基于统计方法的蒙古语依存句法分析模型*

斯·劳格劳，华沙宝，萨如拉
内蒙古大学 蒙古学学院，呼和浩特 010021
E-mail: sloglo@sina.com

摘　要：蒙古语文信息处理已初步完成字、词处理阶段的基本任务，正在步入句处理阶段，并且在国家自然科学基金的资助下构建了蒙古语依存树库 MDTB。本文以 MDTB 为训练和评测数据，设计实现了一种基于词汇依存概率的蒙古语依存句法分析模型。目前，该模型的无标记准确率、有标记准确率和核心词准确率分别达到了 71.24%、61.42%和 93.05%。

关键词：蒙古文；依存语法；句法分析；概率模型

Mongolian Dependency Parsing Model Based on Statistical Methods

S. Loglo, Hua Shabao, Sarula
College of Mongolian Studies, Inner Mongolia University, Huhhot 010021
E-mail: sloglo@sina.com

Abstract: Mongolian language information processing has completed the basic task of word processing stage, is entering the stage of sentence processing. Under the National Natural Science Foundation we have constructed the Mongolian Dependency Treebank MDTB. In this paper, we use MDTB as training and evaluation data, designed and implemented a Mongolian dependency parsing model based on lexical dependent probability. Currently, the model's unlabelled annotation score, labeled annotation score and head word annotation score respectively reached 71.24%, 61.42% and 93.05%.

Keywords: Mongolian; dependency grammar; parsing; probability model

1　引言

蒙古语文信息处理工作始于 20 世纪 80 年代，虽然起步较晚，但发展很快。经过 30 余年的努力，语料库、语法信息词典等基础性建设初具规模，编辑排版系统、办公软件等已实用化，各种蒙古文网络资源也正在稳步增长。从处理层面上看，初步完成了字、词处理阶段的基本任务，现已步入句处理阶段。目前，通过国家自然科学基金项目《现代蒙古语树库的构建》，正在进行树库资源的建设和自动句法分析研究。

纵观各种语言以往的句法标注及分析情况不难发现短语结构语法占据着主流地位，但近年来，依存语法由其形式简洁、易于标注、便于应用等特点受到了研究人员的重视[1]，并在英语、汉语、德语、捷克语等语言句法分析中得到了广泛应用，在被应用的过程中依存语法本身也得到了发展和完善。CoNLL（Computational Natural Language Learning）国际会议从 2006—2009 年连续四次把依存句法分析的评测列入其共享任务[2,3,4]，由此可以看出句法分析和标注采用依存语法是未来的研究热点和发展趋势。

2　基于统计的依存句法分析模型

句法分析的思想是能够根据某种语法 G 给出一个句子 s 的句法分析树 t[5]。在很多情况下，对

* 本文得到国家自然科学基金项目（项目编号：60763003）、国家社科基金项目（项目编号：10CYY022）、教育部人文社会科学研究项目（项目编号：09yjc740045）的资助。

于一个句子会有超过一种句法分析树，我们用 T 表示一个句子所有可能的分析树。在统计句法分析中，建模的目的是寻找一种评价函数，用概率值的大小排列分析结果，并输出最有可能的结果。分析树 t 的概率表示为：

$$P(t|s,G) \text{ 其中 } \sum_{t\in T} P(t \mid s,G) = 1 \tag{1}$$

基于统计的句法分析模型把歧义的消解问题转化为一个最优化的过程，即计算每种分析结果的概率，找出一棵概率最大的分析树 t^*,该过程可以表示为：

$$t^* = \arg\max P(t \mid s) = \arg\max_{t\in T} P(t \mid s,G) \tag{2}$$

随着机器学习方法的快速发展和数据资源的不断丰富，基于统计的依存句法分析也在不断地变换着，当前比较常见的建模方法有生成式依存模型和判别式依存模型。每种模式下也有多种不同的处理技术。本文采用了词汇依存概率模型，它属于一种生成式依存模型。

3 蒙古语词汇依存概率模型

句法分析实质上就是语法歧义的消解过程，而歧义消解是在多种信息的支撑下完成的。消歧所用的各类信息中，词汇本身最具区别能力，因为语言在词汇层面上很少出现歧义现象。但受树库规模限制，基于词汇的统计模型[6]在句法分析中的应用近几年才逐渐开始。

