中 文 信 息 学 报
第19卷 第3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 19 No. 3

文章编号: 1003－0077(2005)03－0061－06

融合丰富语言知识的汉语统计句法分析①

熊德意[1,2], 刘 群[1], 林守勋[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)

摘要: 知识获取一直以来是自然语言处理中的瓶颈, 基于树库的统计句法分析也不例外。树库中潜在隐含的语言知识是非常丰富的, 但它们并不是可以直接得到, 往往需要特定的策略才能将它们融合到模型中。我们的汉语统计句法分析模型从3个方面融合潜在的丰富语言知识: 1) 重新标注树库中的非递归名词短语和非递归动词短语; 2) 设计新的中心词映射表; 3) 引进上下文配置框架以更具体地描述二元依存结构。由于融合了以上三种潜在语言知识, 模型的F1值提高了2.37%, 完全匹配正确率提高了5.36%。

关键词: 人工智能; 自然语言处理; 统计句法分析; 非递归短语; 中心词映射表; 上下文配置框架

中图分类号: TP391 文献标识码: A

Chinese Statistical Parsing with Rich Linguistic Features

XIONG De-yi[1,2], LIU Qun[1], LIN Shou-xun[1]

(1. Institute of Computing technology, the Chinese Academy of Sciences, Beijing 100080, China;
2. Graduate School of the Chinese Academy of Sciences, Beijing 100039, China)

Abstract: Knowledge acquisition is always regarded as a bottleneck in many NLP tasks, such as machine translation, information extraction. Treebank-based statistical parsing is not an exceptant. The latent linguistic knowledge in treebank is very rich, which, however, can't be acquired directly. In our model, the following three ways are used to incorporate such rich linguistic features for Chinese statistical parsing. First of all, non-recursive noun and verb 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 the context configuration frame which provides a stronger representation of bilexical dependency structures. All these three linguistic features gain an improvement of remarkable 2.37% in terms of F1 measure, 5.36% in terms of complete match ratio.

Key words: artificial intelligence; natural language processing; statistical parsing; non-recursive NPs; head percolation table; context configuration frame

1 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。文献[1~3]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。

① 收稿日期: 2004－07－28 定稿日期: 2004－12－31
基金项目: 国家863计划资助项目(2003AA111010; 2001AA114010)
作者简介: 熊德意(1979－), 男, 湖北黄石人, 博士生, 研究方向为统计句法分析, 统计机器翻译.

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],其原因在于直接建立在树库基础上的模型可以获取到树库中的显示知识,但是很难获取隐含在树库中的,丰富的潜在语言知识。显示知识是指树库中被明确标注的知识,这些知识可以通过规则,直接的计数融入到模型中;隐示的知识则指树库中没有被明确标注的,但潜在的信息,如在词汇化的概率上下文无关文法(LPCFG)模型中,中心词的定位是一种很重要的语言知识,但是在树库中,中心词是没有被标注的。如果要获取中心词知识,就需要定义一个启发式的中心词映射规则表。所以,分析能力更强的模型[1,3,4]常常吸收了更多的隐示语言知识。

我们的模型从三个方面来融合和吸收丰富的隐示语言知识:1)重新标注树库中的非递归名词短语和非递归动词短语,非递归短语和递归短语有着非常不同的内部外部分布特性,但宾州树库的标注风格并没有将它们区分开来;2)设计新的中心词映射表,在 LPCFG 模型中,中心词的定位是非常重要的,因而指导这种定位的中心词映射表自然对模型的性能有着一定的影响;3)引进上下文配置框架以更具体地描述二元依存结构,由于 Markov 独立性假设的存在,LPCFG 模型对二元依存结构的描述往往不够充分,许多重要的限制信息都被忽略了,如两个依存的语言成分之间的距离。上下文配置框架可以将这些信息以一种紧凑的方式纳入到模型中。三种隐示语言特征知识的利用使模型的 F1 值提高了 2.37%,完全匹配正确率提高了 5.36%。

