计算机科学 1999Vol 26№ 2

一种新型的面向数据的语言处理技术*)

A new type of Data-Oriented Language Processing Technique

张玥杰 朱靖波 姚天顺

(东北大学信息科学与工程学院计算机系 沈阳 110006)

Abstract Data-oriented language processing technique embodies the assumption that human language perception and production works with representation of concrete past language experiences, rather than with abstract grammar rules. So in the implementation, the modle maintains large corpus of linguistic representations of previously occurring utterances. This paper presents the data-oriented language processing technique with the labeled phrase structure tree.

Keywords Data-oriented language processing, Parsing, Semantic interpretation, Combination operation, Probability computation, Disambiguation

1. 引言

在过去几年中,一种新型的语言处理技术开始出现,并以各种名称为人们所知,如"面向数据的句法分析(Data-Oriented Parsing, DOP)","基于语料库的解释",及"树库文法"等等.统称为面向数据的语言处理或 DOP 方法。该方法由 Scha[1990]提出,并由 Bod[1992-1995]发展,是一种概率的分析策略,其中体现一种假设,即人类对语言的理解与创造,依赖于以往具体的语言经验,而不是抽象的语言学规则. 因此,在实现这种方法的模型中,保留以往出现言语语言学表示的大语料库。当处理一个新的输入言语时,通过组合来自语料库的片段构造该言语的分析。其中片段的出现频率用于估计最可能的分析。

本文将对基于 DOP 的语言处理技术作较为详细的介绍,希望通过这些介绍,使我国进行汉语及外国语言处理的研究者有所借鉴.

2. 面向数据的语言处理

人们过去的语言经验在某种程度上决定他或她的语言分析结果.面向数据方法的基本观点是:这实质上以非传递的方式发生。通常使用具有言语及其分析的大语料库,作为人们过去语言经验的一种表示,分析新输入言语的过程,相当于出现在语料库中的分析片段的构造过程。通过考虑这些分析片段的出现频率,可以判断最可能的构造方式。

按照 Bod[1995],面向数据的语言处理框架可通过四个组件描述:言语分析的形式化描述;言语分析的片段,这些片段可作为组成一个新言语分析的单元;用于组合片段的运算;基于片段在语料库中出现的频率,为一个新言语计算其一种分析概率的方法。因此,DOP 框架允许广泛范围的各种例证,假设可将人类语言处理模型化为一种概率过程,并在过去语言经验表示的语料库中进行运算。但对于语料库中言语分析的表示,这些言语分析结构在处理新输入中所起的作用,及概率计算,将允许存在开放性。

本文描述基于带标短语结构树的面向数据的语言处理模型。实际上,许多模型也采用了带标短语结构树,但是从语料库中抽取子树与概率计算的标准不同,其它一些模型则使用更为丰富的形式来完成语料库标注.

3. 面向数据的句法分析

这里以 Bod[1992,1993]中所介绍的第一个 DOP 模型为例,说明面向数据的句法分析的实现过程,如图 1 所示。

*)本文受国家自然科学基金和国家教委博士点专项基金资助。张玥杰 博士研究生。朱靖波 博士研究生。姚天顺 教授,博士导师。

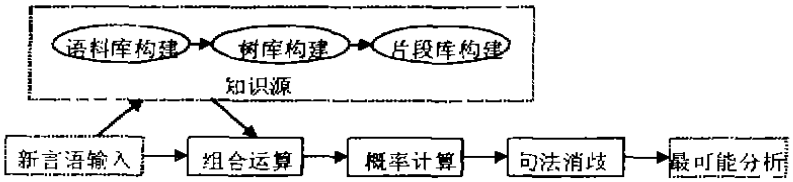

图 1　面向数据的句法分析的实现过程

3.1　树库构建

为选择一种能够提供最适合于语言描述使用者的"句法结构经验"的表示方式，结合现代语言学理论的表示似乎是比较合理的。这里采用简单的描述系统，将言语分析编码为带标树，其中的标记是基本符号。当然这种标注是具有局限性的，不仅不能表达构成的意义，而且忽略了末与表层结构相一致的"深层"或"功能"句法结构，同时也不允许句法特征描述为格、数或性。但它具有两大优势：一是非常简单，二是对于已标注语料库来说，可读性比较好。图 2 即为包含两棵树的树库。

