Chap1. 为什么需要计算思维

学习目的：
（1）为什么计算机能解决客观世界的问题
（2）怎样让计算机解决传播问题
（3）传统的传播与智能传播的区别
（4）汉字为什么能传播 3000 多年

0. 人工智能问题的需求背景

形式系统内的计算：

- 数字控制（数控机床，打印机，计算机操作系统）
- 数学、理论物理中的符号计算（符号微分、符号积分）
- 图形计算
- 物理世界的科学计算问题（数值计算），如流体力学、固体力学的微分方程数值解，用于航空、航天、气象预报等的计算。
- 信息系统中的数据加工（办公室自动化，电子政务，电子商务，税务，企业管理）

客观世界问题的计算（需要首先上升到形式系统，结果要返回客观世界去检验）：

- 各种工程模型（航空中的空气动力学模型，地震能量释放模型，气象预报模型）
- 信息系统中的数据挖掘（规律发现）
- 经济状态和趋势估计（股市，保险，消费，产值）
- 资源保护
- 环境保护
- 图像识别
- 语音识别
- 音乐的识别和理解
- 语言理解
- 语言翻译
- 生理健康
- 心理健康
- 交通流量分析和应对

非形式化问题需要抽象出属性特征及条件，形式化以后，设计算法和程序

这类问题的特点是：

- 有高智能的成份，超出了计算机的能力。
- 有机械化成份，用得上计算机。
- 答案不惟一，评价指标缺乏。

当计算机大量普及，计算机的软硬件性能大大提高以后，社会发展对于计算机的应用就提出了越来越广泛深入的需求，希望计算机能协助解决这类客观世界的问题。

下面具体分析计算机的能力。

1. 人工智能学科的特点

1.1.两类问题

1.1.1.典型计算问题

1.1.1.1. 典型计算问题的特征

可以形式化，可以算法化，计算复杂度不高

1.1.1.2. 概念解释

（1）形式化：

- 有限个符号的集合
- 符号串的构成规则
- 待解问题各状态的表示方法（包括起始状态、停止状态）
- 待解问题的解的表示方法
- 待解问题不同状态之间变换的规则

（2）算法化：

- 确定性：待解问题处于任意一个非停止状态时，最多能找到一个规则，使该状态变到下一个状态
- 可停止性：待解问题从初始状态开始，在有穷次使用规则进行状态变换后，一定能达到停止状态。

（3）算法复杂度不高：

算法复杂度指的是规则使用次数与每条规则计算量的乘积同待解问题规模之间的函数关系。如果每条规则的计算量是恒定的，则算法复杂度就是规则使用次数同待解问题规模之间的函数关系。这个关系不能是指数级的，2 次以上方幂级的也不好，最好是线性的。否则，当待解问题规模相当大的时候，计算机就无法承受该问题的时间和空间的开销。

形式系统中的问题是已经形式化了的。

如：数值计算，符号计算，逻辑演绎

1.1.1.3. 例：自然数加法

（1）形式化：

- 有限个符号的集合：0，1，2，3，4，5，6，7，8，9
- 符号串的构成规则：涉及符号串〈加数符号串 1〉、〈加数符号串 2〉、〈和数符号串〉，都是非 0 符号起始的任意有穷长的符号串。
- 待解问题的状态的表示方法：
  (〈加数符号串 1〉，〈加数符号串 2〉，〈操作位置〉，〈进位数〉，〈和数符号串〉)，其中〈加数符号串 1〉和〈加数符号串 2〉为非空符号串，〈操作位置〉是正整数，〈进位数〉为 0 或 1。起始状态中〈操作位置〉为 1，〈进位数〉为 0，〈和数符号串〉为空。当
  〈操作位置〉> max(|〈加数符号串 1〉|，|〈加数符号串 2〉|)
  时过程停止。
- 过程停止时如果〈进位数〉为 0，则〈和数符号串〉为解，如果〈进位数〉为 1，则〈和数符号串〉左边添上符号 1 为解。
- 待解问题不同状态之间变换的规则：

条件				动作		
〈加数符号 1〉	〈加数符号 2〉	〈操作位置〉	〈进位数〉	〈和数符号〉	〈进位数〉	〈操作位置〉
符号 0-9	符号 0-9	正整数	布尔	符号 0-9	布尔	正整数
0	0	n	0	0	0	n+1
0	0	n	1	1	0	n+1
7	8	n	0	5	1	n+1
7	8	n	1	6	1	n+1

