软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(2):245−257 [doi: 10.3724/SP.J.1001.2011.03809]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

层级分类概率句法分析*

代印唐+, 吴承荣, 马胜祥, 钟亦平

(复旦大学 计算机科学技术学院,上海 200433)

Hierarchically Classified Probabilistic Grammar Parsing

DAI Yin-Tang+, WU Cheng-Rong, MA Sheng-Xiang, ZHONG Yi-Ping

(School of Computer Science and Technology, Fudan University, Shanghai 200433, China)

+ Corresponding author: E-mail: daiyintang@fudan.edu.cn

Dai YT, Wu CR, Ma SX, Zhong YP. Hierarchically classified probabilistic grammar parsing. *Journal of Software*, 2011,22(2):245−257. http://www.jos.org.cn/1000-9825/3809.htm

Abstract: This paper analyzed various existing approaches of structural grammar parsing, and addressed the problem of over-classification and under-classification. Then a hierarchically classified phase structure grammar (HC-PSG) and a hierarchically classified probabilistic context-free grammar (HC-PCFG) parsing are proposed to respond to this challenge. A measure of class clustering is designed to eliminate the classification ambiguity of grammar rules. The HC approach implements a general learning rule from a small number of phrase instances. An instant clustering method is used to disambiguate rules learned from corpus. The HC method is also extended to context sensitive grammar parsing to improve performance. It employs the classification of the context relevancy to handle the problem of corpus sparsity. By all the means, it can leverage the conflicts between under-classification and over-classification.

Key words: phrase structure grammar; probabilistic grammar parsing; hierarchical classification

摘 要: 对已有的句法分析中引入知识的方法进行了归纳分析,认为多种句法分析方法都可被看作是基于特征标记的分类,然后分析了其中的欠分类和过分类问题.在此基础上,提出一种层级分类短语结构文法和一种层级分类概率句法分析方法(hierarchically classified probabilistic context-free grammar),并设计了一种通过对实例进行聚类来消除句法规则的分类歧义方法.还进一步将层级分类扩展到概率上下文相关句法分析方法,利用上下文相关性的层级分类来解决引入上下文相关时的数据稀疏性问题.通过上述一系列方法有效地克服了过分类与前分类之间的矛盾.

关键词: 短语结构文法;概率句法分析;层级分类

中图法分类号: TP391 **文献标识码**: A

句法分析(grammar parsing)又称文法分析,是自然语言理解的基础环节.在自然语言理解中,准确的实体识别、事件识别等信息抽取任务和语义角色标注、问题分类等深入的 NLP 任务都必须基于可靠的句法分析结果,而正确的分词和正确的句法分析树则是相互依赖的.句法分析一般基于特定的文法.短语结构文法(phase structure grammar,简称 PSG)[1]、中心词驱动文法(head-driven phrase structure grammar,简称 HD-PSG)[2]、链接

* 基金项目: 上海市科委、上海市人力资源与社会保障局博士后科研资助计划(10R21421400); 上海市科委项目(075115008)
收稿时间: 2009-04-20; 修改时间: 2009-08-12; 定稿时间: 2009-12-02

文法(link grammar)[3]、词汇功能文法(lexical functional grammar,简称 LFG)[4]、特征合一文法(functional unification grammar)[5]等一大类文法在应用于句法分析时有以下 4 项共同特点:(a) 将一个语句分析为一个句法树结构,认为语句是一个从符号或词语不断构造文法结构的过程.这个句法树的叶节点一般是词语或标点,非叶节点是一个短语或子句.本文称句法树的任何一个树节点为一个语言节点或简称节点.(b) 对每个分析出的语言节点赋予一定的离散的语法和语义特征,节点的离散的特征可以看作是某种分类类型.(c) 从底层节点开始,根据多个下级节点的特征组合,按照一定的规则或模式来构造上级节点.上级节点的特征类型由下级节点的特征类型以及其他上下文因素确定.上级节点的特征类型确定的过程可以看作是一个分类过程.(d) 其中,HD-PSG 和链接文法等以短语的多个子节点中的某个子节点作为核心节点.上级节点从核心子节点继承特征.鉴于这类文法将语句分析成树结构,并且对各级节点进行分类,我们将这一大类文法统称为分类结构文法(classified structural grammars,简称 CSG),其中,特别称概率上下文无关文法为 Classified PCFG.本文研究分类结构文法的句法分析,下文中除非特别说明,文法都是指分类结构文法.一些传统的分类算法可以成功地应用于分类结构句法分析,如,Wang 等人将 SVM、最大熵等分类方法应用于句法分析时取得了较好的效果[6].

分类结构文法的句法分析基于语言节点的离散特征向量.有的文法采用单特征,如 PSG 文法;有的文法采用多特征,如 GPSG.一个多维的离散特征向量总可以投影成一个一维的离散特征值.所以不失一般性,我们只需讨论单特征值的情形,研究特征值的取值如何影响规则的选择和歧义消解.短语结构句法 PSG 就是一种采用一维特征(短语类型)作为特征向量的结构文法.我们以短语结构文法为例,考察 PCFG 句法分析来研究结构句法分析的分类歧义问题及其消解方法.

分类结构文法的一个最大的困难在于句法分析歧义.对于给定的语句,按照给定文法可能产生多个句法树结果.句法分析必须采取措施来消除歧义.概率上下文无关句法(probabilistic context-free grammar,简称 PCFG)通过引入概率对句法分析结果评分,选择高评分的结果[7,8].张浩等人基于 CHART 算法实现了结构上下文相关的概率句法分析[9].Abney 提出了基于有限状态机层叠(finite-state cascades)的分层的句法分析方法[10],通过对文法规则划分优先级组实现了规则优先级消歧.上述统计分析没有改变上下文无关文法的结构,其性能一般只能达到 $F1$ 评分的 70%~75%.

深入研究句法分析歧义的特点和产生原因,这些歧义可以分为两种类型:分类歧义和结构歧义.分类结构文法中的短语作为上级语言节点,其下级短语、词语或标点等各个下级语言节点的特征类型构成一个特征类型序列.上级语言节点的分类类型是根据下级语言节点序列的特征序列的组合模式来分类的.这种根据下级类型序列的分类可以称为一种组合分类.分类结构文法的规则一般是从进行了文法标注的语料库训练学习得到的.因为分类体系的原因和手工标注的原因,语料库的标注类型可能存在歧义.同一个语料库中,同样的特征类型序列组合的节点组合产生的上级节点可能被分类为不同的上级特征类型.如 Penn Chinese Treebank(PCTB)中介词 P 与名词 NN 的序列组合可以得到 PP_DIR,PP_LOC,PP_MNR 等多种标记类型.又如同样的“动词 名词”组合则有可能构成 VP 或 NP,如((执行/V)VP(决议/N)NP)VP 和((合作/V 协议/N)NP)我们称这类歧义为分类歧义.

