大规模语料库中词语接续对的统计与分析

邱超捷　宋柔　欧阳龙根
北京工业大学计算机学院人工智能研究室

摘要：词语接续对的数据库(接续库)对于词语层面的语言处理具有重要意义。为了建立接续库，我们对一个1亿字左右的大规模语料库中的词语接续对进行了统计。本文分析了统计结果，并着重讨论了接续对的数量同语料库规模的关系，接续对的频率分布，以及接续对的可靠性问题。我们的工作说明，语料库规模的增大不仅不能完全解决词语接续对统计不足问题，而且带来了巨大数量的接续对垃圾，从而人的干预是不可省去的。
关键词：词语接续对，接续库，接续对垃圾

Statistical Results and Their Analysis of the Neighboring Pairs of Words on Very Large Corpora

Qiu Chaojie, Song Rou, Ouyang Longgen
AI Laboratory, Computer Institute, Beijing Polytechnic University

ABSTRACT: The Database of the Neighboring Pairs of Words plays the important part in language processing on the word's level. For setting up a DNPW, we dealed with **Very Large Corpora** (size about 200 Megabytes) and got the results based on the **Neighboring Pairs** of **Words** (NPW). In this paper, we analyzed the statistic results, and specially discussed the **relations** between the number of NPW and the size of VLC, the **probability** of NPW, and the **creditability** of NPW. Our works show that the method of increasing the size of the Corpora not only can not completely solve the problem of insufficiency of NPW, but also produce a big amount of the **garbage** of NPW, so that the person's interposing is indispensable.
KEYWORDS: the Neighboring Pairs of Words, the Database of the Neighboring Pairs of Words, the Garbage of the Neighboring Pairs of Words

一 引言

在对汉语文章进行分词时，所能利用的只是字词频率和字词间的相关关系。由于统计数据不足和计算机能力有限，实际上，只利用了相邻两个字词间的相关关系。因为相邻关系有前后顺序，故我们称之为二元接续关系。利用二元接续关系解决分词中的歧义、专名识别、新词提取等问题已有许多成功的例子，校对、整句拼音输入、文字识别输入等应用系统的基本技术之一也是利用了二元接续关系。

我们把一个二元接续关系中的两个词语叫做一个词语接续对。利用词语二元接续关系的应用系统应当有一个词语接续对的数据库（简称接续库），实际运行时用接续库检查被处理对象中的词语是否接续，接续强度如何，从而决定处理策略。建立接续库的常规方法是对大规模语料库进行分词，然后收集其中的全部接续对，并统计同一接续对出现的次数，换算成接续强度，建成接续库。

利用大规模语料库建立接续库，是大规模语料库的重要用途之一。在开发一个实

用系统（计算机辅助校对系统《工智校对通》）的过程中，为了考察这一方法的有效性，我们做了一个建立接续库的实验，语料库规模达1亿汉字。我们分析了接续对的数量同语料库规模的关系，接续对的频率分布，以及接续对的可靠性问题。我们的工作说明，语料库规模的增大不仅不能完全解决词语接续对统计不足问题，而且带来了巨大数量的接续对垃圾。

二 统计方法

1 语料选取

我们对语料未做任何人工整理，对语料的统计顺序也未做特意安排。按照统计顺序，语料情况如下：

① 经济日报１９９２年语料（约１８２０万字）；
② 人民日报１９９３年语料（约２３４０万字）；
③ 人民日报１９９４年语料（约２１８０万字）；
④ 新华社１９９４年语料（约６９３万字）；
⑤ 新华社１９９５年语料（约２５００万字）；
⑥ 新华社１９９６年语料（约６００万字）。

本文所涉及的数据与图表以及最后的统计结果与分析均与语料的统计顺序有关。

2 统计对象

由于汉语词汇极其丰富，所以词的分布就十分稀疏，而词语接续对的分布就更加稀疏。为了统计到足够的词语接续对，需要把词语归成类，统计词语的类间接续关系。但是，有些应用系统，原则上不能使用词语类间接续关系，只能使用词间接续关系，校对系统就是这样。比如，可以说“桌上有本书”，不能说“椅上有本书”，后一句的“椅”应改成“椅子”。如果把“桌”和“椅”归成一类，凡同“桌”能接续的就被看作也能同“椅”接续，这种错误就查不出来了。同样，“大”和“小”不能归成一类，因为一般情况下只说“大好形势”、“大战一回”，不说“小好形势”、“小战一回”，只说“小溪”、“小老头”，不说“大溪”、“大老头”。同大多数多字词相比，一般的单字词和一小部分高频多字词用法比较灵活，专用性比较强，故尤其应研究单字词和高频多字词相互之间的接续关系。基于这一认识，我们利用上述大语料库统计了四千多个词的词间接续关系，这些词包括：

