文章编号：1003-0077(2011)04-0030-05

蒙古语有向图形态分析器的判别式词干词缀切分

姜文斌[1]，吴金星[1,2]，乌日力嘎[1,2]，那顺乌日图[2]，刘 群[1]

(1. 中国科学院 计算技术研究所，中国科学院 智能信息处理重点实验室，北京 100190；
2. 内蒙古大学 蒙古学学院，内蒙古 呼和浩特 010021)

摘 要：蒙古语形态分析中，我们之前的有向图模型取得了较高的性能。这种建模方式以图状结构刻画句中词干和词缀之间的概率关系，从而借助上下文信息为每个词确定最佳的切分标注候选。为每个词尽可能地枚举出所有合法的切分标注候选，是有向图模型有效工作的前提。该文提出了一种基于判别式分类的词干词缀切分策略，与之前基于词干表和词缀表的枚举方案相比，该方法对于词中含有未登录词干的情形具有更好的泛化能力。以 20 万词规模的三级标注人工语料库为训练数据，采用判别式词干词缀切分的有向图形态分析器，对于含有未登录词干的情形，词级切分标注正确率提高了 7 个百分点。

关键词：蒙古语；词法分析；词性标注；词干提取；有向图；判别式
中图分类号：TP391　　文献标识码：A

Discriminative Stem-Affix Segmentation for Directed-Graph-Based Mongolian Lexical Analyzer

JIANG Wenbin[1], WU Jinxing[1,2], Wuriliga[1,2], Nasan-urt[2], LIU Qun[1]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Inner Mongolian University, Huhhot, Inner Mongolia 010021, China)

Abstract: In Mongolian lexical analysis, the directed-graph-based model achieves high performance. This model uses a directed-graph architecture to describe the probabilistic relationship of stems and affixes, thus to determine the best segmented and tagged candidate for each word according to the context. Therefore, it is essential for a directed-graph-based analyzer to enumerate all legal segmented and tagged candidates for each word. This paper proposes a novel stem-affix segmentation model based on discriminative classification method for Mongolian lexical analysis. Compared with the enumeration strategy based on the stem- and affix sets, this method shows better generalization ability for the words with unknown stems. Using the 3rd-level annotated corpus with about 200 000 words as the training data, the directed-graph-based lexical analyzer with discriminative stem-affix segmentation module achieves further 7% improvement on F1 measure(with unknown stems considered).

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

1 引言

形态分析对于黏着语来说，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，形态分析的准确率仍有较大的提升空间[5-11]。这一方面是因为这

收稿日期：2011-04-18　定稿日期：2011-05-25
基金项目：国家自然科学基金资助项目(60736014, 60873167)；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；乌日力嘎(1983—)，女，博士生，主要研究方向为蒙古语信息处理。

些语言的研究起步较晚，另一方面更是因为黏着语本身构词规律的复杂性。

与汉语的字符顺次拼接的构词方式相比，蒙古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词缀组成树状结构，形态分析任务需要解析出词语的词干和词缀构成。我们之前提出了一种针对蒙古语构词特性的形态分析模型。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。为这些转移或生成关系赋以合适的概率形式，则形态分析的过程就是寻找其所有概率乘积最大的有向图。该模型取得了较高的性能，但它存在致命的缺点。模型依据从人工语料库中抽取出的词干表和词缀表，通过递归搜索为每个词枚举所有可能的切分标注候选。显然，该方式无法处理含有未登录词干的词语。

我们为蒙古语形态分析的有向图模型提出了一种新颖的词干词缀切分策略。该方法以判别式分类的思路，将词语的词干词缀切分建模为词中字母的标注问题。这可以和基于字符分类原理的判别式汉语分词进行类比，词中字母串对应到汉语分词的句中字串，词干词缀的切分对应到汉语词语的切分。对每个字母进行分类所依据的特征，是取自邻近窗口内的字母子序列。这使得词干词缀切分模块具有了泛化能力，能够处理词中含有未登录词干的问题。

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。在整个测试集上，采用判别式词干词缀切分的最终模型取得了 95.2%的词级切分标注正确率，与采用基于词干表和词缀表的简单枚举方法的情形持平。而对于测试集中含有未登录词干的词，词级切分标注正确率比采用简单枚举的情形提高了 7 个百分点。

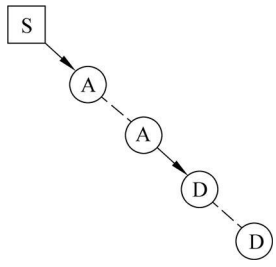
在本文的剩余章节，我们首先介绍之前提出的生成式有向图形态分析模型，然后描述基于判别式分类的词干词缀切分方法，在展示该系统实验结果并进行相应的分析说明后，我们对本文工作给出总结。