研究表明,产生分类歧义的一个重要原因是特征标记的类型较粗[11−13].我们称这类问题为欠分类(under-classification).研究表明,通过细分特征标记可以有效地消除欠分类引起的分类歧义.比如,陈晓辉等人通过将中文中的语法功能词作为特殊节点类型增加规则数量,限制规则适用范围起到减少歧义的作用,提高了准确率[11].Klein 等人通过将上级语言节点的标记作为前缀来扩展下级语言节点的标记类型,改进了句法分析效果[14].高升等人则首先在理论上对所有动词编制规则,得到一个彻底细化的规则库,然后通过规则聚类方法压缩规则数量[12].Liu 等人通过细分动词的类型来提高依存句法分析的精确度[13].词汇化的句法分析方法(如 lexicalized PCFG)可以看作是类型细分方法的极限形式,直接以中心词语作特征分类.如 Collins 提出的 Head Driven PCFG[15]等 Lexical PCFG 则直接以短语的中心词作为特征之一,相当于将类型细分到词语级别.Sun 等人将 Collins 的方法应用到中文句法分析,取得了良好的效果[16].上述细分类型方法缺乏一个如何确定分类层次的系统标准.

在应用类型细分的消歧方法时,所面临的一个核心问题是训练语料库数据稀疏性问题.当类型细分太细时,

特征序列的可能组和数目变得非常庞大,使得语料库中的已标注短语不能覆盖所有可能的特征组合.这样,开放测试或实际应用中的语言现象的特征类型序列组合可能超出语料库的覆盖范围,此时无法获得正确句法分析结果.我们将这种情况称为过分类(over-classification)问题(有的文献称为 over-splitting).过分类还可能造成句法分析算法的规则搜索空间巨大和无效分支过多的问题,带来较大的计算负担.词汇化的句法分析中的未见词(unseen word)也是过分类带来的数据稀疏性问题.

引入语料库以外的外部知识,如分类词表,是解决数据稀疏性问题的有效方法.WordNet[17]、《同义词词林》[18]等都是一种分类词表.Ding 等人基于语义分类,利用《同义词词林》的某个固定层次的分类标记作为 Lexical PCFG 的词语标记[19];Xiong 等人在 HD-PCFG 模型中以词林和 Hownet 结合的语义类替代词语[20].利用同义词或同类词提高了分类的覆盖范围,缓解了过分类问题.这些方法利用同义词或同类词提高了分类的覆盖范围,有效地缓解了过分类问题.Lexicalized PSG,HD-PCFG 以及上述方法的共同特征都是采用某种固定的分类类型级别,可以统称为定级的分类结构文法.在定级分类结构文法中,如果采用的分类层次太高、分类粒度太大,则可能存在欠分类问题;如果分类粒度太小,则过分类问题仍然不能完全解决.因此,找到合适的分类级别就成为定级分类句法分析的重要任务.Ding 等人[19]采用经验方法尝试了《同义词词林》的各个不同级别的分类标记,从而找到一个合适的分类级别.Petrov 等人通过拆分-合并方法[21],通过动态的 EM 方法,优化获得一个最优的分类层级.Ding 和 Petrov 等人定级分类方法的一个缺陷仍然不能完全解决训练数据稀疏问题,特别是当训练所得的规则的分类层级较低时,对于开放测试或应用中的词语,如果没有细分的规则可用,则只能按未见词(unknown word)来处理,丧失了可用的分类相似度信息.

除了分类歧义以外,结构文法中的另一类歧义是一个节点可能参与多个上下文的组合构成上级节点,得到不同的句法树结构.如,典型的“VP NP 的 NP”序列可能有两种不同的解析.这类歧义一般称为结构歧义.如果不考虑基于语义的方法,对于结构歧义,在类型细分的基础上引入上下文相关信息是解决结构歧义的有效方法.Black 等人提出的基于历史的文法考虑父子节点之间的相互影响[22].张浩等人通过在计算概率时考虑上下文相关信息[9].刘挺等人实现了中文的概率依存句法分析,并通过引入结构概率来突破依存文法本身的两节点关联限制[23].当结合细分类型的方法和上下文相关方法时,同样面临训练语料库数据稀疏性问题.

本文提出一种层级分类(hierarchical classification,简称 HC)结构句法分析方法来全面应对过分类和欠分类问题.HC 方法将传统结构分类文法的扁平定级分类空间替换为层级分类空间.本文将 HC 方法应用于改进传统 PCFG 句法分析,提出了层级分类上下文无关概率句法分析方法(hierarchically classified PCFG,简称 HC-PCFG).该方法首先建立一个基于类属包含关系的层次化的标记分类体系,然后在句法分析时赋予句法树的词语、短语等语言节点以层次化的分类特征标记.与定级分类结构句法分析不同,HC-PCFG 对短语文法规则也建立起规则层级体系,使得一个短语或词语一方面能够适用尽量细分的最适合的规则,另一方面在不能适用细分的规则时可以适用最匹配的较粗分的规则,从而达到消除歧义和克服数据稀疏性的双重效果,利用较少的训练语料可以处理较多的语言现象.本文还将层级分类延伸到概率上下文相关文法(hierarchically classified probabilistic context sensitive grammar,简称 HC-PCSG),通过规则关联的层级结构来克服上下文相关性的数据稀疏性问题.HC-PCFG 和 HC-PCSG 在 Penn Chinese Treebank(PCTB)上的实验测试结果表明,HC-PCSG 有效地提高了句法分析的召回率和准确率.文中在说明 HC-PSG 和 HC-PCSG 原理和方法时也在多处以 PCTB 为例来说明.

1 层级分类短语结构文法

传统的 PSG 将语句看作是一系列树形拓扑短语结构,并给每个树节点短语赋予一个标记.所有标记构成一个一维平坦空间.本文对传统 PSG 进行了改进,提出一种层级分类短语结构文法(hierarchical classified phase structure grammar,简称 HC-PSG),利用层级分类标记体系克服结构文法中过分类与欠分类的矛盾.

为了方便后续讨论,这里首先介绍 HC-PSG 中所应用的句法树、语言节点等概念.

定义 1. HC-PSG 中,一个合法的语句可以被分析成一系列短语、词语和符号构成的树形拓扑层次结构,称为句法树.每个短语由若干个短语、词语和标点构成.这种短语层次结构称为一个句法树.每个短语、词语或标

点都称为一个语言节点,简称节点.每个短语由若干相邻并连续的下级短语、词语或符号的有序序列构成,记作 $n=\langle n_1,n_2,\ldots,n_n\rangle$.其中,$n$ 称为上级节点,每个 n_i 称为 n 的一个下级节点.

句法树是 PSG,GPSG,LFG 以及依存文法等结构文法的句法分析结果的共同结构.分类结构文法都认为语言符号流的符号、词语之间存在内部的层次关系.HC-PSG 中同样将语句看作是一种层级短语结构,也给每个语言节点赋予一个特征标记,但这种特征是一种类型.所有标记构成一个层级体系结构.