9	**9**	**n**	**0**	**8**	**1**	**n+1**
9	**9**	**n**	**1**	**9**	**1**	**n+1**

注：

a. 〈加数符号 1〉=getchar(〈加数符号串 1〉,〈操作位置〉),〈加数符号 2〉=getchar(〈加数符号串 2〉, 〈操作位置〉)，其中 getchar(string, n)的值是符号串 string 右起第 n 个符号，如果 n 大于 string 的长度则 getchar(string,n)值为 0。

b. 新状态的〈加数符号串 1〉和〈加数符号串 2〉与老状态相同。

c. 新状态的〈操作位置〉和〈进位数〉为上表所定。

d. 新状态的〈和数符号串〉为 putchar(〈和数符号〉,〈和数符号串〉)，其中 putchar 的参数〈和数符号串〉取自老状态，putchar(s, string) 的值是把符号 s 加到符号串 string 左端得到的新符号串。

e. 假定 getchar、putchar 和加一运算为系统的先天能力。

（2）算法化：

- 当待解问题处于任意一个非停止状态时，只能找到唯一的规则，使得规则中〈加数符号 1〉为 getchar(〈加数符号串 1〉,〈操作位置〉),〈加数符号 2〉为 getchar(〈加数符号串 2〉,〈操作位置〉),〈进位数〉为当前状态中的〈进位数〉。将这条规则施用于这个状态，就可以得到下一个状态。
- 由于每使用一次规则〈操作位置〉都加 1，〈加数符号串 1〉和〈加数符号串 2〉又都是有穷长的，所以有穷次使用规则，一定会使〈操作位置〉从开始的 1 增大到〈加数符号串 1〉和〈加数符号串 2〉的最大长度加 1，从而达到停止状态。

（3）算法复杂度不高：

规则使用次数等于〈加数符号串 1〉和〈加数符号串 2〉的最大长度，每条规则的使用只涉及几个基本操作。

1.1.2.智能计算问题

1.1.2.1. 智能计算问题的特征

不能形式化，或能形式化不能算法化，或能形式化、算法化但计算复杂度太高

1.1.2.2. 实例

（1）不能形式化的问题

客观世界是无法完全形式化的。

自然语言处理不能形式化。因为自然语言的语法是模糊的、不断变化的，语义和语用反映的是整个客观世界和主观世界，而客观世界和主观世界所不包，无法形式化。

以拼音汉字转换为例。

拼音汉字转换问题是：输入一个拼音串，得到对应的汉字串，该汉字串应当是输入者心目中想要的。

例 1 jin tian lai shang ban de shi shi xi sheng

可能转换成

今天来上班的是实习生

今天来上班的事实习生

今天来上班的时实习生

其中后两句语法是错误的。

例 2 ci shi bu hao ban

可能转换成

此时不好办
此事不好办
这两句的对错完全取决于语言环境。
例 3 bei jing quan ju de kao ya dian
可能转换成
北京全聚德烤鸭店
北京泉居得烤鸭店
这个例子的对错取决于知识，无法由机器判断。
例 4 wo zai chuan bo xi xue xi
可能转换成
我在传播系学习
我在船舶系学习
这个例子的对错取决于事实，更无法由机器判断。
又如词汇语义问题：
抓革命：投放力量于革命运动。
抓反革命：拘捕反革命分子。
理论上说，不可能把所有问题形式化，因为如果要把所有问题形式化，就得把“如何形式化”的问题形式化，于是首先要把系统的一切可能的输入都形式化，即把客观世界的万事万物形式化，这是不可能的。
（2）不能算法化的问题
定理证明遵循一个形式系统，其中包括公理、定理、推理规则。可以构造一种方法，使得任何一个可证的命题能在有限步之内被证出，但是无法构造出一种方法，使得不可证的命题在有限步之内被发觉是不可证的。也就是说，当证明过程不断进行下去的时候，无法预知它是否能停下来。
（3）计算复杂度太高的问题
围棋有 361 个落子的位置，每个位置上可能是黑子，可能是白子，也可能是空的，而理论上说黑子与白子的数量应当相等或差 1 个。于是，围棋可能的棋局数目是

$$\sum_{m=0}^{361} C_{361}^{m} C_{361-m}^{[\frac{361-m}{2}]}$$

其中 m 是空位置数。做一个粗略的估计，上式肯定大于 $\sum_{m=0}^{361} C_{361}^{m} = 2^{361} \approx 10^{108}$。假如计算机 CPU 的主频达到了 10G，而且一个节拍就能分析一个棋局，那么要分析这么多棋局，至少需要 10^{90} 年。

