基于局部优先的汉语句法分析方法*

周 强 黄昌宁

(清华大学计算机科学与技术系 北京 100084)
(清华大学智能技术与系统国家重点实验室 北京 100084)
E-mail: zhouq@s1000e.cs.tsinghua.edu.cn

摘要 提出了一种利用局部优先信息对汉语分析算法进行优化的新方法.通过利用从语料库中自动获取的结构优先关系数据作为优先判断依据.此方法使目前的汉语概率分析器的整体效率提高了近30%,显示了很好的应用前景.

关键词 基于优先分析,句法分析,汉语概率分析器,语料库.

中图法分类号 TP18

在图分析(Chart Parsing)算法中,一种很常用的优化技术是Best-First技术.其主要思想是在分析调度器(Agenda)的控制下,每次尽可能选择最佳的成分边进行组合扩展,从而迅速得到句子的最佳分析树,极大地提高了分析效率.在这一过程中,如何选择合适的优先评价机制,对各个不同成分在正确的句法树中出现的可能性进行准确的评估,就成了一个关键问题.近几年来,Magerman & Weir[1]和Charaballo & Charniak[2]开始探索将概率信息引入Best-First技术中,取得了很好的效果.

Best-First技术主要着眼于从全局上把握不同句法成分在分析树中的优先关系.基于优先(Preference-based)技术[3]则更侧重于对局部语境下的歧义结构的排歧处理.一个典型的例子是对英语的介词短语连结(Prepositional Phrase Attachment)问题的自动排歧处理.Hindle和Rooth[4]提出可以利用从自动分析的语料库中获取的介词与中心动词以及中心名词的关联强度的不同来排除PP连结歧义.R. Basili等人[5]进一步利用了语义标记信息,提高了排歧能力.而Collins和Brooks[6]则采用了Backing-Off方法对此问题进行处理.

本文提出一种将以上两种技术有机结合起来的高效分析方法.它将局部优先信息作为Best-First的选择控制机制,通过合理地排除局部优先组合能力较小的句法成分,从而达到提高整体分析效率的目的.其基本分析框架是在目前的汉语概率分析器[7]中采用的匹配分析算法[8],而局部优先信息则利用了从语料库中自动获取的结构优先关系(SPR)数据.目前的实验结果表明,此方法的应用使分析器的整体效率提高了近30%,显示了很好的处理效果.本文第1节简要介绍了汉语匹配分析算法的基本内容,第2节分析了局部优化方法的基本思路,第3节介绍了具体的实现方法,第4节给出目前的一些实验结果,并对此进行了分析,第5节是结束语.

1 匹配分析算法简介

文献[7]中提出的汉语概率分析器对汉语句子的分析主要通过以下3个阶段来完成:① 成分边界预测;② 括号匹配;③ 统计排歧.其中括号匹配处理起着承上启下的作用,它主要用来解决这样一个分析问题:以特征向量$S=\langle WTB, MRR\rangle$作为分析器的输入,如何通过其中左右括号的合理匹配,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).

其中$WTB=\langle W,T,B\rangle$,$W=w_1,w_2,\ldots,w_n$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,

* 本文研究得到国家自然科学基金和中国博士后科学基金资助.作者周强,1967年生,博士,助理研究员,主要研究领域为语料库语言学,机器翻译,机器学习.黄昌宁,1937年生,教授,博士生导师,主要研究领域为计算语言学.

本文通讯联系人:周强,北京 100084,清华大学智能技术与系统国家重点实验室

本文1997-11-20收到原稿,1998-01-23收到修改稿

$B=b_1,b_2,\ldots,b_n$ 则是一串成分边界信息描述，b_i 可取值 0，1 或 2，分别表示词语 w_i 处于某个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，它们是进行括号匹配的基础，并且可以利用现有的成分边界自动预测工具[9]得到. 而 MRR 则是一组匹配限制区间描述，它们将对其间的匹配操作进行有效的限制.[10]

匹配分析算法的实现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. 从直观上看，它可以这样来进行：从左向右扫描句子，直至发现一个匹配右项*，从此成分出发，搜索所有左相邻的匹配左项**，通过匹配操作形成新的句法成分，然后以此新成分为驱动，进行类似的操作. 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.

