新一代语料库的建设与管理

苑春法　黄昌宁　孙致宇　赵　强

清华大学计算机系

摘　要

随着电子出版业的迅猛发展，大规模机器可读文本的获取已成为可能。随之形成了新一代（第三代）语料库的特有风格。本文探讨了新一代语料库的建设与管理，提出管理系统应分为：文本管理模块和专门子语料库生成模块和专门子语料库管理模块。

关键字：语料库　语料库语言学

ABSTRACT:

As the modern electronic publishing is developing rapidly, So huge amount of machine-readable text for corpora becomes available. Now the new-generation (the third-generation) corpora and their characteristic are formed. This paper will discuss the construction and management about the new-generation corpora. We propose that the appropriate management system must include: A module of an archive management system, a module of the special purpose corpus generation system and a module of management system of the special purpose corpus.

KEYWORD: corpus, corpus linguistics

1. 语料库的发展

语料库的建设和语料库语言学的崛起，使计算语言学的发展别开生面。从大规模真实文本中调查来发现并总结自然语言的各种语言事实和语法规律，将有力地支持自然语言的处理。

六十年代初，W.N.Francis 和 H.Kucera在美国 Brown大学建立了世界上第一个存储于数字计算机上的语料库－－Brown 语料库。Brown 语料库收集的是当代美国英语，它按系统性原则采集15类文体的样本共 500个，每个样本不少于 2000 词次，语料库总容量为百万词次。

七十年代初，由英国 Lancaster 大学 G.Leech 教授倡议，由挪威 Oslo 大学 S.Johansson 主持完成，并最后装备在 Bergen 大学挪威人文科学计算中心的 LOB 语料库，是 Brown语料库的姊妹库，它采集的语料是当代英国英语，它的采样原则、样本大小及语料库总容量和 Brown语料库相同。

我们称Brown 语料库和 LOB语料库为第一代语料库，它们的库容量以早期的计算语言学的标准来衡量已可以算是“巨大”的了。在这一时期的语料主要是采用手工输入计算机，因而语料输入计算机的规模不可能很大，但对于研究词法来说已基本满足要求。在当时技术条件约束下，不可能大规模输入语料，而另一方面语料库要求其语料具有代表性，最好的办法就是采用按系统性原则采样的方法，在各种文本中抽取样本，由各种各样的语言片断来组成一个语料库。

进入八十年代之后，由于光扫描文字输入技术(OCR)的发展，这样给语料的计算机输入带来了极大的方便，这样就为形成更大规模的语料库创造了条件。同时人们开始更注重用语料库的方法来研究句法和篇章理解；在这种情况下，扩大语料库的规模是迫切需要的，另外在这一时期人们开始注重录入全文（Text）。这是语料库发展中的二个显著的改变。其中有代表性的是伯明翰大学 John Sinclair 教授主持，得到 Collins出版社赞助的 COBUILD (COllins Birmingham University International Language Database)语料库，规模为 2000 万英语词次。还有 Longman/Lancaster English Language Corpus，规模为 3000 万词次。这样就形成了第二代语料库的风格。

八十年代末和九十年代以来，电子出版业进入了蓬勃发展时期，另外先进的电子通讯系统的发展，使计算机得到大量语料已不再成为困难的问题。但随便的一些语料堆集不能成为一个语料库，一个语料库的语料应反应当代一种或几种（多语语料库）语言的用法，或代表某一领域内某语言的用法，也就是说既然是语料库就一定存在选材问题。然而语料的收集有任意性，这主要取决于获得数据的数据源和数据通道。如何解决得到语料的任意性和组成语料库的语料有选择性的矛盾呢？Birmingham 大学 Sinclair和他的助手们采用的办法是在主语料库之外建立一个监控语料库（monitor corpus），监控语料收集的是能够收集到的未经挑选的语料，然后在监控语料库中选材生成主语料库。这是一种策略，把所有可以得到的机器可读文本基本不加选择地都存储起来建立文档库（archive）。然后我们需要什么样的语料库（如军事领域、经济领域 ..... 等）都可从文档库中抽取这一领域中适当分布的语料来形成。这一时期另一个典型的例子是美国计算语言学学会（The Association for Computational Linguistics）倡议的数据采取计划（Data Collection Initiative）简称 ACL/DCI。这个项目由美国宾西法尼亚大学的 M.Liberman 主持。以 ACL/DCI 为代表形成了第三代语料库。这一代语料库首先对所有可以得到的语料以文本形式存储起来，基本不加改动，甚至还保留着许多版面信息(使用 SGML 语言描述文本），它的容量一般为1亿词次以上，二十一世纪可望达到万亿词次的量级；在这个文档库的基础上，可以抽取适当合理分布的语料形成专门领域的子语料库。