1.2. 人工智能学科的矛盾
迄今为止所有计算机，无论是已经商业化了的还是仅仅出于实验室阶段的，本质上都是一种机械装置，只能处理典型计算问题。
研究如何用计算机解决计算机不可能解决的智能计算问题，这就是人工智能学科的目标，也是人工智能学科内部的基本矛盾。正是这个矛盾推动着这一学科的发展。
1.3. 人工智能学科的基本方法
解决矛盾的关键思想是：并不是建立单纯的计算机系统，而是建立人机结合系统。我们并不追求用计算机完满彻底地解决智能计算问题，而是让人参与进去，追求一种实用的人机

结合的解决方案。这个思想包括两方面的含义：
（1）限制原问题的范围，在该范围内该问题变成典型计算问题，超出该范围的部分由人另想办法处理。这种限制应当只涉及低频的现象，从而对经常性的应用基本上没有妨碍。
典型实例是汉字的键盘输入。非受限汉字的键盘录入系统是不可能实现的，因为古今中外的汉字的数量极其巨大，而且在古籍中有无数自造的字，笔形笔画变化无穷。但是，现代人使用的汉字十分受限，几千个汉字可以覆盖 99.9%的汉字出现。于是，制定了国家便准和国际标准汉字集，集合中汉字数量从几千字到几万字不等。由于是受限的，集合中全部汉字可以一一列举，故可以用编码方式解决键盘录入问题。
（2）限制原问题的功能，从原问题中尽量剥离出典型计算成分让计算机去做，剩下智能计算的核心部分留给人去做。
典型实例：
文本编辑：
写文章要求准确鲜明生动地表达某种思想。计算机无法完成这一任务。
分析写文章的过程，抽象归纳成 7 种操作，其中 5 种是基本操作。

- 插入：位置，字符串。
- 删除：位置，长度。
- 复制：源位置，长度，目标位置（往往缺省）。
- 粘贴：源位置（往往缺省），目标位置。
- 查找：字符串。
- 替换：删除+插入。
- 移动：复制+删除+粘贴。

人给出命令及其参数，计算机执行，人完成辅助动作（换行）。（1970 年代，DOS 系统）
人操作鼠标和键盘，以自然方式给出命令，计算机不但执行命令，而且完成辅助操作，达到“所见即所得”（what you get what you see）的效果。（1980 年代，Windows 系统）
将格式从内容中完全分割出来，有专门的格式设置和调整命令，并可以多任务并发。（1990 年代，Windows 系统）
人机界面更加自然，舍弃键盘，可以在纸上写，字符形式失去了同一性，计算机无法执行查找，但其他功能仍可以执行。（2000 年代，数字墨水）

拼音汉字转换。设要输入 10 个汉字的句子。在 6000 个汉字的集合中，10 个汉字的不同字串的个数是 6000 的 10 次方，这是不可想象的巨大数字。考虑到要符合输入的读音，假定平均一个音节对应 15 个汉字（汉语普通话有大约 400 个音节，相对于 6000 汉字的集合，故平均每个音节 15 个汉字），则符合要求的汉字串数目减少成 15 的 10 次方，降低了 400 的 10 次方倍。进一步统计大规模语料中汉字的邻接关系，真正可能的汉字串常常只剩下两三个。如此大规模地降低转换结果的数量，靠的是计算机的强大的匹配和搜索能力。但是，最后这两三个字串中的选择，需要依靠语法语义语用的分析，依靠对知识的掌握和对事实的了解，依靠对于输入者意图的猜测，这些是计算机无法做的，于是让输入者自己去选择。如此，最大限度地发挥了计算机和人各自的长处，避开了各自的短处，成功地解决了这样一个智能计算问题。
实际上，几乎所有的现实世界中的计算问题本质上都是智能计算问题。我们之所以能在习惯上把它们看成典型计算问题，往往是在范围和功能上限制的结果。比如清点人数，绝大多数情况下是简单的计数问题，但在产房中或在坠机灾难现场，在“人”的概念不清楚的情况下，清点就成问题了。又如长度测量，如果不是允许一定的误差，什么现实物体的长

