词汇化概率句法分析与动词子语类框架获取的互动方法

冀铁亮[1] 穗志方[2]

（1. 北京大学计算语言研究所，北京 100871；2. 北京大学计算语言研究所，北京 100871）

摘 要：概率句法分析器（PCFG Parser）是基于概率规则集的上下文无关文法的句法分析器。规则集主要是针对词类和短语类。然而事实上，词性相同而词汇不同，其所常用的句法规则也通常不同。目前NLP研究的一个趋势和热点就是词汇化的句法分析。针对概率句法分析独立性假设中缺乏词汇化的缺陷，本文将谓语动词的子语类信息与概率句法分析结合起来，提出一种基于动词子语类信息的词汇化概率句法分析方法。论文建立了基于汉语动词子语类框架的统计句法分析模型，并且针对动词子语类框架难以获取的问题，提出一种词汇化概率句法分析与动词子语类框架获取的互动方法。实验利用这种互动的方法获取了汉语中十个常用高频动词的概率化子语类信息，并结合原有的概率句法分析器PCFG实现了一个基于动词子语类信息的概率句法分析器原型系统S-PCFG。实验证明了基于动词子语类信息的概率句法分析对自然语言句法分析的准确率和速度均有所提高。同时分析了新的概率句法分析器的不足之处，为进一步的改进提供条件。

关键词：词汇化概率句法分析；子语类框架；词汇知识自动获取

An Interactive Method for Lexicalized Probabilistic Parsing and Verb's Subcategorization Frame Acquisition

Ji Tieliang[1],Sui Zhifang[2]

(1.Institute of Computational Linguistics, Peking Universtiy, BeiJing,100871; 2.Institute of Computational Linguistics, Peking Universtiy, BeiJing,100871)

Abstract: PCFG Parser is a context-free parser that is based on probabilistic rules. One of the main problems of PCFG is the lack of lexicalized analysis. This paper aims to solve this problem by bring forward a lexicalized probabilistic syntactic analysis method based on verbs' subcategorization frame. We first set up a probabilistic model for syntactic analysis based on verbs' subcategorization frame. Furthermore, in order to deal with the bottleneck of subcategorization frame acquisition of Chinese verbs, we put forward an interactive method for lexicalized probabilistic parsing and verb's subcategorization frame acquisition. After training the probabilistic subcategoriaztion frames for some common-used Chinese verbs, we implemented an S-PCFG parser (subcategorization frame based PCFG parser). By comparing results of both PCFG parser and S-PCFG parser, we prove that S-PCFG parser is more accurate and efficient than PCFG parser.

Keywords: lexicalized probabilistic parsing, subcategorization frame acquisition, lexical knowledge acquisition

1 引言

统计句法分析的基本原理是利用概率评价模型评价每一棵候选句法树存在的可能性，并选择概率值最高的候选句法树为最终的分析结果。随机上下文无关语法模型PCFG是上下文无关语法模型CFG在统计领域的最显而易

本文相关研究得到国家自然科学基金 60503071 和北京市自然科学基金 4052019 的支持

作者简介：冀铁亮（1982-），男，吉林省四平市，在读硕士，E-mail: jtl@pku.edu.cn

见的扩展，即给CFG的规则集中的每一条规则赋予一个概率分布$P(A \to \partial)$，此概率被解释为规则的右部A扩展为左部∂的概率，并且$\sum_i P(A \to \partial_i) = 1$。候选句法树的概率为生成该句法树所用到的所有规则的概率乘积$P(T \mid S) = \prod_{A \in T} P(A \to \partial)$。

PCFG的一个重要缺陷在于其概率模型中未引入词汇化信息。近年来，各国学者提出了不同的词汇化的统计句法分析模型。[1][2]在PCFG的基础上加入了用规则扩展一个成分时句法规则的结构和词汇上下文信息，这些上下文信息包括：当前句法成分的父结点类型以及输入句中以当前规则左端的第一个成分为中心的词性三元组；[3][4]以句法树的中心词之间的依存关系为特征建立统计句法分析的概率模型；[5]在评价句法树的概率时，考虑当前句法成分的中心词及其父结点类型信息[6]。将句法树中的词汇、句法、语义以及结构信息结合到一个统一的概率评价模型之中。与PCFG相比，以上的概率句法分析模型在概率规则的扩展时，考虑了上下文结构以及词汇方面等更多的特征信息，因此，这些模型比PCFG更加准确。