定义 2(类型层级体系). 一个类型层级体系(classification hierarchy)定义为一个三元组 $H=(C,c_0,\subset)$,其中,C 是一个类型集合;$c_0\in C$ 是一个根标记,是一个包含偏序关系;$\subset$是一个包含关系.$\subset$满足以下条件:(a) 归属性:对于任意 $c\in C$ 并且 $c\neq c_0$,存在 c'使得 $c\in c'$;(b) 传递性:如果 $c\subset c'$并且 $c'\subset c''$,则 $c\subset c''$;(c) 无环性:如果 $c\subset c'$,则 $c'\subset c$.

类型(class)在有些文献中称为语义类型(semantic class).典型的类型层级体系是概念分类类型层级体系.类型层级体系的各个节点按照包含关系构成一个有向无环图.本文的实验中实现了树拓扑的概念层级体系.基于层级分类序列的定义,可以定义层级分类上下文无关文法.

定义 3(类型、类型序列、类型结构). 从一个分类层级体系 H 中可以给一个句法树的每个语言节点 n 赋予一个分类类型 $c(n)\in H$,称为 n 的特征类型,简称类型.当一个类型 c 赋予节点 n 时,c 和 c 的所有上级类型直到根类型 c_0 构成的集合$\{c,c',c'',\ldots,c_0\}$就同时赋予了 n,称为 n 的类型继承层级序列,简称 n 的类型继承(class heritage,简称 CH).

一个短语 $n=\langle n_1,n_2,\ldots,n_n\rangle$的所有下级节点 n_i 的类型 c_i 构成一个有序序列 $F=\langle c_1,c_2,\ldots,c_n\rangle$,称为短语 n 的特征类型序列,简称类型序列(class sequence).从 n 的类型 c_n 到 F 的映射 $m:l\rightarrow\langle c_1,c_2,\ldots,c_n\rangle$称为这个短语的类型结构(class structure).

在层级分类中,每个词语、符号或短语这样的语言节点不是只被赋予一个单一的标记,而是被赋予一个类型继承层级序列.如,短语“发言人说,…”中,“说”一词的层级继承被标记为〈[动作]/[信息动作]/[表达]/[告知]〉(本文用“[c]”方扩号表示一个类型,用一个形如“〈[c]/[c']/[c'']〉”的序列表示一个类型继承层级序列).图 1 中演示了一个语言节点的各个下级节点如何标记层级继承.

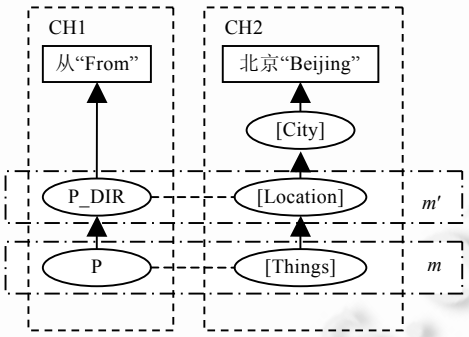


Fig.1 Classification hierarchy of a language node

图 1 语言节点特征类型层级

定义 4(类型序列匹配、类型序列模式). 两个特征类型序列 $F=\langle c_1,c_2,\ldots,c_n\rangle$和 $F'=\langle c'_1,c'_2,\ldots,c'_n\rangle$ 的类型数目相同,如果对应的语义特征满足 $c_i\subset c'_i$,$i=1,\ldots,n$,则称 F 类型序列匹配于 F',记作 $F\subset F'$.对于一组类型序列$\{F_1,F_2,\ldots,F_n\}$,如果有一个类型序列 F^P 使得每个 F_i 满足 $F_i\subset F^P$,则称 F^P 为这组类型序列的类型序列模式(本文对$\subset$符号和$\in$符号多次定义使用,分别表示类型包含、类型序列包含、规则包含等关系.符号的意义取决于符号两边的操作数.这种方式类似于面向对象编程中的运算符重载,是合理的).

句法分析是一个模式识别的过程,一个短语的类型结构的模式一般称为一条文法规则.

定义 5(规则、规则实例). 对于一组短语$\{n_1,n_2,\ldots,n_k\}$,每个 n_i 的类型结构为 $m_i:l_i\rightarrow F_i$,$i=1,\ldots,k$.如果存在类型 l^P 和类型序列 $F^P=\langle r_1,r_2,\ldots,r_n\rangle$使得 $l_i\subset l^P\wedge F_i\subset F^P$,$i=1,\ldots,k$,那么,映射 $m:l^P\rightarrow F^P$ 称为这组短语的类型结构模式,或称

为一条规则.每个 r_j 称为一个角色.每个短语 n_i 称作规则 m 的一个实例,记作 $n_i \in m$.

上述关于规则的定义也同时说明了对于一个语言节点序列,如何根据其类型结构识别其可适用的规则.句法分析的过程就是根据句法规则识别构造句法树的过程.

图 1 演示了语言节点如何提取特征类型层级.语言节点“从”的类型继承 CH1 是[Beijing]/[P_DIR]/[P],“北京”的类型继承 CH2 是[From]/[City]/[Location]/[Things].这样,同一个语言节点序列可能提出不同的特征类型序列,适用不同的规则.如,图 1 的样例可能适用“PP→P Things”和“PP_DIR→P_DIR Location”两条规则.

因为分类层级体系中的类型的包含关系,规则之间也存在包含关系.HC-PSG 中不仅分类标记是层级的,规则也是层级的.

定义 6(规则包含、规则层级体系). 如果两条规则“$m:l\to\langle r_1,r_2,\ldots,r_n\rangle$”和“$m:l\to\langle r_1',r_2',\ldots,r_n'\rangle$”具有同样的右侧项数目并且满足 $l\subset l', r_i\subset r_i'$,则称 m 是 m'的一条子规则,记作 $m\subset m'$.基于分类标记的包含关系的传递性和无环性,规则包含关系同样满足传递性和无环性.这样,一个 HC-PSG 文法系统中的所有规则按照包含关系构成一个格拓扑结构,称为规则层级体系(rule hierarchy).

除了引入分类层级和规则层级以外,HC-PSG 还继承了中心词驱动 PSG(head driven PSG)的原则.不同于在 Head-Driven PSG 中每个短语都有一个头子节点,短语被分为两种类型:有核心词的短语和无核心词的短语.

- 词语、标点和有中心词的短语类型,如 NP,VP,ADJP 等.在每个短语语言节点的所有下级语言节点中,确定其中一个作为中心节点.如果一个短语的中心节点是一个词语,则这个词语作为中心词.短语的分类标记继承自短语的中心词的类型.
- 无核心词的短语类型,如 PP,DNP 等.其分类标记则采用规则的左侧分类标记 $l=LHS(m)$.这类分类标记仍然可以构成层级体系,如 PP_DIR∈PP,PP_LOC∈PP.