度也测不了。
物理、化学等都在理想状态下研究，只是理想状态与实际状态常常十分接近。
如：理想气体要求气体分子的大小以及气体分子间的相互作用忽略不计，
力学研究中忽略速度的影响，才有牛顿定律。
力学应用中许多情况忽略空气阻力。
气象预报的计算采用几百个变量，是对于客观天气现象的一种近似。
同在人造的数学系统中不一样，对于现实世界的问题往往不可能求出全部解、精确解、最优解，能求得大部分情况下的解、近似解、较好的解也就行了。这是我们解决现实世界问题的态度，也是解决人工智能问题的态度。
人工智能方法的最高境界在于：
恰当地限制问题的范围和功能，使得被忽略掉的成分对于实用来说影响最小最小，被减少的功能对于人来讲可以轻而易举地完成，而解决留下来的成分和功能则最能发挥机器的长处。一句话，就是最恰当地找出人和机器各司其职的分界线，从而设计出一个效率最高的人机结合系统。
2. 教学目标
树立一种观念和思想方法，即为解决智能化的问题，需要建立一个适当的人机结合系统。
训练形式化的能力（如何把非形式化的真实世界问题变成形式化的问题）。
学习解决一般性的实际问题的计算方法。
3. 非信息科学类专业的学生学习人工智能应达到的目标
（1）了解计算机的能力及其局限性，不要偏颇。
（2）对于涉及本专业领域的问题，应当有充分的洞察和分析，尽量发现机器可以完成的成分，同时又避免让机器做不可能见效的研究。
（2）在本专业领域内，能遵照已有的形式系统把机器工作中需要使用的数据、规则等知识提炼出来。
（3）在本专业领域内，能通过实例的试验，人工验证已有算法的正确性。

作业：
（1） 在你熟悉的专业领域范围内，举出典型计算问题和智能计算问题的例子。
（2） 举例说明一个智能计算问题如何被一个人机结合系统所解决。
参考书
蔡自兴，徐光祐：人工智能及其应用，清华大学出版社，1996 年 5 月第二版
陆汝钤：人工智能（上册），科学出版社，1995 年 8 月第一版
陆汝钤（主编）：知识科学与计算科学，清华大学出版社，2003 年 1 月第一版
Principles of Artificial Intelligence, Nilsson,N.J.,（机械工业出版社出了影印本和中译本《人工智能原理》）
Artificial Intelligence, Winston, P.H., Second Edition, Addison-Wesley, 1984 （第一版有中译本，《人工智能》，科学出版社）

补充：

人机结合系统示例

一、字音转换（多音字的标音）
1. 多音字问题

单音字：
我，你，他，唱，跑，跳，
多音字：
传（chuan2）达，传（chuan2）球，传（zhuan4）记、阿 Q 正传（zhuan4）。
常用的读音是 chuan2。
开会（hui4），会（hui4）面，会（kuai4）计
常用的读音是 hui4。
单音词：
传达，传记，开会，会计
多音词：
传（chuan2）了一个球，写了一个传（zhuan4）
原子和分（fen1）子，积极分（fen4）子
绝大多数词是单音词。

2. 多音字标音策略

（1） 字的高频音

每个多音字都用它的最常用读音即高频音来标注。此时系统开销最小，但准确率较低。

（2） 词中字的高频音

多音字在特定的词中读音多数是唯一的。
因此建立词表，对文本分词，对单音词中的多音字可以自动给出正确的音，对多音词中多音字标注为该字的高频音。

（3） 根据上下文来确定多音词中的多音字读音。

规则方法：专家针对每个具体的多音词具体总结出一批批规则，为系统自动标音提供依据；
统计方法：收集多音词的各种上下文，人工标注其在不同情况下的读音，将标注结果作为训练材料，使系统在实际使用中能根据上下文对多音词的读音做出判断。
规则和统计两种方法在分词的前提下均可以提高标注准确率，但需要消耗人力来总结规则或准备训练语料，而且这些规则或训练得到的数据还要消耗存储空间。
人机分工
人：
静态工作：
确定多音字集合，给出每个字的所有可能读音，确定每个多音字的高频音
建立词表，在词表中区分单音词和多音词，对词表中单音词中的多音字标音
建立多音词例句语料库，对例句中多音词中的多音字标音
动态工作：
对于机器自动标注的多音词中多音字的音，人工判定并修正。

