《人民日报》1998 年语料库中若干基本语言数据的统计与分析

胡景贺[1]

[1]（北京大学计算机系，北京，100871）

E-mail: hujinghe@sina.com

摘要： 本文汇报了对“北京大学计算机语言所 1998 年《人民日报》语料库”半年语料的统计分析工作。其中统计了语料库中的词频、词在词类上的分布、词类的二元和三元共现、词与二元及三元词类的共现、各种共现在句子首尾端的边界分布。本文还对上述统计结果进行了分析，着重讨论了高频词语词类的分布以及句子的边界情况。这些结论对于该语料库的全面分析提供了重要的基础数据。

关键词: 词频；词类；共现；散列；分布；边界。

引言

“语言知识库是自然语言处理系统（如：机器翻译、文献检索、信息提取等）的必要组成部分。语言知识库的规模与质量是自然语言处理系统成败的关键。汉语形态不发达，适用于自动分析的形式系统不够成熟，尤其需要重视语言知识库的建设。”文献[1]

讨论现代语言中的特点，是分析提取语言要素的前提。对于特殊的有代表性的语料进行统计分析，可以了语言中一个特定种类的语言特色，从而坚定计算机处理自然语言的基础。对语料库进行基本的语言数据分析，是为正确把握语料库中的语言特点做的一项基础工作。本文所总结的工作，仍然停留在底层，以“北京大学计算机语言所 1998 年《人民日报》语料库”（经过切分标注）为分析对象，考察其中的词语使用情况、词类的出现规律、边界出现的词语及词类情况等等。使用了半年的语料，包含词数约 700 万。

1 算法

1.1 期望的统计结果及统计规则

本次统计工作期望得到如下结果：

a). 语料库中的词频；

b). 词在词类上的分布；

c). 词类的二元和三元共现；

d). 词与二元及三元词类的共现；

e). 各种共现在句子首尾端的边界分布;

按如下规则:

a). 统计二元及三元共现时，段落为断点;

b). 单独考虑词组的连接频率(ns, nt, nz, i, l);

c). 合并人名;

d). 词组，数字，数字表示的时间，人名均只存词类的频率，不存具体词。

1.2 数据结构

1.2.1 统计词频用的数据结构

词频统计中存储词语用的数据结构，考虑到语料库较大，使用散列。散列采用词的字符串为关键码。散列中的存储元素为一个整数值（该词出现频率）和一个链表（存储每种词类的出现频率）组成。

存储所有词语的数据结构的表示:

hash_map <const char*, table, MyCompare > hm;

这个模板类有四个参数，我定义其中三个:

其中 const char* 为关键码，为该字符串(词);

散列的每个元素为一个 struct，存储一个词的信息;其中包含一个整数值存储该词的出现频率，以及一个向量用来存储该词每种出现的词类和各词类出现的频率。

struct table{int freq; vector<aStruct> catFreq;};

通过将自定义 MyCompare 类用作 hash_map 模板类的第三参数，使得 hash_map 可以按照字符串关键码存储查找。MyCompare 是一个 traits（特性）模板类，主要需要其中的两个函数对象和一组枚举类型对散列进行设定。MyCompare 是按照原来缺省的 hash_compare 模板类修改而成。以 const char* 和 MyLess 为参数，const char*表示关键码（MyLess 是一个自定义的含有比较函数的类，它为散列提供比较散列元素大小的方法，其实它也是 MyCompare 的一个特性。）在 MyCompare 中可以修改散列的存放单元。在此词频散列中，由于散列非常大，将散列的初始桶数改设为 2048。另外，还需在 MyCompare 中重新定义按照关键码计算散列存放位置和比较元素大小这两个函数。

通过以上操作，就可以生成适应程序要求的散列。这是此应用程序中最主要的数据结构。

1.2.2 关系到词类二元及三元共现所用的数据结构

统计矩阵采用数组和散列数组实现基本构架。因为所有词类的共现数量是固定的，所以采用数组结构存储每种词类共现的出现频率。又因为每种词类共现中具体词数可能极大，且数量不均，因此在使用数组的基础上，用散列来存储对应词类共现的词的出现频率。