有限的词汇构成一个句子，可以表示为：

$$S=\{<w_1,f_1,t_1>,\ldots,<w_i,f_i,t_i>,\ldots,<w_n,f_n,t_n>\}$$

其中，w_i 表示句中第 i 个词（$1\leqslant i\leqslant n$），f_i 表示词 w_i 的形态特征，t_n 表示词 w_i 的词类及子类标注信息。此时，我们可以把上述式（2）表示为：

$$t^*=\arg\max_{t\in T} P(t \mid <w_1,f_1,t_1>,\ldots,<w_i,f_i,t_i>,\ldots,<w_n,f_n,t_n>,G) \tag{3}$$

为了计算 $P(t \mid <w_1,f_1,t_1>,\ldots,<w_i,f_i,t_i>,\ldots,<w_n,f_n,t_n>)$，对句法分析树中的依存弧进行了独立假设，即一条依存弧只与其两个端点的词语有关，不受其他节点和弧的影响。具有 n 个词的句子，其依存树由 n-1 条依存弧构成，按上述独立假设，其依存树的概率为 n-1 条依存弧的概率的乘积，表示如下：

$$P(t|s)=\prod P(A_{ij} \mid w_i,w_j) \tag{4}$$

其中，A_{ij} 为两个词 w_i,w_j（$1\leqslant i,j\leqslant n$；$i<j$）之间的依存弧，其方向由支配词和从属词的位置所确定，如果 w_i 依存于 w_j，用“1”表示其方向，如果 w_j 依存于 w_i,则用“0”表示其方向。

4 依存概率计算

蒙古语具有丰富的形态变化，另外我们在机器词典中设置了词类和细分类特征，消岐算法可以采用词汇本身，也可以采用词类信息、子类信息、语义信息以及结构信息等。独立假设后，每一条依存弧的概率由两个端点 w_i 和 w_j 唯一确定。为了缓解树库资源规模不足导致的词汇信息数据稀疏问题，本文利用词汇本身的基础上充分利用了支配词和从属词的词类特征、子类特征、形态变化特征（主要是格和动词形态变化）计算了每条弧的依存概率。另外，为了一定程度上克服独立假设所带来的信息丢失问题，也考虑了前后词的词类特征。依存弧的概率由

$P_1(A_{ij})=P(A_{ij}|W_i,W_j)$、$P_2(A_{ij})=P(A_{ij}|CAT_i,CAT_j)$、$P_3(A_{ij})=P(A_{ij}|SUBCAT_i,SUBCAT_j)$、$P_4(A_{ij})=P(A_{ij}|CAT_i,MORPH_i,CAT_j)$、$P_5(A_{ij})=P(A_{ij}|CAT_i,MORPH_i,W_j)$、$P_6(A_{ij})=P(A_{ij}|W_i,CAT_j)$、$P_7(A_{ij})=P(A_{ij}|CAT_{i-1},CAT_i,CAT_j,CAT_{j+1})$等 7 项构成。

除两个端点外，依存距离也是影响依存概率的重要因素，恰当的依存区间的划分方法能够提

高整个句法分析的准确率。依存区间怎么划分没有固定方法，在句法分析过程中进行微调，根据分析器的表现最终确定。但总体上讲，划分的区间数目不宜过多，也不宜过少，划分多了产生数据稀疏问题，划分少了降低模型的描述能力。本文经过试验把依存区间 Distance 划分成了如下的三档：

$$\text{Distance}=\begin{cases}1, & \text{如果} abs(j-i)=1\\ 2, & \text{如果} 2\leqslant abs(j-i)\leqslant 4\\ 3, & \text{如果} abs(j-i)>4\end{cases}$$

式中，abs 表示取绝对值，i,j 分别表示两个端点的位置。本文在训练树库上进行统计，然后采用极大似然估计方法分别计算了 P_1 至 P_7。P_k 的计算公式为：

$$\tilde{P}_k(A_{ij}\mid N_i,\ N_j)=\frac{Count(A_{ij},\ N_i,\ N_j)}{Count(N_i,\ N_j)} \tag{5}$$