有中心词的短语的类型就是其中心词的类型.

定义 7(层级分类短语结构文法). HC-PSG 文法系统定义为一个三元组(H,S,M),其中,H 是一个分类层级体系,$S\in H$ 是一个开始符号,M 是一组规则构成的层级体系.

HC-PSG 与传统 PSG 的定义有两点不同:HC-PSG 的标记集是一个标记层级体系;HC-PSG 中不再区分终结符与非终结符.

下面以公理的形式来说明一个短语序列如何适用一条规则.

公理 1. 一个语言节点序列$\langle n_1,n_2,\ldots,n_n\rangle$适用一条规则 $m:l\to\langle r_1,r_2,\ldots,r_n\rangle$的标准是每个节点 n_i 的特征标记 $f(n_i)\subset r_i$.

句法规则的适用判定实际上是一个模式识别的过程,而模式识别根据特征的相似性来判定.短语节点的层级分类标记可以方便地实现特征相似性判定.

推理 1. 对于一个语言节点 n 以及两条规则 m 和 m',如果 $n\in m$ 并且 $m\subset m'$,那么 $n\in m'$.

上述推理说明,一个节点可能适用多条规则,造成歧义.本文第 2.3 节引入基于信息量的局部消歧方法来解决这个问题.

综上所述,HC-PSG 是一种定义了分类标记层级体系、规则层级体系和中心词规则的短语结构文法.

2 层级分类概率上下文无关句法分析

2.1 层级分类概率上下文无关句法分析

基于 HS-PSG 的概率句法分析是对上下文无关概率句法分析 PCFG 的改进,称为层级分类 PCFG.HC-PCFG 采用和 PCFG 一样的概率评分计算公式:

$$p(T,S)=\prod_i p(R_i\mid L_i),$$

其中, $p(R\mid L)=\dfrac{Count(L\to R)}{Count(L)}$.

最后,在所有的分析结果中取 T_{best}=argmax($p(T,S)$)作为分析结果.但是,在HC-PCFG的规则训练中,可以通过规则分类标记层级体系和训练规则层级,这样可以利用细分的规则得到一般的规则.一条规则的实例数目是其所有子规则的实例数目的总和:

$$count(L \mid C_1^R, C_2^R, ..., C_n^R) = \sum_{l \in L, c_i^R \in C_i^R} count(l \mid c_1^R, c_2^R, ..., c_n^R).$$

一个类型短语的数目是该类型及所有自类型的短语的数目的总和:

$$Count(L) = \sum_{l \in L} Count(l).$$

软件实现中容易通过一个分类计数树来计算短语类型的计数和规则实例的计数.

2.2 实例聚类规则精练消歧

应用 HC-PCFG 从语料库训练规则时存在两个方面的问题.一方面,分类结构文法中存在分类歧义问题.规则是从语料库标注过类型的短语实例中提取的,诸如PCTB这样的语料库因为类型不够细分和人工标注等原因存在分类歧义.我们以PCTB中的〈P NN〉序列为例来说明分类歧义和规则精炼过程,如,PCTB的"P NP"类型序列可能构成多种不同的 PP 类型:"m_1:PP_LOC→〈P,NN〉","m_2:PP_TMP→〈P,NN〉","m_3:PP_MNR→〈P,NN〉".一组右侧类型序列相同而左侧类型不同的多个规则称为一个歧义规则组,要消除歧义就是重新确定分类的特征.歧义规则组有一个相同的右侧类型序列 R 和多个歧义的左侧类型 l_i.另一方面,根据 HC-PSG 规则的定义,规则是一组类似的短语的类型结构的模式,也就是短语类型和短语成员角色类型的上级类型.这种上级类型是一个序列,那么问题是到底采用类型序列上的哪个级别的类型作为规则的类型?如果类型太粗,则可能造成欠分类引起的分类歧义;而类型太细,则可能产生过分类引起的数据稀疏性问题.

综上所述,规则训练的目标是:(1) 避免分类歧义;(2) 在目标(1)的前提条件下尽量采用类型层级体系上更高级别的类型来保证规则的覆盖范围,避免数据稀疏性,降低句法分析复杂度.

本文提出一种自适应实例聚类精炼方法,利用决策树分类方案来自动精炼规则.在规则训练时如果发现分类歧义,将启动精炼算法来消除分类歧义.

从语料库中训练 HC-PSG 的规则有两种方式:自底向上聚类(如文献[12]中的方式)和自上而下精炼(如文献[11]中的方式).规则训练的目标是,在保证没有映射歧义的情形下,使得规则的数量尽量少,以优化规则匹配搜索和减少无效分支的数量.我们将自底向上聚类的方法结合 HC-PSG 层级分类设计了一种实例聚类规则精炼消歧方法来消除分类歧义.这里首先介绍此方法所应用的数据结构.

定义 8(角色类型树序列). R 是一个歧义规则集,其中任意一个规则的角色数目为 m.对于每个规则 $r \in R$,r 的一个实例 $n=\langle n_1,...,n_m\rangle, n \in r$,$n$ 的子节点 n_i 称为 R 的一个角色实例节点.n_i 的类型 c_i 称为 R 的一个 i-th 角色类型.一个角色类型可能对应于 R 中多个规则的多个实例的多个角色实例节点.这些所有歧义规则实例节点的集合称为 c_i 的实例集,记作 $N^R(c_i)$.对于同一个角色的两个角色类型 c_i 和 c_i',如果 $c_i \subset c_i'$,则 $n^R(c_i) \subset n^R(c_i')$.

R 的一个角色的所有 i-th 角色类型及这些角色类型的上级类型构成一个层级结构,称为 R 的第 i-th 角色类型树,记作 H_i.R 的 m 个角色类型树构成 R 的角色类型树序列$\mathbb{H}=\langle H_1,...,H_m\rangle$.

图 2 演示了一个 2 角色的角色类型树序列.上述定义实际上也说明了角色类型树序列的构造方法.歧义规则组的角色类型树和角色类型树序列也是一个角色实例的计数工具,主要用于计数一个类型都有哪些实例可以作为歧义规则组的某个角色,据此可以挖掘无歧义的细分的规则.

从$\mathbb{H}=\langle H_1,...,H_m\rangle$的每个 H_i 中任取一个类型 c_i,构成一个类型序列 $R'=\langle c_1,...,c_m\rangle$.$R'$对应的实例集合则是 $\bigcup_i N^R(c_i)$.这些实例的左侧类型可能分别是对于某个歧义类型 L.对于每个歧义规则的左侧类型 L,可以得到一个新的细分的规则 $L \to R'$,称为一个候选细分规则.