对于每种词类的二元、三元共现中的具体词语按位置的出现频率，用散列数组实现，散列数组中每一个单元为一个散列，每个散列都以字符串为关键码，以整数值（字符处在相应位置出现频率）为元素。（hash_map <const char*, int, MyCompare2> hm3[n][n][n][3];//以三维为例。）其中 char*为关键码，为字符串(词)。Int 为该词出现频率。MyCompare2 的定义方法和上面的词频散列中的定义方法类似。

1.3 具体实现

伪码实现主要算法:

读入第一个字符;

```
while(!EOF): {
    读入一行;
    while(!){
```

```
        if (出现词组头){压栈}
        读入一个词并处理;(将词与词类共现放入散列)
        if (出现词组尾){处理词类共现后弹栈}
        准备读下一个词;}
    读入第一个字符;}
```

对边界的处理借助对于三元词类共现的统计，只是在段间加标记从而找出一段单句时句子的开头结尾。以“。”、“；”、“：”、“！”、“？”、“…”六个字符为句子结尾的判定标志。

2 实现中的问题及解决方法

2.1

首先是数据结构上的实现问题，开始时用模板类没有重新定义特性类，因散列的关键码等于是存放字符串的地址。将原有的 hash_compare 模板类中的第二个参数改为 MyLess 后直接使用时，比较函数是没有问题了，但是由于计算元素在散列中的存放位置的函数依然是根据地址的值计算，超出散列初始设定大小时，散列的查找就出了问题。将 hash_compare 重新定义以后，解决了以上问题。散列的存储关键码设定正确。

2.2

第二是读入词和词类的问题。一开始的设计是用许多小函数来实现，但是考虑到调用的复杂程度和速度问题，小函数都很不适用。最后决定使用栈方法。因为只需要一次压栈，不会更多，因此非常容易实现。而且由于是顺序处理，不需要往回走，速度上也要快得多。

2.3

第三是对于特殊词的处理。由于用数字表示的时间，数字，人名都是不计具体词汇的，因此要将其改掉。一开始使用读入之前判断，后来改用读入以后再修改，这样可以保证算法的整齐和连续性。对于合并人名的操作，考虑过读入散列以后再修改，因为操作复杂，速度慢，改用向前判断，读过重复人名的方法。

（以上的问题都是在程序正确运行之前的设计问题。）

2.4

最严重的问题是速度和内存问题。一开始的程序非常慢。而速度问题又是由于内存不足导致使用大量的虚拟存储所致。

改进过程中使用了许多方法优化。包括将一维词频的存放方式由数组改为向量来减小内存用量，以及把移动到下一个词时存储的前面两个词的换位由 string 类型变量之间拷贝转换，变为 char*类型的地址转换等等，但是效果都不是很好。

主要是进行了两种操作后，效果大为好转。一是将读入操作中的每读几个字符就重新修整一次行缓冲 string，改为用整数值纪录读词的位置，每行的缓冲流在当前行中不变；另一个效果更为明显的优化是将一维词及词类散列的初始桶数改大，这样存储时速度明显加快。

这两种优化的效果十分明显，程序运行时间可以减小 50%以上。

2.5

讨论句子边界时发现原始语料中存在语法问题。出现了“？！”共现，导致句头句尾数量上有差别。处理中把这种句子结尾不规范的问题滤掉了。

3 统计结果及分析

3.1 词和词类的分布

“北京大学计算机语言所1998年《人民日报》语料库”半年语料中共有103883个词语。其中出现最多的词语如下：（数词m、名字nr、时间t都是合并修改后的一类词语）

词语	出现次数	出现词类				
，	491773	1	w：491773			
的	358186	2	u：358181	Ng：5		
。	237134	1	w：237134			
数词	189395	1	m：189395			
、	134295	1	w：134295			
名字	117851	2	nr：117586	n：265		
在	78369	3	p：74821	v：1697	d：1851	
了	76214	3	u：67665	y：8075	v：474	
和	72555	6	c：70732	p：1751	v：61	
			Ag：4	Ng：3	q：4	
是	68855	3	v：68792	r：61	Ng：2	
时间	67917	2	t：64969	n：2948		
“	49723	1	w：49723			
”	49655	1	w：49655			
不	31240	1	d：31240			