式中，$1\leqslant k\leqslant 7$，N_i，N_j 分别表示支配词和从属词相关特征。（N_i，N_j）可以取{（W_i,W_j），（CAT_i,CAT_j），（$SUBCAT_i$,$SUBCAT_j$），（CAT_i,$MORPH_i$,CAT_j），（CAT_i,$MORPH_i$,W_j），（W_i,CAT_j），（CAT_{i-1},CAT_i, CAT_j,CAT_{j+1}）}中的一个。Count（A_{ij}，N_i，N_j）表示节点 N_i，N_j 之间存在依存弧 A_{ij} 的次数，设其依存区间为 Distance。Count（N_i，N_j）表示节点 N_i，N_j 以 Distance 距离出现的总次数，包括构成依存关系的和不构成依存关系的。我们用公式（5）分别计算了上面列出的 7 种概率，计算结果存储在统计信息库中。在解码算法中，先从统计信息库查找概率值，然后进行插值平滑获得了依存概率 P。

$$P=\lambda_1P_1+\lambda_2P_2+\lambda_3P_3+\lambda_4P_4+\lambda_5P_5+\lambda_6P_6+\lambda_7P_7+\lambda_8\xi \tag{6}$$

其中，$\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5+\lambda_6+\lambda_7+\lambda_8=1$，$\xi=0.001$。参数的初始值是按经验设置的，在试验过程中根据分析器的性能反馈进行了微调。$\lambda_1=\lambda_4=0.2$，$\lambda_5=0.15$，$\lambda_2=\lambda_3=\lambda_6=\lambda_7=0.1$，$\lambda_8=0.05$。

5 基于统计的依存句法分析

我们在已建立好的统计信息库和依存概率计算公式的基础上研制了一部基于统计的蒙古语句法分析器 MParser。分析算法采用了一种基于分治的局部寻优策略[7]。其中，分治是指分两步分析句中的前焦型依存关系（支配词在前）和后焦型依存关系（支配词在后）。除特殊情形，蒙古语辅助关系和总括关系为前焦型依存关系，其余的依存关系为后焦型依存关系。具体分析时先标注前焦型依存关系，后标注后焦型依存关系。

标注前焦型依存关系时从句子末尾开始顺序读取词语节点，并从位于当前节点左侧的节点中寻找能够构成前焦型辅助关系的词语节点，如果找到，则建立依存关系，否则读取下一个节点。从当前节点左边的节点中搜索满足条件的节点时为了缩短范围，算法把搜索限制在当前片段中，这就是局部寻优的含义。具有唯一核心词的连续出现的一段词语称为句法片段。一个片段可能是一个单词、短语、成分句或分句。两个片段通过之间的依存关系可以构成更大的片段。句法片段的左右边界或其中的一个边界通常可以通过标点、词类以及特定的词汇和结构信息来确定。在蒙古语句法片段的切分中，逗号、动词、连接词（包括联系动词）和语气词是主要标志信息.

分析后焦型依存关系时，依存关系是在两个相邻的子树（包括叶子节点）之间发生的。我们选两个特殊节点 R 和 L。R 位于右侧子树，并且离左侧子树根节点最近。L 位于左侧子树，并且离右侧子树根节点最近。两棵子树之间的依存关系为下列情形之一（参见图 1），要么左侧子树的根节点依存于 R 或 R 的祖先节点，要么右侧子树的根节点依存于 L 或 L 的祖先节点。算法中节点搜索被限制在 R 和 L 的祖先节点中，这也是局部选优的含义之一。当依存关系类型被确定为后焦型依存关系时，我们只需判断左侧子树与右侧子树中 R 或 R 的祖先节点相结合的可能性。

5.1 算法描述

一个具有 n 个词的蒙古语句子经词切分和片段划分后变成如下形式：

$$W_1\ W_2 \ldots W_i \| W_{i+1}\ W_{i+2} \ldots W_k \| \ldots \| W_m\ W_{m+1} \ldots W_n$$

