一种可训练的快速汉语部分句法分析方法*

张益民， 陈玉泉， 陆汝占

(上海交通大学 计算机科学与工程系，上海 200030)

E-mail: yimin.zhang@intel.com; lu-rz@cs.sjtu.edu.cn

http://www.sjtu.edu.cn

摘要：提出一种可训练的快速汉语部分句法分析方法.此方法基于文中提出的扩充多层有限状态自动机模型，其中引入错误驱动的机器学习方法使其正确率可以通过训练得到提高，初步结果表明，该方法是行之有效的.

关键词：快速汉语部分句法分析；扩充多层有限状态自动机；错误修正规则；歧义消解；机器学习

中图法分类号：TP391 文献标识码：A

句法分析是不少自然语言处理系统的重要组成部分，近年来，为了处理大规模真实文本，在效率、鲁棒性方面对句法分析提出了更高的要求，传统的句法分析器是难以满足这些要求的.为此，不少系统已从完全的句法分析转到部分句法分析(partial parsing)上来，这已经成为新一代句法分析技术的一个发展趋势.

本文提出一种可训练的汉语部分句法分析方法.该方法基于我们提出的扩充多层有限状态自动机(augmented cascaded finite state automaton)模型来实现汉语语法及其分析器，在每层中设有相应的修正规则来对某些常见的错误进行修正，并且可利用机器学习方法来学习这些修正规则.

1 扩充多层有限状态自动机模型及句法分析技术

1.1 多层有限状态自动机

多层有限状态自动机(cascaded finite state automaton)是一种适合于快速部分句法分析的技术，已用于英语等多种语言的句法分析之中.S. Abney 于 1990 年首次使用该方法实现了 CASS(cascaded analysis of syntactic structure)系统，未加入任何其他方法就可轻松地达到 85%以上的准确率[1]，而识别主要的短语如主语、谓语的正确率更高达 95%以上[2].从速度上来说也是其他句法分析方法所无法比拟的.据文献[1]中数据显示，在相似的硬件条件下，传统的图算法分析器(chart parsing)的速度为 1w/s(词/秒)，Fidditch 系统中的 Marcus 确定性算法分析器其速度为 1200w/s，而 CASS 则可达到 1300～2300w/s.文献[3]中提出了一种基于双向图算法的快速汉语语法部分分析系统，虽然比传统图算法分析器速度有所提高，但也仅为 200w/min.(词/分钟)(PC486/66 上运行).将多层有限状态自动机应用于汉语句法分析仍存在一定的困难，这主要是因为该方法最初提出时主要是针对英语的，未加入任何歧义处理机制也能达到足够的正确率.但在汉语句法分析中，由于汉语本身的特点：句法灵活；短语、句子的构造规则相似，短语与句子之间没有明显的界限，造成了歧义现象无处不在，因而必须有一种有效的歧义处理机制才能达到足够的正确率.

1.2 扩充多层有限状态自动机模型

为了有足够的能力处理汉语句法中的种种歧义现象，我们在原有的多层有限状态自动机模型基础之上进一步提出了扩充多层有限状态自动机模型(augmented cascaded finite state automaton，简称 ACFSA).它也是由多

* 收稿日期：1999-05-17；修改日期：1999-09-07

基金项目：国家自然科学基金资助项目(69573020)

作者简介：张益民(1972－)，男，湖南长沙人，博士，主要研究领域为语语义理论，自然语言理解；陈玉泉(1968－)，男，江苏徐州人，讲师，主要研究领域为自然语义处理；陆汝占(1940－)，男，江苏苏州人，教授，博士生导师，主要研究领域为语言类型，自动推理技术，汉语语义理论，自然语言理解与处理.

个层次的有限状态自动机串接而成,但在每一层中引入了错误修正规则进行歧义处理,从而有足够的能力处理汉语句法分析中的种种复杂现象.下面给出其形式定义.