而具体的算法则是在以下 3 个基本控制结构上实现的，它们是通过对 LR 分析器[11]和图分析器[12]的有效控制结构的合理吸收和适当改进而形成的.

(1) 括号匹配栈(BMS)：保存了进行句法分析所需的所有边界控制信息，功能相当于 Tomita 算法[11]中的图结构栈.

(2) 压缩共享森林(PSF)：保存了经括号匹配得到的所有句法成分信息，类似于 chart 结构.

(3) 待匹配成分表(PEL)：保存了所有待处理的匹配右项信息，可作为一个分析调度器(Agenda).

有关这一算法的详细内容可参阅文献[8].

2　基于局部优先的优化分析

考虑如下的输入信息片段：$[w_i[w_{i+1}\quad w_{i+2}]w_{i+3}]$(为简便起见，这里省略了词类信息描述)，利用现有的匹配分析算法，将同时得到图 1 所示的两棵分析子树(a)和(b)中的 5 个匹配成分 $A_i, i\in[1,5]$，而实际上，其中只能有一棵子树可以出现在正确的句法分析树中. 从这个角度看，目前的匹配算法中还存在许多冗余的匹配操作，具有很大的可以优化的余地.

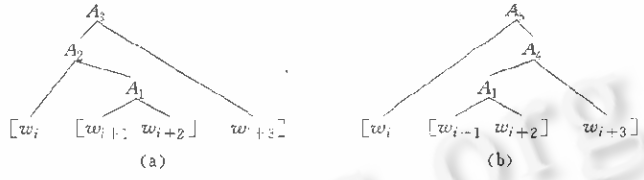

图1　一个输入片段的匹配分析结果

对图 1 深入地分析可以得出，两棵子树优先选择的关键是确定匹配成分 A_1 在局部语境 w_i 和 w_{i+3} 下的优先组合关系：如果左向组合优先，即优先匹配产生句法成分 A_2，则选择子树(a)；如果右向组合优先，即优先匹配产生句法成分 A_3，则选择子树(b)，两者必居其一. 据此，我们可以形成一个利用局部优先信息对现有匹配分析算法进行优化的基本思路：在局部语境信息约束下，只选择优先结构组合进行匹配操作，而排除对不优先结构的匹配操作，从而达到减少冗余匹配，提高分析效率的目的. 与传统的 Best-First 实现机制不同的是，此方法是通过对局部较差情况的排除来间接体现 Best-First 思想的，因为这可以与目前的概率分析器中采用的匹配分析与统计排歧同时进行的分析机制很好地融合在一起.

有多种方法可以用来判断局部语境下的优先组合关系. 目前比较常用的是词关联(Word Association)技术，如相关信息(Mutural Information)测度等. 具体做法是：抽取中间成分的中心词 w_{mh}，分别计算它与左语境中心词 w_{lh} 和右语境中心词 w_{rh} 的关联度 $MI(w_{lh},w_{mh})$ 和 $MI(w_{mh},w_{rh})$，然后选择其中的较大者作为优先组合结构. 前面提到的对英语的介词短语连结问题的处理就是采用了这种方法. 本文则利用了结构优先关系描述项(SPR)

---

* 包括：(1) 具有右边界预测标记的词语($b_i=2$)；(2) 匹配产生的新的句法成分(形如[...]).

** 包括：(1) 具有左边界预测标记的词语($b_i=1$)；(2) 匹配产生的新的句法成分(形如[...]).

信息.其基本形式为〈*IS*,*LP*,*RP*〉,其中 IS 为交段结构,表示为{交段前境 ^ 交段成分 交段后境},而 LP 和 RP 则是交段成分在局部语境 IS 下的左向组合概率和右向组合概率,例如,SPR 项{*p* ^ *np vp*, 0.97, 0.03}就记录了这样的信息:交段成分,即名词短语(np)在局部语境{*p np vp*}下与交段前境,即介词(p)组合的概率为 0.97,而与交段后境,即动词短语(vp)组合的概率则为 0.03.这些信息可以从真实语料文本中自动获取,其基本步骤为:利用概率分析器对真实语料文本进行自动分析,得到每个句子的完全分析树,其中各个成分都带有概率分布信息;然后遍历分析树,发现所有可能的交段结构,并计算分析树中交段左向和右向组合的预期频度.其具体内容将另文介绍.