① ＧＢ２３１２中的６７６３个汉字中大约４０００个单字词；

② 大约５００个高频多字词。少数高频多字词做了归类，如“我们”、“你们”、“他们”、“她们”归成一类；

③ 为了调查专名识别错误对于接续对统计的影响，现代汉语中一些只能用在中外人名、地名中的字也包含在里面，如姓氏用字“刘”、“蔡”，人名用字“淑”、“晔”，外国人名用字“尔”、“斯”等。为了行文简单，本文中把这些字称为含有特定词性的词。

下面所讨论的就是以上词的词间接续关系（接续对）的统计结果。

三 统计数据及分析

1 接续对的数量同语料库规模的关系

我们对２００兆字节（约一亿字）语料进行了统计，得到的接续对约有５９万多个。下面二图是根据统计结果所绘出的接续对数量随语料库规模的变化曲线和增量曲线，用横坐标表示语料大小（单位：兆字节），纵坐标表示接续对数量（单位：个），图１中上曲线是由统计数据绘出的图形，下曲线是可靠的接续对随语料库规模的变化曲线。在图１和图２中，①表示经济日报１９９２年语料；②表示人民日报１９９３年语料；③表示人民日报１９９４年语料；④表示新华社１９９４年语料；⑤表示新华社１９９５年语料；⑥表示新华社１９９６年语料。

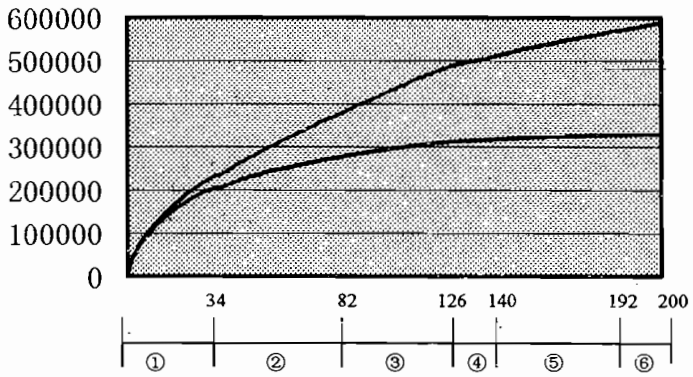

图１　接续对数量随语料库规模的变化曲线

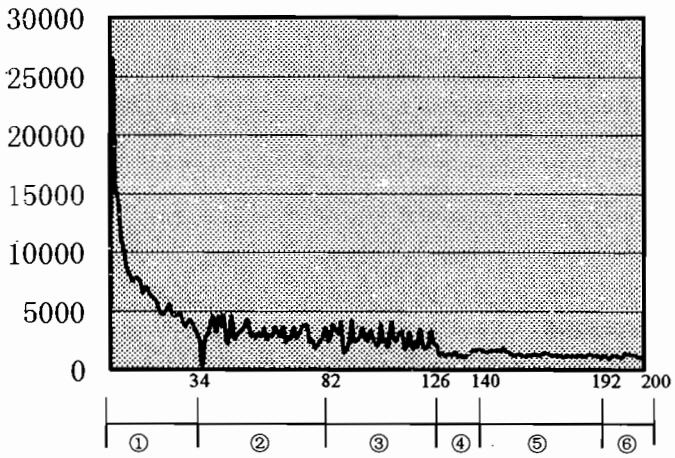

图２　接续对数量随语料库规模的增量曲线

从统计数据及以上图１和图２中分析，它们能有效地反映出真实语料中接续对的

出现规律和语料本身的某些特征。

·语料文体风格的不平衡性。

从图1、图2中知道，人民日报语料（②③）的文体风格变化很大，语料具有很大的不平衡性，而新华社语料（④⑤⑥）就比较平衡，这与新华社语料多为新闻题材文章有关。经济日报语料（①）基本平衡。选取平衡性好一些的语料是建立专业接续库的基础。

