文章编号：1003-0077(2011)05-0094-07

蒙古语词法分析的有向图模型

姜文斌[1]，吴金星[1,2]，长 青[1,2]，那顺乌日图[2]，刘 群[1]，赵理莉[1,3]

(1. 中国科学院 计算技术研究所，北京 100190；
2. 内蒙古大学 蒙古学学院，内蒙古 呼和浩特 010021；
3. 河南师范大学 计算机与信息技术学院，河南 新乡 453007)

摘 要：我们为蒙古语词法分析建立了一种生成式的概率统计模型。该模型将蒙古语语句的词法分析结果描述为有向图结构，图中节点表示分析结果中的词干、词缀及其相应标注，而边则表示节点之间的转移或生成关系。特别地，在本工作中我们刻画了词干到词干转移概率、词缀到词缀转移概率、词干到词缀生成概率、相应的标注之间的三种转移或生成概率，以及词干或词缀到相应标注相互生成概率。以内蒙古大学开发的 20 万词规模的三级标注人工语料库为训练数据，该模型取得了词级切分正确率 95.1%，词级联合切分与标注正确率 93%的成绩。

关键词：蒙古语；词法分析；词语切分；词性标注；词干提取；有向图

中图分类号：TP391 **文献标识码**：A

Directed Graph Model for Mongolian Lexical Analysis

JIANG Wenbin[1], WU Jinxing[1,2], CHANG Qing[1,2], Nasanurtu[2], LIU Qun[1], ZHAO Lili[1,3]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Inner Mongolian University, Huhhot, Inner Mongolia 010021, China;
3. Henan Normal University, Xinxiang, Henan 453007, China)

Abstract: We propose a generative statistical model for Mongolian lexical analysis. This model describes the lexical analysis result as a directed graph, where the nodes represent the stems, affixes and their tags, while the edges represent the transition or generation relationships between nodes. Especially in this work, we adopt three kinds of transition or generation probabilities: a) probabilities of stem-stem transition, affix-affix transition and stem-affix generation; b) the transition or generation probabilities between the corresponding tags; and c) the generation probabilities between stems or affixes and their tags. Using the 3rd-level annotated corpus with about 200 000 words as the training data, this model achieves a word-level segmentation accuracy of 95.1%, and a word-level joint segmentation and tagging accuracy of 93%.

Key words: Mongolian; lexical analysis; segmentation; POS tagging; stemming; directed graph

1 引言

对汉语和许多民族语言来说，词法分析是大多数自然语言处理任务的基础。汉语的词形较为简单，当前的词法分析已经做到实际可用的水平[1-4]，而对于形态复杂的民族语言如蒙古语和维吾尔语，词法分析的准确率仍有较大的提升空间[5-11]。在民族交流与融合需求日益迫切的现阶段，机器翻译技术的重要作用越发凸显。民族语言词法分析作为机器翻译的必备前提，需要得到研究者更多的关注。

与汉语的字符顺次拼接的构词方式相比，蒙古

收稿日期：2010-08-29 **定稿日期**：2011-02-17

基金项目：国家自然科学基金资助项目(Contract60736014)；863 重点项目(2006AA010108)；教育部、国家语委民族语言文字规范标准建设及信息化资助项目(MZ115-038)

作者简介：姜文斌(1984—)，男，博士生，主要研究方向为词法分析、句法分析和机器翻译；吴金星(1987—)，女，硕士生，主要研究方向为蒙古语信息处理；长青(1985—)，女，硕士生，主要研究方向为蒙古语信息处理。

语和维吾尔语等形态丰富的语言构词规律更加复杂。这类语言的词语通常由词干和若干起修饰作用的词缀组成树状结构，词法分析的任务就是解析出词语的词干和词缀构成，并且标定好它们的类别标注。这样一来，在汉语上效果良好的序列标注模型[12-14]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[7-10]，这使得系统的理论价值和实用性大打折扣。另一方面，传统的基于规则的词法分析模式需要专门的语言学人才，往往耗费大量的精力调试搭建后，准确率和稳定性并不尽人意。因此，我们有必要构造更为恰当的统计模型，尽可能准确地描述形态丰富语言的构词规律，从而快速搭建高性能的词法分析系统。