据此,可以这样来对现有的匹配分析算法进行优化:对于一个待匹配成分 *A*,在句子中搜索其左匹配语境 LMC 和右匹配语境 RMC,形成一个局部语境片段{*LMC A RMC*},检索 SPR 表,如在其中发现一个 SPR 项的 IS 与此局部语境片段相同,则可根据它的 LP 和 RP 的差异程度来进行优先选择.下节将介绍具体的实现方法.

3　优化分析算法的实现

在匹配优化算法的具体实现过程中,首先应解决以下几个问题:(1) 如何确定一个待匹配成分的局部语境;(2) 如何保证所利用的 SPR 数据的可靠性;(3) 如何把不同的优化控制机制很好地结合入原来的匹配分析算法中.下面将分别进行详细的讨论.

3.1　局部语境的确定

对于一个待匹配成分 *A*,其局部语境是由其相邻的匹配左项和匹配右项组成的.在目前的自左向右的匹配处理机制中,其匹配左项的确定是很容易的,因为在得到匹配成分 *A* 之前,它左边的所有可能的匹配操作都已完成了,因此,只需搜索 BMS 和 PSF 就可以得到.困难的是匹配右项的获取,因为在当前的分析状态下,成分 *A* 右部的匹配操作还没有进行.

考虑这样一个分析片段:$[w_i[_A\ w_{i+1}\quad w_{i+2}][w_{i+3}\quad w_{i+4}]_]$,对于刚刚匹配生成的成分 *A*,其真正的匹配右项应为词语 w_{i+3} 和 w_{i+4} 匹配产生的成分 *B*,但目前还不能得到,这就影响了对成分 *A* 的优化选择判断.为解决这个问题,我们提出了一种延迟选择机制,具体方法是:增加一个延迟匹配成分表(DCL),每当遇到类似上面的不能确定所需的匹配右项的情况时,就将该成分放入 DCL 中,继续执行自左向右的匹配操作,每当产生一个新成分时,需检查 DCL 中的延迟成分所需的右语境条件是否已满足,若是,则将它从 DCL 中取出,重新插入待匹配成分表(PEL)中.这样,通过 PEL 和 DCL 的相互作用,保证了每个待匹配成分都能获取其局部语境来进行优化选择.

3.2　SPR 阈值的合理选择

目前的 SPR 数据的应用条件设置为:$|LP-RP|>\beta$,其中的阈值 β 反映了在局部语境下交段左向和右向组合概率的差异程度.在此条件下,如果 $LP>RP$,则只进行左向组合匹配;反之,则只进行右向组合匹配.

一般来说,β 越大,优化操作出错的可能性就越小,因为它只是排除了那些在局部语境下极少可能出现的匹配组合情况.但由于它对 SPR 项数的过强限制,使优化效率的提高显得不是很充分,极端情况是 $\beta=1$,此时优化机制将不起作用.反之,β 越小,则满足这一条件的 SPR 项数就越多,从而可以对句子中大量的局部语境进行优化,大大提高了分析效率,但同时可能会排除一些正确的匹配组合情况,从而使最终分析结果的准确度有所下降.因此,合理的做法是通过选择一个合适的阈值 β,在保证所利用的 SPR 数据的可靠性的前提下,尽可能扩大其应用范围,从而在优化效率和分析结果的准确性之间寻找到一个平衡点.

3.3　优化算法的基本控制流程

通过将上两节介绍的优化控制机制有效地结合入原有的匹配分析算法[8]中,我们形成了一个改进的优化匹配算法,下面给出其基本的控制流程.其中,步骤(2)和(3)实现了语境延迟选择机制,而步骤 6 则增加了局部优先条件的判断,据此可以排除大量局部优先性较小的匹配操作.

(1) 从 PEL 中获得一个待匹配成分 *A*;

(2) 在 BMS 中搜索其匹配右项 RMC;

(3) 如果找不到,则将成分 A 移入 DCL 中,转(8);
(4) 在 BMS 和 PSF 中搜索得到一个匹配左项 LMC;
(5) 检查局部语境(*LMC A RMC*)下的优先组合关系;
(6) 如果左向组合优先,则进行匹配操作:[*LMC A*],否则转(7);
(7) 如果还有其他的匹配左项,则转(4);否则转(8);
(8) 如果 PEL 不空,则转(1);否则算法结束.

