汉语匹配分析算法的实现

周　强

清华大学计算机科学与技术系
智能技术与系统国家重点实验室
北京 100084

摘要：本文介绍了汉语匹配分析算法的实现方法，包括算法的基本控制结构、主要控制函数及其处理流程、匹配成分的定性方法等内容，并通过对一个具体实例的分析过程的描述对此算法进行了直观说明。

关键字：括号匹配原理，句法分析，汉语分析器。

Implementation of the Chinese Parsing Algorithm Based on Bracket Matching Principle

Zhou　Qiang

State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science, Tsinghua University, Beijing 100084
zhouq@s1000e.cs.tsinghua.edu.cn

ABSTRACT: This paper introduces some detailed issues about the implementation of the Chinese bracket matching algorithm, including its basic data structures and main controlling functions, labeling methods for the matched constituents and so on, and gives an example to describe the algorithm process.

KEYWORDS: Bracket Matching principle, Syntactic Parsing, Chinese Parser.

一、研究动机

考虑这样一个分析问题：以如下一组特征向量作为分析器的输入，如何通过左右括号的合理匹配，在此基础上组合产生所有可能的句法成分，最终形成输入句子的完整分析树（或森林）？

$$<w_1, t_1, b_1>, <w_2, t_2, b_2>, \ldots\ldots, <w_n, t_n, b_n>$$

其中 $W= w_1,w_2,\ldots,w_n$ 为句子的词语串，$T = t_1,t_2,\ldots,t_n$ 为各词语相应的词类标记串，$B=b_1,b_2,\ldots,b_n$ 则是一串成分边界信息描述，b_i 可取值 0,1 或 2，分别表示此词语处于句法成分的中间位置、左边界（即被赋予左括号）和右边界（即被赋予右括号）位置，它可以利用现有的成分边界自动预测工具[ZQ96]得到。其中将涉及到两个重要的子问题：1) 成分划分问

题，即哪些左右括号对可以相互匹配形成一个可能的句法成分。2) 成分定性问题，即这些匹配形成的成分能标以什么样的句法标记。本文就试图对这些问题给出一个完整的解决方案。

从直观上看，匹配分析算法可以这样来进行：从左向右扫描句子，直至发现一个基本左右括号对（即它们中间没有其他任何括号），对此进行匹配操作，形成一个句法成分，然后以此为驱动，向左寻找相邻的左括号或已匹配成分，向右寻找相邻的右括号，不断匹配形成新的句法成分。这个过程自左向右、自底向上不断进行，直至扫描到句子结束为止。文献[ZQd96]对此算法进行了形式化定义，并且通过括号匹配原理严格证明了这种算法分析结果的完备性，从而为目前的匹配分析算法的实现奠定了坚实的理论基础。

二、基本控制结构说明

通过对 LR 分析器[MT86]和图分析器[TW83]的有效控制结构的合理吸收和适当改进，形成了匹配分析算法的三个基本控制结构：

1) **括号匹配栈**(BMS)

这是控制括号匹配过程的主要结构，其中的每个栈元素为一个四元组<F,B,E,BPS>结构，包含以下信息：① 成分标记(F)：用于标识词元素(='W')和归约生成的短语（即句法成分）元素(='P')。② 成分边界(B)：以一个左右边界对<L,R>保存词或短语在句子中的边界位置，便于匹配控制函数在栈中搜索得到合适的成分边界。③ 边号指针(E)：保存词或短语的相应边号，以便于生成不同短语的子成分信息表。④ 界定预测(BP)：对于词元素，保存自动预测得到的所有成分边界信息，用于控制括号匹配操作的进行。

而对栈的基本控制操作则包括：

① PushBMS：压入一个栈元素。

② PopBMS：弹出一个栈元素，得到有关的信息。

③ GetBMSNodeInfo：搜索得到栈中某个元素节点的信息。

④ ReduceBMSPath：对栈中元素进行归约。

事实上，对一个句子的括号匹配处理，就是交替地执行以下几个操作：压入栈元素、弹出栈元素、搜索特定栈元素的信息、进行括号匹配、在一定条件下归约栈元素，直至处理完句子中的所有词语。此时栈中将形成如下的栈元素排列：

$$[F,<L_1, R_1>, ...], ... , [F,<L_m, R_m>, ...]$$

其中有：$L_1 = 1, R_i = L_{i+1}$ ($i \in [1,m-1]$), $R_m = n$ (n 为句子中的词语总数)

从 BMS 的结构和实现功能来看，它非常类似于 LR 分析器中所用的栈结构，但由于 BMS 中保存了成分边界预测及其位置信息，可以通过对栈元素的检索操作得到相应的匹配成分位置，因此即使对于很复杂的句子的括号匹配处理，也不必采用类似 Tomita 算法[MT86]中的图结构栈结构，从而大大降低了 BMS 的空间消耗。

