

融和丰富语言知识的汉语统计句法分析

熊德意^{1,2} 刘群¹

¹(中国科学院计算技术研究所, 北京 100080);

²(中国科学院研究生院, 北京 100039)

E-mail: dyxiong@ict.ac.cn

摘要: 我们的汉语统计句法分析模型从 3 个方面融合丰富的语言特征知识: 1) 利用非递归名词短语的相对确定性重新标注树库中的名词短语; 2) 设计新的中心词映射表; 3) 引进上下文配置框架。这些语言特征知识使模型的性能提高了 10%。

关键词: 统计句法分析 非递归名词短语 中心词映射表 上下文配置框架

引言

基于树库的统计句法分析是现代句法分析的主流技术, 它的本质就是在有指导训练的前提下, 从树库中自动学习一个映射函数, 将句子从线性序列结构映射到某种正确的句法树结构上。[1][2][3]等模型都是经典的统计句法分析模型, 这些模型的共同特点是不需要人工繁琐地建立各种消歧规则。

统计句法分析面临的一个主要问题是如何发现和利用具有强消歧能力的语言特征知识, 同时保证语言知识的应用不会使模型的参数急剧膨胀而导致严重的数据稀疏问题。人工建立的树库显然是一个较理想的语言知识库, 该知识库中蕴含了各种语法规则, 词汇依存关系, 词类规则等知识。然而直接建立在树库基础上的统计模型分析效果并不是很理想^[2], 分析能力更强的模型^{[1][3][4]}常常吸收了更多的语言知识, 这些语言知识一部分用来改造和扩充树库, 另一部分则内嵌在模型结构中, 构成模型新的参数类型。

我们的模型从三个方面来融合和吸收丰富的语言特征知识: 1) 利用非递归名词短语的相对确定性重新标注树库中的名词短语; 2) 设计新的中心词映射表; 3) 为了利用标点符号, 并列结构, 修饰距离等语言特征知识, 我们提出了上下文配置框架, 该框架能够很容易地将这些语言知识内嵌在模型结构中。三种语言特征知识的利用使模型的性能提高了将近 10%。

本文按如下方式组织: 第 2 节简单介绍我们的统计句法分析模型; 第 3 节介绍模型所利用的语言特征知识, 探讨非递归名词短语的标识, 新的中心词映射表以及上下文配置框架; 第 4 节给出我们的实验结果; 最后是我们的结论和未来的研究方向。

1. 统计模型

我们的统计模型建立在 Collins Model 2 基础上, 该模型提供了一个很好的平台, 能够融合多种语言特征知识, 同时仍保证模型结构的清晰性, 一致性。Collins Model 2 的核心在于它的基于中心词的规则 Markov 分解, 任何一条上下文无关规则 (规则右部为非词类标记), 都可以被看作如下的形式:

$$P[h] \rightarrow L_{n+1}[l_{n+1}]L_n[l_n] \cdots L_1[l_1]H[h]R_1[r_1] \cdots R_m[r_m]R_{m+1}[r_{m+1}]$$

其中 P 为父结点, H 为中心子结点, L、R 为 H 的左右修饰结点 (其中 $L_{n+1} = R_{m+1} = STOP$, 表示

规则右部两端的终结)。方括号内的标记代表对应结点的中心词（由单词以及其词类构成的二元组 $\langle w, t \rangle$ 组成），父结点的中心词从中心子结点的中心词继承而来。按照概率乘法定理，规则的概率为：

$$P(L_n[l_n] \cdots L_1[l_1] H[h] R_1[r_1] \cdots R_m[r_m] | P, h) = P_h(H | P, h) \times \prod_{i=1..n+1} P_l(L_i[l_i] | L_1[l_1] \cdots L_{i-1}[l_{i-1}], P, H, h) \times \prod_{j=1..m+1} P_r(R_j[r_j] | L_1[l_1] \cdots L_{n+1}[l_{n+1}], R_1[r_1] \cdots R_{j-1}[r_{j-1}], P, H, h)$$