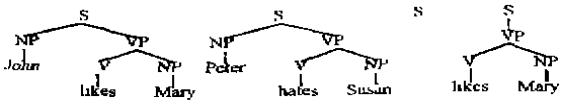

图 2　包含两棵树的树库

图 3　非法子树

3.2　片段库构建

构成片段库的片段相当于子树，在处理中作为分析单元使用。一颗树 T 的子树 t 相当于 T 的一个子图 t，具有如下特征：①t 至少包括一个以上的节点。②t 是连通的。③除了叶节点之外，t 中的任何一个节点必须保持与 T 中相应节点具有相同的儿子节点。

给定一个树库，就可以根据每一棵句法树的构成，从中抽取所有合法子树，构造一个子树包，即片段库。这就是所谓的"分解"操作，即片段生成操作，图 2 中树库的所有合法子树为 34 棵；而图 3 所示的片段为非法子树。

3.3　组合运算

在面向数据的句法分析中，组合运算是一种非常重要的操作，也称之为左侧优先替换，其运算对象为带标树的集合。树 t 与树 u 的组合运算，记为 t∘u，其含义是树 u 的根节点中的标记等于树 t 的最左非终结符叶节点中的标记。如果 t∘u 运算成功，则表明树 u 的副本将替换树 t 的最左非终结符叶节点。这种替换操作，可保证组合运算的唯一性。

以 3 个片段为例说明组合运算的实现，如图 4 所示。在次序上，可将$(t\circ u)\circ v$写作$t\circ u\circ v$。注意该组合运算不符合结合律，即$(t\circ u)\circ v\neq t\circ(u\circ v)$，而是$(t\circ u)\circ v=t\circ u\circ v$。

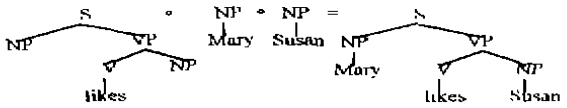

图 4　以三个片段为例说明的组合运算

给定一个片段集合 FC1＝{t1，t2，…，tn}，其组合运算 t1∘t2∘…∘tn，产生树 T，称作为树 T 的一种推导。给定一个片段集合 FC，可以利用集合中各成员，通过多次组合运算生成不包括非终结符叶结点的树集合，称之为由 FC 生成的树语言。这些树所产生的串集合，称之为由 FC 所生成的串语言。

3.4　概率计算

树库可被看作为一套语法，如将串集合定义作为句法分析，这套语法实际上是一套概率语法。如果重新定义句法树的生成过程作为一个随机过程，则该过程考虑了片段库中片段的概率分布。在该过程中，首先以随机方式从具有不同根节点标记（如 S）的子树集合中选择一棵子树，然后继续以随机方式选择一棵能与其进行组合运算的子树，进行组合运算，重复这一步工作直至获取不包括非终结符叶节点的结果树。对于每一棵树及每一个串，都赋予概率权值，该概率权值由这个随机过程产生。

该计算过程不是将树库当作实例来估计随机模型的参数值，而是将由树库所生成的片段库中的子树直接用作为一个随机生成系统，新的输入获得该系统生成的最有可能的分析。

以图 2 中的树库所生成的片段库及图 4 中的组合运算为例，来描述概率计算过程。通过组合运算，利用树库中的部分子树，可以构造图 4 中新输入句子的分析过程，这个推导过程包括了三个随机事件的概率：

- 从根节点为 S 的所有子树中抽取子树 s[NP，vp[v[likes]，NP]]。
- 从根结点为 NP 的所有子树中抽取子树$_{NP}$

[Mary]。

● 从根结点为 NP 的所有子树中抽取子树$_{NP}$[Susan]。

事先假设这些事件都是随机独立的，这个推导的概率计算为：

P(t=s[NP,vp[v[likes],NP]]|root(t)=S) * P(t=$_{NP}$[Mary]|root(t)=NP) * P(t=$_{NP}$[Susan]|root(t)=NP)