我们为蒙古语词法分析建立了一种生成式的概率统计模型。该模型将蒙古语语句的词法分析结果描述为有向图结构，图中节点表示分析结果中的词干、词缀及其相应标注，而边则表示节点之间的转移或生成关系，它们刻画了词干、词缀及其相应标注连接成词的规律。生成式概率统计模型为这些转移或生成关系赋以合适的概率形式，词法分析的过程就是寻找其所有概率乘积最大的有向图。在本工作中我们刻画了词干到词干转移概率、词缀到词缀转移概率、词干到词缀生成概率、相应的标注之间的三种转移或生成概率以及词干或词缀到相应标注相互生成概率。这些转移或生成概率以极大似然估计的方式从训练语料中统计得到。鉴于本工作的意图在于统计建模，在为句中的每个词枚举可能的词语结构候选时，我们并没有利用人工标注词法分析语料库之外的任何语言资源，也没有设计专门的词法和语法知识进行指导，而是依据从人工语料库中抽取出的词干表和词缀表，通过递归搜索穷举所有可能的构词方式。

我们在内蒙古大学开发的20万词规模的三级标注人工语料库（内蒙古大学拉丁语料）上进行实验。我们随机分割出5%和5%的句子分别作为开发集和测试集，剩余的90%的句子全部作为训练集。在测试集上，该模型取得了词级切分正确率95.1%，词级联合切分与标注正确率93%的好成绩。另外，整个系统的训练过程只需要几十秒即可完成，解码过程在PC机上也可达到几百词每秒的速度。而且，由于系统几乎没有借助任何语言学知识，我们相信只需很少的改动就可以应用到其他形态丰富的语言上。

在以下的章节中，我们首先介绍蒙古语词法分析的任务定义，然后描述我们的生成式概率统计模型，在展示该系统实验结果并进行相应的分析说明后，我们与前人工作进行对比，最后是总结和展望。

2 蒙古语词法分析

同其他形态丰富的语言类似，蒙古语的词由词干和可能的词缀组成。不同的是，蒙古语词干与词缀的组合需要服从特有的约束：

a) 词干只能有一个且只能出现在最前面；

b) 分写词缀只能跟在连写词缀之后；

c) 同类词缀中不同词缀须以特定的顺序出现。

约束a)规定一个蒙古语词只能有一个义项中心，这一点与维吾尔语不同；而约束b)和c)规定了不同词缀的特定出现顺序，这一点与朝鲜语又不相同。

以内蒙古大学拉丁语料中的蒙古语词HUUR-NILDU/HU-DU为例，其在特定语境下的一种词法分析结果为：

HUURNI/Ve2+LDU/Fe3+HU/Ft12−DU/Fc21

其中，“+”号和“−”号分别表示后面紧接着的是连写后缀和分写后缀。给定一个蒙古语词，我们可以借助词干表和词缀表，以递归枚举的方式把可能的词法结构罗列出来。在该语料库中，分写后缀和一部分连写后缀在原始词中已经被标识出来。所有的分写后缀都放在词的尾部，且以“−”号分隔，例如词尾的“−DU”；一部分连写词缀位于分写词缀之前，且以“/”号与前面部分分隔开来。这部分连写词缀前面的“/”号是内蒙古大学拉丁语料在初步标注过程中人工加进去的词干与变形附加成分之间的分隔符号，因此我们在系统测试之前将删除输入数据中的“/”号，以模拟真实环境下的蒙古语语句反映词法分析系统的真实性能。

许多蒙古语词拥有不止一种词法分析候选结构，在同时做词干标注和词缀标注的时候，候选结构数量变得更为庞大。如果根据特定的上下文环境为蒙古语词选择正确的词法分析结构，既是歧义排解的问题，也是蒙古语词法分析的难点所在。基于语言学规则的词法分析系统能够为每个单独的词高效地枚举出尽可能精简的候选分析集合，却不擅长于根据上下文环境为每个词选择最恰当的候选。语言的统计建模恰好可以与规则方法实现优势互补。统计方法难以为每个词确定一个精简的合法候选分析