ACFSA 是一个三元组(G,FS,OPS),其中

(1) G 是各层有限状态自动机的序列 $G_1,G_2,\ldots,G_n,G_1,G_n$ 分别为整个模型的入口和出口.$G_i(0<i\leqslant n)$的定义下面将给出.

(2) FS 是各层错误修正规则所公用的特征集.

(3) OPS 是各层错误修正规则所公用的操作集.其中的原子操作有:

$DelPhrase$:取消短语;$Ladd(i)$:将短语的左部扩大 i 个终结符;$Radd(i)$:将短语的右部扩大 i 个终结符;$ChgPhrase(tag)$:将短语标记改为 tag.

另外的一些操作可以由原子操作组合而成.例如,将短语的左、右边界扩大操作是由操作 $Ladd$ 和 $Radd$ 组合而成的.

第 $i(0<i\leqslant n)$层自动机 G_i 的定义是:$G_i=(N_i,T_i,P_i,ECP_i)$,其中

(1) N_i 是非终结符的有限集合.

(2) T_i 是终结符的有限集合.T_i 与低层语法的文法符号集的关系是

$$T_i=\bigcup_{0<j<1} T_j\cup N_j,$$

也就是说,T_i 是所有低层语法的文法符号集的并集.

(3) P_i 是形如 $A\rightarrow\alpha$ 的产生式的有限集合,称为基本文法,其中 $A\in N_i,\alpha\in T_i^*$.α 实际上是个正规表达式,它可以转换为一个有限状态自动机.把一层中所有短语规则所对应的自动机并起来,就形成了一个用于当前层分析的确定性的有限状态识别器,该识别器记为 DFA_i.

(4) ECP_i 是错误修正规则集.它主要用于对当前层分析结果中的错误根据上下文特征进行修正,从而得到正确的结果.它不仅使当前层的分析结果更为正确,也为下一层的分析提供了正确的输入.

每一条错误修正规则的形式是

$$\text{if } A_1=V_1 \text{ and } A_2=V_2 \ \ldots \text{ and } A_m=V_m \text{ then } op.$$

这里,$A_1\sim A_m\in FS$,$op\in OPS$,$V_1\sim V_m$ 是各个特征相应的特征值.以上规则表示,当条件中包含的所有特征值等式均满足时,执行操作 op.

1.3 ACFSA 的语法表示能力

ACFSA 的语法表示能力是一个非常重要的问题,关系到该模型能否正确和有效地表示汉语语法.我们认为,ACFSA 的语法表示能力主要表现在:(1) 可表示语法的层次性.(2) 利用错误修正规则可以进行歧义处理.歧义处理一般是通过上下文特征来进行的,而错误修正规则无疑将上下文特征引入到模型中来了.这一点是原来的多层有限状态自动机方法所无法实现的.(3) 可表示有限层次的中心嵌套.由于语法是分多层放置的,因而可利用层次关系和语法规则的重复来模拟有限层次的中心嵌套.在大规模真实文本中,大多数句子的嵌套层次都不会太大,因而在实际应用中该方法的中心嵌套表示能力已经足够了.(4) 减少了规则的重复和冗余.由于语法分层放置,也便于进行语法的模块化,把一些较为公用的、基本的短语规则放在较低层的有限状态自动机中,而把一些较复杂的短语和句型规则放在较高层的有限状态自动机中.这样,低层的短语规则可被各个更高层次所共享,从而减少了规则的重复和冗余.

由此可见,ACFSA 具有很强的语法表示能力,可以满足处理描述复杂语言的要求.

1.4 基于 ACFSA 的部分句法分析器

利用 ACFSA 可以方便地实现部分句法分析器,下面就实现中几个比较重要的问题:基本文法的构造、错误修正规则集的构造和句法分析算法进行论述.

