用于文本智能处理的电子词典的一种设计方法

李　娜

(南京师范大学控制科学与工程系, 210042, 南京)

[摘要]　提出一种用于文本处理的电子词典的设计方法. 首先给出设计思想, 接着给出实现的方法, 并把这种设计过程用于军用文书文本处理系统的电子词典的实现, 同时检验了这种方法构建词典的可行性.

[关键词]　电子词典, 自然语言处理, 军用文书

[中图分类号] O157. 4,　[文献标识码] A,　[文章编号] 1672- 1292(2003)03- 0031- 04

0　引言

利用计算机来处理文本是一门新学科——计算机语言学. 现在, 研究它的课题已经很广泛, 这些研究属于自然语言处理领域, 这一领域的研究对于计算机的智能化发展是很必要的. 让计算机代替人来处理文本, 相当于让计算机掌握一种语言. 人要掌握一门语言要有一定的词汇量和语法语义知识, 计算机要使用一种语言, 也必须具备这些知识, 那么怎么存储这些知识, 以便于机器的应用, 这就是电子词典研制的内容. 用于自然语言处理的电子词典的编纂所做的工作就是对一种语言的词汇及其使用规律很好的加以组织, 以便于利用.

用于自然语言处理的电子词典不同于供外语学习使用的双语(或三语)电子词典, 后者只存储词的释义、词性、以及对应另一种语言的词汇等, 而前者除了存储这些内容外, 还要存储词汇的语法信息, 语义信息以及语言使用的处理方法和处理规则等. 对于这些信息的管理, 决定了这些信息的使用效率, 所以词典编纂的好坏直接影响着文本处理系统性能.

1　设计思想

设计出一种大而全的词典在目前来说是不现实的, 而且对于特定的应用领域也不一定适合, 反而会显得笨重, 所以, 本文是针对某一特定受限领域来设计一部用于处理文本的电子词典. 尽管我们的应用限制在某一领域——军用文书处理, 但是建立词典的思想及方法具有一定的普遍意义.

设计词典首要考虑的问题是把词库放在什么文件里, 也就是语言知识的存储形式. 词库规模较小时, 可以用几个动态数组来存放各个表, 用序列化函数来读写磁盘文件. 词库规模大时, 为了节省空间, 由两种方法可以选择: 第一, 存放在定义的数据文件中; 第二, 存放在某种数据库文件中. 用自定义数据文件, 需要自己编写大量的代码来管理这些数据, 但可以获得最高的时空效率(访问速度快, 占用空间小). 数据存储在数据库里首先使数据能与其他应用程序共享; 其次, 使数据的安全性、完整性、锁定性好; 再次, 使数据易于维护, 易于管理. 这些都是文本形式存放所不能达到的. 所以, 选择采用数据库的方式建造电子词典.

设计词典其次要考虑的问题是词典构建遵循的原则, 本文的原则是分层建库, 分块管理.

1.1　分层建库

词是具有一定意义的最小语言单位, 也是自然语言处理的基本单元, 本系统将各种语言学知识(语

收稿日期: 2003- 04- 01.

作者简介: 李娜, 女, 1977- , 南京师范大学控制科学与工程系助教, 主要从事智能控制、工业生产过程控制方向的研究.

法、语义、语用)和背景知识体现在词的语法语义范畴和运算规则中,对词进行全方位的信息描写.对词的描述是分类加属性描述,词的分类是否合理直接影响到计算机工作的成效,按语法分类的结构同句法联系紧密,划分简单,但不关心被表达知识的意义,脱离语义.按语义分类的优点是给出了词的深层信息,比较接近人的理解方式,但不利于规则的形式化描述,所以要结合两种方法对此进行分类,第一层按语法分类,第二层按语义分类,基于这两种语言知识的两个层次之上是第三层,语言处理规则的建立.

1.2　分块管理

在分层后的每个层次上,按某些标准化分成若干块,分而治之.就语法层而言,按语法进行分类,按词性分块存储;在语义层则按语义分类,按语义分块存储.同时语法语义库和知识库也分块存储,这样可以对每一个块采取相同的或者不同的结构存储.既适合系统的应用要求,也提高了运行管理效率.

按此原则建立的词典的结构如图1所示.

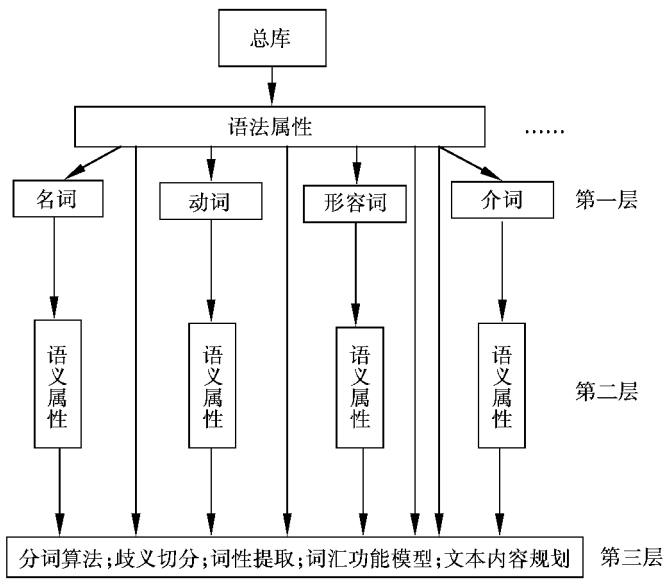

图1　电子词典的结构

在总库的词汇各个属性字段包含了词的词性等语法知识,在这个层次上按照词性分类存储,由此分成了若干块存储了词汇知识.第二层的语义属性层是在第一层分块存储的基础上,也就是名词、动词、形容词等块上建立了语义分类,在这两个层次的基础上建立了第三层.