集，却擅长于高效地为整个句子选择最可能的整体分析结果。本工作的重点即在蒙古语词法分析的统计建模上。

3 有向图概率模型

基于统计的有监督建模总体上可分为两类：生成式统计建模和判别式统计建模。两种模型体现了截然不同的建模思路。生成模型同时考查输入和分析结果，旨在找出产生概率最高的输入与分析结果的组合。因此，它有一部分概率知识用于描述输入语句的生成规律。判别模型则立足于输入考查分析结果，其目的在于找出已知输入的情况下最优的分析结果候选。与生成模型的理念相比，判别模型更符合人们分析解决问题的方式。事实也证明了判别模型的优势。在序列标注的经典问题词性标注上，判别模型在汉语和英语上都比生成模型有明显进步。

然而，判别模型用于蒙古语词法分析还存在不少有待解决的关键问题。一方面，与生成模型不同的是，判别模型的训练要设计大量的判别特征，并通常需要漫长的多轮迭代过程。与形态简单的汉语和英语相比，蒙古语词复杂的形态结构使得模型的搜索空间要大的多，从而需要大的存储占用和更久的训练时间；另一方面，判别模型通常仅适用于为搜索空间结构固定的任务建模，如词性标注和依存分析，它们都有一个固定不变的词语序列。而对于形态丰富语言的词法分析来说则不具备确定的搜索空间结构，因为我们需要枚举各种可能的候选词语结构，并在选择最佳候选结构的同时确定该结构内词干和词缀的标注。

目前，我们对于形态丰富语言词法分析的判别式建模也构想了几套初步可行的方案，相关的研究已经在有序的进行。在本文中，我们仅专注于阐述已经成型的生成式概率统计模型。

3.1 单纯切分的模型结构

同词语形态简单的汉语或者英语相比，词语形态丰富的蒙古语的词法分析更像是一个对树结构进行选择并对树中节点进行标注的过程，而不是一个简单的线性序列标注问题。这里，我们并不马上介绍能够同时进行切分和标注的最终词法分析模型，而是先从较为简单的任务说起，即单纯切分的模型构建。

我们把语句中各词的分析结果定义为链状结构，如图1所示：

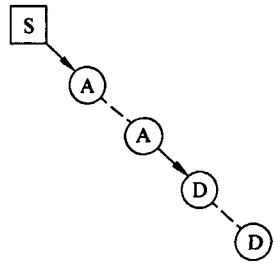

图1 词语内部结构表示

这里，S(Stem)表示词干，A(Adjoin)表示连写词缀，D(Disjoint)表示分写词缀。我们用虚线连接的两个A(或D)表示0或多个连写词缀(或分写词缀)。在词干到词缀之间以及词缀到后续词缀之间，箭头表示生成或者转移关系。对于整个语句，分析结果则可描述为树状结构，如图2所示：

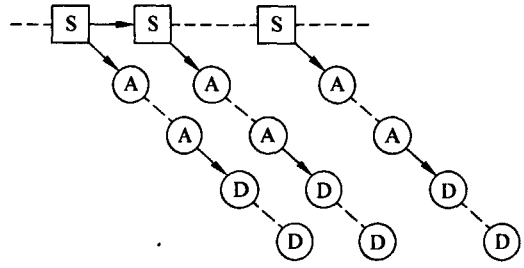

图2 整句的词语内部结构表示

与单个词的分析结果结构相比，整句分析结构中增加了相邻词的词干之间的生成或转移关系，从而在所有词干和词缀之间形成一个拓扑有序的树结构。树中节点即表示词干或者词缀，而节点之间的边则表示词干到词干、词干到词缀以及词缀到词缀的生成或转移关系。

概而言之，无论对于规则模型还是统计模型，建模的本质都是在刻画词干、词缀及其相应标注(如果同时也做标注的话)之间的生成、转移等约束规律。如果我们能为树中的各种不同的边设计相应的权重，这些权重的度量反映了节点之间生成或转移规律的强弱，那么，求解整句词法切分结果的过程，即为在所有可能的候选树中寻找权重之和最高的树的过程。本模型中，我们用类似于隐马模型使用中的转移概率来描述树中边的权重。根据边指向对象的不同，我们设计以下两种转移概率：