·语料内容可能有重复。

语料统计到约34兆字节时，接续对几乎出现了零增长，这表明这部分语料（人民日报语料）与前面的语料（经济日报语料）内容有重复。为了避免语料重复，可以用文本编辑工具（或程序）将这部分重复的语料删除掉。因此建立好的接续库除了要净化语料中的垃圾（关于语料净化有另文发表）之外还要消除重复的文本。

·接续对的离散性很大。

从总的统计趋势看，接续对的增长量（参看图2）除开始阶段变化较大（急剧下降）外，其它时间在缓慢的减少，而且越到后来越缓慢。可以预计，在一个有限的范围内，这个趋势将一直持续下去，增长量虽然趋缓但也保持了一个不可忽略的非零常数。这正好表明了，汉语本身具有相当的灵活性和复杂性，也说明了按照增加语料大小而不考虑语料体裁的平衡性和语料本身的垃圾等问题、不作人工干预的做法来建立完善的接续库是不现实的。

2 接续对的频率分布

在统计完成2000兆语料时，总接续对（不计重复）已达59万多个，而总次数（即计重复的接续对的个数）已达一千万个，平均重复17次。这也反映出尽管语料和接续对具有很大的离散性，但同时也具有一定的收缩性。

这里我们用接续对出现的次数表示它的频率（频度）。图3是接续对的频率分布图（横坐标表示频率从大到小的接续对序号，纵坐标表示频率，由于数据太大，本图只是一个缩略图）。

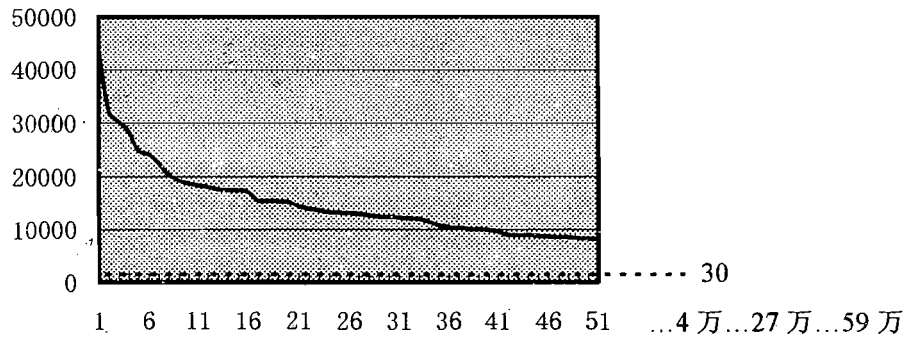

图3 接续对的频率分布图

下表1是频率排在前30位的接续对：

(1)	他们/的 43646	这/是 30022	新/的 24661
	这/一 32097	的/发展 28195	他/说 23924

本 / 报 22267	的 / 问题 17167	(2) 在 / 哪里 13387
发展 / 的 20001	进行 / 了 16918	的 / 重要 13086
上 / 的 18993	党 / 的 15256	的 / 工作 13028
的 / 是 18350	他 / 的 15251	的 / 情况 12951
经济 / 发展 18017	了 / 一个 15063	的 / 新 12513
的 / 一个 17786	工作 / 的 14998	是 / 一个 12747
的 / 经济 17336	举行 / 的 14186	地区 / 的 12209
也 / 是 17274	国家 / 的 13689	的 / 人 12207

(1) 包括“我们/的”、“你们/的”、“她们/的”；(2) 包括“在/那里”、“在/这里”

表 1　高频的接续对

从频率分布的统计数据中可以看出，(1) 频率最高的接续对多为词频较高的词条（如“的”、“了”、“是”、“他”、“上”、“人”、“也”、“发展”等）之间的接续，因此统计数据也一定程度上反映了高频词条的信息；(2) 接续对出现次数越高越可靠。接续对的频率一定程度上反映了它的可靠性程度，在图 3 中我们认为频率在 3 0（虚线）以上的接续对（约有 4 万个）是比较可靠的；(3) 在 2 0 0 兆语料中只出现一次的接续对占半数左右，反映了这些语料中字词的离散性，因而确定接续对是否可靠就具有更大的灵活性，但总的来说，随语料统计的进程新出现的接续对的可靠程度将逐渐降低。