机：
对单音词中的多音字直接标音
对多音词，
如有例句语料库，按统计训练的结果标音
否则标多音词中的高频音

不定因素：词表规模，进行训练的多音词规模，每个多音词的例句规模

多音词分类

高频音所占比例	代表多音词

50-70%	为 重 长 行 地方
70-90%	只 处 传 种 得
>90%	着 还 中 分子 少

语境约束方法实验结果

多音词	高频音	低频音	高频音所占比例%	语境约束方法		
				训练规模（句数）	测试规模（句数）	准确率%
为	wèi	wéi	52.33	8934	4321	93.57
重	zhòng	chóng	65.99	7991	3849	95.19
长	cháng	zhǎng	65.57	10064	4624	95.50
行	xíng	háng	55.61	8452	3325	93.56
地方	dìfāng	dìfang	62.23	8702	4271	95.13
只	zhǐ	zhī	72.05	25564	3753	99.09
处	chù	chǔ	82.15	10953	5110	98.32
传	chuán	zhuàn	86.75	7897	1170	98.55
种	zhǒng	zhòng	89.83	9135	3502	97.40
得	de	dé děi	77.25	8831	4405	94.94
着	zhe	zháo zhuó zhāo	97.48	20539	3567	99.30
还	hái	huán	90.79	9181	4169	99.06
中	zhōng	zhòng	99.52	9066	4182	99.71
分子	fènzǐ	fēnzǐ	93.57	7621	2785	97.88
少	shǎo	shào	93.90	6057	2949	98.17

总体性能代价评估表

人力需求(人天)	存储要求（MB）	字音转换准确率（%）	词表规模（万）
0	0.1	92.53	0
	1.8	97.46	5
	2.5	97.56	10
	3.3	97.57	15
	4	97.58	20
10	3.4	99.02	5
	4.1	99.12	10
	4.9	99.13	15
	5.6	99.14	20
20	4.6	99.27	5
	5.3	99.37	10
	6.1	99.38	15
	6.8	99.39	20
50	7.3	99.35	5
	8.0	99.45	10
	8.8	99.46	15
	9.5	99.47	20

70	9.6	99.40	5
	10.3	99.50	10
	11.1	99.51	15
	11.8	99.52	20
80	10.8	99.43	5
	11.5	99.53	10
	12.3	99.54	15
	13.0	99.55	20
100	12.9	99.45	5
	13.6	99.55	10
	14.4	99.56	15
	15.1	99.57	20
140	17.1	99.47	5
	17.8	99.57	10
	18.6	99.58	15
	19.3	99.59	20
180	20.6	99.48	5
	21.3	99.58	10
	22.1	99.59	15
	22.8	99.60	20
290	31.9	99.50	5
	32.6	99.60	10
	33.4	99.61	15
	34.1	99.62	20
1220	123.3	99.52	5
	124.0	99.62	10
	124.8	99.63	15
	125.5	99.64	20