a) $P(S|S\ \text{ngram})$

词干到词干的转移概率，类似于ngram语言模型。

b) $P(X|S/X\ \text{ngram})$

其他词缀的生成概率，X 代表词缀，即 A 或者 D。S/X ngram 指当前词缀之前的词干或词缀组成的 ngram 历史。

给定一个候选树 T，我们用这些概率的乘积表示该候选的整体生成概率：

$$P(T)=\prod_{S\in T}P(S\mid\cdots)\times\prod_{X=A/D,X\in T}P(X\mid\cdots)$$

为简洁起见，公式中隐藏了两个条件概率的历史条件。容易看出，这可以理解为传统的 ngram 语法模型向树结构的拓展。

3.2　联合切分标注的模型结构

上面的模型仅考虑词语的形态分析而不涉及词干和词缀的标注。当我们也需要词干和词缀的标注信息时，就必须同时对这些标注成分进行概率建模了。事实上，即使我们只需要进行词语形态分析，考虑到人工词法分析语料库规模不会很大，构词元素特别是词干对现实世界中蒙古语语言的覆盖面相当有限，在语料库提供标注信息的情况下，尽可能的对标注建模以利用这些标注信息，也是缓解数据稀疏的重要手段。

对联合切分和标注进行建模的关键在于如何让标注信息有效地参与描述句中各词的形态结构生成过程。本工作中，对应于单纯切分的模型结构，我们为标注信息设计了一个同步树状结构以描述词干和词缀标注之间的生成和转换关系。所谓同步是指树的结构和单纯切分模型的树结构完全一致，只不过树中对应节点，对后者而言是词干或词缀，对前者而言是相应的标注。另外，我们设计两项概率描述两个平行的树结构中节点之间的映射关系：

a) $P(X|t(X))$

X 代表词干或词缀，$t(X)$ 代表其标注。此项概率可类比于隐马模型中状态到观察的生成概率。

b) $P(t(X)|X)$

此项概率代表词干或词缀 X 被赋予标注 $t(X)$ 的概率。此项概率参与建模使得模型倾向于为选择常见的标注。

这两项条件概率在平行树结构的节点之间可表示为不同方向的有边，从而建立起平行树结构之间的映射关系，构建描述能力更强的有向图模型(图 3)。

求解切分和标注结果的过程，即为在候选有向图中寻找概率最大的有向图。有向图 G 的概率定义为：

$$P(G)=P(T)\times P(t(T))\times P(T,t(T))$$

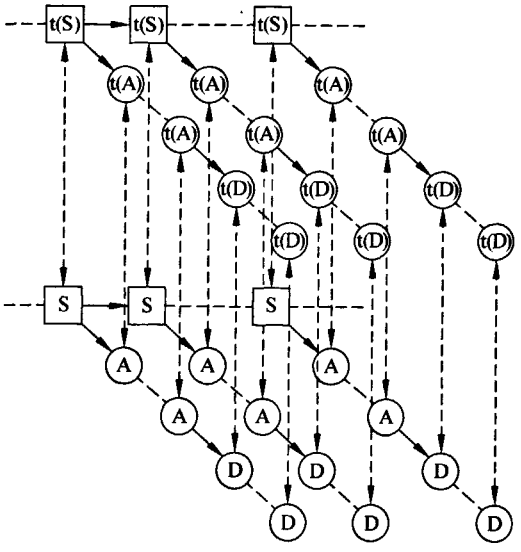

图 3　带词性表注的整句词语内部结构表示

其中，$P(t(T))$ 表示标注树 $t(T)$ 的概率，它和 $P(T)$ 的定义一样，只需把词干和词缀换成相应的标注。$P(T,t(T))$ 表示平行树结构 T 和 $t(T)$ 的映射概率，它定义为平行树中所有节点对的条件概率的乘积：

$$P(T,t(T))=\prod_{X\in T,t(X)\in t(T)}P(X\mid t(X))\times P(t(X)\mid X)$$

理论上，$P(G)$ 的三项乘子概率对于候选有向图的优选可能具有不同的决策力，故为它们赋予合适的相对加权有望提升模型性能。但在本工作中我们暂不考虑乘子加权问题，这相当于所有加权均为 1。

3.3　训练与解码

出现在单纯切分模型和联合切分与标注模型的各项概率，均可以用极大似然估计的方式从人工标注词法分析语料库中统计得来。其中对于词干到词干转移概率、词缀到词缀转移概率、词干到词缀生成概率、相应的标注之间的三种转移或生成概率，可以借助成熟的工具包如 SRI 语言模型工具来实现[15]，这将使我们不必理会概率的回退与平滑，而将精力集中在模型结构的设计上。