1.4.1 基本文法的构造

各层的基本文法 P_i 是由人类专家给出的.具体的文法分层放置策略如下:

(1) 合理安排语法规则的层次.由于语法规则所处的层次不同,它在语法分析中被调用的次序也是不一样的,越是底层的规则越是先被调用,越是高层的规则越是后被调用,也就是说,各项规则实际上因层次的不同而

具有不同的优先级.各项规则的这种优先级的安排当然会直接影响着句法分析的正确率,起到一定的消歧效果.当然,如何确定具体的语法规则的层次安排,主要还是根据各个语法规则在语法体系中所处的层次来决定.对于层次相近的语法规则,要根据它们的使用概率来决定相对的层次,这种使用概率除了凭借语感之外,还可以通过语料库分析来获得更为准确的量化标准.

(2) 适当的冗余规则.虽然在某层自动机中经过修正规则的调用后,有些短语被取消,但我们仍希望能在修正的结果之上识别同类短语,此时,由于语法规则中识别该类短语的有限状态自动机已低于当前分析器所处的层次,因而不会再被调用,要识别该类短语,只有将该类短语的规则在多个层次中重复放置(重置).这样即使过了某个层次也不要紧,因为仍可以在后面的层次中利用重置规则对其进行识别.为了便于理解,下面,我们以下文将要介绍的CCSP(cascaded Chinese syntactic parser)汉语语法中的例子进行具体说明.

在语法的第3层设置了"的"字短语,第7层和第13层之后的某层也设置了"的"字短语.这里,后面两层是作为第3层的缓冲层,当第3层的"的"字短语因为修正规则调用而被取消后,在后面两层仍能识别"的"字短语.

1.4.2 错误修正规则集的构造

错误修正规则集 ECP_i 是各层语法的一个有机组成部分,直接关系到每层乃至整个句法分析的正确率.因此,要使ACFSA充分发挥其效力,达到足够高的正确率,首先应该构造出每层的错误规则修正集.错误修正规则的获取可以通过以下两种途径:

(1) 手工方式.手工方式主要是依赖于人类专家的语言学知识、语感以及对大量语料的观察,充分利用短语内部及上下文的语法、语义特征来进行歧义消解,因而具有很好的概括性和通用性.

(2) 错误驱动的机器学习.除了一些较为明显的歧义消解规则采用手工方式获得以外,我们还利用错误驱动的机器学习方法来自动学习错误修正规则.其显著优点在于,无需过多的人类专家的介入,在效率上比人工获取规则高得多,也提高了系统的可移植性.错误驱动的学习方法的主要思想是,对系统的初步分析结果利用人工校对得到相对正确的结果,从初步分析结果与正确结果的对比中来学习错误修正规则.各层采用统一的方法进行错误修正规则的自动获取,其具体步骤是:

Step 1. 样本准备.调用当前层的识别器DFA对当前层输入进行,得到初步分析结果,对其进行人工校对,得到校对后的正确结果.对每一个经过错误修正的短语提取一个样本,每个样本表示为(*InitialPhrase*,*CorrectedPhrase*,*Environment*,*Op*),其定义分别是:*InitialPhrase*,*CorrectedPhrase* 分别是初始分析结果和校对结果中的短语树形表示(短语结构树).*Environment* 是 *InitialPhrase* 在初始分析结果中的上下文环境,也就是短语所在句子经过当前层自动机分析所得的结果,一般可以表示为一个有序森林(短语结构树的序列).*Op* 是为了得到正确结果而对 *InitialPhrase* 进行的操作.

Step 2. 样本的形式变换.从每个样本的 *InitialPhrase* 和 *Environment* 提取 *FS* 中定义的所有特征的特征值,作为新样本的分量,并将原样本中的操作作为新样本的类别.这种一系列特征再加一个类别的样本表示形式是不少机器学习算法所要求的标准输入形式.

Step 3. 将上一步得到的所有样本作为输入,调用通常的机器学习算法学习样本的分类规则.具体采用哪一种机器学习算法可以根据需要来选定.

