

自动文摘系统中基于全信息词典的复杂语句分析方法及其实现

李 蕾, 钟义信

(北京邮电大学信息工程系 181 # 信箱 B986 班, 北京 100876)

摘 要: 本文介绍了智能型自动文摘系统 Ladies 中汉语复杂语句的化简分析方法——义块组配及其实现. 在使用全信息词典对词的语法、语义、语用信息进行全方位描述的基础上, 义块组配对复杂句进行语义聚合, 将被扩展的各个成分合并成能充分表达语法、语义信息的义块, 从而大大简化了语法语义分析过程, 为进一步的句子语义、语用分析和篇章理解打下了一个良好的基础.

关键词: 自动文摘; 语法语义分析; 全信息词典; 义块组配

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2000) 08-0104-03

CL Based Algorithm Simplifying Analysis of Complex Chinese Sentences

LI Lei, ZHONG Yi-xin

(Department of Information Engineering, Beijing University of Posts and Telecommunications, Box 181, Beijing 100876, China)

Abstract: We present Semantic Chunk Assembling algorithm (SCA) for intelligent abstract system Ladies to simplify the analysis of complex Chinese sentences based on Comprehensive Information Lexicon (CIL) in this paper. CIL describes comprehensive information of words and SCA organizes words of complex sentences into semantic chunks according to the assembling rules. The experimental results indicate that it is quick and effective.

Key words: automatic abstract; grammatical and semantic analysis; CL; semantic chunk assembling

1 引言

Internet 已迅速成长为一个巨大而丰富的信息源, 自动文摘系统则是从中尽快获取所需信息的有效工具. 本文采用基于全信息词典的语句分析模型, 实现了一个面向“神经网络学习算法”领域的智能型文摘系统 Ladies. 它的处理过程主要有三个阶段: 分词和语法语义类标注、基于全信息词典的语法语义分析和文摘生成、润色. 其中语法语义分析是连接分词与文摘理解生成的关键环节, 为了解决科技文献中大量出现的复杂语句的语法语义分析困难, 设计并实现了义块组配方法.

目前国内外的文摘系统以统计抽取为主, 不涉及语句分析; 而受限领域理解式文摘系统中同时实现复杂语句语法语义分析的几乎没有. 国内已经有类似概念层次网络^[6]之类的研究, 但是尚未见到其在自动文摘中的应用报道, 义块组配是这方面的初次尝试.

2 复杂语句分析化简思想的形成

从汉语语法角度来看, 通常用并列或连接的方法把基本语句模式中的词扩展到复杂句中的扩展成分, 如图 1 所示.

使用并列方法可以组成联合、同位等结构短语; 使用连接方法可以组成主谓、偏正等结构短语. 复杂句中的扩展成分有的就是一个短语结构, 有的可能是多种短语结构的再并列或

再连接, 从而形成各种结构多层嵌套的复杂结构. 如果让计算机直接从语法层次上来分析这样的复杂句, 会有以下三个突出的问题导致可能陷入非多项式可解的困境:

- 短语中词的兼类, 尤其是动名词兼类现象严重;
- 短语结构复杂, 层层嵌套关系难以分清;
- 句中有多个动词, 确定主动词并非易事.

目前汉语的深入语法分析系统还没有达到实用的层次, 并且这些问题也不可能单纯依靠语法知识来解决, 必须



图 1 复杂语句的扩展成分分析

进一步站在语义层次上, 把语法和语义知识结合起来才能解决. 从语义角度来看, 无论扩展成分为何种复杂的语法结构, 它总是语句中一个相对独立的意义实体, 我们称之为“义块”. 义块可以看作是带有语义信息的短语, 主要通过词与词之间的语义关系界定, 当然也要满足一定的语法关系. 每个义块语义上相对完整, 在句中充当一个语法角色, 义块之间既有语法规约关系, 也有语义呼应关系.