模型训练完毕之后，解码任务就是一个递归枚举各词的可能分析结果候选，并紧接着进行动态规划搜索确定各词最优候选的过程。枚举过程依据一个词干表和一个词缀表，递归的列举出词语所有可能的词形。需要注意的是蒙古语词的某些字符在特定情境下会发生变形，主要总结为以下两种：

a) 词干词缀划分过程中，若 AYI、EYI、0YI、

VYI、OYI或UYI由非词尾变为词尾，则删掉字符Y。

b) 词干词缀划分过程中，若GA、HA、YA、YE和RE由非词尾变为词尾时，需在中间添加下划线"_"。

动态规划的搜索过程就是自左到右的viterbi解码过程。考虑文章篇幅限制，我们这里对这两个过程不再展开详述。

4 实验

我们在内蒙古大学蒙古学学院开发的20万词规模词法分析语料库上进行实验。该语料库共包括14 115个完整的句子，我们从中随机抽取出各5%的语句分别用做开发集和测试集，各含705句，剩余90%的语句用做训练集，含12 705句。模型各项概率均从训练集中以极大似然估计法统计得来。其中，词干到词干转移概率、词缀到词缀转移概率、词干到词缀生成概率、相应的标注之间的三种转移或生成概率，我们直接借助成熟的语言模型工具包SRILM，以WB平滑方式训练三元模型。

蒙古语的词法分析结果结构远比汉语复杂，传统的正确率、召回率和F值不能直接适用。本工作中我们定义和采纳了多种指标，从不同角度和层面考量词法分析器的性能。这些指标包括：

a) 词级正确率 Pw

以词为单位计量，仅当词内词干、词缀及其标注均正确时，该词才是分析正确的。

b) 词干词缀级正确率 Psa，召回率 Rsa 和 Fsa 值

以词干和词缀为单位计量，仅当词干或词缀及相应标注正确时，该词干或词缀才是分析正确的。因此，词干和词缀可类比为汉语词法分析中的词。此评价标准引自文献[7]。

c) 相应的不考虑标注信息的评测指标：Pw-t，Psa-t，Rsa-t 和 Fsa-t

表1 系统在测试集上的性能/%

		+t	−t
词级	Pw	93.0	95.1
词干词缀级	Psa	92.7	94.3
	Rsa	93.4	95.1
	Fsa	93.0	94.7

表2 不同子模型组合在开发集上的性能/%

	Pw	Psa	Rsa	Fsa
$P(T)$	75.4	80.2	80.6	80.4
$P(t(T))$	83.0	84.6	85.8	85.2
$P(T, t(T))$	75.7	70.9	81.6	75.9
$P(T)+P(t(T))$	86.6	88.7	89.1	88.9
$P(T)+P(T, t(T))$	88.6	89.4	90.4	89.9
$P(t(T))+P(T, t(T))$	92.8	92.8	93.3	93.1
ALL	93.1	93.2	93.6	93.4

表1展示了系统在测试集上以上述几个评测指标考量的最终性能。词级正确率93%意味着系统对测试集中93%的词都能够分析出完全正确的词形结构和标注信息。我们发现，不论对于哪种评测指标，不考虑词性标注都要比考虑词性标注高1个百分点以上。这说明联合词形分析与标注的难度明显高于单纯的词形分析，如何有效地联合利用词干、词缀及其标注信息进行建模值得更加深入的探索。由于联合切分和标注的意义远高于单纯切分，我们在后续的试验中仅报告考虑标注的相关指标分值。

下一步我们验证有向图概率 $P(G)$ 定义的有效性。如本文第2章中描述，$P(G)$由三个子模型概率累乘起来，包括词干词缀树概率 $P(T)$、相应标注树概率 $P(t(T))$以及词干词缀树与标注树之间的映射概率 $P(T, t(T))$。此次试验在开发集上进行，我们分别尝试不同的子模型组合的性能，以验证各个子模型发挥的作用。通过表2我们发现，标注树概率 $P(t(T))$发挥的作用最大，并且它和词干词缀树到标注树的映射概率 $P(T, t(T))$联合使用时，系统性能已经趋近于完整系统了。