1.4.3 基于ACFSA的句法分析算法

当获取了一定量的错误修正规则后,就可以用ACFSA来进行句法分析了.为了下面论述方便起见,首先介绍一下短语结构有序森林的概念.它是短语结构树组成的有序森林,每棵短语结构树表示一个短语的内部层次结构,多棵短语结构树按次序排列组成句子的结构.我们把句法分析看成是一个自底向上的短语结构有序森林的构造过程.最初将句子中的词及其词性标记为输入,送入多层有限状态自动机的最底层.在每层通过该层有限状态自动机的分析将一系列短语结构树组合为更大的短语结构树,并将其作为下一层的输入.这样,一个句子顺次通过各层有限状态自动机,直到通过最后一层自动机的分析,其输出就是对该句子的分析结果.由于是部分句法分析,句子的最后分析结果不一定是一棵树,而可能是一个短语结构有序森林.

利用ACFSA对一个输入句子进行句法分析的算法描述如下:

输入:CurResult=组成句子的词序列.

输出:句子的句法分析结果.

算法步骤:

Step 1. $i=1$;

Step 2. 调用基本分析算法(下面将给出定义)对 *CurResult* 进行识别,得到的分析结果仍保存到 *CurResult* 中,作为当前层 ECP 应用算法的输入;

Step 3. 调用 ECP 应用算法(下面将给出定义)对 *CurResult* 进行错误修正,得到的结果仍保存到 *CurResult* 中,作为下一层的输入;

Step 4. if $i<=n$ goto Step 2;

else return.

基本分析算法的描述如下:

输入:上一层的分析结果.它是一个短语结构有序森林,记为 $P_1,P_2,\ldots,P_n$.

输出:当前层的分析结果 *CurResult*.它也是一个短语结构有序森林.

算法步骤:

Step 1. curpos=1;它代表在输入短语结构树序列中的当前位置,如 P_i 的位置为 i.初始时定位在最左边的短语 P_1;

CurResult=*NULL*;初始时结果为空;

Step 2. 调用当前层的识别器 DFA_i,从位置 curpos 开始进行短语识别.若识别出短语,得到的新短语记为 *NewPH*,其右边界的位置记为 endpos.否则,将当前位置的短语结构树加入 *CurResult*,转到 Step 4.在识别短语时,当两条或多条规则均适用于该位置时,优先选用长的规则.这种"最长规则优先"的策略可以起到歧义消解的效果,因为从实际应用中所得到的经验是,在其他条件均等的情况下,往往应优先选用长的规则;

Step 3. 为 *NewPH* 生成一棵新的短语结构树,其子结点为从 curpos 到 endpos 的所有短语结构树.将这棵新的短语结构树加入 *CurResult*;

Step 4. if endpos<n

curpos=endpos+1;

goto Step 2;

else return.

ECP 应用算法的描述如下:

输入:当前分析结果 *CurResult*.它是一个短语结构有序森林,记为 $P_1,P_2,\ldots,P_n$.

输出:经过错误修正的分析结果.

算法步骤:

Step 1. $i=1$;

Step 2. 对 P_i 提取 FS 中定义的所有特征的特征值;

Step 3. 在当前层的 ECP 中搜索 P_i 所满足的错误修正规则,并对 P_i 执行规则中的操作;

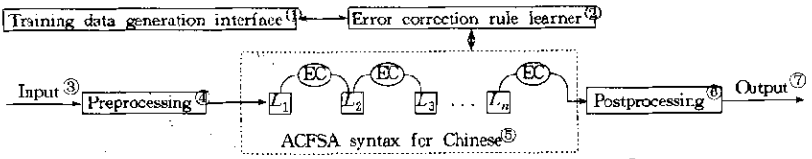
Step 4. if $i<=n$ goto Step 2;

else return.