准确率对比图

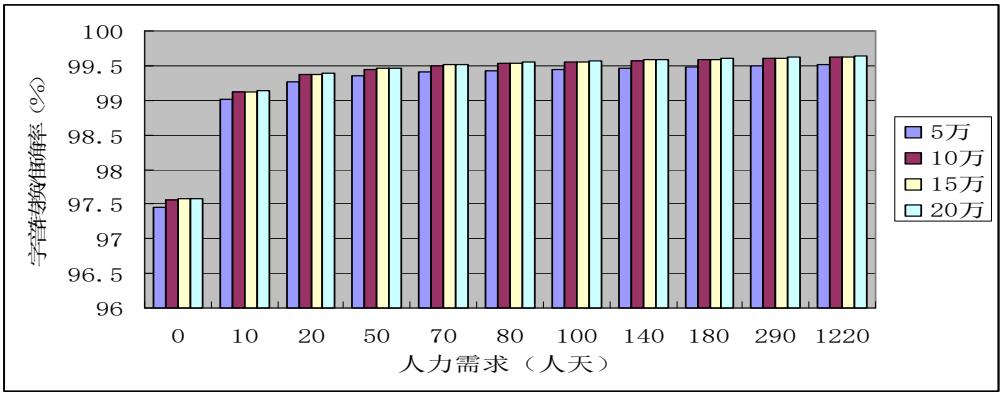
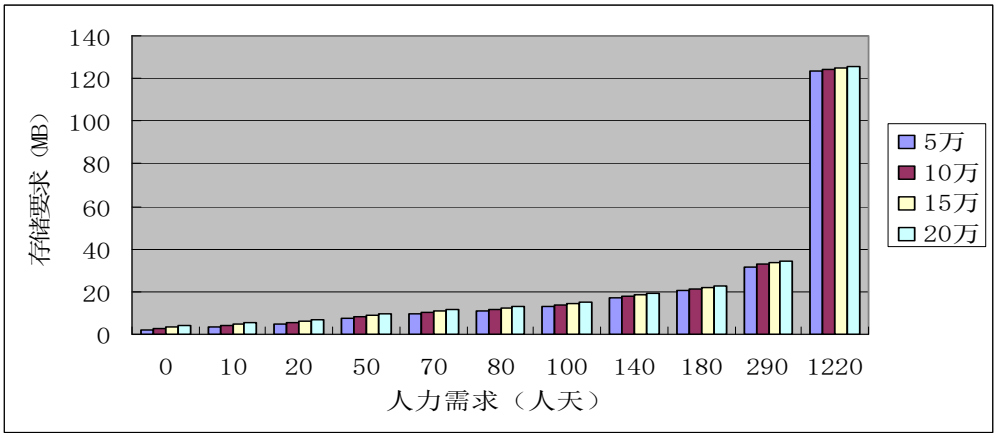