有　　30484　　2　　v：30479　　m：5

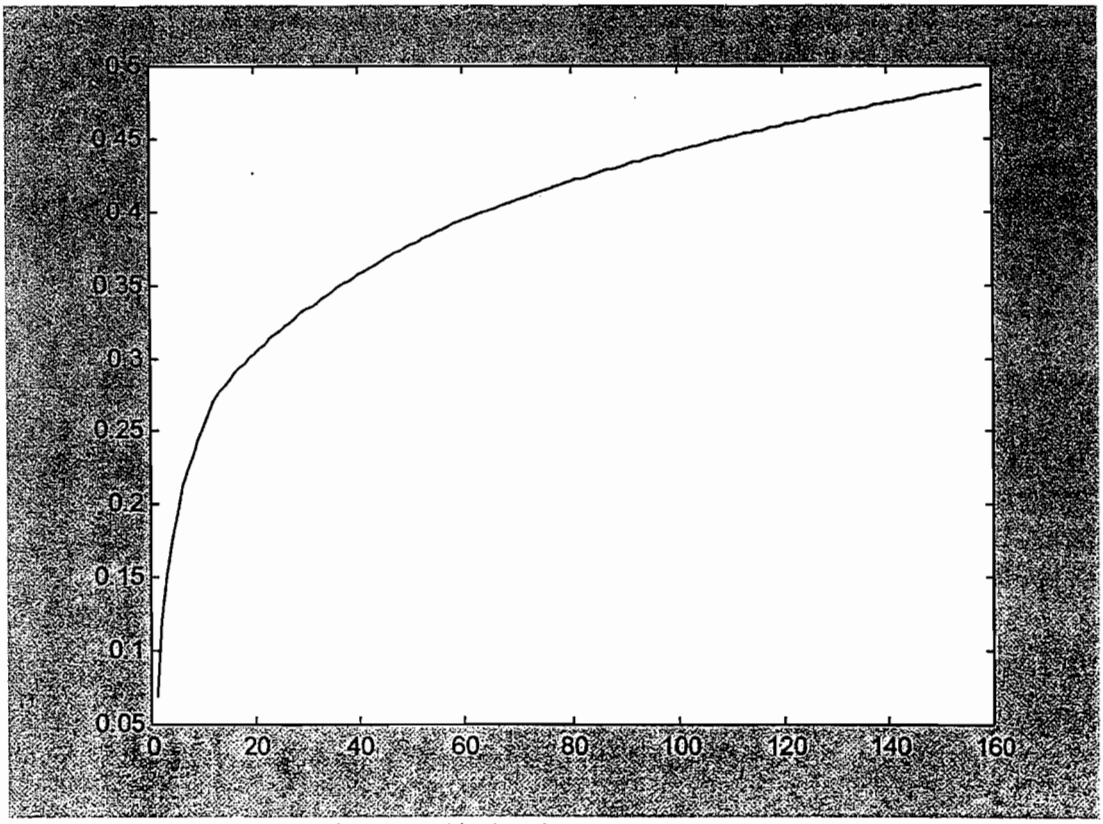
单个词语最多出现7种词类，有以下4个：

词	次数						
当	2829	v：1195	p：1623	Ng：6	Ag：1	o：2	Bg：1	d：1
分	2000	v：802	q：1084	n：62	Vg：2	Ng：38	b：11	j：1
多	18318	m：9647	a：6685	v：331	d：836	ad：732	Ng：83	j：4
少	1680	a：1183	v：204	ad：173	j：5	d：108	Ng：5	Ag：2

词类中含有词数和每种词类的出现词数分别为（前5种）：

名词n出现1527530次，包含词语46204个；动词v，1193935次，16761词；助词u，489601次，38词；副词d，320783次，1295词；动名词Vn，301407次，8080词。另外，标点出现1103547次，共有73种。

这一结果就已经说明了该语料库中在词和词类的使用上的特点。而对于统计结果分析以后，发现少数词语占据了大部分的出现次数。以下这张图表示出现频率最大的前n个词语在总出现次数中的比重。横坐标表示词数，纵坐标表示比重。共103883词，总出现数7187878次。



不过这张图中考虑词语的包括标点。去掉标点（占总出现次数的1/7）后曲线会缓和一些。但是仍可看

出出现最多的一百词占总数的 1/3 以上，这说明汉语中词语的使用是比较“密集”的。

3.2 二元和三元共现

二元和三元共现的统计结果分为两部分。一是一张频率表，表示每两种（或三种）词类间共现的次数。二是一张词表，表示每种共现中出现的词语。凭借这两张表，还可以知道在固定两种词类后，出现某一特定词语的概率，也就是词与二元及三元词类的共现。