我们称概率 P_l 和 P_r 的条件部分为 *history*，并将它们统一表示为：

$$P_m(M_i, w_i, t_i | history, direction)$$

通常要对 *history* 作等价类划分，以减少模型参数的个数；我们的模型基于 2 阶 Markov 独立性假设：

$$P_m(M_i, w_i, t_i | history, direction) = P_m(M_i, w_i, t_i | M_{i-1}, P, hw, ht, direction) = P_{m_n}(M_i, t_i | M_{i-1}, P, hw, ht, direction) \cdot P_{m_w}(w_i | M_{i-1}, M_i, t_i, P, hw, ht, direction)$$

整个句子中心词的概率为：

$$P_{top}(H, ht, hw | P = TOP) = P_{top_n}(H, ht | P = TOP) \cdot P_{top_w}(hw | H, ht, P = TOP)$$

对于词类规则 $t \rightarrow w$ ，其概率为 $P_{pos}(w | t)$ ¹。一棵分析树的概率是上面四种参数类型

($P_h, P_m, P_{top}, P_{pos}$) 对应的所有概率的乘积。

2. 语言特征知识

这一节介绍我们的模型如何从三个方面融合多种语言特征知识，即非递归名词短语的标识，新的中心词映射表和上下文配置框架。对每种语言特征知识，我们将说明引进它们的原因，以及模型是如何利用这些知识来改善的本身的性能的。

2.1 非递归名词短语

我们定义非递归名词短语为子结点中不包含任何名词短语 NP 的短语，如(NP (NN 国务院) (NN 发展)(NN 研究) (NN 中心))中的 NP 是非递归名词短语，而(NP (NP (NR 浦东)) (NP (NN 开发)))中的第一个 NP 就不是非递归名词短语。引进非递归名词短语的主要原因是：非递归名词短语由于本身结构的内聚性（宾州树库的标注风格使得非递归名词短语的内部结构不再被细分），其两端的界往往比较明显，始端常常有形容词性的修饰语。因此引进非递归名词短语，模型在短语边界处计算的终结概率 $P_m(STOP | history, direction)$ 将会足够大，这样分析器不会错误地将“担任总统期间”

¹ 在 Collins 模型中，此推导被认为是确定的，即概率为 1。

中的“总统期间”判为一个名词短语。

基于以上的分析，我们对宾州汉语树库的名词短语重新标注，将那些不含其他名词短语的非递归名词短语的标记 NP 改为 BNP，同时为了树库标注的一致性的原则，对那些重新标注为 BNP 的结点，如果其父结点的标记不是 NP，则在其上增加一个父结点 NP，如(LCP(NP(NT 今年))(LC 底))重新标注为(LCP(NP(BNP(NT 今年)))(LC 底))。在测试分析器性能时，所有的 BNP 又被重新修改为 NP，同时那些新增加的 NP 将被去掉。非递归名词短语的标识使我们的模型性能 F1 值增加了近 8%。

2.2 中心词映射表

由于树库并没有标注中心子结点，中心词驱动统计句法分析模型一个首要的任务是要寻找任何父结点对应的中心子结点。通常的作法是人为构造一个中心词映射规则表，然后由算法将规则表强加在树库上。映射规则表中的规则类似如以下的形式：

$$P \text{ direction } \langle h_1, h_2, \dots, h_n \rangle$$

P 为父结点， h_i 为 P 可能的中心子结点，direction 为映射方向，如对映射方向为 left 的规则，算法将从父结点的最左子结点开始，自左向右，将所有子结点与该父结点的中心子结点列表匹配，匹配上的子结点则为该父结点的中心子结点。

一个直观的看法是，即使中心词映射规则表不是模型的关键，也会对模型的性能产生一定的影响。[4]的汉语句法分析器采用的是[6]中定义的中心词映射表，我们修改了该表中的两条规则，一条是关于 CP（由标志语“吗”、“的”、“的话”等引起的单句，见[7]）结构的，另外一条是关于 UCP（并列短语结构，但并列成分类型不同，见[7]）结构的，修改情况见表 1。

表 1

父结点	原来的映射规则	新的映射规则
CP	CP right CP IP	CP right DEC SP
UCP	UCP right UCP	UCP left PU CC