汉语方面已有多家专门的文本语料库，如深圳大学的红楼梦语料库，台湾的二十五史文本库等。清华大学近年来为按系统性原则采样的语料库收集了大量机器可读文本，总容量在 5000 万字左右。为了充分发挥这 5000 万字语料的效益，和进一步收集更多的语料迫切需要建立一个管理系统，把所有这些文本语料库管理起来，同时要求这个系统有自动生成许多专门子语料库的功能。也就是说需要按第三代语料库的风格来建设语料库。

2. 语料库管理系统

新一代的语料库和以往的语料库不同，它不再是选好的文本或文本片断没有结构地存放。它收集大量原始文本，这些文本的存放应有结构，以便于生成最终可用的子语料库。

新一代的语料库区别于一般的文档库，它不仅需要存有文本(text)的属性信息，而且需要存有文本的结构信息如章名、节名及其定位，这些信息的作用不只是服务查找某一文本，而且在于文本内部可以按一定原则抽取样本生成专门的最终可用的子语料库。

由上所述，新一代语料库管理系统必须具有如下三个模块：

(1) 文本管理模块，它的功能包括文本入库，文本删除、以及文本查询。

(2) 专门子语料库生成模块，它的功能是在某一特殊领域，按某种特殊需要，从文本当中选取文本或选取文本中的片断以生成一个子语料库。

(3) 专门子语料管理模块，它的功能应和传统语料库管理系统的功能相同，即包括：关键字检索、字频、词频统计、以及分词、词性标注、句法标注以及语义标注等工具软件。新一代语料库管理系统的框图应如图 2.1所示。

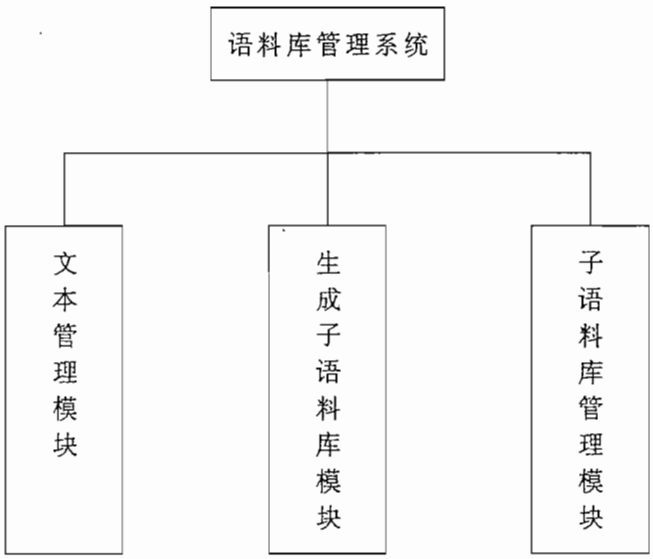

图 2.1 语料库总体框图

在这种语料库管理系统中，最重要的是文本管理，如果文本管理这个模块做得好，那么生成子语料库就很方便。这里关键问题是文本中那些信息需要抽取出来？以什么样的结构放这些信息？本文主要针对这些问题展开讨论。

3. 文本管理

在新一代语料库中，文本是基本的存储单位。对于每一个文本可以从二个特性来描述，一是基本属性，二是层次结构，而每一特性都还有一定的层次。

3.1 文本的属性

文本属性可以分四个方面，一是有关文本本身内容的特性，包括：文章类型，是政治经济、自然科学 ……，还是文学艺术等；文章体裁，是小说、散文 ……，还是新闻报导等；文本加工情况，是原始文本、分好词的文本或是有句法标注的文本。二是有关出版的特性，包括：出版社名称、出版年、月。三是有关作者的特性，包括作者数目和作者性名。四是有关文本名称特性，包括文本名称及源文件名称。所有这些有关文本的属性可以用一棵子树来表示，如图 3-1所示。

