分析汉语句子的解释学习方法

李晓黎　郭炳炎
（山西大学计算机系，太原 030006）

摘　要：我们研制了一个基于解释学习的汉语句子分析系统，试图把解释学习这一机器学习技术应用于汉语处理，其目的是为了解决汉语句子分析中的效用性问题。通过对 250 个例句的运行，得出了具有 120 条学习规则的规则集。这样，对大部分句子可绕过常规的分析器，而使用学习规则集进行处理，速度提高 1/3，从而大大提高了分析效率。

THE EBL APPROACH OF PARSING CHINESE SENTENCES

Li Xiaoli　　Guo Binyan
(Dept. of Computer Science ,ShanXi University 030006)

ABSTRACT: This paper contributes to establish a Chinese sentences parsing system based on explanation—based learning. The system applies EBL to Chinese processing inorder to solve utility problem in Chinese analysis. A liberary of 120 learned rules are acquired through running 250 training examples in experiment. For most sentences, the system uses rule method rather than normal syntatic and sementic processing to parse, this can win a great deal of efficiency.

1　引　　言

自然语言理解是知识信息处理的核心课题，也是智能计算机系统的主要研究领域之一。基于解释学习（简称 EBL）是机器学习领域中非常活跃的一个分支[1-3]，我们试图将其应用于汉语句子分析，以解决分析中的效用性问题，这是一个崭新而又困难的问题。

众所周知，汉语是丰富的，但句型却是有限的，对汉语处理而言，大量时间是以相同方式处理同一类型的句子。如果我们能够加速处理这种"典型句型"的有限集，那么，将可节省大量的计算时间。基于这一想法，我们试图由用户给出的训练例句，通过 EBL 自动抽取学习规则，构成规则集，对大部分输入句子可绕过常规分析器使用学习规则集进行处理。这样，可大大加快处理过程，提高分析器效率。尤其是对大规模自然语言系统更能显出其优越性。

EBL 应用于汉语句子分析可非形式化地描述为[4]：

(1) 目标概念 (Goal Concept)：对汉语句子的一般化描述。假设这种描述不满足可操作准则，不能直接用于汉语句子的识别。

(2)训练例子(Training Example)：一个句子经分词之后得出的符号串。

山西省自然科学基金资助项目

(3)领域知识(domain Theory):即语法规则和词典,这是学习的依据,用来解释训练例子为何是目标概念的一个例。

(4)可操作性准则(Operationality Criterion):规定输出句子的表达形式,限定其所使用的谓词及词汇。

我们的基本思路是:首先选取一定数量的例句进行预处理(分词、兼类词处理),然后对处理结果进行解释概括,通过学习规则的抽取算法来获取学习规则,接着建立学习规则的索引以加快对学习规则的查找,最后用模式匹配的方法对待分析句子进行处理。

2 分析器的建立

为适应 EBL 在汉语句子分析中的应用,我们以 DCG 文法[5]为基础,用逻辑程序设计语言 PROLOG 建造了一个汉语句子分析器,该分析器主要由词典、文法库和分析机制构成。

分析器中的词典被描述为一组外部数据库中的事实。并以 B+树作为外部数据库的索引,所以词典的查询非常快。查询是由 PROLOG 的合一功能自动完成的。

词典中的词条主要包含有词法、句法、语义及有关个性规则等信息。它可以表示为:

dic(entry,lexcat,syn,cy,sem,rule),

其中,entry 表示词条名;lexcat 表示词类;syn 表示该词所要求的主、宾语性质;cy 表示该词的语义分类(如有生命、类别、抽象等);sem 对动词而言表示施事格,受事格,时间格,处所等格的必须、可选及不允许,对名词而言,给出该词的候选格,对介词而言,给出介词短语的可能格关系;rule 是仅适用于该词的用于处理一些常用情况的个性规则。

文法库中的规则可描述成 Horn 子句的形式,每个非终结符都含有两个额外参数,分别表示到目前为止正在处理的句子串和处理后剩余的句子串,这一点是从分析效率上考虑,以减少回溯,加快处理速度。下面我们给出一条规则。

```
predicat(predicat(adv(LAD),v—head(LH),comp(LCM),obj(LOB)),S0,S):—
adve(LAD,S0,S1),!,headv(LH,S1,S2,OBJ1),!,comp(LCM,S2,S3),!,
n—adjunct(OBJ2,S3,S),!,append(OBJ1,OBJ2,LOB).
```