2) **压缩共享森林**(PSF)

由一组边组成，保存了经括号匹配而得到的全部短语。其中的每条边包含以下信息：① 成分标记：用于标识短语边和词边。② 成分边界：标识词或短语在句子中的起始位置和终止位置。③ 句法标记：保存匹配短语的句法标记信息。④ 压缩子节点表：保存此短语成分的所有歧义结构组合信息。每个结构组合是由所有子成分的边号组成的 PSF 边号路径，以节省空间消耗。⑤ 最佳路径标记：保存经排歧处理得到的最佳结构组合路径的指针信息。

其中信息项 ③、④、⑤ 只对短语边有意义。

与传统的活动图分析方法[TW83]不同的是，这里采用了压缩共享森林结构[MT86]，它有以下几个好处：① 通过节点信息的压缩，节省了大量存储空间，提高了短语成分的检索速度。② 由于在压缩节点中保存了分析过程中遇到的所有结构歧义现象，可以很方便地对此进行统计排歧和剪枝处理，便于从中选择一个最佳的分析结果。

3) 待扩展路径表(PEL)

保存了在检索 BMS 过程中所发现的所有待扩展路径的信息。通过对其中路径的排序处理，可以控制不同短语的组合优先性。

三、匹配分析算法描述

目前的匹配分析是通过逐词扫描输入句子完成的：首先将词语的有关信息压入 BMS 中，如果某词语的界定预测输出中包含右括号，则检索 BMS 中的路径信息，进行匹配处理，产生短语成分加入 PSF 中，当所有的匹配操作完成后，在适当的条件下，可以对 BMS 进行归约。然后再处理下一个词，如此不断进行，直到句子结束。下面给出算法中的三个主要控制函数的详细描述。此算法的复杂性为 $O(n^3)$。

1). 匹配操作总控函数

输入： S=<W,T,B>

背景知识： 树库统计数据和语言规则描述。

基本流程：

① 初始化 BMS， PSF 和 PEL；

② 顺序处理句子中的所有词语；

- 创建一条词边，将一个词元素压入 BMS ；
- 如果界定预测串包括右括号，则加一条待扩展路径入 PEL，进行括号匹配 MatchBrkt( )；

2). 括号匹配函数： MatchBrkt( )

处理目标： 匹配所有可能的左右括号，生成相应的短语边加入 PSF 中。

输入： 待匹配处理的右括号位置(rbpos)。

背景知识： 树库统计数据和语言规则描述。

基本流程：

① 顺序处理 PEL 中所有的待扩展路径(其终止位置=rbpos)；

- 自待扩展路径的起始位置起，检索得到它在 BMS 中的前一个栈元素；
- 如果是短语元素，则组合短语边 CombinePEdges( )；
- 否则(是词元素)，针对该词语的不同界定预测分别进行以下处理：

a). $b_i = 0$，则通过加入一个新的短语成分，形成一条新的扩展路径，并加入 PEL 中。

b). $b_i = 1$，则：

- 通过匹配成分定性分析（见下节），检查其句法合理性，并确定合适的句法标记（或标记串）；
- 如果是合理的匹配成分，则：
  - 创建一条或多条短语边，加入 PSF 中；

■ 将这些短语边作为新的待扩展路径加入 PEL 中；

c). $b_i = 2$，则组合短语边 CombinePEdges( )；

② 归约 BMS （若满足归约条件）；

3). 组合短语边处理函数： CombinePEdges( )

处理目标： 将一个短语或一组短语成分与其它短语组合成新的更大的短语。

输入： 向左检索得到的第一个右括号位置(fbpos)，
待匹配处理的右括号位置(rbpos)，
待扩展的子节点路径。

背景知识： 树库统计数据和语言规则描述。

基本流程：

① 检索得到 PSF 中所有终止位置等于 fbpos 的短语边；

① 顺序处理所有这些短语边；

■ 通过在待扩展的子节点路径中加入短语边边号，形成可能的新短语的子节点表；

■ 通过匹配成分定性分析（见下节），检查其句法合理性，并确定合适的句法标记（或标记串）；

■ 如果是合理的匹配成分，则：

■ 创建一条或多条短语边，加入 PSF 中；

■ 将这些新短语边作为新的待扩展路径加入 PEL 中；

四、匹配成分的定性分析

对匹配成分的定性分析，即确定它们的句法功能标记，可以充分利用其内部的结构组合信息和外部的语境约束信息对它们进行语法合理性检查，以排除那些不合语法的匹配组合。它可以通过以下几个步骤来完成：

首先，根据语言学上的句法成分组合规律，在现有的词类标记集和句法标记集基础上，列举出常见的错误组合情况，并总结有用的错误句法结构判定规则，据此可以排除大量不合语法的匹配成分。