一个候选细分规则 $L \to R'$的召回率定义为 $L \to R'$的实例数和 $L \to R$ 的实例数之比:

$$recall_{R'}^{L}=\frac{count(L\rightarrow R')}{count(L\rightarrow R)}.$$

精确度定义为 $L\rightarrow R'$实例数与 R'的所有实例数之比:

$$precision_{R'}^{L}=\frac{count(R'\rightarrow L)}{count(R')}.$$

如果召回率和准确率大于一定的门限,就可以认为得到一条无歧义的细分规则.

上面定义了相关数据结构和评判方法,下面总结介绍实例聚类规则精炼消歧方法:

(1) 统计歧义规则组的角色类型树序列和每个角色类型的实例集.

(2) 构造候选规则.对每个角色的每个分类逐个构造角色分类标记序列 R'.对于每个 R'和每个歧义 L 的组合(R',L)构成一个候选精细规则 $L\rightarrow R'$.

(3) 计算候选细分规则的召回率/准确率.当某个组合(R',L)的召回率和准确率大于特定阈值时,认为消歧成功.该分类序列 R'被称为 L 完备识别序列.

(4) 符合标准的候选规则作为无歧义规则添加到规则集.

图 2 显示了前述歧义规则组的所有实例聚类得到的一个角色类型树序列,其中,根节点代表角色序号.66 个短语实例按角色分配到每个角色的角色类型中.然后可以构造细化的类型序列,如,〈[方式介词],[事物]〉或者〈[对象介词],[人]〉.

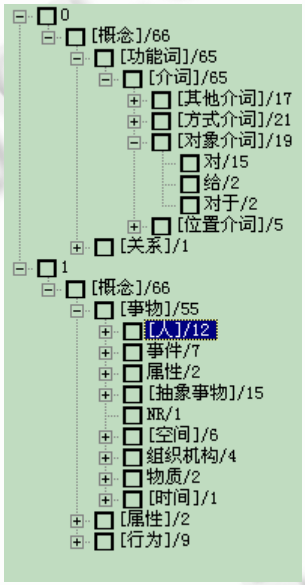


Fig.2 Classification clustering tree of the roles of the instances

图 2 实例角色聚类树

对于 n 个角色的歧义规则组,如果每个角色的分类树的所有类型集合为 H_i,那么这 n 个角色的类型集合的笛卡尔集构成一个类型序列的空间,每个候选的细分类型序列是这个 n 维空间中的一个向量.以上实例聚类规则消歧的过程可以看作是在这个空间按照召回率和准确率标准执行的搜索过程.如果每个角色 r_i 的分类树节点数为 N_{r_i},那么总的搜索空间为 $\prod N_{r_i}$,这可能是一个巨大的搜索空间.实际实现中,一些启发式搜索措施可以用来减小搜索空间.首先,如果一个角色聚类树的一个类型节点 C 下只有 1 个子节点 C',节点 C 将不参加分类标记序列构造;其次,分类序列构造总是逐个角色执行,逐层次执行.实际实现中,大多数歧义分类仅仅通过一个单独角色的分类树就可以找到完备识别序列.

2.3 基于信息量的局部消歧

HC-PSG 句法分析时,每个语言节点都被标记一个分类层级体系.一个节点序列可能被识别出多个类型序列,应用多条规则.这也是一种分类歧义,称为局部分类歧义.图 1 就演示了这样一种局部分类歧义的情形.本文提出一种基于信息的局部消歧方法.该方法首先定义类型和规则的信息量,然后取信息量最大的规则作为有效分析结果.

HC-PSG 的层次标记体系中,每个标记的作用是区分节点的类型,每个不同的类型包含不同的下级类型数.一个类型包含的下级类型越少,说明其歧义度越低,所包含的信息量越高.所以,一个类型标记 c 的信息量 $I(C)$定义为该类型下辖的下级类型点 $c\subseteq C$ 的数量和分类层级体系 H 中总的概念数量之比的反对数:

$$I(C)=\log_2\left(\frac{count(H)}{count(c,c\subseteq C)}\right).$$

显然,当 $c\subset c'$时,$I(c)>I(c')$.这个定义符合信息量关于信息确信度的描述特性.

然后,每个规则 $m:l\rightarrow\langle r_1,r_2,\ldots,r_n\rangle$都被赋予一个识别信息量:

$$I(m)=\sum_i I(r_i).$$

同样,显然,当 $m\subset m'$时,$I(m)>I(m')$.这样,对于一组语言节点序列,其适用的规则所定义的类型越细,则短语的信息量越高.最后,取 m_{best}=argmax($I(m)$)作为节点序列所适用的规则,完成局部消歧.信息量的计算是在规则训练时就完成的,并完成排序.这样,在执行句法分析时只需根据规则信息量之间的排序就可以迅速完成局部分类歧义的消歧.

3 层级分类上下文相关概率句法分析

当分类歧义消除后,仍然存在另外一种歧义:由不同的语义节点序列组合构成的多个语义节点对应同一个语言节点.这样,一个语句或语句中的一段连续的符号可能被分析为不同拓扑结构的语言树,对应多个语言节点.这种歧义一般称为结构歧义.

引入上下文信息的概率上下文相关文法(probabilistic context sensitive grammar,简称 PCSG)是消除结构歧义的一种有效方法[9,22].但是,在规则未细分的欠分类概率句法中,在歧义严重的情形下,上下文相关文法的效果被歧义所削弱.而 Lexicalized PCFG 等过分类的概率句法分析则由于数据稀疏性问题使得上文相关句法难以适用,引入上下文相关关联信息使语料库数据稀疏性问题变得更为严重.例如,对于 N 条规则的文法,其不考虑位置 2-gram 的上下文相关关联的组合数量级为 $O(N^2)$,对于 PCTB 这样的数千条规则的句法树库,可能的上下文关联将达到百万数量级.语料库中包含的短语实际上小于 10 万个(如 PCTB 语料库包含约 50 000 短语),不可能覆盖这些关联现象.所以,Collins 在研究细分到词语的 HD-PCFG 时采用了独立性假设,不再考虑上下文关系[15].

本文基于层次化的分类规则分类集提出一种层次化的上下文相关概率句法分析方法,将上下文相关现象实例按层级分类规则进行聚类,以应对上下文相关性时的数据稀疏性问题,可以有效缓解这一现象.这里,先明确定义规则上下文关联.

定义 9(上下文关联). 如果一个适用规则 M 的短语的第 *role* 个子节点也是一个短语并适用规则 m,则称三元组(m,*role*,M)为一个上下文关联(context relevancy),简称关联.m 称为前规则,M 称为后规则,*role* 称为关联角色.

特别地,规则共现关联关系体现了规则规约的顺序.利用上下文相关信息来对句法树进行评分时,规则的概率不再是独立的.句法树的评分公式为

$$P(t)=\prod P(m\mid M,role),$$

其中,M 是 m 的后规则,$P(m|M,role)$是上下文关联的关联概率:

$$p(m\mid M,role)=\frac{p(m,role,M)}{p(M)}=\frac{count(m,role,M)}{count(LHS(m,role,M))}\Big/\frac{count(M)}{count(LHS(M))}=\frac{count(m,role,M)}{count(M)}.$$

关联概率体现了一个下位短语作为上位短语的角色的可能性.