2 CCSP——基于 ACFSA 的快速汉语部分句法分析器

我们基于 ACFSA 实现了一个原型系统 CCSP.CCSP 系统的处理流程可如图 1 所示.其中 L_1 到 L_n 分别是汉语句法的 ACFSA 表示中的各层自动机.输入为一个句子的切分标注结果,经过预处理送入汉语句法对应的 ACFSA,依次经过每层自动机及其该层的错误修正模块处理,然后通过必要的后处理得到各个小句的句法结构.其中,错误修正模块主要是利用该层的错误修正规则对该层的输出结果中的错误进行修正.错误修正规则学习器则用于从训练样本中自动学习错误修正规则,该学习器中还包含一个用于用户生成训练样本的人机界面.错误修正规则学习器只有 1 个,由各错误修正模块所共享.

下面,我们主要从预处理、汉语句法的 ACFSA 表示、错误修正规则集的获取、后处理这 4 个部分对 CCSP 系统中的主要技术进行介绍.



EC: Error correction module⑧; L_i: The ith level automata⑨.

①训练样本生成界面,②错误修正规则学习器,③输入,④预处理,⑤汉语句法的ACFSA表示,⑥后处理,⑦输出,⑧错误修正模块,⑨第i层自动机.

Fig. 1 The flowchart of CCSP system

图1 CCSP系统的处理流程

2.1 句法分析的预处理

句法分析的预处理的主要功能是划分小句和词性标注的错误修正.小句划分采用逗号、分号等标点符号作为小句的分隔标记,这当然会带来一些问题,因为有些小句的各组成部分是被逗号分开的,因而分出的单位有些本身不是小句,而只是小句的一个部分,这些错误将在后处理中予以解决.

2.2 汉语句法的ACFSA表示

我们在CCSP系统中给出了一个汉语句法的ACFSA表示,这是系统用于分析汉语句子的最基本的语言知识.

根据前面提到的文法分层放置策略,短语规则的层次安排大致如下:(1) 基本名词短语,各类准短语,数量短语.(2) 带动词的基本名词短语,偏正式形容词短语,简单"所"字短语.(3) "的"字短语,联合式形容词短语,方位短语,"地"字短语.(4) 偏正式名词短语.(5) 同位名词短语,联合名词短语.(6) 动宾短语,兼语结构,述宾式介词短语.(7) "的"字短语.(8) 偏正式名词短语,联合式介词短语.(9) 动补短语,述补式形容词短语,述宾式形容词短语,比拟短语.(10) 状中式动词短语,状中式形容词短语.(11) 联合动词短语或连动结构.(12) 动词谓语句,形容词谓语句.(13) 主谓谓语句.

第13层以下各层可根据中心嵌套的层数要求,与第3~13层类似来设置规则,只不过其中短语的某些组成部分可以是动词短语或小句.目前,我们实现的语法只设置了两层嵌套,已能处理大多数真实文本中的句子,根据需要还可以加入更多层数的嵌套.

最后层:这是语法中最后的一个层次,其中的语法规则主要包括关于ZS(主谓句)、NZS(非主谓句)、EXISTS(存现句)的规则,并且在小句结构中加入了句尾的标点符号.

具体的语法表示这里不再列出.

2.3 错误修正规则集的获取

错误修正规则根据来源分为手工规则和习得规则两类.手工规则是由人类专家手工加入系统的,习得规则则是利用机器学习方法得到的规则,这里主要论述习得规则的获取.

特征选取是学习算法中非常重要的一环.如果选取的特征数目太少,缺少某些重要特征,则无法对某些错误学习到其相应的修正规则,另外也容易得到过多的错误规则.如果选取的特征数目太多,也会使算法的时空复杂度提高,影响学习算法的实用性.这里,我们主要采用了一些对短语结构有预测性的特征,包括语法特征和语义特征(主要是语义类型),除了局部特征即短语结构及其相邻短语的特征外,还包括较远的上下文特征(如从当前短语到句末所出现的介词、动词等).

具体的机器学习算法采用了CN2算法[5],这是由Clark提出的一种规则列表学习算法(rule list learning).该算法的输入是一个例子集,其中每个例子列出了它的类别和主要特征,学习算法的任务就是从例子集习得对例子进行分类的规则列表.采用该算法而不采用一般的判定树学习方法(例如C4.5算法)的原因是,该算法可输出无序的规则列表,这种形式的规则便于进行人工检验.