4 实验结果分析

我们采用了文献[10]中开发的汉语树库(Treebank)作为实验语料,它由以下两部分组成:(1) 汉英机器翻译研究的测试题库(语料 *A*).语料的规模为 1 434 个汉语句子,约 11 821 个词,汉字总数为 17 058,平均句长为 8.243 词/句;(2) 新加坡小学语文课本语料*(语料 *B*),总规模为 4 139 个句子,约 52 609 个词,汉字总数为 72 434个,平均句长为 12.711 词/句.

通过对实验语料的均匀抽样,形成了 11 个测试样本,平均每个样本包含约 507 个汉语句子.然后取其中的前 10 个组成训练语料,共包含 5 071 个句子,用于训练得到进行局部优先分析所需的 SPR 数据和其他统计数据.最后的第 11 个样本作为测试语料,共包含 506 个句子,用于检测优化算法的分析效果.

实验的主要目的是检查局部优先信息的运用对汉语概率分析器的分析效率和分析结果准确度的影响.其中,对分析结果准确度的评估主要依据了以下几个性能指标:① 括号召回率(MR),② 括号正确率(MP),③ 交叉括号数(CBs),④ 标记正确率(LP).有关它们的详细定义可参阅文献[7,10].而分析效率则是通过以下几个数据体现出来的:

(1) 匹配成分总数(MCSum):经匹配操作而加入 PSF 中的所有匹配成分的数目,它是与所进行的匹配操作的数目一一对应的.

(2) 分析树总数(PTSum),PSF 中所有的完整分析树**的数目,它从一个侧面反映了对句子分析结果进行排歧处理的难易程度,其具体计算方法可参阅文献[10].

(3) CPU 时间:为分析输入句子而花费的所有 CPU 时间(包括进行统计排歧所需的时间).

首先进行了 SPR 阈值选择实验.通过设置 0.1~0.9 之间的 9 个不同阈值,并记录不同阈值作用下的优化分析数据,我们得到了如表 1 所示的实验结果.其中同时列出了原来的未经优化的匹配分析算法的相应数据($\beta=1$ 的行),以作为对优化效果的判定依据.

表 1 不同 SPR 阈值作用下的优化分析结果

β	MR(%)	MP(%)	CBs	LP(%)	MCSum	PTSum	CPU 时间(s)
1	89.93	90.01	0.87	95.17	1.19×10^4	6.53×10^6	228
0.9	89.87	90.00	0.87	95.21	8.55×10^3	1.91×10^5	184
0.8	90.10	90.24	0.84	95.22	8.39×10^3	9.44×10^4	166
0.7	90.08	90.21	0.85	95.27	8.25×10^3	6.92×10^4	179
0.6	90.31	90.44	0.82	95.31	8.13×10^3	4.46×10^4	173
0.5	90.40	90.53	0.82	95.31	8.08×10^3	4.02×10^4	170
0.4	90.57	90.70	0.80	95.30	8.03×10^3	3.57×10^4	158
0.3	90.50	90.64	0.81	95.32	7.94×10^3	3.11×10^4	166
0.2	90.46	90.58	0.81	95.29	7.93×10^3	3.10×10^4	176
0.1	90.48	90.60	0.81	95.24	7.82×10^3	2.03×10^4	17

带阴影行($\beta=0.4$)为最佳阈值

从表 1 中可以看出,随着阈值 β 的不断降低(从 0.9 到 0.1),优化算法的分析效率的变化趋势基本上是与我

* 此语料的电子版本由国立新加坡大学顾金定博士提供,在此表示感谢.

** 一棵完整的分析树是指覆盖输入句子的所有词语的分析树.

们的预期估计相一致的:即随着对 SPR 项的限制条件的不断放宽,可以对越来越多的局部语境下的匹配操作进行优化,从而使分析器产生的匹配成分总数和分析器总数不断下降.尽管由于统计排歧过程的影响,使得 CPU 时间的变化趋势显得不是很明显.但从总体上看,分析效率还是在不断提高的.