假设从语料库中抽取子树是一个随机过程，因而条件概率的计算如下：

P(t=s[NP,vp[v[likes],NP]]|root(t)=S)=#(s[NP,vp[v[likes],NP]])/#(t|root(t)=S)

其中#(x)表示类型为 x 的子树出现的个数，因此图 4 中的概率计算为：

#(s[NP,vp[likes],NP]])/#(t|root(t)=S) * #($_{NP}$[Mary])/#(t|root(t)=NP) * #($_{NP}$[Susan])/#(t|root(t)=NP)=1/20 * 1/4 * 1/4=1/320。

由此得出，推导 $t_1 \circ t_2 \circ \cdots \circ t_n$ 的概率计算为：

$$P(t_1 \circ t_2 \circ \cdots t_n)=\prod_i \#(t_i)/\#(t|root(t)=root(t_i))$$

一棵分析树的概率计算需要考虑其所有推导，例如，图 4 中的分析树还存在如下两种推导，如图 5 与图 6 所示。

由此可见，采用片段库中不同的片段，一棵分析树可能存在许多不同的推导。每一种推导都具有其自身概率，计算过程与上面图 4 中分析树的概率计算相同。

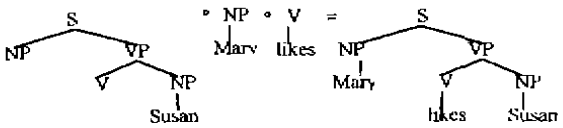

图 5　图 4 的一个不同推导

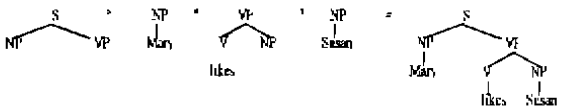

图 6　图 4 的另外一个不同推导

一棵分析树的概率应该等于其任何一种推导生成树的概率。因此，一棵分析树的概率等于其所有推导的概率之和。如果一棵分析树具有 k 种不同推导：$(t_{11} \circ \cdots t_{1j} \circ \cdots)$，$(t_{21} \circ \cdots t_{2j} \circ \cdots)$，…，$(t_{k1} \circ \cdots t_{kj} \circ \cdots)$，则其概率为：

$$\sum_i \prod_j \#(t_{ij})/\#(t|root(t)=root(t_{ij}))$$

因为许多句子具有歧义性，即它们具有一些不同的句法分析树。因此，生成一个句子的概率为它的所有句法分析概率之和。

3.5　句法消歧

句法消歧实际上是计算一个输入言语的最可能句法分析树的过程，首先通过上面的处理，为输入句子生成一个句法分析森林；之后通过消歧，从森林中选择最可能的句法分析。

3.5.1　句法分析森林构建。为一个输入句子构建一个句法分析森林的算法，可从针对上下文无关文法的算法中获得。该方法利用一个图表或合式的子串表，采用上下文无关重书写规则的集合及一个句子作为输入，并产生一个带标短语的图表作为输出。类似图表的句法分析森林，可通过从一种类型到其它类型的指针获得。

在形式上，面向数据的语言处理模型可以看作为一种随机树替换文法(Stochastic Tree-Substitution Grammars，STSG)。同时，如果将 STSG 的每一棵基本树看作为一个上下文无关重书写规则，则可将图表句法分析方法应用于利用 STSG 的句法分析。因此，下面以利用 STSG 的句法分析为例说明句法分析森林的构建过程。为获得一个输入句子的句法分析森林，不仅利用各个短语的句法类型进行标注，而且利用其完整的基本树进行标注，在所生成的类似图表的森林中，生成相同树的不同推导并不冲突。

图 7 给出 STSG 的一个样本，图中各树为其基本树。采用[Kay，1980]中的描述形式表示，即图表中的每一条目(i，j)，由跨越句子第 i 个与第 j 个单词的一条边表示。利用已连接的基本树标记每一条边，这些基本树构成句子的子推导。