图 2　存储要求对比图



二、文本校对

1. 背景

传统校对的任务：

校异同（对照打印稿同原稿）

校正误（脱离原稿，检查打印稿中的错误）

计算机文字处理系统中排版的任务：

校正误。

（1） 词语错误

（1.1）别字和非规范字

例 1.1. 元霄节

正确：宵，错误类型：同音别字

例 1.2. 暗渡陈仓

正确：度，错误类型：同音别字

例 1.3. 以後怎麼办

正确：后 么，错误类型：繁体字

例 1.4. 买鸡旦

正确：蛋，错误类型：同音别字，废止的简化字

（1.2）漏字，多字，错字

例 1.5. 北京国安足球俱部

正确：俱乐部，错误类型：漏字

例 1.6. 科研工作的的成果

正确：的，错误类型：多字

例 1.7. 用水果刀刺中对方

正确：刺，错误类型：近形错字

例 1.8. 新型的睡眠伟感器

正确：传，错误类型：近形错字

（2）句法错误

例 2.1.通过这次活动，使大家受到了很大教育。
正确：通过这次活动，大家受到了很大教育。或 这次活动使大家受到了很大教育。

错误分析：符合错误模式“通过……，使……”，其错误性质是缺主语。

例 2.2.李明洗完衣服。
正确：李明刚洗完衣服。或 李明洗完衣服了。

错误分析：这句话是典型的“主语-动词-宾语”结构，语义搭配也完全正确，但仍然站不住。

例 2.3.一分到工厂，她就被一排排的热处理炉所吸引了。
正确：一分到工厂，她就被一排排的热处理炉吸引住了。

错误分析：“所+V”是一种合法的短语形式，但 V 后不能有“了”。

例 2.4.他写了一篇论文，他获得了一等奖。
正确：他写了一篇论文，获得了一等奖。

错误分析：第二句的主语必须省略。

例 2.5.他总是想起父亲的那句话，就想大哭一场。
正确：他总是一想起父亲的那句话，就想大哭一场。

错误分析：“一……就……”是搭配使用的副词，但“就”的用法很活，很多情况下单独使用并不错（他小学毕业没多久，就去山里修路了。）。

例 2.6.出发的时正下着小雨。
正确：出发的时候正下着小雨。或 出发时正下着小雨。

错误分析：“时候”是名词，“时”一般用作时间方位词，前面不能加“的”。

例 2.7.最近几乎世界各大公司成群结队参加在北京举办的国际汽车展览会。
正确：最近，世界各大汽车公司几乎都成群结队地来北京参加国际汽车展览会。

错误分析：这个例子中的语法错误不止一处，难以用规则的方式表示出错误所在和修改方法。

（3）语义错误

例 3.1.建没有中国特色的社会主义。

正确：建设有中国特色的社会主义。

错误分析：形近的错字引起语义错误。

例 3.2.他宁可不加班，也要多休息。

正确：他宁可不要加班费，也要多休息。

错误分析：连词意义错误。

例 3.3.董事会有罢免经理的权利。

正确：董事会有罢免经理的权力。

（4）标点错误

例 4.1.《人民日报发表社论

正确：《人民日报》发表社论

例 4.2.他不知道上不上课？只好去问同学。

正确：他不知道上不上课，只好去问同学。

例 4.3.只有深化改革。我们大家才有出路。

正确：只有深化改革，我们大家才有出路。

（5）知识性错误

例 5.1.河北省青河县

正确：河北省清河县

错误分析：青河县在新疆，清河县在河北。

例 5.2.南美洲有袋类进化后形成的食肉种群称为鬣狗类。

正确：南美洲有袋类进化后形成的食肉种群称为袋鬣狗类。

错误分析：鬣狗同袋鬣狗是两类不同的动物。

例 5.3.1937 年“九一八”事变

正确：1931 年“九一八”事变

错误分析：时间和事件不相配。

例 5.4.全国人大常委会副委员长许家璐

正确：全国人大常委会副委员长许嘉璐

错误分析：人名错误

（6）事实性错误

例 6.1.他在路上遇见一个河南人。

正确：他在路上遇见一个海南人。

错误分析：地名混淆。

例 6.2.海河下游 80 个县 540 万亩农田被淹。

正确：海河下游 80 个县市 450 万亩农田被淹。

错误分析：数字混淆。

（7）格式错误

如各级标题的字体字号，段落开始处缩进的深度，目录、脚注、参考文献的格式，科技公式中各类字母的字体等等。这类错误需要结合排版系统中格式的表示法来检查。

2. 汉语校对系统的策略

（2.1） 词的接续关系模型

L 是语言中合法句子的集合。

$L=\{W_{i_0}W_{i_1}\cdots\cdots W_{i_n} \mid (W_{i_k}, W_{i_{k+1}}) \in R, (0 \le k \le n-1)\}.$

R 的实际意义是：$(W_i, W_j) \in R$，当且仅当 W_i 和 W_j 在大规模语料库中邻接，邻接的顺序是 W_i 在左、W_j 在右。我们把这种关系称为二元接续关系，简称为接续关系。

因此，一个词串是合法句子，当且仅当任何相邻的两个词具有接续关系。至于任意给定的有序词对是否具有接续关系，则可以从大规模语料库中统计得来。

根据这个语法，校对系统的工作方式是：把所给文章看作一个词串，如果其中有两个相邻的词并不具有接续关系，则对这两个词给出警告，如果所有相邻的词都具有接续关系，则给予通过。

关于任意两个词 W_i 和 W_j 的接续关系，有两种评价函数。一种是看语料库中 W_i 和 W_j 的接续次数，即对于预定的阈值 θ，是否

$f(W_i, W_j) \ge \theta.$

采用接续关系的文法作为校对系统的语言模型，其原因是：

- 简单可行。所需要的软件工作只是分词和词的接续次数统计。虽然分词准确率达不到百分之百，但可以达到 98%左右，对于校对系统来讲已是基本上可以使用了。分词的速度可以达到每秒几十万字，已经可以满足实用要求。
- 基本反映实际情况。以汉语为母语的人，其文章的电子文档中的错误，主要不是语义错误或句法结构错误，而是因为各种偶然因素导致个别的词语错误。表现在形式上，就是有错误的词同其左右两边的词不具有接续关系。比如：“元宵节”错成“元霄节”，后者经过分词后切分成“元/霄/节”，但通过大规模语料的统计，发现“元”同“霄”、“霄”同“节”都不具有接续关系，所以系统对于“霄”给出警告。再如，“科研工作的的成果”切分后进行二元邻接关系检查，发现两个“的”不能邻接，于是对于“的的”给出警告。