中心词映射规则表的修改

修改 CP 结构映射规则的一个主要原因是原来的映射规则导致概率 P_h ， P_m 计算的稀疏性，因为以 IP 作为中心结点，其中心词 hw 就可能千变万化，而以标志语 DEC、SP 作为中心子结点，则中心词的范围是可控的，基本上集中在“吗”、“的”、“的话”等几个单词上；同时 DEC 和 IP 在 CP 结构中出现的概率基本上相等，这样原来的映射规则使得概率 P_h ， P_m 的 history 部分呈现多样性，而新规则使 history 内敛，从而在一定程度上削弱了数据稀疏。而修改 UCP 映射规则的一个原因除了与修改 CP 映射规则的原因相同外，另一个原因是原来的规则基本上不起作用，因为 UCP 出现递归嵌套的可能性很少。新的映射规则表使模型性能增加了 1.3%。

2.3 上下文配置框架

逗号、顿号等标点符号对成分的边界有预示作用，同时它和连词 CC 连接的成分往往处在分析树的同一个层次上。但是对模型的输出结果分析，我们发现分析器在处理附着结构时常出错，如本来应该在同一个层次的附着，却分析成较高层次上的附着或较低层次上的附着。可能原因是我们的模型独立性假设过于强硬，在计算概率 P_m 时，模型将 history 映射为中心子结点和临近的前一个兄弟结

点，也就是说当前修饰结点的概率只依赖于这两个已经扩展的结点。实际上，子结点的概率可以依赖于以前扩展的任何结构，但是这样做会引入大量的参数；一个好的做法是这些后来扩展的结构的概率不是直接依赖于以前扩展的结构，而是依赖于这些结构的某个函数（即 history 的等价类函数）。

上下文配置框架就是一种等价类函数，它的思想来源于 Collins 模型的 distance 函数，但是将它推广了。上下文配置框架是对当前扩展结构的上下文的一个简单描述，它是一个各分量取值 0/1 的向量 $\langle \tau_1, \tau_2, \dots, \tau_n \rangle$ ，向量中的每一个分量对应上下文的某个特征，该特征在上下文中出现，则对应分量取值 1，否则为 0。即：

$$\tau_i = \begin{cases} 1 & \text{if } f_i \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

显然这和最大熵模型的特征结构非常类似。我们的配置框架（下面我们称为 CCF）采用了 3 个特征，第一个特征是 direction，即修饰结点在中心结点的哪一边（我们将它作为特征，是想让模型更紧凑），第二特征是修饰结点是否与中心结点临近（即中间不存在任何其他结点），第三个特征是在中心结点和当前修饰结点之间是否存在逗号、顿号或连词 CC。引入配置框架后，概率 P_m 按如下方式计算：

$$P_m = P_{m_n}(M_i, t_i | M_{i-1}, P, hw, ht, CCF) \cdot P_{m_w}(w_i | M_{i-1}, M_i, t_i, P, hw, ht, CCF)$$

CCF 的应用使得模型性能增加了 0.5%，虽然效果不很明显，但是 CCF 的优越性是非常吸引人的，这将是我们的未来的一个研究方向。

3. 试验设计和结果

我们的分析器采用基于 chart 的自底向上算法，运用 beam 搜索策略。其中 beam width 设置为 10^9 ，即每个 cell 中的边的概率只有超过该 cell 目前最大概率的 10^9 才会被保留，否则搜索算法将会剪掉该边。同时我们硬性规定每个 cell 中最多容纳 50 条边，以加快分析速度。这些参数的设置和[4]中的一致，实际上，我们的试验设计也采用了他们的配置，即将树库按照 8:1:1 的大致比例划分为训练集，调试集，测试集；1-270 为训练集，含句子 3477 个，271-300 为测试集，含句子 348 个，301-325 为调试集，含句子 350 个。所有的数据都经过了正规化处理，即去掉无用的标点符号，如“、”、“《、》”等（这些标点符号对分析没有正作用，同时在树库设计时，它们往往也是在最后才标注的），去掉空结点，去掉递归的一元规则，即形如 $n \rightarrow n$ 的规则。表 2 是我们的实验结果，baseline 模型采用[6]中的中心词映射规则表。由于我们的参数设置、试验数据的分配和[4]中的一致，因此我们也和他们的模型进行了比较，比较结果见表 3。我们的模型性能优于他们的 PCFG 模型，劣于 TAG 模型，这一定程度上说明我们的模型吸收的语言知识还不够丰富，仍有改进的空间。