然后，进行以下几个阶段的成分定性分析：

① 利用如下形式的一组句法标记归约规则：

句法结构组合 ::[成分特征约束] --> {句法标记，归约概率}+

可以为绝大多数合乎语法的成分组合标上合适的句法标记。这些规则可以通过以下两个途径得到： A) 人工总结， B) 树库(treebank)统计，它们具有各自的优势。人工总结可以为不同的句法结构给出丰富的特征约束信息，从而得到很准确的归约标记；而树库统计则可以从大规模的标注语料库中获得比较客观的归约标记概率分布信息，从而为多个归约标记的选择提供一定的依据。两者的结合就能取得最佳的处理效果。

② 利用边界标记分布数据：

词类 t_i 界定预测('[' 或 ']') 词类 t_{i+1} → {句法标记，分布概率}+

可以得到匹配成分的两个边界在其局部语境下的标记分布集，若它们的交集不为空，就可以此作为整个成分的句法标记。类似的，这组数据也可以通过人工总结和树库统计两条途径得

到。一个具体的处理实例是：

为了确定句子片段“... 在/p [一/m 九/m 七/m 八/m] 年/q ...”中的匹配成分“[一/m 九/m 七/m 八/m] ”的句法标记，首先检索边界标记分布信息表，得到以下的两组边界分布数据：

p [m --> mbar 0.97, tp 0.03

m] q --> mp 0.35, np 0.44, sp 0.11, mbar 0.06, tp 0.04

然后计算两个集合：{mbar,tp}和{mp,np,sp,mbar,tp}的交集，就可以得到此匹配短语的句法标记集： mbar-tp 。

③ 对剩余的句法成分，指派一个不在句法标记集中出现的特殊标记，通过对它们的分析统计可以发现一些新的结构组合形式。

五、一个具体处理实例

本节给出了对一个具体实例的匹配分析过程。选择了成分边界预测结果“[安装/v [在/p [桌子/n 上/f] 的/u 灯/n]”作为分析器的输入，通过记录分析过程中 BMS 和 PSF 的变化来显示匹配算法的具体流程。为简单起见，对 PSF 中的边信息描述进行了以下简化：对于词边，只记录词性标记和词语信息；对于短语边，则记录句法标记和它的压缩子节点表信息。

BMS

‘W’ <3,4> 4 {2}
‘W’ <2,3> 3 {1}
‘W’ <1,2> 2 {1}
‘W’ <0,1> 1 {1}

PSF

1 [v 安装]
2 [p 在]
3 [n 桌子]
4 [f 上]

图 1 具体实例的分析状态图

分析处理的初始状态是 BMS 和 PSF 都为空，然后匹配算法从左向右逐词扫描句子，检查每个词语的界定预测情况，图 1 显示了刚刚检查完词语“上/f ”后的分析器状态。（注意 BMS 中的词语位置从 0 开始计算）

由于词语“上/f ”的界定预测输出为{2}，因此启动了匹配操作 MatchBrkt, 其间对 BMS 进行了归约，图 2 显示了匹配处理完成后的分析器状态。

BMS

‘P’ <2,4> 5 /
‘W’ <1,2> 2 {1}
‘W’ <0,1> 1 {1}

PSF

1 [v 安装]
2 [p 在]
3 [n 桌子]
4 [f 上]
5 [sp (2,4)]
6 [pp (2,5)]
7 [vp (1,6)]

图 2 具体实例的分析状态图（续一）

然后继续移进两个词语“的/u ”和“灯/n ”，形成图 3。

BMS

‘W’ <5,6> 9 {2}
‘W’ <4,5> 8 {1}
‘P’ <2,4> 5 /
‘W’ <1,2> 2 {1}
‘W’ <0,1> 1 {1}

PSF

1 [v 安装] 5 [sp (2,4)] 9 [n 灯]
2 [p 在] 6 [pp (2,5)]
3 [n 桌子] 7 [vp (1,6)]
4 [f 上] 8 [u 的]

图 3 具体实例的分析状态图（续二）

最后在词语“灯/n ”上启动匹配操作，形成了图 4的分析结果。

BMS

‘W’ <5,6> 9 {2}
‘W’ <4,5> 8 {1}
‘P’ <2,4> 5 /
‘W’ <1,2> 2 {1}
‘W’ <0,1> 1 {1}

PSF

1 [v 安装] 6 [pp (2,5)] 11 [tp (6,8,9)]
2 [p 在] 7 [vp (1,6)] 12 [np (6,8,9)]
3 [n 桌子] 8 [u 的] 13 [np (7,8,9)]
4 [f 上] 9 [n 灯] 14 [vp (1,12) (1,11) (1,15)]
5 [sp (2,4)] 10 [np (5,8,9)] 15 [pp (2,10)]

