北语学人书系

第一辑

# 宋柔
# 语言工程论文集

宋柔 著

北京语言大学出版社
BEIJING LANGUAGE AND CULTURE
UNIVERSITY PRESS

北语学人书系

第一辑

# 宋柔
# 语言工程论文集



宋柔 著

北京语言大学出版社
BEIJING LANGUAGE AND CULTURE
UNIVERSITY PRESS

**图书在版编目（CIP）数据**

宋柔语言工程论文集 / 宋柔著. -- 北京 : 北京语言大学出版社，2012.8
（北语学人书系. 第1辑）
ISBN 978-7-5619-3333-6

Ⅰ. ①宋 Ⅱ. ①宋 Ⅲ. ①语言学－文集
Ⅳ. ①H0-53

**中国版本图书馆CIP数据核字(2012)第178243号**

书　　名：宋柔语言工程论文集
责任印制：姜正周

出版发行：北京语言大学出版社
社　　址：北京市海淀区学院路15号
邮政编码：100083
网　　址：www.blcup.com
电　　话：发行部　82303650/3591/3648
　　编辑部　82301016
　　读者服务部　82303653/3908
　　网上订购电话　82303668
　　客户服务信箱 sevice@blcup.com
印　　刷：保定市中画美凯印刷有限公司
经　　销：全国新华书店

版　　次：2012年8月第1版　　2012年8月第1次印刷
开　　本：787毫米×1092毫米　1/16　　印张：17.75
字　　数：303 千字
书　　号：ISBN 978-7-5619-3333-6／H·12115
定　　价：43.00元

# 出版说明

北京语言大学是一所颇具特色的学校。在这里，聚集了数百名语言教学和研究人员，语言学研究队伍极为庞大。近年来，随着中国语言文学和外国语言文学两个一级学科博士点的建立，中、外语言文学已然成为北京语言大学的两大支柱学科。依托这两大学科，一批学科带头人和学术骨干脱颖而出，其中有的已成为本专业领域的领军人物。在汉语国际教育、汉语研究、外语研究、语言信息处理、中国文学研究、比较文学研究等领域，北语学人已成为一支不可或缺和不可忽视的力量。

倏忽之间，北语建校已经五十周年。五十年来，代有才人。然而，学校一直未能对北语学人积累下来的珍贵的学术遗产进行系统的梳理。为弥补此缺憾，值此建校五十周年的特殊时刻，学校决定设立“北语学人书系”，收录北语优秀学人的优秀论文，每人自成一册，不定期陆续出版。因为时间仓促，本辑只约请了已退休的博士生导师和现任博士专业学科带头人，以便能赶在校庆期间见书，初步展示北语学人的学术风貌。今后，我们仍将继续组织征集优秀书稿，以“北语学人书系”的名义分辑出版，以体现北语学术的全面性和延续性。

为在短时间内完成这批高质量书稿的征集和编辑工作，校科研处做了大量的组织宣传工作，各位作者积极甄选论文、认真校对，北京语言大学出版社的领导高度重视，编辑们付出了大量辛勤的劳动，最终使第一辑书系得以如期出版。这正是北语精神的具体体现，亦当记录并彰扬也。

北京语言大学

2012 年 6 月

# 目 录

## 语言工程方法



## 语言工程技术



## 语言工程知识

# 语言工程方法

# 统计和规范中的误区*

## 一、关于统计方法

随着计算机硬件的飞速进步以及文本数量的海量增长，自然语言处理中的统计方法表现出了不可替代的优势，但盲目使用统计公式而不问其适用条件的倾向却是一种误区。

在关于自然语言处理统计方法的论文、教科书中，我们常常看到一些概率演算公式。比如，讲解汉语统计分词的原理时，概率演算过程如下：

（1）假设待切分的字串为C，问题描述为求词串W，使得P(W|C)为最大，即求：

ArgMax P(W|C)

* 本文得到国家自然科学基金课题（项目编号：60272055）和国家863计划（项目编号：2001AA114111）的资助。本文写作过程中作者曾得到中科院系统所冯士雍研究员的指教，并在同南京师范大学陈小荷教授、北京语言大学刘贵龙教授和荀恩东副教授的讨论中获益，特此致谢。