值得注意的是分析结果准确度的变化趋势.它经历了一个逐步提高,直至达到最高点($\beta=0.4$),然后又逐步下降的变化过程.这是因为当 β 值较大时,SPR 项所反映的局部优先知识比较可靠,因此利用它们来对匹配算法进行优化,不但不会对整体的排歧机制产生危害,反而是有利的,因为它们可以尽早排除在局部语境下极少可能出现的句法成分组合,并且这些被排除成分的数目将随着 β 的减少而不断增加,从而使统计排歧算法可以从可能性较高的成分组合中更快、更好地选择出一棵概率意义上最佳的分析树.但当阈值 β 下降到一定程度时,SPR 信息的可靠性将大幅度降低,此时再利用它们来进行匹配优化,就可能把大量正确的成分组合在局部语境检查时就排除掉了,从而使分析结果的准确度逐步下降.

综合以上分析,我们认为,针对目前所用的局部优先数据 SPR,选择 $\beta=0.4$ 作为阈值是比较合适的.在此条件下对原有的匹配分析算法进行优化,得到了这样的结果:汉语概率分析器的 CPU 时间、匹配成分总数和分析树总数分别下降了 30.7%,32.5%和 99.5%,并且最终分析结果的准确度也有所提高,显示出很好的优化效果.

图 2 和图 3 进一步显示了在优化前和 $\beta=0.4$ 的优化条件下,测试语料中具有不同长度的句子集上的平均匹配成分总数和平均分析树总数的分布特点.从中可以看出,对于简单的句子(其中的词项(包括句子中的词语和标点符号数)<20),局部优化效果并不是很明显;而对于复杂的句子(其中的词项数≥20),新算法则显示了很强的优化能力.这主要是因为,在复杂句子中,各种可进行优化的局部语境出现的频度很高,同时,早期对局部语境下的某个句法成分组合的排除,往往会导致对整个分析森林中包含这个成分的成千上万棵分析子树的排除.这可以从测试语料中最长的句子(词项数=61)在优化前后的分析性能的变化中清楚地看出来,见表 2.

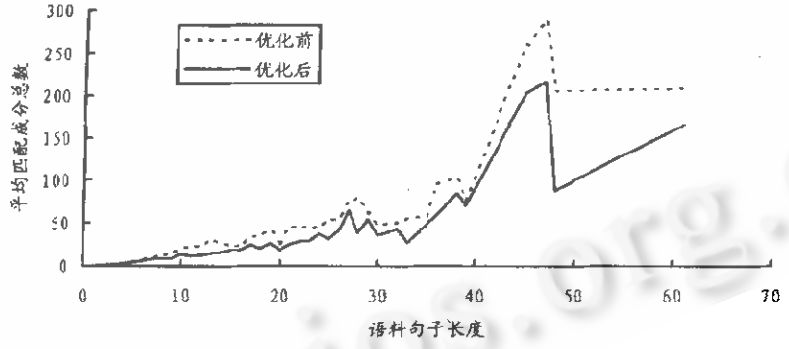

图 2　语料句子长度与平均匹配成分总数关系

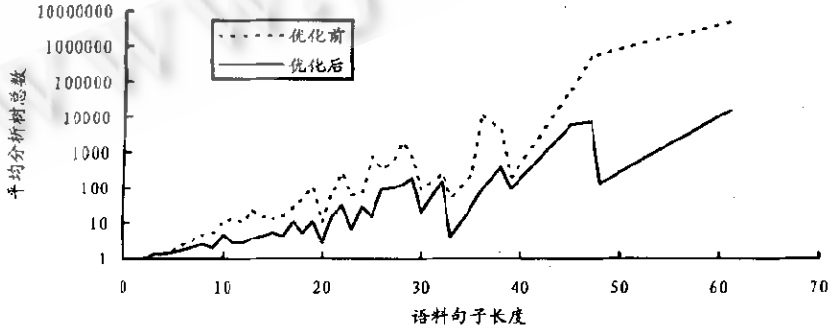

图 3　语料句子长度与平均分析树总数关系

表2 一个复杂句子(词项数=61)的优化分析结果

β	MR(%)	MP(%)	CBs	LP(%)	MCSum	PTSum	CPU时间(s)
优化前	80.00	76.60	9	94.44	2.09×10^2	5.22×10^5	4
优化后	82.22	78.72	8	94.59	1.66×10^2	1.61×10^4	4