2　有向图形态分析模型

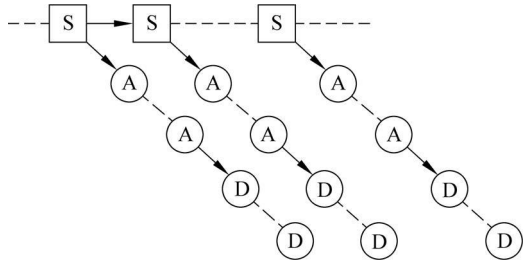
2.1　单纯切分的模型结构

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，我们先从较为简单的任务说起，即单纯切分的有向图模型。

有向图模型把语句中各词的分析结果定义为链状结构：



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：



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。

有向图模型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。这样，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。有向图模型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有如下两种转移概率：

a)　$P(S \mid S\ \text{ngram})$　词干到词干的转移概率，类似于 ngram 语言模型。

b)　$P(X \mid S/X\ \text{ngram})$　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, 有向图模型用这些概率的乘积表示该候选的整体生成概率:

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

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 N-gram 语法模型向树结构的拓展。

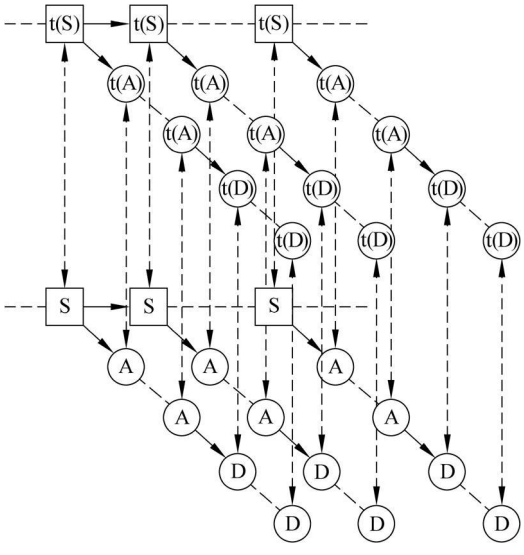
2.2 联合切分标注的模型结构

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。对应于单纯切分的模型结构, 有向图模型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, 有向图模型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) $P(X \mid t(X))$ X 代表词干或词缀, $t(X)$ 代表其标注。此概率类似于隐马模型中状态到观察的生成概率。

b) $P(t(X) \mid X)$ 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:



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:

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

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, 故为它们赋以合适的相对加权有望提升模型性能。但在本工作中暂不考虑乘子加权问题, 这相当于所有加权均为 1。

2.3 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的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]。

解码过程首先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的最优候选。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。基于对训练语料的观察和分析, 我们对之前工作所用的变形规则进行了更改和扩充:

a) 词干词缀划分过程中, 字母串 AYI、EYI、OYI、VYI、OYI 和 UYI 中间的字符 Y 在特定情形下会丢掉。

b) 词干词缀划分过程中, 字母串 GA、HA、YA、RA、MA、YE、RE 和 OS 在特定情形下, 需在中间添加下划线。

实际解码过程中我们采用简单枚举的方案, 在每一处可以应用变形规则的地方, 我们分别尝试应用和不应用两种选择, 从而为待分析词枚举出所有可能的变形状态。每个变形状态都将用于候选分析结果的生成, 这些候选分析结果由接下来的动态规划解码过程进行排歧。动态规划的解码就是自左到右的 viterbi 搜索, 考虑到文章篇幅的限制, 这里不再详述。

3 判别式词干词缀切分

接下来我们介绍基于判别式分类的词干词缀切

分策略。词干词缀切分用于词法分析器解码过程的第一阶段，即词语的候选分析结果枚举。

对于给定的待分析蒙古文词或者其变形形态：

$$W = C_1 C_2 \cdots C_n$$

其中 $C_i(1<=i<=n)$ 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：

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

其中，$e_m = n$，字母序列 $C_{1:n}$ 划分为 m 个子序列。第一个子序列 $C_{1:e_1}$ 是词干，剩余的字母序列是连写词缀或分写词缀。

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。我们将其与基于判别式字符分类的汉语分词进行类比，将每个蒙古文字母 Ci 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

当对整个蒙古文词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者词缀，相应地我们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。我们所用的特征模板列在下面表格中。其中，C0 表示当前考察的字母，C_i/Ci 表示 C0 左边/右边的第 i 个字母。借助这些特征模板，我们从训练语料中抽取字母分类实例，然后用张乐开发的最大熵工具包[①]训练字符分类器。