相比英语、汉语和其他资源丰富的语言来说，当前蒙古语词法分析人工标注语料规模要小得多。通过模型改进带来的性能提升毕竟有限，要想大幅度提高蒙古语词法分析的准确率，必须有更大规模的人工标注语料支持。在语料库扩建之前，我们可以先探索一下性能提升和语料规模扩大的关系。为此，我们固定开发集和测试集不变，而从训练集中每次提取不同规模的子集以训练系统，并考查该系统在测试集上的表现。整个训练集含12 705句标注语句，我们从中随机选取一系列不同规模的子集，分别含有6 000，3 000，1 500，800，400，200和100个语句，并按照由小到大的次序画出系统性能随训练数据增加的变化曲线。

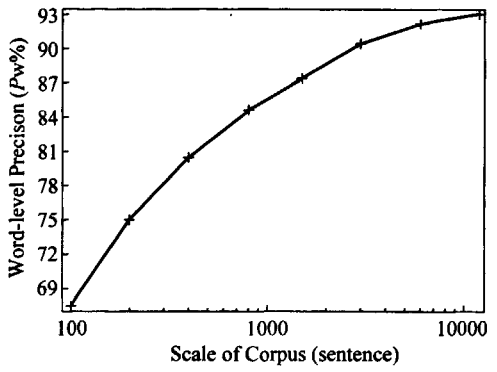

图 4　训练集规模—系统性能曲线图

通过图 4 所示的训练集规模—系统性能曲线我们发现，随着训练集语句数量的增加，系统性能持续变化。这在训练语料规模较小的时候尤其明显，例如，训练集从 100 句扩大到 200 句时，系统的词级正确率从 67.5%上升到 75%。随着语料规模的继续扩大，系统性能的提升幅度趋于缓和，例如，训练集从 6 000 句扩大到 12 000 句时，词级正确率从 92.2% 提升到 93.1%。这带给我们两方面的启示：其一，鉴于现在蒙古语词法分析语料规模仍然较小，通过进一步扩建语料库以提升性能仍然是有意义的和必要的，毕竟接近 1 个百分点的绝对增长量也是一个相当可观的性能提升；其二，语料规模继续增加到一定程度后，统计模型通过语料扩建提升性能的方案的性价比会越来越低，此时，引入语言学知识来增强统计模型可能是最有希望的研究思路。

5　相关工作

蒙古语词法分析的主要工作都是基于语言学规则的。统计知识的引入，是从侯宏旭等人借助语言模型对规则系统的提供的候选结果进行择优排歧的工作[7]开始的，该工作取得了 94%的词切分准确率。而后又有一些工作也取得了较好的结果，例如，赵伟等人[8]取得了 99.2%的词切分准确率，丛伟[9]取得了 97.1%的词切分准确率，艳红和王斯日古楞[10]取得了 96.8%的词性标注准确率。但上述工作都将任务限定为粗切分或标注，且采用的数据集与我们不同，因此和我们目前工作缺乏可比性。这些工作一般直接借用现成的线性序列标注模型，既没有考虑黏着语的构词特性，也没有采用严整精细的切分标注标准，从而使得系统的理论价值和实用性有所限制。与之相比我们工作的优越性如下：

第一，我们的工作同步地实现了词形分析和词干词缀的标注，所采用的词干、词缀拆分标准也更加细致复杂；之前工作多专注于切分而很少给出标注信息，并且词语切分的粒度也很粗略。相比而言，我们解决的任务更为严整复杂，系统也相应的更具实用价值。

第二，我们的工作针对词干词缀间的连接特性，建立更贴合黏着语构词规律的树状生成模型；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，我们对蒙古语词法分析的建模更加科学有效。

第三，我们为蒙古语的联合切分和标注任务建立了高度形式化的，基于由同步树结构组成的有向图的概率生成模型。这是针对黏着语构词特性的崭新的建模方式。因此，与以前工作相比我们的工作具有更好的扩充性和提升空间，相应地也更具理论价值。

6　总结与展望

本工作为蒙古语词法分析建立了一种生成式的概率统计模型，将蒙古语语句的词法分析结构描述为有向图结构，图中节点表示分析结果中的词干、词缀及其相应标注，而边则表示节点之间的转移或生成关系。整体上，有向图由同步的词干词缀树和标注树以及树间的映射关系组成，分别描述词干词缀的生成转移关系、相应标注的生成转移关系以及词干词缀与标注间的生成关系。最终系统在内蒙古大学开发的 20 万词规模的人工语料库做到了较好的水平，词级切分正确率为 95.1%，词级联合切分与标注正确率为 93%。