5 结束语

本文提出了一种基于局部优先信息的优化句法分析方法.初步的实验结果显示,即使只利用基于词类标记和句法标记描述的比较简单的结构优先关系数据进行优化处理,在适当的SPR阈值作用下,也可以使分析器的整体效率提高约30%,并且仍保持很高的分析精度.其主要缺陷是,对于交段左向和右向组合概率相差较小(即小于SPR阈值)的局部组合,优化机制将不起作用.为此我们设想,在今后的研究中,可以进一步利用词汇优先信息补充SPR数据描述的不足之处,从而更好地发挥不同层次的局部优先信息的综合优化效能.作为基础性的分析方法的探索研究,本文的成果将为汉外机器翻译、汉语信息检索和信息抽取等应用领域的研究提供有力的支持.

参考文献

1 Magerman D M, Weir C. Efficiency, robustness and accuracy in Picky chart parsing. In: Church K ed. Proceedings of the 30th Conference of ACL (Association of Computational Linguistics). Newark, Delaware, 1992. 40～47

2 Carballo S A, Charniak E. New figures of merit for best-first probabilistic chart parsing. Technical Report, Brown University, Nov. 26, 1996

3 Church K W, Mercer R L. Introduction to the special issues on computational linguistics using large corpora. Computational Linguistics, 1993,19(1):1～24

4 Hindle D, Rooth M. Structural ambiguity and lexical relations. Computational Linguistics, 1993,19(1):103～120

5 Basili R, Pazienza M T, Velardi P. Semi-automatic extraction of linguistic information for syntactic disambiguation. Applied Artificial Intelligence, 1993,(7):339～364

6 Collins M, Brooks J. Prepositional phrase attachment through a backed-off model. In: Yarowsky D, Church K eds. Proceedings of the 3rd Workshop on Very Large Corpora. Cambridge, Massachusetts: Massachusetts Institute of Technology, 1995. 27～38

7 Zhou Qiang. A statistics-based Chinese parser. In: Zhou Joe, Church K eds. Proceedings of the 5th Workshop on Very Large Corpora. Beijing: Tsinghua University Press, 1997. 4～15

8 周强.汉语匹配算法的实现.见:陈力为,袁琦编.语言工程.北京:清华大学出版社,1997.194～200
(Zhou Qiang. Implementation of Chinese parsing algorithm based on bracket matching principle. In: Chen Li-wei, Yuan Qi eds. Language Engineer. Beijing: Tsinghua University Press, 1997. 194～200)

9 周强.一个汉语短语自动界定模型.软件学报,1996,7(增刊):315～322
(Zhou Qiang. A model for automatic prediction of Chinese phrase boundary location. Journal of Software, 1996,7(supplement):315～322)

10 周强.汉语语料库的短语自动划分和标准研究[博士学位论文].北京大学,1996
(Zhou Qiang. Phrase bracketing and annotating on Chinese language corpus [Ph. D. Dissertation]. Beijing University, 1996)

11 Tomita M. Efficient Parsing for Natural Language — a Fast Algorithm for Practical System. Boston: Kluwer Academic Publishers, 1986

12 Winograd T. Language as a Cognitive Process. Vol. 1. Syntax. Reading, MA: Addison-Wesley, 1983. 116～129

An Improved Approach for Chinese Parsing Based on Local Preference Information

ZHOU Qiang HUANG Chang-ning

(*Department of Computer Science and Technology Tsinghua University Beijing* 100084)
(*State Key Laboratory of Intelligent Technology and Systems Tsinghua University Beijing* 100084)

Abstract In this paper, a new technique based on local preference information is proposed to improve the efficiency of Chinese parsing algorithm. By using the statistics of structure preference relations as the figures of merit, the overall efficiency of the current Chinese probabilistic parser has been improved 30% by this method, which shows good application prospects.

Key words Preference-based parsing, parsing, Chinese probabilistic parser, corpus.