我们的思路就是通过义块组配完成复杂句中各个扩展成分义块的语义聚合, 用义块把上面完全语法分析的难题屏蔽

掉,从而简化分析过程。在 Ladies 完成后才有机会阅读了黄曾阳先生的 HNC 理论。其中把语义块作为句子分析的基本单位,虽然具体实现与我们不同,但其出发点与我们是完全相同的。可见语义分析是必然的趋势。

义块的语法和语义信息是由组成它的各个词的信息共同决定的,下一节将介绍全信息词典对词的语法和语义信息的描述。

3 全信息词典

“全信息”是词的语法、语义和语用信息的总称^[1]。全信息词典把三种信息有机地组织在一起,对词语进行全方位的描述。语法信息涉及词的语法功能,语义信息描述词的语义内涵。与概念层次网络^[6]相比,全信息词典里最小意义单位是词,而且词的各层次信息除了常识性定义之外,还能够针对某一特定领域进行扩展和细化,因此为实用、灵活和具有针对性。我们采用语法和语义分类相结合的方法,上层按语法分类,下层按语义分类,允许低层次的概念继承它们祖先的特征。具体的语法分类参照国家标准 GB13715 定义,共 13 种。

语义分类主要是针对神经网络算法类文章所包含的文摘信息制订的,如名词又可细分为算法类名词、技术类名词等。这样的语义类别共有 76 种。对于其它词汇都视为普通语义词。语用信息反映的是词语对文摘目标的效用,它直接引导文摘信息的提取。义块组配主要处理语法语义分类,本文限于篇幅,对语用信息不再详述。

Ladies 的全信息词典对 4 万 8 千余条词的语法语义类别都作了标注,义块组配就是在这些标注的基础上进行的。

4 义块组配

4.1 义块的语法语义分类

为了表示义块的语法语义信息,必须对其进行分类描述。一个义块无论其结构多么复杂,它在句中的作用都跟一个词相当,所以我们采用的义块分类体系和词相同。

4.2 义块组配规则及其形式化描述

义块组配规则是一套完整的语义聚合规则。在词的语法语义分类基础上,义块组配规则对句子进行分析组配,最终将复杂句的扩展成分从义元序列聚合成具有语法功能的义块,并对义块进行语法语义分类和语用规则号标记。规则中有普适规则和特殊规则。普适规则描述的是汉语中普遍适用规则,如助词粘附于动词、常用搭配等等。特殊规则是针对被处理领域的语料特点和领域概念进行的扩展。但考虑到领域移植的要求,对所有规则都进行了形式化描述,规则的统一形式语言如下:

RULE::=No./K/p₁+p₂+...+p_n=r(l₁%l₂%...%l_n)/F

其中, No. :当前规则记录的序列号;

/ :分隔符; K 表示操作类别;

p_n :匹配入口项 n(可以为汉字或语法语义类);

+ :匹配入口项分隔符;

= :结果项提示符;

r :结果项(一般为语法语义类);

() :限制项标识符;

l_n :限制项 n(可以为汉字或语法语义类);

% :限制项分隔符;

F :有无下一条规则记录

用户可按照自己的需要任意添加修改规则。扩展领域时只需要重做分类和相关规则。目前系统中的形式化规则总数为 120 多条。

4.3 义块规则的具体制订

常用搭配是一类普适规则,可以实现时间、处所、方位、目的、对象等扩展状语义块的组配。汉语在使用常用搭配上重复频率比较高,我们的规则就是把这些搭配都合并起来,作为一个状语义块存在,使句子的语法语义更加清晰。这样做还可以把一些难以处理的语言现象,如词的兼类在一个更高的层次上屏蔽掉。让我们看下面的例子:

在 ROTH0 我们 NSUG0 对 ROTH0BP 算法 NALG0 的 ZOTH0 初步 JOTH0 研究 VDIS20 时 NOTH0,BCOM0 采用 VWAY11 的 ZOTH0 就是 VSCR18 将 VOTH0 单一 NOTH0 的 ZOTH0 相对 NTECO 形式 NOTH0 误差函数 NTECO 作为 VOTH0BP 算法 NALG0 的 ZOTH0 误差 NOTH0 公式 NOTH0,BCOM0 计算机 NTECO 仿真 NRES0 的 ZOTH0 实践 NOTH0 也 AOTH0 证实 VVER9 了 ZOTH0 这一 POBJ0 点 NOTH0.BPER0