词汇化句法结构规则的建立对于词汇化概率句法分析是至关重要的。子语类框架表示的是动词与它可带的句法成分之间的搭配模式。它描述的是每一个动词可以具有的基本句法行为，可以解决句法层面上存在的大量歧义问题。带有概率信息的动词子语类框架是词汇化概率句法分析可以利用的重要基础资源。本文将谓语动词的子语类信息与概率句法分析结合起来，提出一种基于动词子语类信息的词汇化概率句法分析方法。论文建立了基于汉语动词子语类框架的统计句法分析模型，通过词汇化概率句法分析与动词子语类框架获取的互动方法，利用这种互动的方法获取了汉语中十个常用高频动词的概率化子语类信息，进而，结合原有的概率分析器PCFG实现了一个基于动词子语类信息的概率句法分析器原型系统S-PCFG。实验证明了基于动词子语类信息的概率句法分析对自然语言句法分析的准确率和速度均有所提高。

2 基于动词子语类信息的概率句法分析

2.1 PCFG Parser

一个PCFG G包括：一个终结符集合，$\{w_k\}$,k=1，…，V、一个非终结符集合，$\{N_i\}$，i=1，…，n、一个指定的初始符，N_1、一个规则集合，$\{N_i \to \zeta_j\}$(在这里ζ_j指的是一个终结符或非终结符序列)、一个对应的规则概率集合：$\forall i \sum P(N_i \to \zeta_j) = 1$

随机上下文无关文法PCFG的一个重要缺陷在于其概率模型中未引入词汇化信息。在PCFG中，短语结构规则VP□V NP NP的概率与规则右部的具体动词V是无关的。这种假设是不合理的。就如前面所提到的，如果规则右部的具体动词V是双宾动词“给”，则该规则的概率应该比规则右部的具体动词V是单宾动词“吃”的概率大得多（P(VP(Head=’给”)□V NP NP)> P(VP(Head=’吃”)□V NP NP)）。这类问题可以通过在概率句法分析中引入动词的子语类框架信息来解决。

2.2 子语类框架的概念

什么是子语类框架？子语类框架由英文“Subcategorization Frame”翻译而来，它表示的是（作为谓语的）动词与修饰它的必需的句法成分之间的搭配模式。例如：

“给”的子语类框架为：给 NP NP （给 你 一本书）；

从某种意义上看，子语类框架可以看作是词汇化了的短语结构规则。进一步，动词的子语类框架可以看作以该动词作为中心词的动词短语的短语结构规则。例如：

VP(Head=给)□V NP NP； VP(Head=喜欢)□V VP； VP(Head=喜欢)□V NP。

这种词汇化的短语结构规则可以在语法词典中动词的子语类框架属性中加以标明。随着HPSG以及Lexicalized Tree Adjoining Grammar等一系列词汇主义语法理论的发展，子语类框架已经成为动词的词汇知识的

最基本的表示形式。

2.3 基于动词子语类信息的概率句法分析 S- PCFG Parser

2.3.1 建立基于汉语动词子语类框架的渐进的统计句法分析模型

如何将动词的子语类框架信息与概率句法分析结合起来？核心是建立基于汉语动词子语类框架的统计句法分析模型。如果固定为自顶向下，从左向右，深度优先的推导策略，一个句子S的句法树T唯一对应一个推导过程D，D可以用重写规则序列r1, r2 ...rm来表示，所以对于输入句S，一个句法树的概率可以被定义为在给定输入句S为的条件下，生成当前句法树的推导过程的概率。

$$P(T \mid S) \approx \prod_{i=1}^{n} P(r_i \mid [r_1, r_2, ...r_{i-1}, S]) \approx \prod_{i=1}^{n} P(r_i)$$

在PCFG的基础上，基于动词子语类框架的统计句法分析模型在扩展规则时，可以考虑更多的词汇信息。

基于动词子语类框架的统计句法分析模型可定义为：

$$P(T \mid S) = \prod_{A \neq VP} P(r_i \mid A) \bullet \prod_{A=VP} P(P(Head \mid A) \bullet P(r_i \mid A, Head))$$

