式分类的训练实例。采用和基于字符分类的形态拆分模型同样的特征模板，并同样适用最大熵工具包训练分类器。为提高形态还原的召回率，使用类似于立方体剪枝的策略，为待分析词生成多个最佳形态还原候选。

4 相关工作

黏着语形态分析目前已有很多工作。Cahill 提出了基于音节切分的韩语形态分析技术[7]，该方法使用音节作为最小的切分单位，在韩语中有较好的效果。Lee 和 Rim 运用最大似然概率的方法进行统计语料中出现的词干、词缀以及音节等，然后按照词素、音节等粒度进行韩语形态分析工作[8]。维吾尔语的词法分析研究起步较晚。早克热等研究并设计了维吾尔语名词构形词缀有限状态自动机[9]。阿孜古丽详细探讨了维吾尔语动词附加语素的复杂特征[10]。还有一部分工作采用统计方法[11, 12, 13, 14]，但所选择的特征主要以字符为主。侯宏旭等利用语言模型对切分结果进行重排序[15]，Ablimit M 等将形态切分看作序列标注，运用条件随机场进行有监督学习，设计特征来表述形态切分的规律[17, 18]。

总的来说，形态丰富语言的词法分析主要借助规则和生成式方法优化切分结果，对 OOV(未登录词)的处理能力较弱。在相关的判别式统计方法中，更多是从词本身考虑，考虑的上下文信息有限。以句为单位将形态分析建模为字母的序列标注问题，也是较为直观的方式。但这种方式只能考虑特定窗口内的上下文信息，无法刻画更大范围内形态元素之间的语言学关联。与之相比，本文的工作具有更高的理论和实用价值。针对词干词缀间的联接特性，建立了更贴合黏着语构词规律的图状模型。而之前工作则通常借用现成的序列标注模型，将句中所有词干和词缀视为单一线性的序列结构。因此，本文对黏着语词法分析的建模更加科学有效。基于无向图参数结构的判别式模型，是针对黏着语构词特性的崭新的建模方式。与以前工作相比，这一模式具有更好的扩充性和提升空间，相应地也更具理论价值。

5 实验

本文以韩语和维吾尔语为例进行形态切分实验。韩语的形态切分语料来自网络和人工标注语料，共有 54358 句，各随机抽取 1%的句子作为开发集和测试集，剩余的全部作为训练集。维吾尔语语料来自新疆大学多语种重点实验室标注的维吾尔语百万词词法分析语料库，共有 72739 句，类似地，各随机抽取 1%的句子用作开发集和测试，剩余的语句用以训练。

关于形态切分的评价指标，前人的工作中已提出多种方案，有基于整词的评估模式，也有将词干和词缀分别对待的模式。本文采取最严格的评测方式即以整词为单位进行评估。对于一个黏着语词语，只有当其词干和词缀全部识别正确时才算是分析正确的。整词正确率P_w定义为分析正确的词占全部词语的比例。考虑到有些应用更关注词干识别精度，采用词干正确率P_s定义为正确识别出词干的词占全部词语的比例。

音变还原和形态拆分为上层的图状模型提供形态拆分候选，首先验证这两个模块的性能。对于音变还原和形态拆分抽取的训练实例，采用最大熵工具包 MaxEnt[16]训练分类器。高斯先验设置为 1.0，剪枝阈值设置为 0，其余参数保持默认，进行 200 轮 L-BFGS 迭代训练。

由实验得到图 5、图 6，分别显示了音变还原和形态拆分分类器输出 N-best 时的 Oracle 情况，即 N-best 列表包含正确答案的可能性。发现当 N 取较小的值时，音变还原和形态拆分分类器的 Oracle 就已经非常高了。在后续图状模型的实验中，将这两个模块的 N-best 数量都设置为 3，即针对每个词语，音变还原模块输出最多 3 个候选，随后的形态拆分模块为每个音变还原候选生成最多 3 种拆分方案。

图状模型的性能列在表 4 中。将音变还原模块和形态拆分模块串联起来，即为最直观和最常用的基于序列标注的形态切分模式，这可以作为图状建模策略的基线模型。相比基于字母的序列标注基线模型，基于形态元素的图状模型取得了显著的精度提升，超过了前人工作[2]在相应语言上的性能。