该系统采取由顶向下与自底向上相结合的策略[6],其分析过程是:第一步,系统自底向上数据驱动,汉语中一些词或短语的结合能力较强,它们都可根据规则一次匹配成功,这样可以避免在分析中作出许多与句子毫不相干的假设。第二步,由顶向下期望驱动,不断选择 DCG 定义的规则,匹配输入语句,一些词可以提供预示信息,可利用它们制导对周围成份的分析,直到 DCG 规则全部无一遗漏地覆盖住输入的语句,合一为最终目标为止。为了减少大量的回溯过程,采取"早决策"的方法,把判断放在头部,若匹配则继续执行该语句,否则寻找另一子句。

3 学习规则的获取

定义 1:谓词集 P 为系统所能用于描述的谓词集合。集合 $AP=\{dic(WORD_1,LEX_1,SYN_1,SEM_1,\ldots),\ldots,dic(WORD_n,LEX_n,SYN_n,SEM_n,\ldots)\}$为可操作集。

领域知识 T 形如:

M←N 的一些公式集,其中 $M\in P$,

$N=C(q_1,\ldots,q_n)$,$n\geq 1$,$n\in I$,$q_i\in P$,$i=1,\ldots,n$,C 为定义在∧、∨上的 逻辑函数。

S 为训练例子集。

目标概念 GC 为 Analyse(sentence(SUBJECT,PREDICATE),e,[])

其中 e∈S,sentence(SUBJECT,PREDICATE)为未来的分析结果。

定义 2:谓词 A 的关联集 link(A)为:

(1)Φ,若 A∈AP。

(2)$\{B_1,\ldots,B_m\}$,$m\geqslant 1$,$m\in I$。

其中,$B_i=b_1\wedge,\ldots,\wedge b_k$,$k\in I$,$b_j\in P$,$j=1,\ldots,k$。

若 T 中存在 $A\leftarrow B_1$,$A\leftarrow B_2,\ldots,A\leftarrow B_m$。

定义 3:令 TREE 为定义在 T 上的一棵树。TREE 扩展后的新树为 E(TREE)。

(1)E(TREE)结束。对于 TREE 的全部叶结点 $leave_i$,满足 $link(leave_i)=\Phi$。

(2)对任一叶结点,

若 link(leave)≠Φ,取 L∈link(leave),若 L∈P,则令 L 直接做为 leave 的子节点,若 L 为形如 $b_1\wedge b_2\wedge\ldots\wedge b_n$ 的形式,则令 $b_1,b_2,\ldots,b_n$ 均为 leave 的子节点,这样就形成一棵新树 E(TREE)。

若 link(leave)=Φ,则进行变量替换入栈保存,同时将当前分析结果存入表中。

定义 4:任一结点 A 为真的条件:

若 $A=b_1\wedge b_2\wedge\ldots\wedge b_n$,则当 $b_i\in AP$ or $b_i\in ture$,$i=1,2,\ldots,n$ 时,A 为真。

若 $A=b_1\vee b_2\vee\ldots\vee b_n$,则任一 $b_i\in AP$ or $b_i\in ture$,$i=1,2,\ldots,n$ 时,A 为真。

定义 5:任一结点 A 为假的条件:

若 $A=b_1\wedge b_2\wedge\ldots\wedge b_n$,则只要任一 $b_i\notin AP$ or $b_i\neq true$,$i=1,2,\ldots,n$ 时,A 为假。

若 $A=b_1\vee b_2\vee\ldots\vee b_n$,则 $b_i\notin AP$ or $b_i\neq true$,$i=1,2,\ldots,n$ 时,A 为假。

定义 6:任一结点 A 不可判断的条件:A 的子结点不可判断真假。

在上述定义的基础上,可给出算法如下:

BEGIN

循环 while S 中仍有未处理的句子时,

取 x∈S,

Tree←GC,

循环 { while GC 不可判断时,

Tree←E(TREE),

enddo }

若根结点为真时,说明该例在该领域知识内为可解释的,

取 GC 的第一个参数为规则的结论 GOAL,

堆栈 STACK 中的元素 Li 的合取为学习规则的条件,

构成学习规则 $R_i=GOAL:-\&L_i$,

将 Ri 加入规则库中。

否则,跟踪解释链修改领域知识,重新对该例加以解释。

enddo

END

例如,对输入语句"他昨天在教室里借了我一本管理体制的书"得到的学习规则为:

sentence(Subject(n—head(pron(P10)),adj([])),predicate(adve(adv(P11),

preph ([prep(P12),pn(P13),loc(P14)])), v—head(verb(P15,交接),adx(P16)),

comp([]), doubobj([nounph(pron(P17))], [nounph([adju(num(P18),cl(P19),
verb(P20),noun(P21),adx(P22))] , n－head(P23))])), 施事(P10),客体(P21),
时间(P11),处所(preph(prep(P12),pn(P13),loc(P14))))
:－dic(P10,"pron"...,"人"),
dic(P11,"tn"...),
dic(P12,"prep"...),
dic(P13,"pn"...),
dic(P14,"loc"...),
dic(P15,"verb",["人","物"],"交接",...),
dic(P16,"adx"...),
dic(P17,"pron"...,"人"),
dic(P18,"num"...),
dic(P19,"cl"...),
dic(P20,"verb",...),
dic(P21,"noun",...,"抽象",...),
dic(P22,"adx"...),
dic(P23,"noun",...,"物体",...).

由此可以看出:EBL 通过对训练例子的解释和概括,将句子的不可操作的描述转化为可操作的描述,其学习过程并不产生新知识,只是对现有知识进行有选择的重组,以提高系统的分析效率。规则的充分条件中除了词类信息外,还包括词的语法信息,如及物动词加表抽象的名词形成的定中结构,动词"借"具有"交接"类动词的性质,同时也包括词的语义信息,代词"他"表示人,名词"书"表示物等。最终学到的规则是可以用于识别同类型句子的。

4 规则判定树的建立

在学习规则库形成后,如何快速查找学习规则就成为极待解决的问题。为此,我们建立了规则判定树以作为对学习规则的索引[7],[8]具体算法如下:

(1) Tree←root.

Table←nil.

(2) 若输入规则已穷尽,则本算法结束。

(3) 令输入规则为 RULEi,

若 Table≠nil,则查 Table

3.1 求 maximatch(RULEi,Subrule,Node),其中 Subrule 为 RULEi 与 Table 中全部规则求最大匹配后剩余的部分。Node 为 Table 中最大匹配规则在最大匹配点对应的结点。

3.2 从 Node 开始将 Subrule 加入 Tree 中。

3.3 将 RULEi 的前件与其对应的结点加入 Table 中

否则,

从 root 开始将规则 RULEi 的前件全部加入到 Tree 中。将 RULEi 的前件与其对应的结点加入 Table。

(4) 转(2)。

说明:Table 为存放到当前所有规则的前件及其对应结点的表。Tree 为当前所生成的

树。

5 系统的结构

系统的总体结构可分为两部分，即学习规则库的生成部分与应用部分，分别如图 1、图 2 所示。

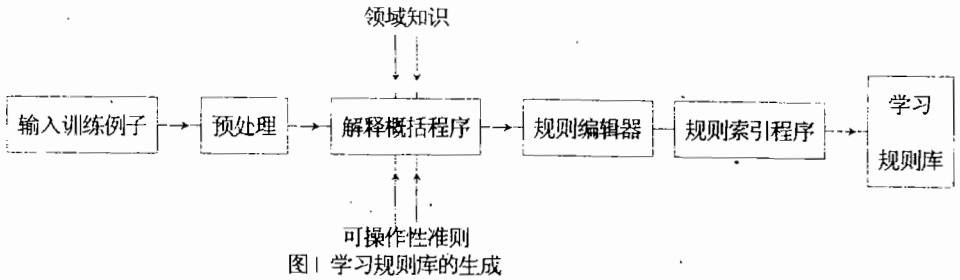

图 1 学习规则库的生成

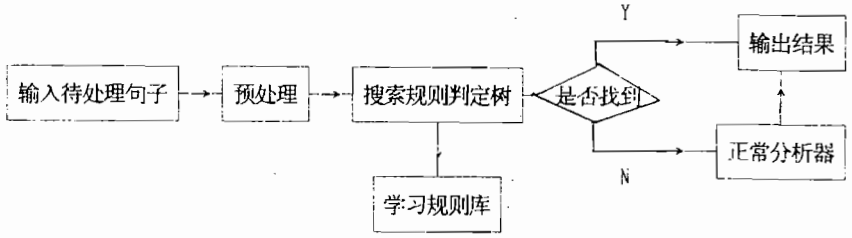

图 2 学习规则库的应用

在图 1 中，对于要分析的训练例子，首先进行预处理(分词、兼类词处理)，接着由解释概括程序得出一条学习规则，然后由规则编辑器对新获得的学习规则进行处理。处理过程是：先检查规则库中是否有规则与新规则相同或包含新规则，在这两种情况下新规则都不被加入规则库；如果新规则包含旧规则，则把新规则加入，把旧规则删除。所谓规则 A 与规则 B 相同是指两个句子产生的规则完全相同。所谓规则 A 包含 B 主要是指：A 与 B 的基本成份相同，但 A 比 B 仅多一些修饰成份。这是因为在汉语的句子中，主、谓、宾的位置比较固定，而且是句子的主要成份。其它成份如定、状、补语为句子的修饰成份且位置比较灵活。所以，如果在规则库中的规则仅比新规则多一些修饰成份，则我们仅把规则库中的修饰语注明可缺省，新规则就不加入到规则库中。例如规则库中的某规则仅比新规则多状语，我们仅注明新规则中状语可缺省，不把规则加入规则库中。最终经规则索引程序将其加入学习规则库。