其中P(Head|A)代表在所有的动词中具体的各个中心词动词所占的比率，P(ri|A,Head)表示作为中心词的动词的各个句法分析规则的概率。

基于动词子语类框架的渐进的统计句法分析模型可定义为：

$$P(T \mid S) = \prod_{A \neq VP \vee Head \notin SFS} P(r_i \mid A) \bullet \prod_{A=VP \wedge Head \in SFS} P(P(Head \mid A) \bullet P(r_i \mid A, Head))$$

3.2.1 2.3.2 S- PCFG Parser 定义

一个S-PCFG 包括：

一个PCFG 、一个针对具体动词的子类框架集合， $\{ N_i(v) \rightarrow \zeta_j \}$(v这里指的是具体的一个动词)、

一个针对具体动词的子类框架的概率集合 $\forall v \sum P(N_i(v) \rightarrow \zeta_j) = 1$ 。可以看出，基于动词子语类信息的PCFG比通常意义的PCFG增加了针对具体动词的子类框架及其概率信息。

S- PCFG Parser的基本原理是：在进行句法分析时，如果所用的规则右部出现动词，则对于该动词，如果系统中存在V的子类框架规则，则优先用这类规则；如果关于当前动词V，系统目前还没有V的子类框架的概率信息，则使用默认的规则。

3 词汇化概率句法分析与动词子语类框架获取的互动方法

由2.3.1可知，基于动词子语类框架的统计句法分析模型可定义为：

$$P(T \mid S) = \prod_{A \neq VP \vee Head \notin SFS} P(r_i \mid A) \bullet \prod_{A=VP \wedge Head \in SFS} P(P(Head \mid A) \bullet P(r_i \mid A, Head))$$

其中P(Head|A)代表在所有的动词中动词Head所占的比率，P(ri|A,Head)表示作为中心词的动词Head的各个句法分析规则所占的比率。在未经过句法分析的实际语料中，这两个比率都是难以得到的，而目前树库的规模还不足以支持子类信息的获得。针对这个问题，我们提出了词汇化概率句法分析与动词子语类框架获取的互动方法。

3.1 基本思想

基本思想是：利用一个现有的概率句法分析器对句子进行分析，在分析结果的基础上提取动词的子类信息，进而利用获得的子语类框架信息对概率句法分析器进行修正；再利用修正后的概率句法分析器对句子进行分析，得到新的动词子语类框架信息的分析结果；经过多次迭代分析，可以将该中心语动词的最常用的规则概率放大。

从而实现在提高概率句法分析器准确率的同时，较为准确地提取动词的子语类框架信息。

3.2 具体实现

首先利用概率句法分析器分析句子，得到中心语动词的句法分析规则的概率，然后利用中心语动词的句法分析规则对句子进行重新分析，在得到的分析结果中选取概率最大的五个作为正确的分析结果，对其中含有中心语动词的句法分析规则进行计数，从而得到新的中心语动词的句法分析规则的概率。重复上述互动步骤，直到得到的分析结果中概率最大的五个句子的句法分析与此次迭代之前的分析结果相同，则停止互动，将得到的最后一次的中心语动词的句法分析规则的概率作为中心语动词最终的句法分析规则的概率。

4 实验

4.1 实验语料的准备

利用北京大学中文系和北京大学计算语言所开发的概率句法分析原型系统作为论文所用的PCFG parser；现有的句法规则集共有规则628条，各种规则总的应用次数为12016次，可以近似的代表句法分析规则在实际当中的应用比例。本论文实验从北京大学计算语言所的1998年人民日报标注语料集中提取含有常用的十个动词（成立、从事、代表、发表、反应、欢迎、说明、通过、争取和支持）的带有词性标注的句子，其中含有每个动词的句子各1000句，总计10000个测试语料句。每一个句子都进行了基础切词以及词性标注。可以近似模拟每个动词在实际语言中各种应用的比率。

4.2 S-PCFG parser 的实现

4.2.1 针对每个具体动词的概率化子语类信息 P(ri|A,Head)的获得

由于实际当中，并没有时间和充足的人力对训练集中所有句子的分析结果人工筛选正确结果。论文采取一种近似的方法对分析的结果进行处理：利用改进后的概率句法分析器对词性标注句子进行句法分析，得到的分析结果的概率最大的五个分析结果中，含有正确的分析结果的概率很大。因此，就近似选取分析结果中概率最大的五个用来记录中心动词所应用的语法分析规则记录。随后，将在词汇化的概率句法分析器上进行动词子语类框架获取和句法分析的互动，以便更准确地获得的动词的子语类信息。