下面是该算法所习得的一个变换规则.

规则1:

if Tag=noVP|zVP and PrevTag=DENP

then

Tag=NPBS1; //改变短语标记

该规则表明,如果当前短语为某种动词短语(noVP 或 zVP),并且前一短语为"的"字短语,则将当前短语标记变为名词短语(NPBS1).例如,对于句子"你们完得成今年的生产任务吗?",某层的初步分析结果为:

[rn 你们][vgo 完][usdf 得][vc 成] [DENP[NPT[t 今年]][usde 的]] [noVP[vg 生产][NPBS[ng 任务]]] [??]

经应用该变换规则后,其中"生产任务"的短语标记变为 NPBS1,结果如下:

[rn 你们][vgo 完][usdf 得][vc 成] [DENP[NPT[t 今年]][usde 的]] [NPBS1[vg 生产][NPBS[ng 任务]]] [??]

2.4 后处理

后处理阶段的主要任务是修正预处理中小句划分给句法分析带来的种种错误.前面提到,在预处理中以逗号作为划分小句的重要形式特征之一.但由于逗号用法的复杂性,会使句法分析得到的结果出现错误,有些分析出的非主谓句实际上是小句的成分,有些则是不完整的小句.后处理的任务就是将不完整的部分组合成真正的小句.

3 性能测试

对 CCSP 的性能测试,我们采用类似于 CASS 系统中的方法进行了初步测试.在未采用训练方法时,该系统的分析正确率已达 70%,这说明我们的多层语法层次设置是较为合理的.而在进行了一定量的训练之后,系统的正确率有明显的提高,达到 80%以上.随着训练量的加大,该方法的正确率将会有进一步的提高.

目前,我们采用的特征主要是局部特征,而有些问题必须在句子的全局范围内才能得到解决.如何利用更多的全局特征及如何根据全局特征来对分析结果进行修正,将是今后进一步研究的课题.

References:

[1] Abney, S. Partial parsing via finite-state cascades. In: Carroll, J ed. Proceedings of the ESSLLI'96 Robust Parsing Workshop. Prague, Czech Republic: ESSLLI, 1996. http://www.sfs.nphil.uni-tuebinger.de/~abney/96h.ps.gz.

[2] Abney, S. Rapid incremental parsing with repair. In: Proceedings of the 6th New OED Conference: Electronic Text Research. Waterloo, Ontario: University of Waterloo, 1990. 1~9. http://www.sfs.nphil.uni-tuebingen.de/~abney/90j.ps.gz.

[3] Ye, Dan-jin. Fast partial parsing for large scale real Chinese texts [MS Thesis]. Shanghai: Fudan University, 1998 (in Chinese).

[4] Clark, P. Rule induction with CN2: some recent improvements. In: Yves, K ed. Machine Learning——Proceedings of the 5th European Conference (EWSL'91). Berlin: Springer-Verlag, 1991. 151~163.

附中文参考文献:

[3] 叶丹瑾.大规模真实文本的快速部分汉语语法分析[硕士学位论文].上海:复旦大学,1998.

A Trainable and Fast Partial Parsing Method for Chinese

ZHANG Yi-min, CHEN Yu-quan, LU Ru-zhan

(*Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai* 200030, *China*)

E-mail: yimin.zhang@intel.com; lu-rz@cs.sjtu.edu.cn

http://www.cs.sjtu.edu.cn

Received May 17, 1999; accepted September 7, 1999

Abstract: In this paper, a trainable and fast partial parsing method for Chinese is presented. This method is based on the augmented cascaded finite state automaton proposed in this paper, which uses error-driven machine learning to achieve high accuracy. The preliminary result shows that the proposed method is promising.

Key words: fast partial parsing for Chinese; augmented cascaded finite state automaton; error correction rule; ambiguity resolution; machine learning