在图 2，对待处理的句子进行预处理后，搜索规则判定树，若找到可应用规则，就执行并输出结果。否则就用正常分析器处理。

6 性能研究

我们从中学地理课本中随机地选取了 250 个例句，其中 200 个句子用来产生学习规则，50 个句了用来测试，结果共产生了 120 条规则。该实验是在 386 微机上实现的。

实验 1 顺序匹配与规则判定树的比较

为了在规则库增大的情况下，进行顺序匹配与规则判定树匹配的比较，首先将系统设置

为学习规则的生成模式，分别产生 20、40、...、120 条学习规则；然后将系统设置为应用模式。一方面对输入句子进行顺序匹配，另一方面利用规则判定树对输入句子进行处理。实验结果如图 3 所示。

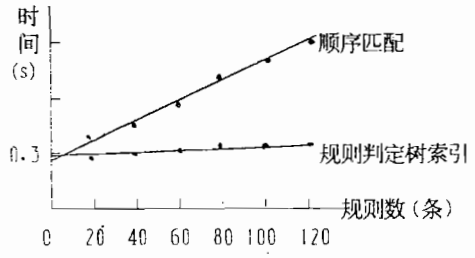

图3 顺序匹配与规则判定树的比较

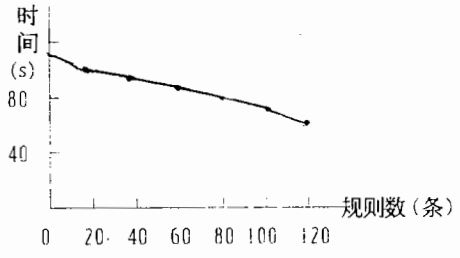

图4 应用EBL分析汉语句子的性能

实验表明：顺序匹配随规则数增加，查找时间线性增加；而规则判定树方法仅略有增加。

实验 2 应用 EBL 分析汉语句子的性能

在规则数为 0—120 的范围内，分别以 7 次处理 50 个句子。其结果如图 4 所示。

在没有学习规则时，处理 50 个句子的时间为 106 秒，有 20 条规则时处理时间为 92.5 秒，...，有 120 条规则时处理时间为 72.3 秒，所以处理时间越来越快。结果表明，随着学习规则库的不断扩大，EBL 处理句子的速度会越来越快，EBL 分析句子的性能会大大提高。

7 结束语

本文对 EBL 用于汉语句子分析进行了有益的尝试。实践表明：应用 EBL 这一机器学习技术的确可以大大提高汉语句子的分析效率。同时也体会到，如果没有一个好的领域知识，这一工作也是很难进行的。因此，我们进一步的工作将是应用 EBL 与归纳学习相结合的机器学习方法自动修正领域知识，这对汉语信息处理无疑是有益处的。尽管我们的课题仍在研究之中，但相信机器学习技术用于汉语处理将会有光明的前景。

参考文献

[1] Mitchell, Explanation—based Generalization: A Unifing View, Machine Learning, 1(1), pp. 47—80, 1986

[2] Dejong & Mooney, Explanation—Based Generalization: A Altemative View, Machine Learning, 1(2), pp, 145—146, 1986

[3] Minton, Explanation—Based Learning: A Problem—Solving Prespective, Artifical Intelligence, (40), 11—62, 1989

[4] 石纯一等，基于解释的机器学习方法，AI 基础理论研讨会，1992 年 1 月于哈尔滨

[5] Pereira, F. N. C and Shieber. S. Prolog and Natural Language Understanding, CSLI Lecture Notes, University of Chicago Press, 1985

[6] 刘开瑛，郭炳炎，自然语言处理，北京，科学出版社，1991

[7] Christer Samuelesson & Manny Rayner, Quantitative Evaluation of Explanation—Based Learning as an Optimization Tool for a Large—scale Natural Language System, IJCAI, 609— 615, 1991

[8] 郭炳炎，李晓黎，基于解释学习的汉语句子分析研究，智能计算机基础研究 '94，清华大学出版社，1994, 4