（2）两次使用条件概率公式，得到：

$$P(W|C)$$
$$=P(WC)/P(C)$$
$$=P(C|W)P(W)/P(C)$$

因此

$$\text{ArgMax } P(W|C)$$
$$=\text{ArgMax } P(C|W)P(W)/P(C)$$

（3）考虑到 P（C）在比较中不起作用，并考虑到对于任意 W 都有 P（C|W）=1，上式就简化成：

$$\text{ArgMax } P(W)$$

（4）反复使用条件概率的定义，有：

$$\text{ArgMax } P(W)$$
$$=\text{ArgMax } P(w_1w_2\cdots w_n)$$
$$=\text{ArgMax } P(w_1)P(w_2\cdots w_n|w_1)$$
$$=\text{ArgMax } P(w_1)P(w_2|w_1)P(w_3\cdots w_n|w_1w_2)$$
$$=\text{ArgMax } P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_n|w_1w_2\cdots w_{n-1})$$

（5）引入 1 阶（有时为 2 阶）马尔科夫模型假设，即假定一个词的出现仅与它左邻的一个词相关，则上式简化为：

$$\text{ArgMax } P(w_1)P(w_2|w_1)P(w_3|w_2)\cdots P(w_n|w_{n-1})$$

这里有几个问题：

（1）上面的推导中用到了 P（W）。但是，当把 W 解释做词串（通常是句子）的时候，P（W）是没有意义的，因为句子的样本空间是无法界定的。任意给定一个实用问题，我们无法找到一个关于该问题的具有足够代表性的句子集合。即使能把关于该问题的所有已经说过的句子都找出来，还有几乎无穷无尽的从未说过但可以说的句子。比如，可以把中文信息处理领域所有论文中的句子都找出来，但是本文中的句子大概绝大多数都是新句子。事实上，绝大多数的句子在我们可以得到的无论多么大的语料库中要么出现 1 次，要么出现 0 次（用基于 PatArray 的字串索引库很容易验证这一点），这种数据没有统计意义。

（2）通过引入1阶马尔科夫模型，我们把词串概率的计算归结为词的二元转移概率的乘积。如此，原来没有意义概率变成有意义且可计算的了。但是，从理论上说，引入马尔科夫模型是割断了历史，把准确的算式变成了不准确的。于是，

$P(W)=P(w_1w_2\cdots w_n)\approx c(w_1w_2\cdots w_n)/Ns$

$P(W)=P(w_1w_2\cdots w_n)$

$\approx (c(w_1)/N_w)*(c(w_1w_2)/c(w_1))*\cdots*(c(w_{n-1}w_n)/c(w_{n-1}))$

这两个算式到底应当相信哪一个呢？（其中Ns表示语料库中句子的总数，Nw表示语料库中词的总数，c（w）表示语料库中词（词串）w出现的次数。）

（3）事实上，这两个算式哪一个都不能相信，没有人这样计算句子$w_1w_2\cdots w_n$的概率。但是，用

$ArgMax(c(w_1)/N_w)*(c(w_1w_2)/c(w_1))*\cdots*(c(w_{n-1}wn)/c(w_{n-1}))$

来确定某句子的分词结果$w_1w_2\cdots w_n$，却被实践证明是有一定效果的。因此，这个式子（它说的是“字串的正确切分结果往往是导致词的转移概率乘积最大的词串”，它并不是说“字串的正确切分结果往往是概率最大的词串”）应当被看成是一个经验公式，而不是基于概率理论的逻辑推理的结果。概率演算只是得到这一公式的一种启发性步骤。

通过以上分析，就不会再对以下事实感到奇怪：对字串进行全切分，无遗漏地列出所有理论上可能的分词结果，再用上式找概率最大者，其分词准确度并不是很高；不作全切分，使用一些启发式规则滤掉一批理论上存在、事实上基本不可能存在的词串后再行处理，反而准确性会高一些。

另一个类似的例子是校对。有人企图用$P(w_1w_2\cdots w_n)>\alpha$来给正确的句子和错误的句子划界线，甚至穷举某种替代的词串，找概率最大者作为可能错句的修改建议，结果无一见诸实施。