本文按如下方式组织:第二部分简单介绍我们的统计句法分析模型;第三部分介绍模型所利用的隐示语言知识,探讨非递归名词和动词短语的标识,新的中心词映射表以及上下文配置框架;第四部分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。

2 统计模型

我们的统计模型建立在 Collins Model 1[1] 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。模型的核心在于它的基于中心词的规则 Markov 分解,任何一条词汇化的上下文无关规则(规则右部为非词类标记),都可以按如下的形式分解成更小的语言对象:

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

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$<Hhw, Hht>$组成,也称为对应非终结符的词汇项),父结点的中心词从中心子结点的中心词继承而来。相应地,规则的概率也分解成三个部分。第一部分是在给定父结点句法标记 P,中心词 Hhw 以及对应的词性 Hht 条件下,生成中心子结点句法标记 H 的概率:

$$Pr_H(H \mid P, Hhw, Hht)$$

中心子结点扩展之后,它的每一个的左/右修饰结点按照下面两个步骤扩展得到:首先是在给定父结点(P),中心子结点(H, Hhw, Hht),前一个已扩展的姊妹结点$(M_{i-1}, M_{i-1}ht)$,以及方向信息 dir 条件下,生成句法标记 M_i 和对应的中心词词性 M_iht:

$$Pr_M(M_i, M_iht \mid P, H, Hhw, Hht, M_{i-1}, M_{i-1}ht, dir)$$

第二步生成新修饰结点的中心词 M_ihw:

$$Pr_{M_w}(M_i hw \mid P, H, Hhw, Hht, M_{i-1}, M_{i-1}ht, M_i, M_i ht, dir)$$

对以上三个分布概率,模型采用 Witten-Bell 插值方法来平滑。对于概率 Pr_M,与 Collins[1] 不同的是,我们建立了六层 back-off 结构,而不是三层,原因是六层 back-off 结构极大地提高了模型的 F1 值。另外一个不同之处在于,我们的模型是基于二阶 Markov 假设的。模型在计算由当前修饰结点 M_i 和中心子结点 H 构成的二元依存结构的概率 Pr_M 和 Pr_{M_w} 时,还必须考虑前一个已扩展的修饰结点 M_{i-1}。虽然依存结构中已包含了很多的信息,但是如上文所述,仍有些重要信息被忽略了,下节将要介绍的上下文配置框架的主要功能就是重新捕获这些信息。

句法分析器的所有输入是已经分词但未作词性标注的句子。对于未登陆词,模型根据该词的第一个字来估计概率 $Pr(unword \mid tag)$,类似于 Levy 等人[5] 的方法。如果第一个字也没有出现,则采用绝对减值法估计。为了加快句法分析器搜索的速度,我们采用文献[6] 中推荐的方法,将完全边的 beam 阈值设为 9,非完全边的阈值设为 7。

3 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、动词短语的标识,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本身的性能。

3.1 非递归名词、动词短语

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。非递归动词短语的定义与此类似。引进非递归短语的主要原因是非递归短语和递归短语有着非常不同的内部外部分布特性。就名词短语而言,非递归名词短语内部结构中不包含任何 NP 标记,所有名词都是以词性标记而非句法标记出现,也就是说,非递归名词短语,要么本身只有一个名词,要么有多个名词但内部结构基本上不再做句法层上的分析,因此非递归名词短语的内部结构是非常平坦的。递归名词短语则相反,其内部结构有着明确严格的句法层次,不同的修饰成份,中心成份处在不同的句法层次上。就外部结构而言,非递归名词短语和递归名词短语也是非常不同的,非递归名词短语外部修饰结构比较稳定,一般由量词性短语,形容词性短语充当;递归名词短语的外部环境,相对而言,要灵活得多,复杂多变得多。这就使得非递归名词短语的边界相对于递归名词短语的边界要明显的多。总而言之,非递归名词短语内部结构趋于平坦,外部环境比较稳定,边界相对确定;递归名词短语内部结构有很强的层次感,外部环境多变,边界模糊。虽然这两种短语有明显的区别,宾州树库并没有区分它们,而是用统一的标记 NP 来标示它们。这样,来自于递归名词短语的计数信息和非递归名词短语的计数信息混合在一起,模型很难加以区别利用。在训练集中,经过我们的初步统计,非递归名词短语有 23522 个,递归名词短语有 11500 个,显然,如果不区分它们,大量的计数信息因为混合在一起而失去了判别作用。