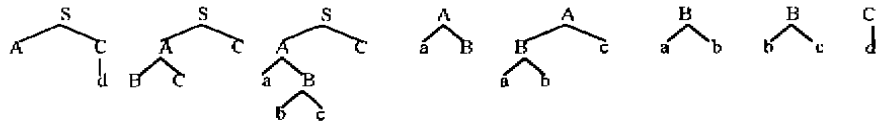

图 7　一个 STSG 样本的基本树集合

对于字符串 abcd，具有如图 8 所示的推导森林。森林中的一些推导生成相同的树，通过彻底分解森林，获得四种不同推导，它们可生成两棵相同的树，即这两棵树通过不同推导（可能具有不同概率）分别生成两次。

3.5.2 消歧。一个输入句子可能具有指数增长的大量句法分析，并且这些句法分析可具有指数增长的大量推导。因此，为找到最可能的句法分析，通过彻底分解图表，并不是一种有效途径。即使为决定一个句法分析的概率，加入该句法分析的所有推导的概率也不是有效的。Bod 提出基于蒙特卡罗算法的消歧策略，该方法的基本观点是，通过随机抽取推导样本来估计最可能的句法分析。

采用随机优先搜索，可从推导森林中生成一个随机推导，该推导提供基于子推导概率的随机选择。通过重复生成大量随机推导，可估计最可能的句法分析作为由这些随机推导最为频繁产生的句法分析。随机样本数目越大，则估计最可能句法分析的精确率越高，根据大数定律，最频繁生成的句法分析，收敛于最可能的句法分析。蒙特卡罗方法即通过随机抽取样本，来估计一个事件的概率。

实现中，以自底向上的方式完成随机推导的选择。在图表中每一个共享节点处选择一个随机子推导，一旦到达 S 节点，可通过考虑在每一共享结点处所做的选择，来修正整个句子的随机推导。延迟样本抽取直至 S 节点，以便直接从所有 S 推导的分布中抽取样本。这将花费指数时间，因为对于一个句子，存在指数增长的大量推导。通过在歧义出现的每一节点处自底向上抽取样本，将每一共享节点处的不同子推导的最大数目约束为一个常量。因此，输入句子的一个随机推导的时间复杂度，等于找到最可能推导的时间复杂度。

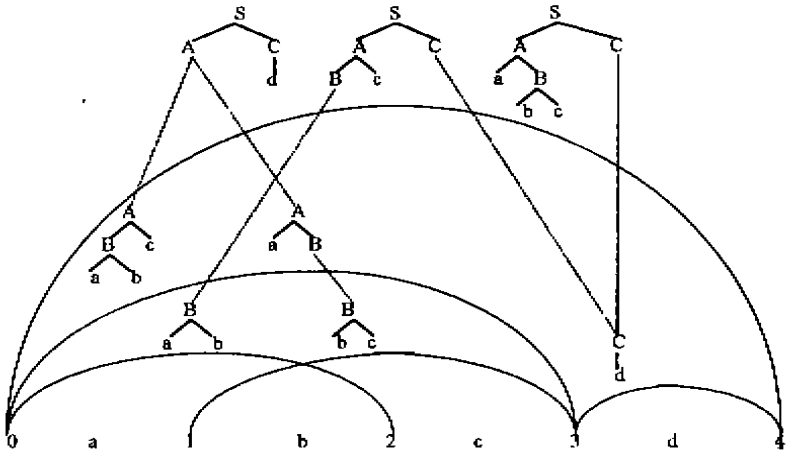

图 8 字符串 abcd 的推导森林

4. 面向数据的语义解释

类似于面向数据的句法分析的处理过程，面向数据的语义解释需要四个步骤：①确定一种形式化描述，用于表达句子意义和表层结构。②采用语义表示来标注句子语料库及其表层构成。③建立一种方法，根据语料库中任意子树的抽取及其组合运算，来推导意义表示。④概率权值的计算。下面以 Bod [1996]所提出的面向数据的语义解释方法为例说明其实现过程。

4.1 语义形式化描述

实际上，语义的形式化描述具有很大的任意性。如果能够定义一个比较好的模型理论，则在该领域内足以表达句子意义和相关表层即可。一种比较著名的标准形式化描述为：扩展类型理论。该理论是一种高度有序的逻辑语言，它组合抽象的希腊字母：连接词和量词。