3 接续对的可靠性

3.1 接续对垃圾的来源

在统计到的接续对中有许多的接续对是不可靠的，我们把这些不可靠的接续对叫做接续对垃圾，而把产生接续对垃圾的语料称为语料垃圾。那么，接续对垃圾是如何产生的呢？我们在统计过程中，对语料在 5 0 兆、1 0 0 兆、1 5 0 兆、2 0 0 兆时抽样考察了当时新出现的接续对和它们所在的上下文，下表 2 是统计到 2 0 0 兆语料时部分接续对和它们所在的上下文：

(a1)	块 / 达到	最高的田块到达百分之八十。	（“田块”未收录为词条）	
(a2)	级 / 并不	黄河洪水量级并不大……	（“量级”未收录为词条）	
(a3)	药 / 种	——坚持送药种到……	（“药种”未收录为词条）	
(a4)	同 / 问题	对格式化合同问题进行……	（歧义段“格式化合同”分词错误）	
(a5)	李 / 润	“雷锋模范”李润虎……	（“李润虎”未识别为人名）	
(a6)	紫 / 溪	楚雄市紫溪山东……（“紫溪”非词而“山东”是词，分词错误）		
(a7)	应 / 金	郭政民应金某某的邀请……	（“金某某”未识别为人名）	
(a8)	让 / 丁	横下一条心，要让丁烟村……（“丁烟村”未识别为地名）		
(b1)	有利于 / 肯	将在所有有利于肯中……	(b2)	关于 / 澳　中澳草签关于澳在……
(b3)	反 / 叙	则谴责以色列反叙歇斯底里…	(b4)	津 / 甘　…项目，推动津甘的……
(b5)	成为 / 菲	…百分之五十七，成为菲国民…	(c1)	步 / 当　许多人或以步当车……
(c2)	价 / 保	……元的保价保利……	(c3)	税 / 取得　依法治税取得进展……
(c4)	界 / 庆	香港米业界庆筹……	(c5)	受 / 侵　反映老人权益受侵而积极……
(d1)	防 / 手	技术细腻、防手稳固…	(d2)	北 / 滞　南濒长江、北滞黄河、…

表 2　语料在 2 0 0 兆时新出现的接续对和它们所在的上下文

从抽样得到的接续对的上下文看，不可靠的接续对产生的原因大致归纳为如下几点：

⑴ **分词错误**。这是在统计了大量语料之后新出现的接续对中出现接续对不可靠的主要原因，而分词错误的原因在于专名识别错误、收录词条不足和歧义切分错误（这是任何分词系统都难以避免的），如例子(a1)、(a2)、(a3)、(a4)、(a5)、(a6)、(a7)、(a8)；

⑵ **缩略语**。如例子(b1)中的“肯”和“中”、(b2)中的“中”和“澳”、(b3)中的“叙”、(b4)中的“津”和“甘”以及(b5)中的“菲”等是国家名称或省市名称或其它名称的简称，显然这些都是不正常的接续对；

⑶ **文言成分**。如例子(c1)中的“以步当车”、(c2)中的“保价保利”、(c3)中的“治税”、(c4)中的“米业界”、(c5)中的“受侵”等字词都是不规范或不常见的用法；

⑷ **原文错误**。很难避免语料本身的错误，这也是产生接续对垃圾的重要原因。如例子(d1)中“防手”应为“防守”、(d2)中的“滞”应为至。

从这些原因可以看出，这些垃圾产生的原因很大程度上在于汉语的特殊性。英语的专名和缩略语有词法特征，汉语却没有；文白相杂也是现代汉语所特有的问题。所以，接续对垃圾问题在英语语料库统计中可能并不突出，但在汉语中就十分严重，是汉语语料库语言学必须花大力气解决的问题。

3.2 接续对垃圾同语料规模的关系

从不同阶段的抽样得到的新增加接续对可以看出，在新增加的接续对中接续对垃圾所占比率随统计进程越来越大。在统计到５０兆时，垃圾已占５０％以上；在统计到１００兆时，垃圾约占６８％；在统计到１５０兆时，垃圾约占８０％；在统计到２００兆时，垃圾约占９０％。由此可见，随着语料库规模的增大，在新增加的接续对中垃圾逐渐占据了其中的大部分甚至绝大部分。