表 2

	Len. <= 40			Len. <= 100		
	LR	LP	F1	LR	LP	F1
Baseline	60.9	72.8	66.3	58.6	70.8	64.1
Baseline + BNP	69.6	79.5	74.2	67.4	77.5	72.1
Baseline + BNP +NHT	70.9	80.6	75.5	68.8	78.8	73.5
Baseline + BNP + NHT + CCF	72.3	80.0	76.0	70.2	78.2	74.0

融合不同语言特征知识的模型的试验比较结果。LR 和 LP 分别表示标记召回率和正确率；+BNP 表示用非递归名词短语重新标注了树库中的名词短语；+NHT 表示模型采用了新的中心词映射表，即 Xia（1999）中的映射表经过表 1 修改后的结果；+CCF 表示模型采用了上下文配置框架。

表 3

	Len. <= 40		
	LR	LP	F1
Bikel & Chiang 2000 BBN Model	69.0	74.8	71.8
Present work	72.3	80.0	76.0
Bikel & Chiang 2000 TAG Model	76.2	77.2	76.7
Chiang & Bikel 2002 TAG Model	78.8	81.1	79.9

我们的模型的与[4]的模型比较。BBN 模型是基于词汇化的 PCFG 模型，TAG 模型是基于树粘接语法的统计模型。

4. 结论和未来方向

我们的试验结果表明语言特征知识（即与语言相关的特征）对统计句法分析有很大的影响，这从一个侧面指出了汉语统计句法分析研究的一个方向：从语言学角度寻找更多的特征知识。长期以来，自然语言处理中存在两种不同的研究思路，一种是采用知识丰富型方法（knowledge-rich approach）解决各种自然语言处理任务，另一种是试图依赖计算机强大的计算能力来解决问题，而不考虑语言上的特征。显然从统计句法分析的角度来看，一个好的计算模型加上丰富的语言特征知识才是上选。

我们下一步将继续沿着语言特征知识的方向探索，从概率和语言学角度寻找具有强消歧能力的语言特征知识，同时保证语言知识的利用不会使模型的参数急剧膨胀而导致严重的数据稀疏问题。我们将继续丰富上下文配置框架，考虑各种可能的特征。

参考文献

- [1]. Michael Collins. Head-Driven Statistical Models for Natural Language Parsing. PhD thesis, University of Pennsylvania, 1999.
- [2]. Charniak Eugene. 1996. Tree-bank Grammars. AAAI/IAAI, Vol. 2.
- [3]. Dan Klein, Christopher D.Manning. 2003. Accurate Unlexicalized Parsing. In Proceedings of the 42th Association for Computational Linguistics.

- [4]. Daniel M. Bikel and David Chiang. 2000. Two statistical parsing models applied to the chinese treebank. In Proceedings of the Second Chinese Language Processing Workshop, pages 1-6.
- [5]. David Chiang and Daniel M. Bikel. 2002. Recovering Latent Information in Treebanks In the proceedings of COLING 2002.
- [6]. Fei Xia. Automatic Grammar Generation from Two Different Perspectives. PhD thesis, University of Pennsylvania, 1999.
- [7]. Nianwen Xue and Fei Xia. 2000. The Bracketing Guidelines for Chinese Treebank Project. Technical Report IRCS 00-08, University of Pennsylvania.

作者简介: 熊德意 (1979-), 男, 湖北黄石人, 直博生, 目前的研究方向为统计句法分析, 统计机器翻译; 刘群 (1966-), 男, 江西萍乡人, 博士, 副研究员, 主要研究领域为机器翻译, 自然语言处理与中文信息处理。

Chinese statistical parsing with rich linguistic features

Devi Xiong^{1,2} Qun Liu¹

¹(Institute of Computing technology, the Chinese Academy of Sciences, Beijing, 100080);

²(Graduate School of the Chinese Academy of Sciences, Beijing, 100039)

E-mail: dyxiong@ict.ac.cn

Abstract: Rich linguistic features are incorporated into our model for Chinese statistical parsing by the following three ways. First of all, non-recursive noun phrases are annotated in the Penn Chinese Treebank because of their strong mark of boundaries. Second, a new head percolation table is designed based on Xia's table. The last linguistic feature our model uses is context configuration frame which builds a platform for incorporating knowledge about commas, coordination constructions and so on. All these three linguistic features give about 10% improvement of our model.

Key words: statistical parsing; Non-recursive NPs; head percolation table; context configuration frame