其中，W_i 表示词语，‖表示片段切分。分析是从位于最右边的两个节点开始的，经过多步分析后一个句子变成如下形式（如图 1 所示）：

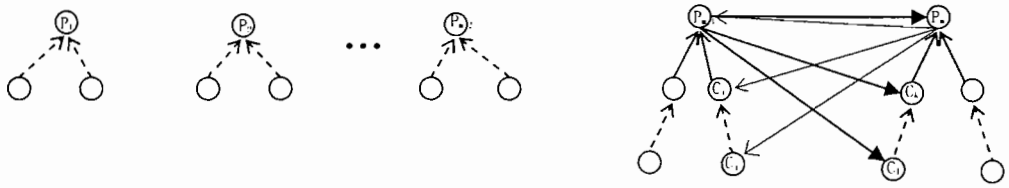

图1　分析算法示意图

图中，虚线表示两个端节点之间有 n 层节点，P_1，P_2，…，P_{m-2}，P_{m-1}，P_m 分别表示各个子树的根节点，C_i，…，C_j，C_k，…，C_l 表示两棵子树中与当前分析相关的子孙节点，这些节点均位于两棵子树的邻接面上，也就是说，在以 P_{m-1} 为根的子树中，这些子孙节点均为其父节点最右侧的孩子节点，在以 P_m 为根的子树中，这些子孙节点均为其父节点最左侧的孩子节点。下一步的分析将在 P_{m-1}、C_i、…、C_j 和 P_m、C_k、…、C_l 之间进行，如同图中的箭头所示。可能产生依存关系的节点组合有：$P_{m-1}\rightarrow C_l$；$P_{m-1}\rightarrow C_k$；$P_{m-1}\rightarrow P_m$；$P_m\rightarrow C_i$；$P_m\rightarrow C_j$；$P_m\rightarrow P_{m-1}$；那么到底哪两个节点之间产生依存关系，取决于两个节点之间的结合能力。计算每一组可能结合的概率，再按概率值进行排序，得分最高的一组建立依存关系，本次分析结束。经过上面的分析，P_{m-1} 和 P_m 被合并为一棵树，合并后的树再与 P_{m-2} 合并。以此类推分析完所有子树为止。

5.2 算法示例

下面通过一个具体例子来说明算法的解码过程。设输入的句子为：ENE BAGVDAL-VN EMUN_E NIGE BUDUGUN VDA BVI, TEGUN-U DEGER_E HEDUN JAGVN VRAN SIBAVHAI 0ND0RCV BAYIN_A.

（1）词法分析：词法分析包括复合词识别、词类标注和构形附加成分的识别等三个任务。这个过程是通过蒙古语词法分析软件实现的[8]。在每个词语后面的花括弧里标注了词法信息。标注时分别用“/”和“;”隔开了同类和不同类信息。特别说明的是对主格没有标注形态信息。ENE{R/Rj} BAGVDAL-VN{N;Fc1} EMUN_E{0/0n/c1/c4} NIGE{M;Fm0} BUDUGUN{A/Ac}VDA{N} BVI{S/Sb}, TEGUN-U{R/Rj;Fc1} DEGER_E{0/0n/c0/c1} HEDUN{R/Ra/mRa} JAGVN{M/Fm0} VRAN=SIBAVHAI {N/Nn}0ND0RCV{V/Ve/Ve2;Fn1} BAYIN_A {V/Vz/Vz1;Fs2}.

（2）片段切分：按照定义将例句切分为“ENE BAGVDAL-VN EMUN_E NIGE BUDUGUN VDA BVI,”和“TEGUN-UDEGER_E HEDUN JAGVN VRAN SIBAVHAI 0ND0RCV BAYIN_A.”两个片段。

实验表明，如果一个句子中只有前焦型依存关系或者只有后焦型依存关系，那么这个句子的分析难度会大大降低。真实文本中这种句子很少，但句法分析的某个阶段子树之间的依存关系可以变为纯前焦型或纯后焦型关系。为了达到这种目的，我们在句法分析算法中先标注了前焦型依存关系。分析时从句子或片段右端开始两两配对，如果两个节点的属性特征满足某条规则的条件，则其间建立依存关系。为了便于描述，我们用词语在句中的位置代替了词本身。分析过程如图 2、3、4 所示。