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。与 Collins 不同的是,我们并没有引入额外的 NP 层①,我们的做法更类似于 Klein[3] 等人的做法。重标注时,模型不需要做任何变动,只需要将训练集中的非递归名词短语标记由 NP 改为 BNP,原来属于 NP 的一部分计数就会自然地分给 BNP;而

① Collins 重新标注 BNP 时,为了保持树库标注的一致性,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,分析器标注的所有 BNP 又被重新修改为 NP。实验结果表明,非递归名词短语的标识使模型性能有了显著的改善。

非递归动词短语的标注类似于非递归名词短语。最初我们的想法是非递归动词短语的重新标注同样可以提高分析器的性能,但是实验结果表明,恰恰相反,模型性能下降了一点。我们分析,这可能是动词短语的结构并没有名词短语那么复杂,标记区分的过细,反而会使得计数信息分解到不同的标记上而导致数据稀疏。但是如果同时标注非递归动词短语和非递归名词短语,试验结果表明,分析器 F1 值总体上仍得到了改善,原因现在还不是很清楚,我们认为可能是非递归动词短语和非递归名词短语之间有某种共现特性,更深入的分析将是我们未来的工作。

3.2 中心词映射表

由于树库并没有标注中心子结点,中心词驱动的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以下的形式:

$$P\ direction < h_1, h_2, \cdots, h_n >$$

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。

一个直觉的看法是,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。文献[4]的汉语句法分析器采用的是文献[8]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[9])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[9])结构的,修改情况见表 1。

表 1 中心词映射规则

父结点	原来的映射规则	新的映射规则	#Train	#Test	%Train	%Test
CP	CP right CP IP	CP right DEC SP	2203	226	63.4	64.9
UCP	UCP right UCP	UCP left PU CC	55	5	1.6	1.4

#Train 和 #Test 分别代表该父结点在训练集和测试集中出现的次数,%Train 和%Test 分别代表父结点在训练集和测试集中的覆盖率。我们可以看到,在训练集中,每个句子平均含 0.634 个 CP 结构,测试集中每个句子平均含 0.649 个 CP 结构。显然 CP 结构的覆盖率是非常高的,其对应映射规则的变化足以引起分析器性能的改变。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 Pr_H, Pr_M, Pr_{M_w} 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 Pr_H, Pr_M, Pr_{M_w} 的条件部分(我们称之为 history)呈现多样性,而新规则使 history 内敛,从而在一定程度上消弱了数据稀疏。UCP 的覆盖率虽然很低,修改它的原因主要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。

3.3 上下文配置框架

如上文所述,我们的模型是基于二阶 Markov 假设的。实际上,这种独立性假设仍过于强硬,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整数的向量$<\tau_1, \tau_2, \cdots, \tau_n>$,向量中的每一个分量对应上下文的某个特征,取值可以根据实际需要来定义。

我们的配置框架(下面我们称为 *CCF*)采用了两个特征,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),第二个特征是修饰结点与中心结点之间的距离,这种距离按照下面的方式定义:

1. 如果是 *STOP* 结点,并且与中心结点之间无任何其他结点,则距离为 0;如果间隔一个结点,则距离为 1;如果间隔二个或二个以上的结点,则距离为 2;