4.2 语义标注

可以在以前句法标注的树库基础上进行语义标注，该标注树已经描述表层组成结构。可以对每一个具有意义的句法结点增加一个逻辑类型符号，表示相应的表层构成的意义。对于图 9 中的两棵句法分析树，其语义标注如图 10 所示。

树库中句法分析树的语义标注分为以下两个方

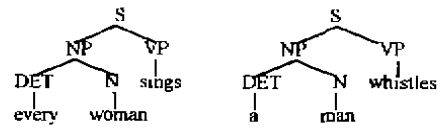

图 9 "Every woman sings"与"A man whistles"的句法标注

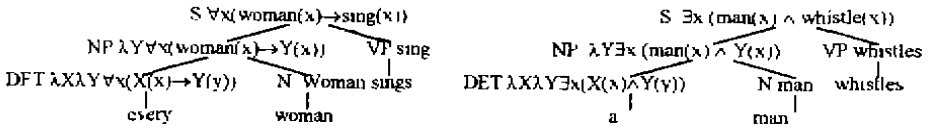

图 10 "Every woman sings"与"A man whistles"的语义标注

面：一是对于每一个具有意义的词汇节点，利用一个类型逻辑公式进行标注，以表示其意义，二是对于每一个具有意义的非词汇节点，利用一个公式模式进行标注，并且该模式可以通过子女节点的公式组合来表达词汇节点的意义。

在标注过程中，模式中的变量 d1 描述最左侧子女构成的意义，d2 描述第二个子女构成的意义，依此类推。通过这种标注，图 10 中的语义标注可转换为如图 11 中的形式。

很明显，这种标注假设表层构成的意义表达可以通过对其子构成的意义表达组合而成，这种假设是没有问题的。

4.3 子树的意义及其组合运算

类似于面向数据的句法分析，仍然可以考虑用所有推导来计算语义分析的概率。推导过程相当于从语料库中抽取子树，采用组合运算产生分析树。但是与句法分析相比较，在分解子树操作与子树组合运算两个方面有所不同。如果从一棵树中抽取一棵子树，必须利用一个相同类型的统一变量来代替新叶节点的语义。相应地，当将一棵子树与该节点组合时，必须利用用于替换子树的模式来描述该变量。这就需要该模式的语义类型与统一变量的语义类型相匹配。