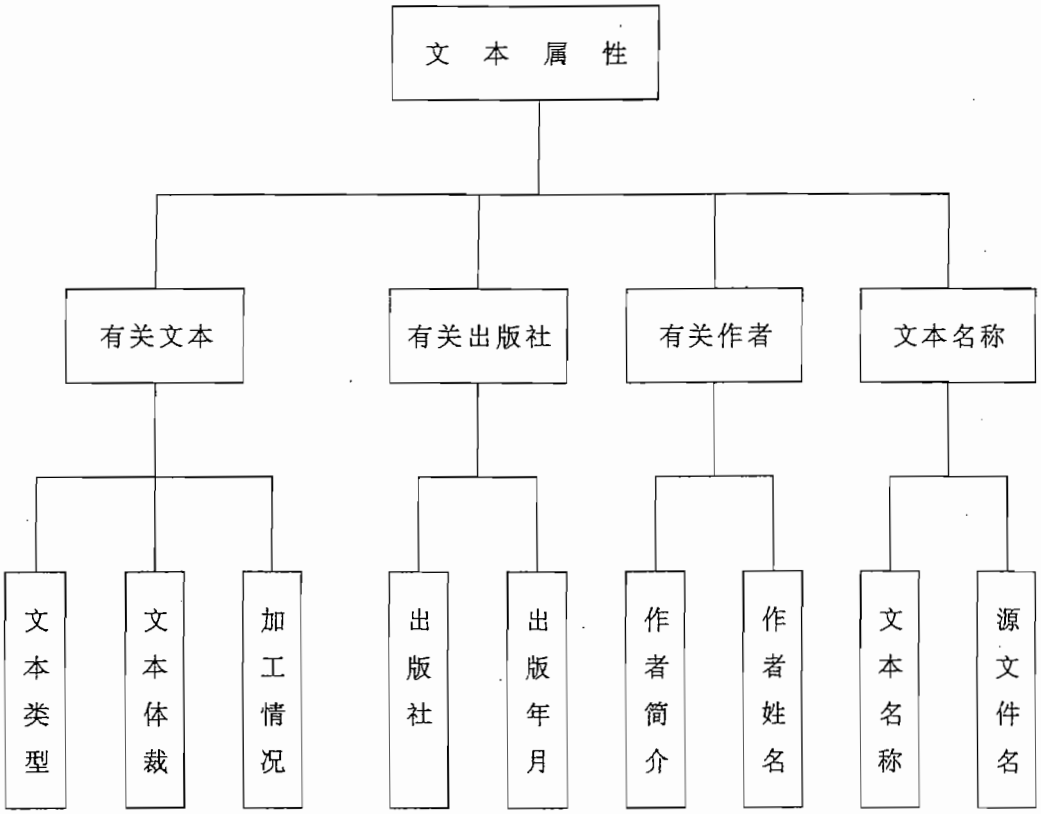

图 3-1 文本属性结构

3.2 文本层次结构

一个文本一定存在着一定的层次，比如一本书应分有若干章、每章又分若干节、每节又分若干小节 ..... 。这些信息，联同它在文本中的位置，若能从文本中取出来，那么对于检索和生成专门子语料库将是十分有用的。一个文本的层次结构可以用图 3.2来表示。

如果某文本是一本书，那层一结构节点一般为章，层二一般为节，.....。而对于每一个结构节点，都附有节点的名字、节点的层次序号、指向兄弟节点和子节点的指针，以及节点位置的相对偏移量。