定义 10(上下位关联、关联层级体系). 如果两个关联(m,i,M)和(m',i,M)满足 $m'\subset m$ 和 $M'\subset M$,则称(m,i,M)是(m',i,M')的上位关联,后者是前者的下位关联,记作$(m',i,M')\subset(m,i,M)$.所有的上下文关联的上下位关系构成一个层级体系,称为关联层级体系(relevancy hierarchy).

上下位关联的关联角色必须是相同的.一个上位关联的样本计数是其下位关联的样本计数的总和:

$$count(m,role,M)=\sum_{m'\in m,M'\in M} count(m',role,M').$$

实现中,容易通过一个关联聚类计数树和本文第 2.1 节中类似的分类计数树来计算关联概率.

HC-PCSG 的关键在于,在句法分析时,如果一个关联在训练得到的关联层级中不存在,则查找其最接近的关联的概率作为其关联概率.这样,对于语料库中没有的关联现象也能找到最近的近似.我们以 PCTB 中的短语和规则为例来说明如何应用上下文关联和关联聚类.以下为 PCTB 中的 5 条规则:

- M':VP→⟨PP_TMP,VV⟩;
- M:VP→⟨PP,VP⟩;
- M:PP→⟨P,NP⟩;
- m':PP_TMP→⟨P,NT⟩;
- m'':PP_LOC→⟨P,NP_PN_LOC⟩.

因为 PP_TMP⊂PP,PP_LOC⊂PP,NT⊂NP,从而满足 $m'\subset m$ 和 $M'\subset M$,因此,$(m',1,M')\subset(m,1,M)$.假设训练语料库中只存在关联$(m',1,M')$的短语实例,根据聚类可以得到 $p(m|M,1)=count(m',1,M')/count(\underline{M})$.在测试语料库中,在计算关联$(m'',1,M)$的评分时,因为 $m''\subset m$,所以$(m'',1,M)\subset(m,1,M)$,就可以直接应用 $p(m|M,1)$作为其评分.这就是说,若训练语料库中没有出现过关联,则以其类似的关联聚类获得的上位关联来计算.

上下文相关概率的另一个意义是一条规则的某个角色的可扩展性,度量是可扩展的/不可扩展的.仍然以 PCTB 为例来说明如何通过关联概率判定一个角色是否可以扩展.例如,"M:VP→⟨VV,了/u⟩"规则中,$count(m,2,M)\equiv 0$,即 $p(m,2,M)\equiv 0$,这说明,M 的第 2 个角色是不可展开的,可以确定不能参与新的组合.

Abney 提出了基于有限状态机层叠(finite-state cascades)的分层的句法分析方法[10],通过将规则划分优先级组实现规则优先级消歧.但对于包含数千条规则的大规模语料库,难以人工识别规则之间的层叠关系.HC-PCSG 中对关联关系进行挖掘,还可以得到规则的层叠优先级.利用层叠的规则集分别进行句法分析,可以实现与组块分析类似的效果,减小规则搜索空间,提高句法分析速度和精度.

4 实 验

4.1 实验语料库与评测标准

实验在 Penn Chinese Treebank(PCTB)上进行,参照相关工作的通行做法,选取文章 1~270 约 3 000 多个语句作为训练及和封闭测试,选取文章 271~300 约 300 多个语句作为开放测试.测试用通行 ParseVal[24]作为评价体系.在相同的测试集和相同的评价标准下,实验结果与文献[6,9,16,19]这些工作进行了对比.实验中,我们参照文献[9,15]的工作实现了基于线图(chart)算法的句法分析器,并运用我们提出的 HC-PCFG 和 HC-PCSG 模型对分析出的句法树进行评分消歧.

语料库语言学中,语料库标注的质量和语法语义信息的深度对语言处理的效果至关重要.目前,PCTB 语料库的标注中存在一些问题,如定级分类标注、不一致的规则拆分、语义角色标注不彻底等问题,需要对其进行一系列的改造.实验中,我们根据 HC-PSG 文法模型改造了 PCTB,并尽量消除语料库中的不一致标注.

4.2 PCTB上的HC-PSG规则训练与精炼消歧

PCTB 是按照通常的短语结构文法来进行标注的,而 HC-PSG 文法标注是类型层级序列.要应用 HC-PCFG/PCSG 时必须将 PCTB 语料库改造为 HC-PSG 标注,在评测时则将 HC-PCFG/PCSG 句法分析产生的 HC-PSG 句法树重新映射为 PCTB 语料库进行比较测算.在 PCTB 上训练 HC-PSG 规则集首先要重新标注为 HC-PSG 文

法标注,按照以下步骤完成:

(1) 学习训练初始 PSG 规则集.参照文献[6]的方式,实验中对语料库的语句进行了清洗,消除了句法树中的单目的转换和空节点.然后训练初始的文法规则集.PCTB 是一个手工标注的研究语料库,其中一些长短语其实可以拆解为一系列较短的短语的上下文关联.如 m="VP→PP_LOC ADVP VP"可以拆解为 m_1="VP→ADVP VP"和 m_2="VP→PP_LOC VP"构成的上下文关联(m_1,1,m_2).在标准学习方式下,从 PCTB 训练得到 3 400 条规则,通过规则拆解,成功拆解 400 多条长规则.

(2) 短语中心词解析.我们参考文献[15,16]中的中心词识别规则解析中心词,利用规则中心词角色学习每个短语的中心词和短语的构成词序列.不同于文献[15],我们将规则划分为两类:有核心词的规则(如"NP,VP,ADJP 等")和无核心词的规则(如左侧项为 PP,DNP 等).训练语料库 270 篇文章共 3 400 个语句可以分解为约 54 000 个短语,其中约 34 000 个短语可以解析出核心词.不能解析出核心词的短语则直接以其短语标记类型作为核心词.

(3) 映射 PCTB 的 PSG 标记到 HC-PSG 类型.对于词语,首先根据分类词表解析词义,得到每个词语的分类标记.对于标点符号,直接以该标点为其标记.对于短语标记,或者映射为 HC-PSG 的本体类型,或者在 HC-PSG 中.

(4) 精炼规则集,消除歧义,得到 HC-PSG 规则集.如果一组规则之间出现结构歧义,则根据本文第 2.2 节的规则分类精练消歧方法,对歧义规则组的所有实例按构成词序列规则进行精炼.精炼是一个扩散的过程,一个短语精炼后,作为上级短语的构成角色将使上级短语的规则也得到精炼.利用精炼的规则集对开放测试语料库语句进行句法分析.当一个短语节点序列适用多个规则时,根据规则之间的细化关系消除歧义节点.