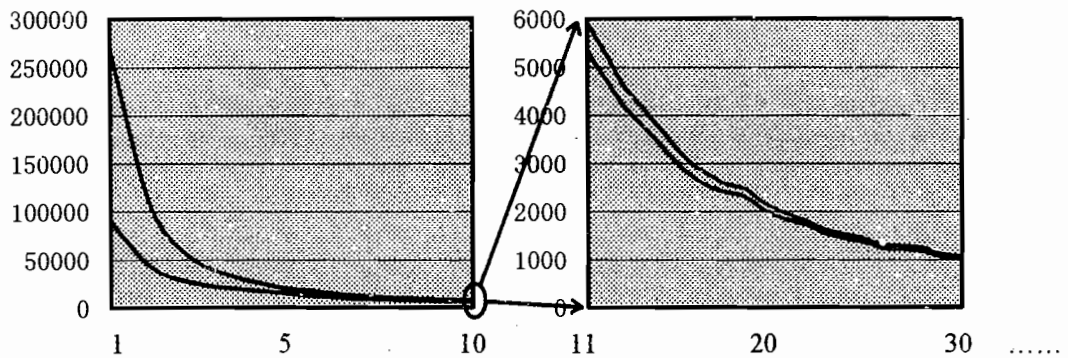

图４　接续对和可靠的接续对的变化曲线

3.3 接续对垃圾同接续对频度的关系

此外，我们还对２００兆语料中出现次数（频度）的各阶段可靠的接续对所占比率进行了统计，如图４，横坐标表示频度，纵坐标表示接续对的个数，上曲线为统计得到的接续对的个数，下曲线为可靠的接续对的个数，右图是将横坐标和纵坐标的刻

度适当调整之后的延续图形。

统计数据表明，在出现1次的接续对中，可靠的接续对约占三分之一；在出现2次的接续对中，可靠的接续对约占五分之二；在出现3次的接续对中，可靠的接续对约占二分之一；在出现10次的接续对中，可靠的接续对约占十分之九；出现30次以上的接续对基本上是可靠的（图4中上下二曲线基本重合）。

但是，在出现高频（30次以上）的接续对中，仍然有很小一部分不可靠的接续对。如接续对“埃/博”出现了196次，显然这是不可靠的，经调查它出现的上下文发现，全是由于“埃博拉”（病毒名）未收录为词条（或识别错误）所致。

3.4 接续对垃圾同特定词性的关系

为了表述的方便，我们在下面用G，H，L分别表示高频多字词、含有特定词性的单字词、不含特定词性的单字词，用D表示单字词即D＝H∪L。我们认为，在统计到的接续对中，(1) G与G之间的接续对（约占8%）是可靠的；(2) G与D、L与L之间的接续对（约占30%）基本可靠；(3) H与L之间的接续对（约占38%）的可靠程度很低，但由于其数量巨大，也不能一概否定；(4) H与H之间的接续对（约占24%）基本不可靠。

四 统计结果的提示

如何用大规模语料库建立接续库，是一个值得深入研究的课题。本文的工作给出了一些提示：

(1) 要收集到足够的接续对，必须使用大规模语料库；

(2) 即使语料库的规模很大，仍然有一些可靠的接续对统计不到；

(3) 用大规模语料库统计接续对，会统计到大量的接续对垃圾。随着语料库规模的增大，新增加的接续对中的垃圾逐渐会占大部分甚至绝大部分。垃圾主要分布在统计到的低频度接续对中，主要来源是分词中专名识别错误；

(4) 为了建立一个比较准确使用的接续库，应使用大规模语料库（100兆以上），但必须作到以下几点：① 用人工或程序对语料库事先进行清理，尽量清除语料垃圾；② 花大力气改进分词系统，特别是提高专名识别的准确性；③ 对统计到的接续对进行人工鉴别，重点是低频接续对和由专名用字构成的接续对。

本项研究得到国家自然科学基金、国家863计划、北京市自然科学基金、北京市教委开发基金的支持。

参考文献

[1] 北京语言学院语言教学研究所，现代汉语频率词典，北京语言学院出版社；

[2] 宋柔 邱超捷 欧阳龙根 徐绿兵 王鑫，二元接续关系及其在分词和校对中的应用，ICCC'96, National University of Singapore, Singapore, 96.6；

[3] 夏莹等，中文字字同现概率统计及应用，计算语言学进展与应用，清华大学出版社；

[4] 欧阳龙根 宋柔 邱超捷，汉语校对系统的功能定位，HKBJICC'97。