2. 如果是非 *STOP* 结点,并且与中心结点之间无任何其他结点,则距离为 0;如果间隔一个或一个以上结点,则距离为 1。

引入配置框架后,二元依存结构的概率按如下方式计算:

$$Pr_M(M_i, M_i ht \mid P, H, Hhw, Hht, M_{i-1}, M_{i-1} ht, CCF)$$

$$Pr_{MW}(M_i hw \mid P, H, Hhw, Hht, M_{i-1}, M_{i-1} ht, M_i, M_i ht, CCF)$$

在定义特征时应避免数据稀疏,以及与 history 中已有的信息重叠。过多的特征会引起数据稀疏问题,因此一定要寻找最重要的特征。Collins 的 distance 函数中考虑了二元依存结构之间是否间隔有标点符号和动词性成分,我们的 *CCF* 当然也可以考虑这些因素,但是这些信息和 history 中的前一个扩展结点 M_{i-1}有一定的交叉。试验表明,如果去掉 history 中 M_{i-1}结点信息,*CCF* 引入这两种特征会提高模型的性能;相反,如果保留 M_{i-1}结点信息,*CCF* 引入这两种特征会降低模型的性能。但是在总体性能上,保留 M_{i-1}结点信息,分析器结果会更好,所以我们选择了保留 M_{i-1}结点信息,但 *CCF* 不引入这两种特征。

4 试验设计和结果

表 2 融合不同特征的实验结果

	累计值				单个语言特征引起的性能增加值	
	F1	DF1	CM	DCM	DF1	DCM
Baseline	75.92	—	27.42	—	—	—
BNP	77.58	1.66	30.77	3.35	1.66	3.35
BVP	77.75	1.83	30.43	3.01	−0.16	0.67
CCF	78.01	2.09	32.44	5.02	0.22	1.01
NHT	78.29	2.37	32.78	5.36	0.26	0.34

我们的试验设计采用了文献[4]中的配置,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,所有结果都是在句子长度小于等于 40 个单词的基础上得到的,baseline 模型采用文献[8]中的中心词映射规则表。由于我们的参数设置、试验数据的分配和文献[4,7]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,劣于优化

的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。

融合不同语言特征知识的模型的试验比较结果。F1 为标准的标记召回率和正确率的平均值,DF1 为 F1 值的增加值,CM 为完全匹配正确率,DCM 为 CM 的增加值。BNP 表示用非递归名词短语重新标注了树库中的名词短语;BVP 表示用非递归动词短语重新标注了树库中的动词短语;CCF 表示模型采用了上下文配置框架;NHT 表示模型采用了新的中心词映射表。

表 3　现有结果与其它模型结果的比较

	Len ＜ 40		
	LR	LP	F1
Bikel & Chiang 2000 BBN Model	69.0	74.8	71.8
Bikel & Chiang 2000 TAG Model	76.2	77.2	76.7
Present work	78.0	78.6	78.3
Chiang & Bikel 2002 TAG Model	78.8	81.1	79.9

我们的模型与文献[4, 7]的模型的比较。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。

5　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。我们将进一步研究非递归短语标注的作用,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 [D]. PhD thesis, University of Pennsylvania, 1999.

[2] Charniak Eugene. 1996. Tree-bank Grammars [A]. AAAI/IAAI [C], Vol. 2.

[3] Dan Klein, Christorpher D. Manning. 2003. Accurate Unlexicalized Parsing [A]. In: Proceedings of the 42th Association for Computational Linguistics [C].

[4] Daniel M. Bikel and David Chiang. 2000. Two statistical parsing models applied to the chinese Treebank [A]. In: Proceedings of the Second Chinese Language Processing Workshop [C], 1—6.

[5] Roger Levy, Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? [A]. In: Proceedings of the 42th Association for Computational Linguistics [C].

[6] Deyi Xiong, Qun Liu and Shouxun Lin. 2005. Lexicalized Beam ThresholdingParsing with Prior and Boundary Estimates [A]. In: Proceedings of CICLing 2005[C], Mexico.

[7] David Chiang and Daniel M. Bikel. 2002. Recovering Latent Information in Treebanks [A]. In: Proceedings of COLING 2002 [C].

[8] Fei Xia. Automatic Grammar Generation from Two Different Perspectives [D]. PhD thesis, University of Pennsylvania, 1999.

[9] Nianwen Xue and Fei Xia. 2000. The Bracketing Guidelines for Chinese Treebank Project [R]. Technical Report IRCS 00—08, University of Pennsylvania.