4.3 分类词表构造

分类词表是层级分类句法分析的重要依据.这里简要介绍分类词表的构造方法.

(1) 手工构建层级分类体系.该类型的分类体系接近 WordNet 和中文知网的分类体系.图 3 演示了实验中构造的层级分类体系.

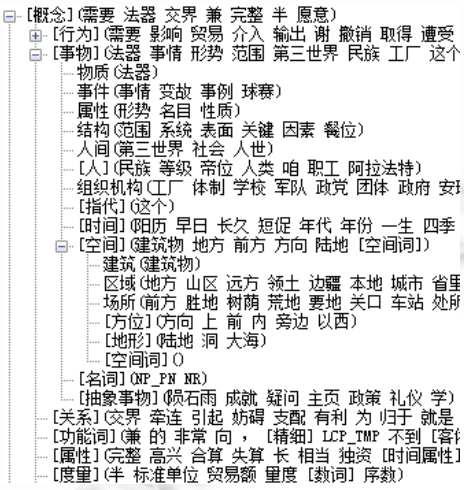


Fig.3 A section of the classification hierarchy

图 3 层级分类体系

(2) 将《同义词词林》的类型分配到 HC 类型.这样,汉语中的主要词语都具有了层级分类类型.

(3) 学习 PCTB 中的所有词语的词表,共计 10 000 多个.

(4) 进行 PCTB 词表到层级分类体系类型的映射,词义映射时还要进行词义消歧和词义合并.PCTB 的词语

都标记了 NR,VV 等词类,按照词类与层级分类体系的映射关系解析 PCTB 词表的类型.

(5) 手工调整部分词类.《同义词词林》的词类按照一种相关性来分类,有的语用不同的词语被分为同一类型,这样,在封闭测试时就会发生错误.根据错误类型进行语言学分析,调整成正确的类型.

4.4 PCTB与HC-PSG的标记的映射

HC-PSG 句法分析标注结果是中心词的层级类型标记或者无中心词短语的细分类型.在评测句法分析结果与 PCTB 比较时,需要逆映射为 PCTB PSG 标注.在逆映射和评测中我们做如下处理:

- HC-PCFG 和 HC-PCSG 的分析结果是细分的类型,如[组织机构]、[区域]等.我们将 PCTB 的类型也建立了类型层级,如 NP_PN 和 NP_TMP 都是 NP 类型的子类型,并与 HC-PSG 的层级进行映射.句法分析完成后要重新映射到对应的 PCTB 的最细标记类型,然后与语料库的句法树进行比较.如[组织机构]、[区域]和[人]类型都映射为“NP_PN”.
- 有的标记类型不属于句法分析范畴,如 IP_HLN 表示作为标题的语句,我们将其作为与 IP 等同.
- PCTB 的标注较粗,有的应该标注为 NP_PN 或 NP_TMP 等细分类型的短语则仅仅简单地标注为 NP.在统计 LP,LR 等指标时,我们按照向下匹配的原则来计算,即分析结果逆映射为类型细分类型的短语(如 NP_TMP),如果 PCTB 标注为上级类型(如 NP),则仍然认定为一致;反之,则认定为不一致.
- Penn Chinese Treebank 不仅是简单的句法分析树库,其中已经包含了相当丰富的语义知识信息.PCTB 虽然不像 Penn Chinese Propbank 那样详细标注所有语义角色,但其 Treebank 的很多标记都包含了一个短语在上级短语中的语义角色的信息,如 NP_PN_SBJ 标记的短语在上级短语中作为 Subjective 角色.一个短语的角色类型是由其在上级短语的规则中的角色位置决定的.我们对 Treebank 的标记进行了分解,分成类型和角色两个部分.这样,在句法分析时,HC-PSG 的每个短语都被赋予角色.在评测结果时,我们采用与前述类型细分问题类似的向下匹配原则,即分析结果有语义角色而 PCTB 没标注语义角色的,仍然认为正确.

4.5 实验结果

实验中,参考文献[9,15]的工作实现了一种优化的线图算法句法分析器,运用训练和精炼获得的规则对测试语句进行句法分析.最后,将 HC-PCFG 和 HC-PCSG 的结果与已有的一些研究进行了对比,结果见表 1.相关对比工作的性能数据直接摘自相关参考文献.HC-PSG 的句法分析结果为层次标记,若与 CTB 的标准句法标注进行比较评估,则还需要根据 CTB 和 HC 之间的映射完成逆影射后在与 CTB 句法树进行比较,计算 LR,LP,CB 等各项指标.

Table 1 Test results

表 1 对比实验结果

Model	LP	LR	CBs	0CB	2CB
Baseline: PCFG[9]	75.71%	70.27%	3.42	24.81%	40.24%
Class based PCFG[19]	82.56%	79.60%	2.27	47.10%	70.52%
Head-Driven PCFG[16]	86.4%	85.9%	–	–	–
SVM[6]	86.9%	87.9%	–	–	–
HC-PCFG	84.2%	82.8%	2.19	51.5%	70.2%
HC-PCSG	91.7%	90.3%	2.02	57.3%	75.2%

HC-PCFG 不考虑上下文信息,其 LP,LR 结果略高于 Class based PCFG.可以认为,层级分类方法比定级分类方法提高了约 2%的效果.HC-PCSG 引入上下文信息后,召回率、准确率有较大幅度的提高.

我们分析了测试中 HC-PCFG 和 HC-PCSG 分析结果与 PCTB 句法树的差异,发现其中一部分在于对句法树怎样分解的不同理解.PCTB 的句法标注因为手工标注的原因还存在标注不一致的问题.如,第 298 篇文章中的“俄罗斯完成从波罗的海三国撤军工作”同一个短语在两个语句中有不同的句法树标注.如果排除上述因素,则评测成绩还有一定的提升空间.

5 结 论

数据稀疏性问题是引入分类和知识进行句法分析消歧方法中的核心问题.本文研究表明,句法分析可以看作是一个模式识别过程,而分类是模式识别的主要方法之一.句法分析过程可以看作是一种组合分类,分类的方式和粒度则可以有效地弥补语料库样本数量限制.此外,通过规则细分和信息量消歧,在提高分析准确率的同时大幅度减少了无效状态数,实际上提高了分析效率.

本文提出的 HC-PSG 则采用了层级的分类,通过必要而足够精细的分类,在保证消除歧义的同时,又能避免数据稀疏性问题.在语料库具备精细知识时能够充分利用这些精细知识,如果待分析的目标语句中的词语和语法现象在语料库中没有对应的精细知识,则可以利用尽量接近的较粗的知识.