(a) 第一片段中标注前焦型依存关系后　　(b) 第二片段中标注前焦型依存关系后

图2　前焦型依存关系标注

(a) 第一片段中标注后焦型依存关系后　　(b) 第二片段中标注后焦型依存关系后

图3　后焦型依存关系标注

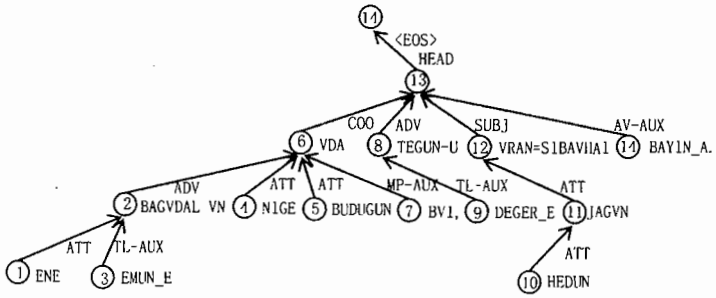

图4　片段合并后生成的依存树

6　实验及分析

实验数据采用了内蒙古大学蒙古语文研究所建设的中学蒙古语文依存树库 MDTB（Mongolian Dependency Tree-Bank），该库共有 461，240 个词，31，722 个句子，每个句子的平均长度为 14.54 个词。全部试验中，以第二册至第十一册为训练语料，包含 26,737 个句子，402,432 个单词，平均句长为 15.05 个词。以第一册和第十二册为测试语料，包含 4985 个句子，58,808 个单词，平均句长为 11.80 个词。我们以无标记准确率（只有依存弧，没有类型标记，Unlabeled Annotation Score，简称 UAS）、有标记准确率（有依存弧，也有类型标记，Labeled Annotation Score，简称 LAS）和核心词的准确率（Head-word Annotation Score，简称 HAS）为指标评测了 MParser。

由于句子长度对句法分析的准确率影响很大,我们统计了测试集中句子的长度分布,并根据不同词长的句子分别进行评测，下面是评测结果，如表 1 所示。

表 1　MParser 的评测结果

Length	1—5	5—10	10—15	15—20	20—25	25—30	30—35	35—40	40—45	45—50	>50	平均
比例	21.35	32.62	21.94	11.53	5.65	3.26	1.48	0.79	0.61	0.43	0.34	11.08 词
UAS	88.39	76.92	71.90	68.61	67.35	65.97	64.00	64.42	61.73	60.48	54.58	71.24
LAS	78.07	66.15	61.91	59.31	57.77	57.10	54.26	57.07	52.31	52.36	44.28	61.42
HAS	96.13	94.83	94.08	92.63	92.45	92.13	91.69	91.73	90.92	91.68	89.43	93.05

从实验结果看，核心词查准率非常高，这是因为蒙古语是核心词后置的语言，句中除一些辅助成分，核心词一般出现在句末。LAS 比 UAS 低十个百分点左右，这是由同型结构引起的。例如，与格形式的名词有时充当状语，有时充当间接宾语，很难判断，但其依存对象的判断相对简单一些，因此 UAS 和 LAS 的分数出现了较大的差距。通过扩充训练语料可以提高 LAS 值。

参 考 文 献

[1] 刘海涛. 依存语法和机器翻译，语言文字应用，1997(3)，89-93.

[2] Jan Hajič, Massimiliano Ciaramita, Richard Johansson et al. The CoNLL-2009 Shared Task on Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 2009. 1-18.

[3] Mihai Surdeanu, Richard Johansson, Adam Meyers et al. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning*, 2008. 159-177.

[4] Joakim Nivre, Johan Hall, Sandra Kübler et al. The CoNLL-2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL2007*, 2007. 915-932.

[5] Eisner, J. M.. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of ACL-1996*, 1996. 340-345.

[6] M. Collins, Three Generative, Lexicalized Models for Statistical Parsing. *Proceedings of the 35th annual meeting of the association forcomputational linguistics*, 1997. 16-23.

[7] 马金山，基于统计方法的汉语依存句法分析研究，哈尔滨工业大学博士学位论文，2007. 78-90.

[8] S. Loglo, HuaShabao and Sarula, Research on Mongolian Lexical Analyzer Based on NFA,In *Proceedings of 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems (Volume 2)*, 2010. 240-245.