图 4 具体实例的分析状态图（续三）

六、算法的分析与讨论

句法分析算法的设计主要研究如何充分利用各种语言学知识，从输入句子中自动分析出正确的句法结构树，其间包括成分边界的识别和句法结构的组合等问题。在这一过程中，不同的语言知识描述形式对分析算法的形成和设计起着重要的作用。

传统的基于规则的句法分析技术主要有两种：移进--归约分析(shift-reduce parsing)技术和图分析(chart parsing)技术。以此为基础的 LR 分析器[MT86]和图分析器([TW83], [SFI89], [SS94])主要利用了产生式规则中的句法成分描述信息，通过生成 LR 分析表及指导图边的组合顺序，来对分析过程进行控制。尽管它们可以达到很高的分析效率，但规则描述的局限性限制了其鲁棒性和灵活性的提高。

近几年来，随着语料库语言学的不断发展和标注语料库规模的不断扩大，许多研究人员开始尝试直接利用语料库中的标注信息进行句法分析，通过对输入句子进行状态变换搜索和统计优化处理，选择得到最佳的分析树。如： R. Bod 提出的面向数据分析(Data Oriented

Parsing)方法([RB93],[RB92])中的子树替换和 Monte Carlo 搜索技术，文献[SD91]中采用的模拟退火(Simulated Annealing)分析方法， David M. Magerman 的概率型判定树(statistical decision-trees)分析模型([DM95], [DM94])和动态规划(Dynamic programming)搜索方法等。由于利用了语料库作为主要的知识源，大大提高了分析器的鲁棒性和灵活性，但其分析效率一般不够高。

作者的研究则试图吸收以上这些分析方法和技术各自的处理优势，其主要特点是： 1) 通过进行成分边界自动预测的预处理，正确地识别出了绝大部分句法成分的边界位置。这实际上提供了移进--归约及边组合的边界控制信息，从而把句法成分的识别问题转化成为括号匹配问题，即哪些左、右括号对能够互相匹配。而在这个过程中就可以利用传统分析方法中的有效控制机制和处理技术，达到最佳效果。 2) 通过采用多层次的句法成分定性方法，可以充分利用从树库语料中自动获取的各种知识，使此方法具有更好的适应性和开放性。

以此算法为基础，并辅之以匹配区间限制、概率评分模型和基于优先排歧等有效处理机制，作者已开发出了一个性能较好的汉语句法分析器[ZQd96]，它充分利用了传统分析技术中的高效控制机制和目前比较常用的语料库知识获取和有效的统计优化技术，具有很强的灵活性和鲁棒性，可以较好地适应对大规模真实语料文本进行自动句法分析的需要。

七、致谢

本文是作者的博士论文一部分研究工作的总结，得到了两位导师：姚天顺教授和俞士汶教授的悉心指导和北京大学计算语言学研究所许多老师和同学的热情帮助，这里一并表示感谢。本项研究得到国家自然科学基金资助，项目号为 69483003 。

参考文献

[DM94] David M. Magerman. (1994). *Natural Language Parsing as Statistical Pattern Recognition, Doctoral dissertation, Stanford University, Stanford, USA.*

[DM95] David M. Magerman. (1995). "Statistical Decision-Tree Models for Parsing", In *Proc. of ACL-95*, 276-303.

[MT86] M.Tomita. (1986). *Efficient Parsing for Natural Language --- A Fast Algorithm for Practical System.* Kluwer Academic Publishers.

[RB92] Rens Bod. (1992). "A Computational Model of Language Performance: Data Oriented Parsing", In *Proc. of COLING-92, Nantes.*

[RB93] Rens Bod. (1993). "Using an Annotated Language Corpus as a Virtual Stochastic Grammar", In *Proc. of AAAA-93, 778-783.*

[SD91] Clive Souter & Time F. O'Donoghue. (1991). "Probabilistic parsing in the COMMUNAL project", In Stig Johansson and Anna-Brita Stenstrom (eds.) *English Computer Corpora : Selected papers and Research Guide.* 33-48.

[SFI89] O. Stock, R. Falcone, & P. Insinnamo. (1989). "Bi-directional charts: a potential technique for parsing spoken natural language sentences", *Computer Speech and Language*, 3, 219-237.

[SS94] G. Satta, & O. Stock. (1994). "Bi-directional context-free grammar parsing for natural language processing", *Artificial Intelligence*, 69, 123-164.

[TW83] T.Winograd. (1983). *Language as a cognitive process. Vol1, Syntax*, 116-129.

[ZQ96] 周强. (1996). "一个汉语短语自动界定模型"，《软件学报》第 7 卷，增刊： 315-322

[ZQd96] 周强(1996). "汉语语料库的短语自动划分和标注研究"，博士学位论文，北京大学计算机系， 1996.6.