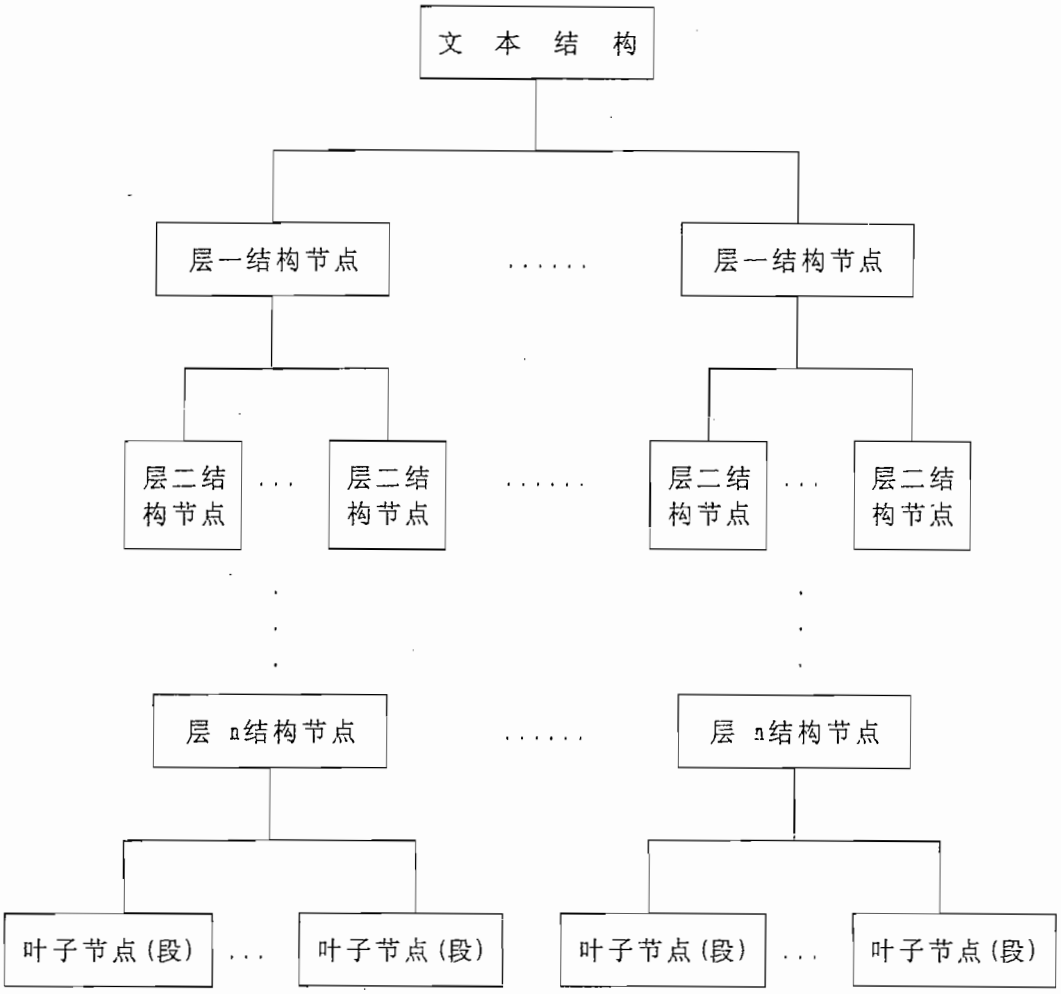

图 3-2 文本层次结构

3.3 文本结构层次的自动识别

文本的结构层次信息早已经存在于文本之中，如果能从文本中自动识别出这些信息就可以大大减少文本入库时的开销。我们针对“书”这类文本，设计了利用版面信息自动识别书的层次结构的程序，实验结果令人满意。识别文本层次结构主要是识别出文本中的标题，其次是识别出该标题所属层次。

识别文本中标题可供利用的版面信息有：

(1) 行前的空格数。作为标题行，行前空格数一般大于 4 。

(2) 行长度。做为标题行，每行长度小于一般行长度。

(3) 标点符号。在标题行一般不可能出现句子结束符如“。”“！”“？”“；”。

(4) 关键字。有些关键字的出现的行，是标题行的可能性较大，如“第××章”，“第××节”，“一”，“一.”，“(一)”，……等。

由以上信息相互结合可以初步形成一个判断某行是否为标题行的规则。

有关标题层次判别，主要是依靠对每个标题行的版面特性的记忆。相同层次的标题行版面特性应是一样的，逐层比较以判别该标题属于那一层。其次依靠关键字，比如“第二章”与“第一章”同属一个层次，“1.1.1”开头的标题与“1.1.2”开头的标题同属一个层次。当计算机无法判别该标题层次时，弹出一个窗口提示由人来选择。

4. 结束语

随着语料库发展的日新月异，语料库的结构与管理系统的研究越来越重要。本文我们对新一代语料库的结构与管理做了初步探索，许多问题尚待进一步研究。

参 考 文 献

[1]. 黄昌宁：国外语料库述评，陈肇雄主编《机器翻译进展》，电子工业出版社，北京 1992.9.

[2]. Geoffrey leech: The State of the Art in Corpus Linguistics, In English Corpus linguistics, Edited by Karin Aijmer & Bengt Altenberg, 1991.