句法分析仅仅解决了符号、词语和短语之间的组合关系,深入的自然语言理解还必须理解各个构成角色之间的关系,即语义角色标注.本课题实验中将句法分析和语义角色标注结合的方法对句法分析与自动语义角色标注的结合也提供了可行的途径.在精细化的层级分类规则集中,已经体现了丰富的语义知识信息.利用 WordNet,Hownet 等框架语义知识库,可以进一步弥补语料库数据稀疏性缺陷.引入外部知识是分类消歧方法的重要特点,而外部分类词表分身的可靠性和覆盖范围就成为句法分析性能最直接的影响因素.本工作中构造的分类词表目前主要针对 PCTB,所以收纳的词汇有限.除了手工分类的层级分类序列以外,我们认为还可以通过机器学习方法,根据词义共现组合等特征来挖据学习层级分类序列.而句法分析则能够提供语句中词语的语义相邻关系.这样,通过句法分析和分类的递进求精才能最终获得实用的句法分析系统.

致谢 感谢复旦大学黄萱菁教授、陶小鹏教授帮助审阅论文并提出具体的指导意见.感谢哈尔滨工业大学刘挺教授的指导帮助.本文实验中的分类词表参考了哈工大信息检索研究室语言技术平台共享包 v1.5.0 中的《同义词词林》的设计.本文研究部分受 EMC 中国研究院语义 Wiki 项目支持.在此,一并表示感谢.

References:

[1] Chomsky N. Syntactic Structure. Berlin: Mouton Press, 1957.

[2] Pollard CJ, Sag IA. Head-Driven Phrase Structure Grammar. Chicago: University of Chicago Press, 1994.

[3] Sleator D, Temperley D. Parsing English with a link grammar. In: Proc. of the 3rd Int'l Workshop on Parsing Technologies. 1993.

[4] Dalrymple M. Lexical functional grammar. In: Keith B, ed. Proc. of the Encyclopedia of Language and Linguistics. Elsevier, 2006.

[5] Kay M. Parsing in functional unification grammar. In: Dowty DR, Karttunen L, Zwicky AM, eds. Proc. of the Natural Language Parsing: Psychological, Computational, and Theoretical. Cambridge: Cambridge University Press, 1985. 251−278.

[6] Wang MQ, Sagae K, Mitamura T. A fast, accurate deterministic parser for Chinese. In: Proc. of the 21st Int'l Conf. on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2006. 325−432. [doi: 10.3115/1220175.1220229]

[7] Magerman D, Marcus M. Pearl: A probabilistic chart parser. In: Proc. of the 5th Conf. on European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1991. 15−20.

[8] Briscoe T, Carroll J. Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. Computational Linguistics, 1993,19(1):25−60.

[9] Zhang H, Liu Q, Bai S. Structural context-related probabilistic grammar parsing. In: Proc. of the SWCL 2002. Beijing: Beijing University, 2002 (in Chinese).

[10] Abney S. Partial parsing via finite-state cascades. Natural Language Engineering, 1995,1(1):1−8. [doi: 10.1017/S1351324997001599]

[11] Chen XH, Zhou YY, Yuan CF, Wu GS. An efficient probabilistic syntactic analysis algorithm for Chinese. Application Research of Computers, 2006,(1):141−143 (in Chinese with English abstract).

[12] Gao S, Wang XL. A study on semantic rules in Chinese sentential input system. Computer Engineering and Application, 2003,(4): 80−82 (in Chinese with English abstract).

[13] Liu T, Ma JS, Zhang HP, Li S. Subdividing verbs to improve syntactic parsing. Journal of Electronics (China), 2007,24(3): 347−352. [doi: 10.1007/s11767-005-0193-8]

[14] Klein D, Manning CD. Accurate unlexicalized parsing. In: Proc. of the ACL 2003. Stroudsburg: Association for Computational Linguistics, 2003. 423−430. [doi: 10.3115/1075096. 1075150]

[15] Collins M. Head-Driven statistical models for natural language parsing. Computational Linguistics, 2003,29(4):589−637. [doi: 10.1162/089120103322753356]

[16] Sun HL, Jurafsky D. Shallow semantic parsing of Chinese. In: Proc. of the Human Language Technology Conf. of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. Stroudsburg: Association for Computational Linguistics, 2004.

[17] WordNet. http://wordnet.princeton.edu/

[18] Mei JJ, Zhu YM, Gao YQ, Ying HX. Tongyici Cilin. Shanghai: Shanghai Dictionary Press, 1983 (in Chinese).

[19] Ding HF, Zhao TJ, Li S. Parsing Chinese text based on semantic class. In: Proc. of the 2007 Int'l Conf. on Machine Learning and Cybernetics. 2007. 3377−3380. [doi: 10.1109/ICMLC.2007.4370731]

[20] Xiong DY, Li SL, Liu Q, Lin SX, Qian YL. Parsing the Penn Chinese treebank with semantic knowledge. In: Proc. of the 2nd Int'l Joint Conf. on Natural Language Processing (IJCNLP 2005). 2005. [doi: 10.1007/11562214_7]

[21] Petrov S, Barrett L, Thibaux R, Klein D. Learning accurate, compact, and interpretable tree annotation. In: Proc. of the 21st Int'l Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2006. 433−440.

[22] Black E, Jelinek F, Lafferty J, Magerman DM, Mercer R, Roukos S. Towards history-based grammars: Using richer models for probabilistic parsing. In: Proc. of the 5th DARPA Speech and Natural Language Workshop. 1992. 134−139. [doi: 10.3115/981574. 981579]

[23] Liu T, Ma JS, Zhu HJ, Li S. Dependency parsing based on dynamic local optimization. In: Proc. of the CoNLL 2006. 2006.

[24] Black E, Abney S, Flickenger D, Gdaniec C, Grishman R, Harrison P, Hindle D, Ingria R, Jelinek F, Klavans J, Liberman M, Marcus M, Roukos S, Santorini B, Strzalkowski T. A procedure for quantitatively comparing the coverage of English. In: Proc. of the DARPA Speech and Natural Language Workshop. Stroudsburg: Association for Computational Linguistics, 1991. 306−311.

附中文参考文献:

[9] 张浩,刘群,白硕.结构上下文相关的概率句法分析.见:第一届学生计算语言学研讨会(SWCL 2002).北京:北京大学,2002.

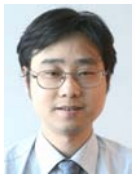
[11] 陈晓辉,周源远,袁春风,武港山.一种有效的汉语概率句法分析算法.计算机应用研究,2006,(1):141−143.

[12] 高升,王晓龙.语句级汉字输入系统中语义规则研究.计算机工程与应用,2003,(4):80−82.

[18] 梅家驹,竺一鸣,高蕴琦,殷鸿翔.同义词词林.上海:上海辞书出版社,1983.

代印唐(1971－),男,四川自贡人,博士,高级工程师,主要研究领域为语义网络,自然语言处理,知识智能.

马胜祥(1983－),男,工程师,主要研究领域为机器学习,信息获取,信息检索.



吴承荣(1971－),男,上海人,副教授,主要研究领域为信息安全.

钟亦平(1953－),女,上海人,教授,博士生导师,主要研究领域为网络安全.