2　设计方法

2.1　机用词汇描述方法

词汇的语法语义信息一般从两个侧面进行刻画:分类方法与属性方法[1].分类法与属性描述法是机用词典常用的两种信息表述方法.在理论上,这两种方法对于认识事物是等价的,在实际操作中,各有优缺点.首先分类是研究事物的基础.为了研究词汇也必须对词汇进行分类.分类法简单明了,上下层级之间属性易于传递和继承,但分类标准的交叉矛盾难以克服.相比之下,属性描述法更便于掌握,只要填写出具体词语的语义属性,而不必很费劲地去归类,但是这种方法容易遗漏信息,一致性不好保证.本文采用二者结合的方法,利用二者的优点,回避它们的缺点.

首先,对词汇进行分类.从语法和语义两个角度进行分类,在这个角度上还可以按照不同的应用目的和领域进行不同的分类.本文针对特定的领域是军用文书一类文本的处理,所以使用在通用分类基础上进行修改得到的一种特定的分类体系,由于篇幅有限,在此不予列出,参见文献[2].

接着,对词的属性进行描述.由于信息词典主要是服务于计算机分析与生成汉语句子的,因此属性项目的确定首先取决于自然语言处理程序的需要,一方面帮助分析程序理解自然语言固有的或机械处理所引发的歧义,另一方面也帮助生成程序产生通顺的汉语句子.计算机分析汉语句子所依据的语法理论与算法可能是各种各样的,但其基本步骤都包括:(1)将连续书写的句子切分成单个词;(2)给词标注词性;(3)逐层将词组合成短语直至句子;(4)确定词和短语在更大的结构(短语或句子)中的句法功能或语义角色.语法信息词典的词语属性就要为这些基本步骤提供尽可能丰富的知识.因此在词典的词汇的属性里面设置了词的词法信息(比如词的重叠情况)、句法信息(比如词汇修饰语的情况)、担任句法成分情况信息(是否可以用作主语等)以及词汇的前后照应信息.

2.2 语言处理方法的组织

贯穿词典设计始终的设计目的就是便于文本处理系统的应用, 上面我们也说到处理文本一般都涉及到几个连续的处理过程, 也就是先分词, 再词性标注, 然后确定词在句子中的作用或者成分. 可知这些步骤中每一步都涉及一些处理方法, 提炼好的处理方法, 存储在词典中, 使词典具有这些处理功能, 减轻了文本处理系统的负担, 同时可以将来把词典扩展为知识库, 这在文献[4] 中有论述. 这不失为一种好的设计词典的方法, 同时这也是对词典应用的检验.

本文初步尝试了在词典中实现最大分词算法、词性标注、三字链歧义消除等功能, 这些功能通过利用数据库的存储过程来实现.

3 用于军用文书智能处理的电子词典的实现

基于以上思想和方法, 对特定领域军用文书的智能处理 ——文书的理解和生成, 构建和实现了用于军用文书智能处理的电子词典.

军用文书是军队内部用于传达上下级命令, 通知或通报等文件. 在作战的情况下, 用人来写下或传达文书内容在时间的浪费上已变得越来越不能忍受了. 智能处理作战文书的研究已经越来越迫切了. 一个文书处理系统的简要处理流程如图 2 所示:

设计的电子词典就是为了能够给这个系统提供丰富的词汇和语言知识, 以及部分处理功能. 首先词典的内容包括了这一领域的大部分词汇, 以及这些词汇的语法语义知识, 然后按照图 1 的结构构建这部词典. 其中总库中包含各个分库则是为其它文本的处理提供知识. 包含的属性项目和数据类型如表 1 所示:

表 1 总库结构

char	词语
int	同字词
varchar	全拼音
char	虚实
char	体谓
char	词类
varchar	同形
char	切分歧义
char	分词标识
varchar	组成离合
int	分类序号

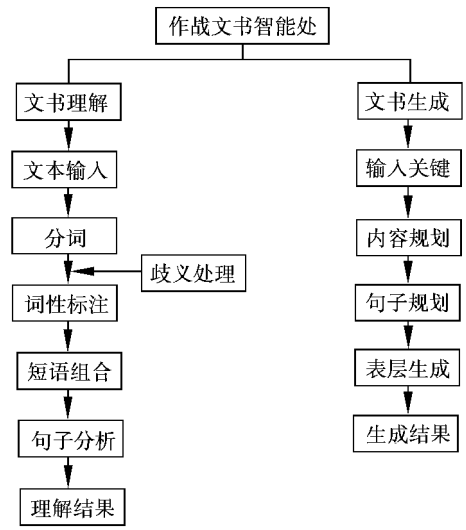

图 2 文书智能处理系统

以下是总库部分属性字段语法语义含义及填写规范:

3.1 同字词

本词典中的同字词包括以下 3 种情况: (a) 汉字相同但读音不同的词, 如: “重(chong2) ”和“重(zhong4) ”、“合计(he2ji4) 和“合计(he2ji5) ”. (b) 汉字、读音皆同, 但词类不同的词, “制服(名词) ”和“制服(动词) ”“巩固(动词) ”和“巩固(形容词) ”. (c) 汉字、读音、词类皆同但词义不同的词(包括同形词和多义词) , 如“拐弯”的“拐”和“拐骗”的“拐”. 这些同字词各作为一个词语看待. 这样总库中就有两个“制服”、巩固”、“拐”, 每个记录的本字段均填“2”.

3.2 同形

词类相同的同形词中, 全拼音不同或者词项不同的, 分别注以 A, B, C; 词项相同而义项不同的, 则填 1, 2, 3, 字母与数字同时存在时, 则将字母置于数字之前, 如 A1, A2, A3, B1, B2.

3.3 分词标识

词语左边有典型的分词标识, 填"左", 右边有典型的分词标识, 填"右", 左右都有分词标识的, 填"左右", 无典型分词标识的不填. 具体地说:"继续"中"继"一般前置, 本字段就填"左", 续就填"右".

3.4 组成离合

有些离合词拆开使用的时候, 前后两个语素可能相距很远, 有时后一个语素甚至可以提到前边. 如"打仗"一词, 可以说"打败仗", 对于这类离合动词, 为了给分析器提供信息, 在"打"与"仗"这两个记录中, 本字段都填上"打仗". 例子还有:"革命、打架"等.

其它各分库的数据结构就不在此列出, 参见文献[2].

从各个分库的属性字段我们可以看到包含了用于军用文书处理的词汇的丰富语法语义知识. 同时也实现了最大匹配分词、词性标注功能、词汇功能模型的建立和内容结构模型建立的功能. 其中最大匹配分词的运行结果如图 3 所示, 上面窗口是文书文本中的'指挥所开设位置', 计算机进行文书理解时前先进行分词, 得到下面的分词结果. 然后根据分析结果到词典里面提取词性, 进行标注, 会得到标注结果'名词/动词/名词'(篇幅有限不在此列出运行结果), 再接下来进行语法分析等. 经过一系列处理后, 得到文书理解结果.

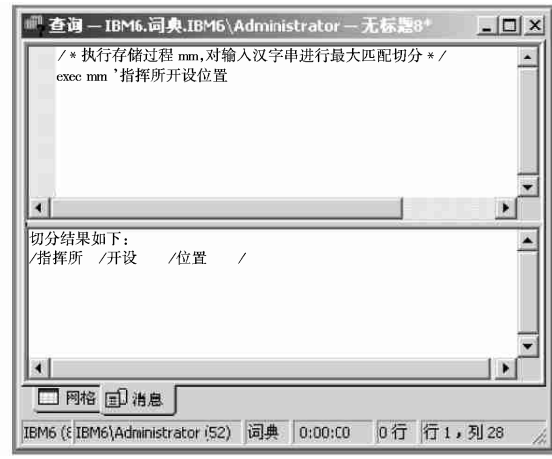

图 3　词典的分词功能

此词典为军用文书处理系统提供了丰富的语言知识和一些功能, 同时合理的词典结构为处理系统提供了有效的支持.

4　结束语

本文研究的是用于文本处理的电子词典编撰的思想和方法. 从使用和管理等角度出发, 给出了词典设计的一种方法, 并在特定的受限领域应用. 此词典给军用文书智能处理奠定了良好的基础, 同时这一应用也证明了这种构建词典方法的可行性.

[参考文献]

[1] 俞士汉. 现代汉语语法信息词典详解[M]. 北京: 清华大学出版社, 1998: 111~358.

[2] 李娜. 用于作战文书智能处理的电子词典的研究与实现[D]. 硕士论文. 南京: 南京理工大学自动化系, 2003.

[3] 姚天顺. 自然语言解释: 一种让机器懂得人类语言的研究[M]. 北京: 清华大学出版, 1995.

[4] 陆雪莹. 文本生成系统中知识库构造设想[J]. 电子器件, 1997, 20(1): 5~10.

A Method of Designing Electronic Dictionary Applied to Intelligent Processing of Text

Li Na

(Department of Control Science and Engineering, Nanjing Normal University, 210042, Nanjing, PRC)

Abstract: This paper introduces a method of designing the electronic dictionary applied to intelligent processing of text. Both the idea of the dictionary design and the method for the implementation are dealt with. This method has been used to design the dictionary of operation document processing system, with the feasibility of the method for the dictionary design demonstrated.

Key words: electronic dictionary, NLP, operation document　　[责任编辑: 刘健]