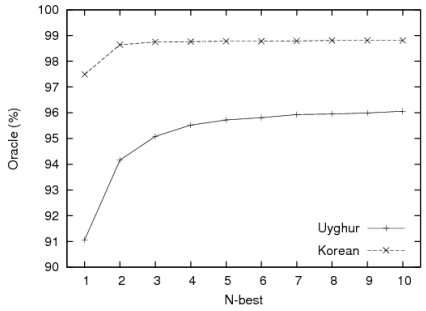

图 5 音变还原分类器 Oracle 曲线

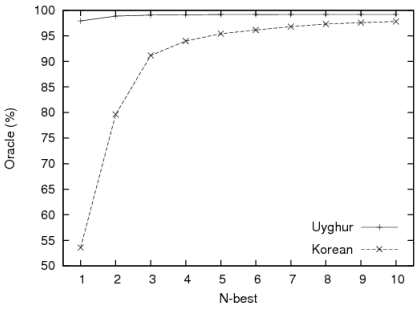

图 6 形态拆分分类器 Oracle 曲线

表 4 形态分析图状模型性能

模型	维吾尔语		韩语	
	P_w	P_s	P_w	P_s
基线模型	91.26	92.32	91.01	93.64
图状模型	95.67	96.85	93.79	95.88
前人工作				
麦热哈巴 et al.	94.39	95.74	91.83	93.65

6 结论

本文针对黏着语的词句法特性，为形态分析任务建立了一种图状结构的判别式模型。该模型将黏着语语句的形态分析结果建模为以形态成分为节点的图状结构，并通过丰富的特征描述词语内部形态成分之间以及分属相邻词语的形态成分之间的关联约束。同时，将词语形态拆分和音变还原都建模为基于字符分类的序列标注问题，摆脱了对语言学规则的依赖，从语料库中自动学习到更具覆盖度和泛化性的形态切分和音变还原规律。在韩语和维吾尔语上的实验表明，图状建模的形态分析比线性建模方式取得了显著的性能提升。

参考文献

[1] Deok-Bong Kim, Sung-Jin Lee, Key-Sun Choi et al.A TWO-LEVEL MORPHOLOGICAL ANALYSIS OF KOREAN[C]// In Proc. 15th Int. Conf. Computat. Linguist., 1994:535–539.

[2] 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.

[3] Lawrence. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition [C]// In Proceedings of IEEE, 1989:257--286.

[4] John Lafferty and AndrewMcCallum and Fernando Pereira.Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]// In Proceedings of the 18th ICML, 2001:282—289.

[5] McCallum, A., Freitag, D., &Pereira, F. Maximum entropy Markov models for information extraction and segmentation [C]//In Proceedings of ICML, 2000:591–598.

[6] Michael Collins.Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms [C]//In Proceedings of EMNLP,2002, 410(21-23):1-8.

[7] Lynne J Cahill. Syllable-based morphology [C]// In Proceedings of CCL,1990,3:48–53.

[8] Do-Gil Lee and Hae-Chang Rim. Probabilistic modeling of korean morphology [C]// IEEE Transactions on Audio, Speech, and Language Processing, 2009,17(5):945–955.

[9] 早克热.卡德尔，艾山.吾买尔，吐尔根.依布拉音，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.

[10] 阿孜古丽.夏力甫. 维吾尔语动词附加语素的复杂特征研究[J]. 中文信息学报，2008，22(3):105–109.

[11] Mijit Ablimit, Mihrigul Eli, and Tatsuya Kawahara. Partly supervised Uyghur morpheme segmentation [C]// In Proceedings of COCOSDA, 2008:71–76.

[12] B. Aisha and Maosong Sun. A statistical method for Uyghur tokenization [C]// In Proceedings of NLP-KE，2009:1-5.

[13] B. Aisha. A letter tagging approach to Uyghur tokenization [C]// In Proceedings of IALP, 2010:11 –14.

[14] A. Wumaier, Z. Kadeer, P. Tursun et al. Maximum entropy combined FSM stemming method for Uyghur [C]// In Proceedings of COCOSDA, 2009:51–55.

[15] 侯宏旭，刘群，那顺乌日图，等. 基于统计语言模型的蒙古文词切分[J]. 模式识别与人工智能，2009,22(1):108–112.

[16] Le Zhang. Maximum Entropy Modeling Toolkit for Python and C++ [OL] , http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

[17] Ruokolainen T, Kohonen O, Virpioja S et al. Supervised morphological segmentation in a low-resource learning setting using conditional random fields [C]// Proceeding of

the Seventeenth Conference on Computational National Language Learning. Sofia, Bulgaria: Association for Computational Linguistics, 2013: 8-9.

[18] Ablimit M, Kawahara T, Pattar A, et al. Stem-affix based Uyghur morphological analyzer [J]. International Journal of Future Generation Communication and Networking, 2016, 9(2): 59-72.

徐春（1977—），博士研究生，副教授，主要研究领域为自然语言处理、无线光网络。

E-mail:xuchun@mails.ucas.ac.cn

蒋同海（1963—），博士，研究员，主要研究领域为自然语言处理、人工智能。

E-mail:jth@ms.xjb.ac.cn

阿布都热合曼·卡的尔（1975—），博士，副教授，主要研究领域为人工智能。

E-mail: ar@xjufe.edu.cn

作者联系方式：徐春 北京市海淀区中关村科学院南路 6 号中国科学院计算技术研究所智能信息重点实验室 邮编：100190 电话：15099065915

电子邮箱：xuchun@mails.ucas.ac.cn