表 **1**　字符分类采用的特征模板

一元组	二元组	三元组
C_2	C_3 ◦ C_2	C_4 ◦ C_3 ◦ C_2
C_1	C_2 ◦ C_1	C_3 ◦ C_2 ◦ C_1
C0	C_1 ◦ C0	C_2 ◦ C_1 ◦ C0
C1	C0 ◦ C1	C_1 ◦ C0 ◦ C1
C2	C1 ◦ C2	C0 ◦ C1 ◦ C2
	C2 ◦ C3	C1 ◦ C2 ◦ C3
	C_1 ◦ C1	C2 ◦ C3 ◦ C4

考虑到词干词缀切分的歧义性，我们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。N 最佳切分方案可以采用类似于立方体剪枝[16]的策略高效地求得。借助词干和词缀的词性列表，我们可以为每一个词干词缀切分候选枚举出所有可能的词性标注方案，从而得到待切分蒙古文词可能的候选分析结果集。

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% 的语句用作训练集，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。我们沿用之前工作所用的评测指标，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 **2**　变形规则改进和判别式词干词缀切分带来的整体性能提升/%

设　　置	Pw＋t	Pw－t	Fw＋t	Fw－t
原有变形规则	93.0	95.1	93.0	94.7
改进变形规则	95.2	97.7	95.9	97.6
＋判别式词干词缀切分	95.2	97.8	96.1	98.0

对比表 2 的第 1、2 行，变形规则的改进带来了大幅度的整体性能提升[②]。这说明，通过增加有用

① http://homepages.inf.ed.ac.uk/s0450736/maxent-toolkit.html.

② 之前工作中我们不对分写词缀和连写词缀进行区分。本文的形态分析器则区分两种词缀，但仍沿用之前的评测标准。

的变形规则模板和改变变形规则的应用模式,我们更有可能为待分析蒙古文词找到其正确的变形形态,虽然这将产生更多的变形形态候选并进而导致更大的候选分析结果集,但后续的排歧过程仍能有效地找出最佳候选分析结果。然而,在改进变形规则的基础上进一步采用判别式的词干词缀切分,分析精度的提升并不明显,如表 2 的 2、3 行所示。

表 3 判别式词干词缀切分对于词干未登录情形的性能提升/%

设 置	Pw+t	Pw-t	Fw+t	Fw-t
改进变形规则	2.7	45.1	20.5	45.0
+判别式词干词缀切分	9.9	59.3	25.1	60.4

我们认为,采用判别式词干词缀切分策略,其优势更加体现在词中含有未登录词干的情形。当待分析词的词干和词缀都在训练语料中出现时,基于词干表和词缀表的简单枚举方法就能找到正确的分析结果候选。对于蒙古语来说,词缀的数目是有限的,训练语料的数据可以轻易地覆盖全部词缀。词干的情况则复杂得多,新生词和外来词随着社会发展会不断涌现。当待分析词的词干在训练语料中不存在时,简单枚举方式无法找到正确的分析结果候选。而判别式的词干词缀切分策略则可能具有良好的泛化能力,如同汉语分词中的情形。表 3 的实验数据验证了我们的假设。对于含有未登录词干的词,判别式的词干词缀切分策略带来了大幅度的性能提升。

5 总结

本文为蒙古语形态分析的有向图模型提出了一种新颖的词干词缀切分策略。该方法以判别式分类的思路,将词语的词干词缀切分建模为词中字母的标注问题。与基于词干表和词缀表的简单枚举方式相比,基于判别式分类的词干词缀切分策略具有良好的泛化能力,能够有效处理词中含有未登录词干的问题。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。对于测试集中含有未登录词干的词,判别式词干词缀切分策略使得词级切分标注正确率提高了 7 个百分点。

参考文献

[1] Hwee Tou Ng and Jin Kiat Low. Chinese part-of-speech tagging: One-at-a-time or all-at-once? Word-based or character-based? [C]// Proceedings of EMNLP, 2004: 277-284.

[2] Wenbin Jiang, Liang Huang, Yajuan Lv, and Qun Liu. 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 and Hongkui Yu. 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] 那顺乌日图, 淑琴. 面向信息处理的蒙古语规范化探究[J]. 中央民族大学学报(哲学社会科学版), 2007.

[7] 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.

[8] 赵伟, 侯宏旭, 从伟, 宋美娜. 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.

[9] 丛伟. 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.

[10] 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.

[11] 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.

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

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

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

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

[16] Huang Liang and David Chiang. 2005. Better k-best parsing [C]// Proceedings of the IWPT, 2005: 53-64.