它的第一个分句若从词的语法语义分类来看是:ROTH \$ NSUG \$ROTH \$NALG \$ZOTH \$JOTH \$VDIS \$NOTH \$BCOM 即:介词\$提示类名词\$介词\$算法类名词\$助词\$形容词\$讨论类动词\$普通名词\$逗号,可谓种类繁多。若从语法结构来看,将相邻的两个或多个词组合起来都不能正确判断其语义。例如,“ROTH + NSUG”即“在我们”和“NALG + ZOTH + JOTH”即“算法的初步”根本无实在意义。再如“JOTH + VDIS”即“初步研究”,一种可能是形容词修饰动词,另一种可能是动词属动名词兼类,此处为形容词修饰名词。可见,这样的分析是很难得出正确结果的。但如果使用“在……时”的常用搭配规则,很快就能把整个分句准确地组配成为一个状语义块,这样所有的问题都迎刃而解了,可得到组配结果:

在对 BP 算法的初步研究时 AOTH0,BCOM0 采用的就是 NOTH0 将 VOTH0 单一的相对形式误差函数 NTECO 作为 VOTH0BP 算法的误差公式 NOTH0,BCOM0 计算机仿真的实践 NOTH0 也证实了 VVER9 这一点 NOTH0.BPER0

显然,由于使用了“在……时”组配规则,把原本极为复杂的一个分句聚合成一个状语义块,大大简化了对于该句的分析,动词“研究”活用做名词的兼类现象也处理了,不会对后续分析形成干扰。

后缀类名词的组配是一类特殊规则。科技文献中大量出现组合型的专业词汇,其中很多是我们所需要的文摘信息,但不可能把这些词汇都添加到词库中去。为此,设计了多种后缀类名词、技术类名词和后缀类名词组配规则。例如,为了将文中的算法名称义块分析出来,制定了算法类后缀名词及其组配规则,该规则与相关规则联合作用,对“算法/方法”这类后缀义块和技术类名词义块敏感,可以把算法名称分析出来。例如,在领域语料中,“方法”和“学习方法”皆为算法类后缀名

词,当“方法”前有“模糊分类”、“规则”、“学习方法”前有“有导师”、“共振”之类的界定时,它们都是技术类名词,所以都可以合成算法名词义块,再应用名词序列规则、数量义块规则和破折号规则,就可以合成算法名称义块.因此对下面的句子

本文 NSUG 提出 VPUTI 了 ZOTH0 - UOTH0 种 QOTH0 学习 NTECO 模糊分类 NTECO 规则 NTECO 的 ZOTH0 方法 NAPO0 - - BBRO0 有导师 NTECO 共振 NTECO 竞争 NOTH0 学习方法 NAPO 0 (BYZU0SRCL)BYOO. BPERO

组配结果为:

本文 NSUG 提出了 VPUTI 一种学习模糊分类规则的方法 - - 有导师共振竞争学习方法 NALG. BPERO

只从词的语法语义分类来看是: NSUG \$VPUT \$NALG \$BPER, 即: 提示类名词 \$ 提出类动词 \$ 算法类名词 \$ 句号. 可见, 组配后的句子成了非常典型的主谓宾结构的简单句, 易于分析. 句中的宾语就是所需要的算法名称, 提取非常方便.

“的”字规则也是一类特殊规则. 汉语中“的”字只表示语法意义, 不表示词汇意义, 但其结合面宽, 具有使谓词性成分体词化的功能.“的”字短语常修饰名词或代替名词. 在科技文献的复杂句中由于修饰、限制关系而存在的“的”字更是普遍, 因此必须准确处理. 为此, 我们制定了有关“的”字的一套规则, 描述如下:

sentence:: 当前处理的一句话;

length:: 该句中当前的义块数

word(n):: 第 n 个义块, 可以为词;

word(n) → character:: 第 n 个义块的汉字

word(n) → type:: 第 n 个义块的语法语义类别;

word(n) → number:: 第 n 个义块的语用规则号

sentence(length):: = word(1) | word(2) | | word(n) | |

word(length)

step1: 逆向扫描 sentence(length), 如果找不到, 返回; 如果找到, 将其位置记为 pos;

step2: 若 word(pos + 1) → type 为标点, 则合并 word(pos - 1) → character + “的”, 令 word(pos - 1) → type 为 JOTH;

step3: 若 word(pos + 1) → type 为动词, 令 word(pos + 1) → type 为 NOTH, word(pos + 1) → number 为 0;

step4: 若 word(pos - 1) → type 为动词, 则合并 word(pos - 1) → character + “的” + word(pos + 1) → character, 令 word(pos - 1) → type 为 word(pos + 1) → type, word(pos - 1) → number 为 word(pos + 1) → number;