由于这些表过于庞大，只能举一些比较引人关注的数据。以下是名词后出现词类的次数：a:25193，ad:8260，an:3095，Ag:251，b:5316，Bg:4，c:53141，d:78197，Dg:159，e:8，f:49527，h:8，i:5004，j:8854，k:5007，l:5352，m:25023，Mg:7，n:261799，nr:34721，ns:3650，nt:691，nz:934，Ng:4591，o:60，p:39248，q:415，Qg:0，r:9230，Rg:2，s:2321，t:5708，Tg:67，u:134353，Ug:5，v:208375，vd:1343，vn:85311，Vg:1358，w:393744，x:0，y:1945，Yg:2，z:1895，词组 ns:122，词组 nt:962，词组 nz:40，词组 i:5，词组 l:2。而 a-n 共现只有 69631 次，说明虽然汉语中形容词前置是基本方式，但形容词后置也很常用。两个名词的共现是非标点共现中最多的，另外，v-n:216981，u-n:207792，d-v:198689，v-v:193857，v-u:155635 也是比较高频的共现。

3.3 边界分布

边界是值得关注的。

句首第一词出现了 18606 个，也就是说，有 1/5 的词可以放在句首。词类在句首出现的次数（主要的）：a：4025，c：17979，d：11254，j：4200，m：20202，n：37247，nr：19240，ns：14336，nt：4379，p：33425，r：46215，t：22971，v：33045。可见代词最容易出现在句首，然后是名词、介词和动词。

句首第二词：22108 个，词类出现次数（主要的）d：15396，m：12821，n：57695，ns：11405，p：17078，r：15084，t：13563，u：11381，v：68405。句子的第二个词最可能是动词和名词。

句子末尾倒数第一词共有 28930 个，主要词类出现次数：a：13298，n：125117，q：11187，v：58205，vn：32588。

倒数第二词：23981 个，主要词类出现次数：a：21340，d：18044，m：18888，n：59206，q：10126，u：56351，v：54532，vn：19707。

由此可见，在《人民日报》语料库中，句子开头中，名词和代词、动词和介词最多，而结尾则是一名词（动名词也应算在内）为多，然后是动词。

4 总结和进一步打算

本文只是总结了对于“北京大学计算机语言所 1998 年《人民日报》语料库”第一阶段的统计工作。但是这些已有统计结果已经可以显示出该语料库中语言的一部分特点。接下来，我们准备继续统计词语的二元和三元共现。这样才能更好的理解词语之间的使用特点。而已有的统计结果，还有待于进一步的分析，找出真正有意义的规律。

参考文献：

[1]《大规模汉语标注语料库》俞士汶，段慧明，朱学锋，孙斌

[2]北京大学计算机科学技术系，北京大学计算机语言学研究所《计算语言学文集》第 4 集，俞士汶，朱学锋

致谢 感谢北京大学计算机语言所孙斌老师的指导。这是我的第一篇语言学方面的论文。做这项工作之

前，我对计算机语言学几乎还一窍不通。感谢孙老师的帮助。

作者简介：　胡景贺（1980—），男，北京人，本科，对于计算机语言学还在学习阶段。

Statistic and analysis of some fundamental linguistic data in a 1998 People's Daily corpus

HU JINGHE[1]

[1](BEIJING University, Beijing 100871, China);

E-mail: hujinghe@sohu.com

Abstract: This paper reports the statistic and analysis of half-year corpus from "1998 People's Daily Corpus of The Institute of Computational Linguistics (ICL) of Peking University". The statistic work includes the word frequencies, word distribution on the word categories, the binarycooccurrence and the triplecooccurrence of word categories, the binarycooccurrenced or triplecooccurrenced word categories, the distribution on the beginning and the end of the sentences and the binarycooccurrences of words. This paper also analyzes the statistic results, with emphasis on the high-frequency words or word categories, and the distribution on the beginning and the end of sentences. The results may have possible importance for the farther processing of similar corpus.

Key words: frequency; category; cooccurrence ; hash; distribution; edge