然而，当前模型还很初步，许多重要的方面仍有待改进。首先，关于模型构建，鉴于判别式模型普遍优于生成式模型，如何为形态丰富语言建立有效的判别式词法分析模型并设计相应的特征表示，将是我们接下来的重要探索方向之一。再者，我们目前只是根据从训练集中自动抽取出的词干表和词缀表，为每个待分析词递归地穷举可能的候选结构，这导致过多的非法候选，以致引入无谓的歧义。如何利用语言学规则约束候选生成甚至解码过程，也是我们未来要进行的重要研究内容。

参考文献

[1]　Hwee Tou Ng, Jin Kiat Low. Chinese part-of-speech

tagging: One-at-a-time or all-at-once? Wordbased or character-based?[C]//Proceedings of EMNLP, 2004: 277-284.

[2] Wenbin Jiang, Liang Huang, Yajuan Lv, et al. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging[C]//Proceedings of the 46th ACL, 2008: 897-904.

[3] Huaping Zhang, Qun Liu, Xueqi Cheng, Hao Zhang, et al. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model[C]//Preceedings of Second SIGHAN workshop affiliated with 41th ACL, 2003: 63-70.

[4] 米海涛，熊德意，刘群. 中文词法分析与句法分析融合策略研究[J]. 中文信息学报，2008，22(2)：10-17.

[5] 那顺乌日图，雪艳，叶嘉明. 现代蒙古语语料库加工技术的新进展—新一代蒙古语词语自动切分与标注系统[C]//第十届全国少数民族语言文字信息处理学术研讨会，2005.

[6] 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，2009，22：108-112.

[7] 赵伟，侯宏旭，从伟，等. 基于条件随机场的蒙古语词切分研究[J]. 中文信息学报，2010，24(5)：31-35.

[8] 从伟. 基于层叠隐马尔科夫模型的蒙古语词切分系统的研究[D]. 内蒙古大学硕士毕业论文，2009.

[9] 艳红，王斯日古楞. 基于HMM的蒙古文自动词性标注研究[J]. 内蒙古师范大学学报(自然科学汉文版)，2010，39(2)，206-209.

[10] 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.

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

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

[13] McCallum, A., Freitag, D., Pereira, F. Maximum entropy Markov models for information extraction and segmentation[C]//Proc. ICML, 2000：591-598.

[14] Stolcke, Andreas. Srilm—an extensible language modeling toolkit[C]//Proceedings of the International Conference on Spoken Language Processing, 2002：311-318.

(上接第93页)

semantic orientation of adjectives[C]//Proceedings of the Eighth Conference on European Chapter of the Association For Computational Linguistics. European Chapter Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 1997: 174-181.

[6] Yi, J., Nasukawa, T., Bunescu, R., Niblack, W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques [C]//The Third IEEE International Conference on Data Mining, November 2003,. IEEE Computer Society Press, Los Alamitos, 2003：427-434.

[7] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques[C]//Conference on Empirical Methods in Natural Language Processing. 2002：79-86.

[8] L. Mangasarian, D. R. Musicant. Lagrangian support vector machines[J]. Journal of Machine Learning Research, 2001, 1：161-177.

[9] 徐琳宏，林鸿飞，杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报，2007，21(1)：96-100.

[10] Aizerman M, Braverman E, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning[J]. Automation and Remote Control, 1964, 25：821-837.

[11] Mercer J. Functions of positive and negative type and their connection with the theory of integral equations [J]. Philosophical Transactions of the Royal Society of London, 1909, A209：415-446.

[12] Chapelle O, Vapnik V N, Bacsquest O, et al. Choosing multiple parameters for support vector machine [J]. Machine Learning. 2002, 46：131-159.

[13] Cucker F, Smole S. On the mathematical foundations of learning [J]. Bulletin of the American Mathematical Society, 2001：1-49.

[14] Smithsgf, Jordaanem. Improved SVM regression using mixtures of kernels [C]//Proceedings of the 2002 International Joint Conference on Neural Networks. Washington, DC: IEEE, 2002, 3: 2785-2790.

[15] Turney P, Littman M. . Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003, 21(4)：315-346.