step5: 若 word(pos - 1) → type 不为动词, 向前扫描, 直到找到“的/了”或标点或动词, 记此时的位置为 stop, 合并 word(stop + 1) → character + word(stop + 2) → character + + “的” + word(pos + 1) → character, 并检查其中是否有 NTEC 类, 若有令 flag 为 1, 否则为 0;

step6: 若 word(pos + 1) → type 为 NOTH/ 后缀名词且 flag 为 1, 令 word(stop + 1) → type 为 NTEC/ 对应的语义类名词, 若 flag 为 0, 令 word(stop + 1) → type 为 word(pos + 1) → type;

step7: 若 word(pos + 1) → type 为其它词, 令 word(stop + 1) → type 为 word(pos + 1) → type, 调整指针, 转 step1 继续向前扫

描下一个“的”字.

使用“的”字规则不仅会使复杂句的语法语义结构骤然清晰, 而且对于解决多个动词等问题也是很有有效的. 例如:

充分 AOTH0 利用 VWAY 11 了 ZOTH0 专家 NOTH0 给出 VPUTI 的 ZOTH0 样本 NTECO 分类信息 NTECO, BCOM0 使 VOTH0 学到 VOTH0 的 ZOTH0 模糊分类 NTECO 规则 NTECO 更加 AOTH0 合理 JOTH0. BPERO

句中每个分句都有两个动词(带下划线的词), 确定主动词并非易事, 需要考虑极其复杂的情况, 测试多种搭配环境. 尤其是第一个分句中“利用了”和“给出的”皆为动词后接助词, 更难分辨. 使用“的”字规则后, step4 就会把“给出的样本”和“学到的模糊分类”合并成为技术类名词义块, 再与名词序列规则连用便可得到下面的组配结果: 充分利用了 VWAY11 专家给出的样本分类信息 NTECO, BCOM0 使 VOTH0 学到的模糊分类规则 NTECO 更加合理 JOTH0. BPERO

此时句子的语法语义已经变得非常清楚, 排除了限制性动词“给出”、“学到”对主动词“利用”、“使”的干扰.

5 计算机实现

在 Ladies 系统中, 所有的义块组配规则都用程序来实现, 并特别处理了各种短语结构的多层嵌套问题, 取得了良好的结果.

6 结束语

目前, 复杂语句的句法语义分析是实用系统的瓶颈之一. 通过义块组配来跳出繁杂的词与词之间的关系, 站在义块的高度来看句子, 从而能得到句子的清晰结构和意义. 该方法已经向进一步的语义分析和语用分析, 甚至篇章理解迈开了坚实的一大步. 除了自动文摘系统之外, 这种方法还可以用于其他的汉语处理系统中. 今后的研究一方面是继续完善组配规则, 另一方面是深入研究更深层次的语义和语用分析和表示方法.

作者简介:



李 蕾 北京邮电大学信息工程系在读博士研究生, 主要研究方向为自动文摘、自然语言处理.



钟义信 北京邮电大学副校长, 博士生导师, 长期从事信息科学、人工智能和神经网络的教学与研究, 已在国内外发表多篇著作和论文.

(下转 109 页)

i 在 t 处若为上升沿,则 $X(t) = 1$;若为下降沿,则 $X(t) = -1$;否则, $X(t) = 0$;

ii $(XY) = X \cdot Y + Y \cdot X = X \cdot Y + Y \cdot X$;

iii $(X + Y) = X + Y - 2(XY)$,

其中 X^- 表示 $X(t)$, Y^- 表示 $Y(t^-)$.

定理 5.2 波形函数 $Y = F(T_{s_{11}} X_1, T_{s_{12}} X_1, \dots; T_{s_{21}} X_2, \dots; \dots)$ 在各原始输入波形中在 x_0 处仅有 $|X_i|_{x_0} = 0$, 设

$$F(\dots, T_{s_{1i}} X_i(x, t), \dots) = f + T_{s_{1i}} X_i g \quad (7)$$

其中 f, g 中均不含因子 $T_{s_{1i}} X_i$, 若存在区间 $T = (x_0 + s_{1i}^-, x_0 + s_{1i}^+)$, $s_{1i} > 0$, 使得在此区间内,

$$f = 0, \quad g = 0, \quad g = 1 \quad (8)$$

则通路 $X_i(x_0^-, x_0^+)$ 称为可单敏化的真通路.