4.2.2 P(Head|A)的获得

在实际的自然语言当中，每一个动词的使用概率是难以统计的。由于论文针对的语料是1998年人民日报标注集，根据其本身的特点论文假设该标注集为自然语言的全集，而对标注集中的所有动词进行了计数统计，然后又对十个常用动词进行了计数统计。将各常用动词的统计数除以总的动词统计数。就可以得到各动词在总的动词当中所占的比率。

具体的统计结果如下表：

表1 中心语动词比率统计表

Tab 1 Approximate rate of the ten commonly used verbs among all of the verbs

动词总数	成立	从事	代表	发表	反映	欢迎	说明	通过	争取	支持
257627	4595	2012	2127	4476	2960	2356	1530	3032	1348	5838
比率	0.018	0.008	0.008	0.017	0.011	0.009	0.006	0.012	0.005	0.023

3.2.2 4.2.3 实现 S-PCFG parser

在获得针对每个具体动词的概率语法规则集P(ri|A,Head)以及P(Head|A)之后，可以利用这些资源实现S-PCFG parser。对待分析的词性标注句子进行句法分析时，首先在切词的过程中，如果发现有十个中心语动词，对规则的匹配时，首先要在具体动词的概率语法规则中进行查找。如果匹配成功，那么就返回该规则，同时将规则的概率P(ri|A,Head)与该中心动词的在总的动词中所占的比率P(Head|A)相乘得到的结果作为对规则的概率参数修正，

将参数修正与原规则的概率相加得到新的规则的概率，如果没有匹配成功就在总的概率规则集中进行匹配，以后的功能就与PCFG的原理相同。

4.3 实验结果

在1998年人民日报标注集中针对十个动词各选取十个句子作为测试语料，利用原有的概率句法分析器原型进行句法分析，然后在词汇化的概率句法分析器上进行动词子语类框架获取和句法分析的互动，对得到的结果进行比较，验证词汇化对概率句法分析的准确性和效率的提高。

论文只选择了正确率和运行时间两个关键参数作为PCFG parser、初始S-PCFG parser以及最终S-PCFG parser性能比较的主要指标，同时对动词子语类框架获取的互动进行了迭代计数，其结果如下表所示：

表2 PCFG与S-PCFG性能比较

Tab 2 Comparison among the PCFG initial S- PCFG and final S-PCFG

中心语动词	PCFG Parser		初始S-PCFG Parser		最终S-PCFG Parser		
	正确率	运行时间（毫秒）	正确率	运行时间（毫秒）	正确率	运行时间（毫秒）	迭代次数
成立	0.5	129591	0.5	94848	0.6	94472	4
从事	0.7	86830	0.7	71677	0.7	72937	4
代表	0.4	6540	0.6	5282	0.6	5308	5
发表	0.5	103842	0.5	91123	0.5	92838	4
反映	0.5	14982	0.5	12906	0.7	13339	6
欢迎	0.4	7105	0.7	6598	0.7	6610	3
说明	0.5	8079	0.6	8306	0.6	8216	5
通过	0.5	6054	0.5	4977	0.5	5079	4
争取	0.7	22253	0.7	24809	0.8	24552	4
支持	0.5	7749	0.5	7295	0.7	7343	4
平均	0.52	39302	0.58	32782	0.64	33069	4.7

S-PCFG Parser在准确率和总体运行时间上比PCFG Parser有相对明显的提高。其中正确率由原来的平均52%提高到58%，正确率提高了6个百分点，提高了将近百分之十，同时，运行速度也有较大提高。而经过动词子语类框架获取的互动迭代后的S-PCFG Parser正确率达到了64%。

4.4 实验结果分析

S-PCFG Parser在准确率和性能上比PCFG Parser的提高，论文分析主要有以下两点原因：

（1）针对不同的动词应用具体的概率使得概率规则词汇化，降低了概率句法分析独立性假设对词汇化的弱化。在PCFG句法分析中，完全没有利用词语之间的词汇依存关系。最显著的特点就是缺乏词汇化，vp扩展成为一个动词跟随两个名词短语的概率与参与的具体动词之间是独立的，而实际当中支持或者表示之类的双宾语动词要比其它动词的可能性大得多。而经过S-PCFG用中心语来实现词汇化进行改进以后，在分析句子的时候，不再是面向所有的动词的平均值，而只是针对具体一个动词的概率规则，使得这种缺陷在一定程度上得以弥补。

