基于解释学习的汉语句法分析初探

郭炳炎　李晓黎
山西大学计算机科学系

1、引言

我们知道，自然语言是丰富的，但是句型却是有限的。句型是句子的结构类型，是从形式上对具有相同组合关系的句子的概括和抽象。句子是处于动态的话语的运用单位，而句型则是处于静态的能生成句子的语言模式。在多数情况下，自然语言的用户使用一定数量的句型。Manny Rayner 认为 100 个句型占到某一自然语言用户所用句型的的 80% 是合理的。如果有一种简单的方法能自动识别这些问题类型，那么除了少数特殊的句子外，绕过正常的分析机制去赢得高效率将是可能的。而对这些特殊的句子来说，我们仍可用常规的句法分析来处理。基于这一思路，我们认为：利用基于解释方法（Explanation－based Learning，简称 EBL）对汉语句子进行分析，以提高分析效率，应当是很有希望的。

2、EBL 在汉语处理中的非形式化描述

EBL 是近年来很受人们关注的一种机器学习技术，它是一种从单例出发利用领域知识进行分析学习的方法。它对汉语句子分析的解释学习问题可描述为：(1)目标概念：对汉语句子的一般化描述，假设这种描述不满足可操作性准则，不能直接用于汉语句子的识别。(2)训练例子：一个句子经分词之后得出的符号串。(3) 领域知识：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：规定输出句子的表达形式，限定其所使用的谓词及词汇。

给定上述信息之后，EBL 可分两步进行学习。第一步，利用领域知识构造一种解释，说明为什么训练例子满足目标概念，这个解释从概念上讲以树的形式构造，而树的叶结点满足可操作性准则。第二步对此解释做概括，以获得一个一般的充分条件以便用于同类句子的处理。

3、学习规则的提取：

领域知识主要包括文法库和词典，我们将文法库中的语法规则描述成 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，减少了回溯，加快了处理速度。词典中主要包含有词法、句法、语义及有关搭配信息，它可以表示为：　DIC(entry，lexcat，syn，cy，sem(Y，N，Y，O)，dapy，rule)，entry 表示词条，lexcat 表示词性，句法 syn 中表示了该词所要求的主、宾语性质；cy 表示了该词的语义分类(如有生命等)；语义 sem 中分别表示施事格、受事格、时间格、处所格的必须、可选及不允许(这主要是对动词而言)；rule 是仅适用于该词的用于处理一些常用情况的个性规则。现以"他昨天在教室里借了我一本词典" 为例来说明。该句分析的结果为：

sentence(subject(n－head([pron("他"),person("他")]),adjunct([])), predicat(adverbial([adv("昨天"),preph([prep("在"),pn("教室"),loc("里")])]),v－head([verb("借"),adx("了")]),comp([]),doubobj([nounph(["我"]),person("我")],[nounph(derph([num("一"),cl("本")]),nouph([n("词典")]))])),施事("他")，客体("词典")，时间("昨天")，处所("preph([prep("在"),pn("教室"),loc("里")])) :－ pron("他"),person("他"),tn("昨天"),prep("在"),pn("教室"),loc("里"),verb("借"),adx("了"),pron("我"),person("我"),num("一"),cl("本"),n("词典") 除谓词 pron(X)外，person(X)等也为满足操作性准则的谓词。

在句子分析的基础上，我们再经过概括，便可得出学习规则。

即：sentence(subject(n－head(pron(A)),adjunct([])), predicat(adverbial([adv(B),preph([prep(C),pn(D),loc(E)])]),v－head([verb(F),adx(G)]),comp([]),doubobj([nounph([H]),person(H)],[nounph(derph([num(I),cl(J)]),nouph([n(K)]))])),施事(A),客体(K),时间(B),处所("preph([prep(C),pn(D),loc(E)])) :－ pron(A),person(A),tn(B),prep(C),pn(D),loc(E),verb(F),adx(G),pron(H),person(H),num(I),cl(J),n(K)。

4、EBL 模型的设计：

EBL 模型可由两部分组成，即学习规则的形成部分和运行部分。形成部分主要是从单例中获取学习规则以及将学习规则进行组织，而运行部分则利用学习规则对句子进行快速处理。