例 考虑如下电路,其中时滞标记在门符号内(门内的 + 表示或)

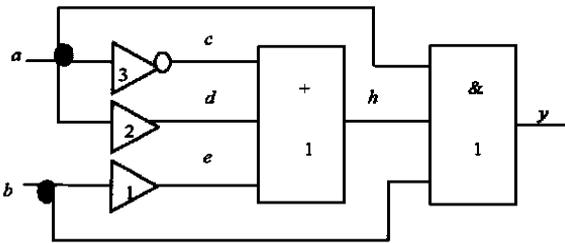


图 2

很容易得出

$$y = T_1 a T_1 b (1 + T_5 a (T_4 a + 1 + T_3 b + T_3 b T_4 a)) = T_1 a T_1 b + T_5 a T_1 a T_1 b (T_4 a + 1 + T_3 b + T_3 b T_4 a)$$

所以只要令 $g = 1$ 即可,即

$$T_1 a T_1 b (T_4 a + 1 + T_3 b + T_3 b T_4 a) = 1$$

显然,我们取 $a = w_2, b = w_3$ 时 $y = 1$

即通路 $T_5 a$ 可以敏化,这和文[2]是一致的.

6 结论

本文定义了波形空间,并在波形空间中定义了距离,极限,运用布尔运算推导出了波形空间是 Banach 空间等许多好的性质,好的结论.从而从数学角度将布尔过程这一理论进一步严格化,完善化.定义在此理论基础上的延迟算子与实际相

符,波形的极限和差分的定义反映了电路中的通路敏化,说明了布尔过程论有广泛的应用前景.

感谢闵应骅教授对本文的指导以及研讨班各位老师和同学的帮助,在此表示谢意.

参考文献:

- [1] Min Y. Boolean process—An analytical approach to circuit representation [A]. Proc IEEE Third Asian Test Symposium Japan, 1994, 249 - 254.
- [2] 闵应骅,李忠诚,赵著行. boole 过程论 [J]. 中国科学(E辑), 1996, 26(6) :541 - 548.
- [3] 赵宇虹,李忠诚,闵应骅. 带时间参数布尔函数的符号表示及其在计算电路延迟中的应用 [J]. 计算机学报, 1997, 20(10) ,908 - 917.
- [4] 赵著行,闵应骅,李忠诚. 布尔过程对路径敏化的应用 [J]. 计算机学报, 1996, 19(8) ,568 - 575.
- [5] Z. Zhao, Z. Li, and Y. Min. Waveform Polynomial Manipulation Using BDDs [A]. IEEE Fifth Asian Test Symposium, hsinchu, Taiwan, Nov., 1996:136 - 276.

作者简介:



尤志强 硕士,1972 年生,1995 年获湖南大学应用数学系学士学位,现为湖南大学计算机系研究生,主要研究方向是容错计算.



张大方 1959 年生,博士,教授,湖南大学计算机科学系主任.主持国家自然科学基金项目和 863 计划项目若干项,发表学术论文 100 余篇,主编教材 4 本.现任中国计算机学会容错计算专委会委员兼测试与诊断学组副组长,全国计算机继续教育研究会副理事长,IEEE 会员,全国高等学校计算机教育研究会理事,湖南省政协常委.

(上接第 106 页)

参考文献:

- [1] 钟义信.从“统计”到理解,从“传输”到“认知” [J]. 电子学报, 1998, 26(7) :1 - 8.
- [2] 杨晓兰,钟义信.基于文本理解的自动文摘系统研究与实现 [J]. 电子学报, 1998, 26(7) :155 - 158.
- [3] 陈桂林,王永成. Internet 网络信息自动摘要的研究 [J]. 高技术

- 通讯, 1999, 2:33 - 36.
- [4] 伍谦光. 语义学导论(第二版) [M]. 长沙:湖南教育出版社, 1995.
- [5] 房玉清. 实用汉语语法(第四版) [M]. 北京:北京语言学院出版社, 1996.
- [6] 黄曾阳. HNC(概念层次网络)理论 [M]. 北京:清华大学出版社, 1998:3 - 13.