这些例子说明，对于统计公式（包括各种其他的数学公式）不能迷信。一般来说，数学方法是一套建筑在公理基础上的严格的逻辑系统，本身并无问题。但是，现实问题是否适用某个数学方法，则取决于现实问题本身的性质同数学方法要求的条件是否匹配。当这种匹配性可疑时，就需要通过有一定规模的真实环境

下的实验去验证。用人为选择的数据做几个小规模的试验，虽然可能发表几篇论文，但最终可能还是自欺欺人。

## 二、关于语料库的统计

现在，从国家部门到课题组，很多人都在建各种各样的语料库。语料库建设确实重要，但在大规模地投入人力财力之前，若未能慎重地进行一些可行性研究，恐怕是很有问题的。

其中一个重要问题是语料库的规模得有多大才能满足统计的需要。

投掷硬币（质地不一定均匀），会出现正面朝上、正面朝下两种状态，试验几十次肯定能看出这两种状态各自的概率。

英文字母有 26 个，它们的出现是不独立的（例如 q 后面几乎都是 u，w 后面几乎不出现 u）。为统计英文字母在英文文本中出现的概率，大约需要几千字母的文本。

常用汉字有几千个，相互不独立。为统计常用汉字的概率，有百万字量级的语料库就可以了。

常用的汉语词大约有几万到十几万，相互不独立。为统计词的概率，往往需要数亿字的语料库。

为使用概率公式进行分词，需要使用词对相邻共现频次。如果词有几万到十几万，词对就有十几亿到几百亿。为了保证统计数据有一定的可信度（比如 95% 或 99%），需要多大的语料库呢？

在分词的同时进行词性标注，处理对象是带有词性的词。比起不带词性的词，数量多了几倍（对于兼 a、b 两个词类的词 w，需要把［w a］和［w b］看成两个不同的对象）。这时，为保证统计结果达到某个预定的可信度，需要多大的语料库？

树库的情况更复杂。树上的每一个词是一个叶节点，带有从根开始的路径信息（包括路径上各节点的名称及各节点通向子节点的分支序号）。各叶节点的可能的路径数大约是十几个到几百个。同样的问题是，为了保证统计结果达到某个

预定的可信度，需要多大规模的树库？

与此相关还有两个问题：

首先，满足一定质量要求的语料库，其规模不可能无限扩大。电子文本的总量毕竟有限，质地比较均匀的语料库（面向某个特定历史时期、特定地域、特定领域、特定人群、特定应用的语料库）更是规模有限。

其次，人对于语料库的加工能力也是有限的。语料库用于训练和测试，必须达到一定的精度。这种精度肯定是机器的自动加工能力所不能满足的（否则我们就不必要训练和测试了），从而需要大量的人工投入。但是，人在单位时间内加工语料的数量是有限的。随着人员的增加和时间的延长，语料加工的一致性和可靠性肯定会打折扣。而且，如果人力加工语料库的速度接近或低于语言发展变化的速度，这种加工就失去意义了。因此，在预定精度的情况下，人力加工的语料库规模恐怕是有上限的。而且加工深度越深，上限就越低。

由上可知，问题的复杂性要求被统计的语料库规模不断扩大，但语料库的本性和人的能力又使得语料库的规模不可能无限制地扩大。这样一对矛盾约束了语料库方法所能达到的性能极限。我们应当就这个问题进行基础性的研究，并在作语料库建设的规划时对这一限制有所顾及。

# 三、关于语料库标注规范

为了避免重复建设，需要制定语料库加工规范，使得语料库加工结果能为社会共享，这显然是一个好主意。但是，如何制定规范，却是不容等闲视之的问题。这里主要谈词性标注问题。

语料库建设的基本问题之一是词性标注。但这个任务的语言学方面的基础理论是有问题的。陈小荷教授（1999）有明确的论述，这里作一些具体分析。

## 3.1 兼类问题

在一些语法书（俞士汶等，1998；朱德熙，1985）中，用图 A 说明词的兼类是什么意思。如果词类 b 是具有某些属性的词的集合（图 A 中左边的圆圈），词

类 d 是具有另一些属性的词的集合（图 A 中右边的圆圈），所谓 b 和 d 的兼类词就是既具有 b 的属性又具有 d 的属性的那些词（图 A 中两个圆圈相交的部分）。比如，“自动”就是区别词与副词的兼类词：“自动步枪”中的“自动”算做区别词，但“自动发射”中的“自动”算做副词。