（2）具体的针对中心语动词的概率句法规则在一定程度上起到了快表的作用。由于论文所做的实验所用语料都是含有中心语动词的句子，在句法分析的规则匹配过程中，程序先是对具体的概率规则进行匹配，匹配失败然后再对概率规则集进行匹配，这个过程与快表的原理很相似，由于具体的概率规则集很小，查找消耗的时间比查找总的概率规则集要快得多，因此一旦匹配成功，就会节约很多时间。同时测试语料中都含有中心语动词，因此在具体规则集中的匹配成功概率很高，从客观上提高了整个分析过程的速度。

5 结论

本文将谓词动词的子语类信息与概率句法分析结合起来，提出一种基于动词子语类信息的词汇化概率句法分

析方法。论文首先建立了基于汉语动词子语类框架的统计句法分析模型，并且，针对动词子语类框架难以获取的问题，提出一种词汇化概率句法分析与动词子语类框架获取的互动方法。实验利用语料训练集产生汉语中十个常用高频动词的概率化子语类信息，并结合原有的概率句法分析器PCFG实现了一个基于动词子语类信息的概率句法分析器原型系统S-PCFG。实验证明了基于动词子语类信息的概率句法分析对自然语言句法分析的准确率和速度均有所提高。同时分析了新的概率句法分析器的不足之处，为进一步的改进提供条件。

目前S-PCFG只是针对10个常用动词获得了它们的子语类框架信息，在一定程度上解决了概率句法分析在词汇化方面的缺陷。由于动词的数量巨大，如果针对每一个动词产生一个词汇化概率规则集，则需要相当大的时间和资源代价。动词子语类信息的有效获取是一个亟待解决的问题。关于这个问题，国内外都有一些相关研究[8] [9] [10] [11]，但至今为止，还没有一部从语料库中提取的大规模的子类框架词典。在子类框架资源的建立方面还需要投入更多的努力。

致谢：

感谢北京大学计算语言所詹卫东老师提供概率句法分析规则及概率句法分析原型系统，感谢张化瑞老师、段惠明老师提供人民日报标注集。

参考文献：

[1] Magerman, D.M. and Marcus, M.P., "Pearl: a probabilistic chart parser", in Proceedings of the European ACL Conference, Berlin, Germany.1991

[2] Magerman, D.M. and Weir, C., "Probabilistic prediction and Picky chart parsing", in Proceedings of the February 1992 DARPA Speech and Natural Language Workshop, Arden House, NY.1992

[3]David M. Magerman, "Statistical decision-tree models for parsing", in Proceedings of the 33th Annual Meeting of the ACL.1995

[4] Collins. M. J, "A new statistical parser based on bigram lexical dependencies", in Proceedings of the 34th Annual Meeting of the ACL.1996

[5] Collins. M. J, "Three generative lexicalised models for statistical parsing". in Proceedings of the 35th Annual Meeting of the ACL.1997

[6] Charniak. E, "Statistical parsing with a context-free grammar and word statics", in Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI Press/MIT Press, Menlo Park.1997

[7] Black. E, Jelinek, F.,Lafferty, J.,Magerman, D.M., Mercer, R. and Roukos, S. "Towards history-based grammars:using richer models of context in probabilistic parsing" in Proceedings of DARPA Speech and Natural Language Workshop, Arden House, NY. the February 1992

[8] Briscoe T, J Carroll, "Automatic extraction of subcategorization from corpora" in Proceedings of the fifth Conference on Applied Natural Language Processing, Washington, DC.1997

[9] Ushioda, A., Evans, D., "The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora, In Boguraev, B. and Pustejovsky, J. eds", in Proceedings of SIGLEX ACL Workshop on the acquisition of lexical knowledge from text, Columbus, Ohio.1993

[10] Anna Korhonen. "Improving SubcategorizationAcquisition using Word Sense disambiguation", in Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sappora, Japan. 2003

[11] Xiwu Han, Tiejun Zhao "SubcategorizationAcquisition and evaluation for Chinese verbs", in Proceedings of Journal of Software Feb. 2006.17(2):259-266