以图 11 中"A man whistles"的语义标注为例可清楚阐述实现过程，如图 12 所示。

在图 12 中，一棵树被分解为两棵子树，断点 N：man 的语义由一个变量替代。图 12 中例句的分析生成过程，如图 13 所示。

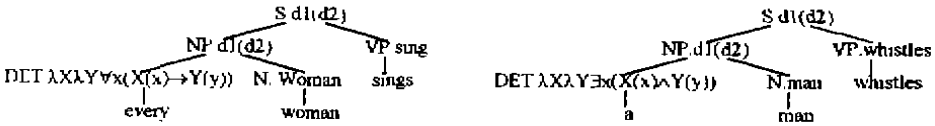

图 11 与图 10 中语义标注相对应的转换形式

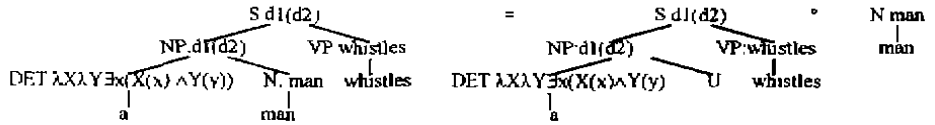

图 12 利用统一变量将一棵树分解为子树

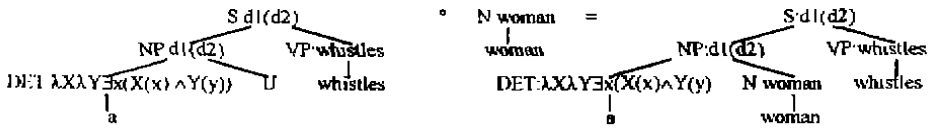

图 13 生成句子"A woman whistles"的语义分析

（下转第 77 页）

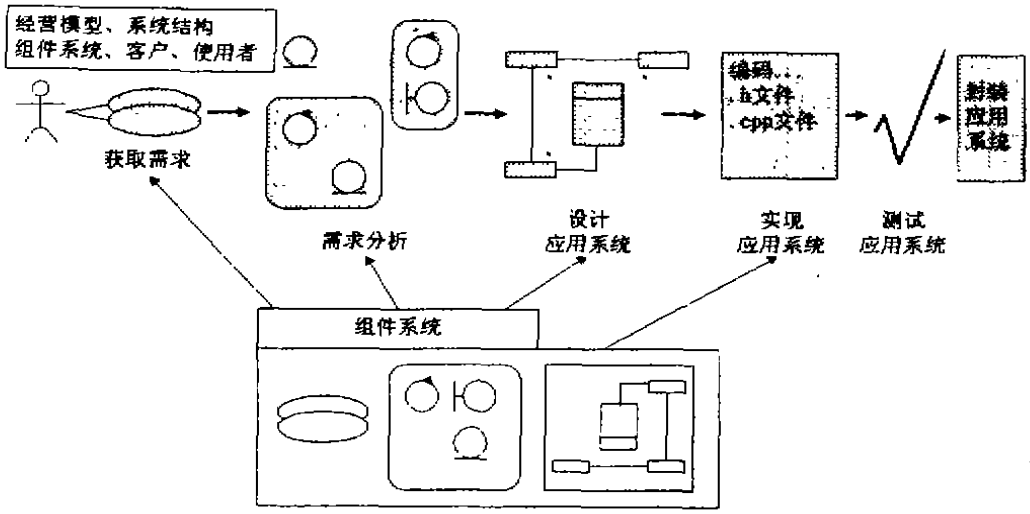

图 4　应用系统工程经营过程步骤

在获取需求阶段可以采用 Use Case 方法[5]、从用户的直接要求出发建立系统的需求模型；在分析与设计阶段可以采用 Booch 方法[6]，由类图、场景图等组成系统设计的分析模型。采用这些面向对象技术是软件系统工程经营过程方法的基础。

结论　柔性软件系统是未来计算机软件发展的必然趋势，其中综合并系统化了当前软件技术中的诸多好的方法和思路，因此具有广阔的研究前景。同时计算机硬件技术的飞速发展使得对应用软件的需求不断增加，尤其我国目前正处于计划经济向市场经济转轨的过程中，涉及到管理体制、经营过程、机构设置、人员调动等多方面根本性的变化，开发更灵活、更具柔性的应用软件系统无疑将更好地适应这种需求。当然柔性软件系统在实际应用中存在高费用、高风险与高收益并存的现象，但令用户长远意义上的获益匪浅将产生广阔的应用前景。

参 考 文 献

1　Bradshaw J M. Software Agents. First edition, New York: The MIT Press, 1997

2　Jacobson I, et al. Software Reuse. The ACM Press, 1997

3　Scholz-Reiter B、Stichel E. Business Process Modeling. Springer, 1996

4　范玉顺，等．"制造业 CIMS 应用集成平台总体设计与原型系统开发"项目总体设计报告：[技术报告] 北京，1997

5　Jacobson I. Object-Oriented Software Engineering: a Use Case Driven Approach. Addison-Wesley Pub, 1992

6　Booch G. Object-Oriented Analysis and Design with Applications. Second edition, Addison-Wesley Pub, 1994

（上接第 61 页）

4.4　面向数据的语义解释的统计模型

给定一个已标注树库，就可以定义一个输入字符串语义解释的概率。正如上面描述的分解过程，树库中的一棵树可以分解成包含所有子树的多个集合。利用一棵子树替换左侧类型为 C 的叶节点，其概率即为在多个集合中，从所有类型为 C 的子树中选择该子树的概率。该概率值等于这棵子树的出现次数与所有类型为 C 的子树数目的比率。

一个串的推导是一组子树、它们的组合产生一棵树，生成这个串。一个推导的概率由这些子树的替换概率产生。一个串的推导产生的树称为该串的一个分析。一个分析概率相当于其任何一个推导产生结果的概率，等于其所有推导概率之和。

一个串的解释是一个模式，相当于这个串分析（树）的根节点的语义标注。一个具有解释 I 的串 S 的概率，等于该串 S 根节点被标注为 I 模式的所有分析概率之和。

（参考文献共 18 篇，略）