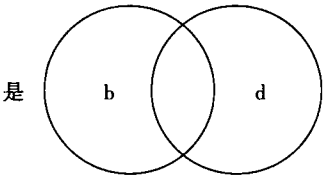


图 A

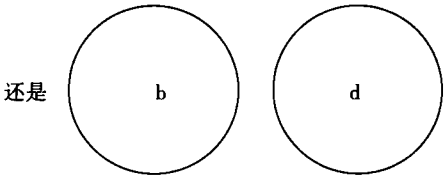


图 B

这在逻辑上是荒谬的。

按照传统语法的定义，副词是“**只能做状语（修饰动词和形容词）的词**”，而区别词是“**只在名词或助词‘的’前边出现的黏着词**”。注意，它们的定义中都有一个“只”字。

根据以上定义，可以画一个表格表示词类同属性的关系：

| 属性 / 词类 | 做状语 | 在名词前出现 | 在“的”前出现 | 带宾语 | …… |
|---|---|---|---|---|---|
| 副词 | √ | × | × | × | × |
| 区别词 | × | √ | √ | × | × |
| 副区兼类 | / | / | / | × | × |

由于副词与区别词在上表中前三个属性中表现为相斥，它们不可能有相交的部分（只要有一个表示内涵的属性相斥，其外延就不可能相交）。或者说，这两个词类的关系应当如图 B 所示，是分离的。

那么为什么又说“自动”是兼类词呢？仔细研究，发现其中作了概念偷换。副词的定义被偷换成“**能做状语（修饰动词和形容词）的词**”，而区别词的定义被偷换成“**在名词或助词‘的’前边出现的黏着词**”，都把“只”字丢掉了。上面的表格换成了下面的：

| 属性 / 词类 | 做状语 | 在名词前出现 | 在“的”前出现 | 带宾语 | …… |
|---|---|---|---|---|---|
| 副词 | √ | ? | ? | ? | ? |
| 区别词 | ? | √ | √ | ? | ? |
| 副区兼类 | √ | √ | √ | ? | ? |

如此一来，“自动”既能做状语，又能在名词前和“的”前出现，所以就兼类了。

其实，这不仅是对具体词类的定义偷换了概念，而且是对词的分类原则偷换了概念。一个词属于哪个词类是这个词的句法能力的反映，不应随它的动态出现的上下文而变（俞士汶等，1998）：

> 划分词类的依据只能是词的语法功能。所谓词的语法功能，概括地讲是指词在句法结构中的位置与分布，具体地讲是指以下两类功能：（1）词在句法结构中充当句法成分的能力，（2）词与特意选择的某类词或某些词组合成短语的能力。

“自动”之所以被看成兼类，是因为出现在“自动步枪”中和“自动发射”中的两个“自动”意义上没有什么差别，没法看成两个词，但两次出现的句法角色不同，一个是定语，一个是状语。可见，这一看法的基础是把词类划分标准偷换成动态角色，即一个词定成什么词类取决于它在具体的语境中充当什么句法角色。

如此偷换了概念，词性排歧就成了句法角色（甚至是语义角色）的认定。但是，如果目标是认定句法角色，干脆就直接标成定语和状语，何必在词性分析中标成区别词和副词，到句法分析中再费一番手脚，把它们分别改成定语和状语呢？在工程实践中，未尝不能把词处理、句法分析和语义分析放在一起做，但这样做的前提是有一个能自圆其说的理论，有一个清楚的目标，不能糊里糊涂地做一些不清不楚的事情。

注：我们这里反对流行的兼类词概念，但同时承认一些语义基本无关只是形式相同的词应看成不同的词，如“小张和小王分别了十年，现在分别在北京和上海工作”中的两个“分别”应看成两个不同的词。词法处理的一个重要任务就是区分这些“同形异质词”。

## 3.2 面向人的规范和面向机器的规范

汉语语料库加工中的词性标注使用的基本上是语言学中面向人的词类体系。如上所述，这套体系是有问题的。如果机器能理解副词和区别词的定义并能进行推理，它肯定说“自动”既不是副词也不是区别词，而属于另一种词类。

面向人的词类体系不仅给语料库标注带来明显的矛盾，还带来了大量难以摆脱的缠结。

以下是某语料库词性标注的结果：

（1）我国/n 安装/v 信函/n 自动/d 分拣/v 设备/n 的/u 城市/n

（2）建构/v 起/v 产品/n 结构/n 的/u 自动/d 调整/v 机制/n

（3）电脑/n 控制/vn 的/u 音响/n 便/d 自动/d 停止/v 扩音/n

（4）率先/d 推出/v 全自动/b 除/v 霜/n 功能/n

（5）世界/n 上/f 先进/a 的/u 卫星/n 自动/b 校时钟/n

（6）右边/f 是/v 自动/b 开关/n 的/u 电热水壶/n

（7）电子/n “/w 警察/n”/w 就/d 是/v 指/v 市/n 交通/n 监控/vn 中心/n 系统/n 和/c 冲/v 红灯/n 自动/b 照相仪/n

以上7个例子中，前3个例子“自动”标为副词d，后4个例子“自动”标为区别词。标注的依据是“自动”修饰的是动词还是名词。但是，这个问题实际上常常是说不清楚的。比如，第五句，“校时钟”是名词，所以“自动”是区别词。但是，严格地说，这个例子中的短语结构应该是：

（（卫星（自动 校时））钟）

即用卫星来自动校时的钟。这样看，“自动”是动词状语，在这里应当被标为副词。

最后的例子中，“自动”修饰的是“照相”还是“照相仪”，这是一个答案模糊的问题，因而“自动”到底算状语还是定语也就难以确定了。

在这些例子中，另一个问题是关于动词v和名动词vn的区别：

（1）信函/n 自动/d 分拣/v 设备/n

（2）自动/d 调整/v 机制/n

（3）电脑 /n　控制 /vn　的 /u　音响 /n

（4）交通 /n　监控 /vn　中心 /n

上面的斜体词为什么有些是 v 有些是 vn？其中有的可能是误标，有的可能有一大套理由或硬性规定。无论如何，这样的标注体系恐怕很难说清为什么这种标注方法比别的方法好，并很难在标注中保持一致性，从而很难说能对语言现象统计分析有多大的积极意义。

语料库词性标注的大部分精力，大概都花到这种很难说有什么意义的“排歧”中去了。那么，为什么许多人的语料库加工还要采用这一套自相矛盾、纠缠不清的体系呢？有一种说法是，面向机器的规范应当同面向人的规范一致，以便人机共享。

真的应当这样吗？机器和人的能力、工作方式非常不同，二者的规范应当不同才对。例如，人的识别，面向人的时候采用姓名，因为有助记忆；面向机器的时候采用身份编码、账号等，以便互不混淆。汉字输入码和机内码要加以区分也是这个道理。

在面向人进行语法教学时，由于人的“心领神会”的能力，概念不确切并无伤大局，重要的是不把词类搞得太多，否则不便记忆。但对于机器就不同了。计算机的机械式记忆和匹配的能力远远超过人，多一些词类对于计算机来说毫无问题。因此，设计语料库词性标注规范时，可以摆脱旧有的词类体系，完全按照机器处理的需要，设计新的体系。当然，面向人的词类体系中一些合理的成分应当保留，不适合机器处理的部分（逻辑上不一致的，意义上不确切的，对于上层应用没有意义的，等等）则应当大胆破除。举例如下：

（1）可以增加新的词类。比如，“自动步枪”和“自动发射”中的两个“自动”可以看成属于一种新的词类。它的分布特点是能且只能充当如下句法角色：在动词前用作状语，在名词前用作定语，后面加“的”用作主语或宾语。与它同类的有“必然”、“长期”、“定期”、“非法”、“高速”、“共同”、“临时”、“书面”、“双向”、“永久”、“主要”、“专门”等。

（2）可以考虑另外一些属性作为划分词和词组类别的标准。比如，“忽然”是副词，“突然”是形容词，只是强调“忽然”不能做谓语、不能受“很”修饰。