采用这个方法至少需要解决以下两个问题。

- 要有一个准确高效的分词系统。分词不准确就会使计算机自动学到的接续知识的可靠性大打折扣。比如，“李晓华来了”这句话应当切分成“李晓华/来/了”，计算机将统计到人名同“来”的一次接续，“来”同“了”的一次接续。但如果计算机不能识别“李晓华”是人名，就会把这句话切分成 5 个单位：“李/晓/华/来/了”，从而“李”和“晓”、“晓”和“华”、“华”和“来”各有一次接续。错误切分的情况多次发生，将使计算机把许多“接续垃圾”当作正确的知识学进去。当然，绝对准确的分词系统是不存在的，也就是说“接续垃圾”是免不了的，因此必须有一道清理垃圾的工序，保证计算机学到的接续知识有较高的可靠性。此外，分词系统应当不断改进，以便尽量从源头杜绝“接续垃圾”。这就又对分词系统提出了高效率的要求。因为每一次改进了分词系统以后，都需要让计算机对巨大规模的语料库重新分词，以便重新学习接续知识。如果语料库的规模是几亿字，计算机的分词速度是每秒几万字，则完成一次学习所用的时间是几个小时；但如果计

算机的分词速度是每秒几百字，则完成一次学习所用的时间是十几天甚至几十天，从而使校对系统开发者无法承受。

- 要解决统计数据稀疏的问题。自然语言的特点之一是词汇多而大多数词汇的出现频率很低。统计资料说明，汉语的几十万个或几万个词中，词频超过万分之一的词只有 1300～1400 个。也就是说，绝大多数词的频率低于万分之一，因而它们发生接续的频率一般来说会低于亿分之一。或者说，在一个 2 亿字规模（约 1 亿词）的语料库中，绝大多数词的接续知识是统计不到的。如“对别人要求严，对自己要求宽”中，“要求”和“宽”在 7 年人民日报全部内容工程的语料库中只接续 3 次，但我们的阈值θ设为 10，于是“要求宽”被警告。又如，“要正确处理敢闯”中，“处理”同“敢”在上述语料库中接续次数为 0，所以“处理敢”被警告。

（2.2）词与词类混合的接续关系模型

针对校对系统的特点，解决统数据稀疏问题的方法是：

中低频词用词类代替词本身统计接续次数。例如：：“要求明确”、“要求简单”、“要求彻底清查”、“要求切实执行”等，都是“要求”同形容词接续的例子。这类例子很多，所以“要求”和形容词能接续，“要求宽”不会被警告。但“要求宽”是错的，应当是“要求宽一些”或“要求宽大”

（2.3）错误驱动的反面规则

权利，右环境：机关→权力

既，右环境：使 s /，也不→即

的，左环境：形容词，右环境：认为→地

通过 s /，使→通过 s /,

青河县，左环境：河北省→清河县

嘉靖，左环境：清朝→嘉庆

渡，左环境：暗，右环境：陈仓→度

这种反面规则的一般形式是：

规则::=〈错误词串〉[,〈左环境说明〉][,〈右环境说明〉]→〈正确词串〉

〈左环境说明〉::= 左环境：〈词串〉

〈右环境说明〉::= 右环境：〈词串〉

〈错误词串〉::=〈词串〉

〈正确词串〉::=〈词串〉

其中左环境说明或右环境说明可以不出现。若不出现，表明这一规则与环境无关。此外，s 表示不含句末标点符号的词串。

这种规则不仅指出了错误发生的条件，而且给出了修改方案。

（2.4）特殊规则

特殊规则是基于特殊的语言学知识或外部知识构造的。

比如各种引号和括号应当成对出现，一旦发现没有左引号（括号）的右引号（括号），或发现左引号（括号）后面未出现右引号（括号）就又出现左引号（括号），系统就给出警告。

又如系统可以建立“单位-职务-人名”对照表，当发现三者连续出现，其中两者相符而一者不符时，系统就给予警告并提示正确答案。如建立知识库“北京大学 校长 许智宏；清华大学 校长 王大中；……”当发现有“北京大学校长许志宏”时，系统对于“许志宏”提出警告，并建议改成“许智宏”。

（2.5）校对系统的工作流程

系统首先对待校文章进行分词并标注词性，然后用接续关系库中的接续数据、反面规则和知识库表示的各种特殊规则进行检查，发现可疑之处并在屏幕上标出出来。当用户要求修改建议时，根据错误类型给出修改建议，由用户确定是否修改和如何修改。系统的工作流程如下：

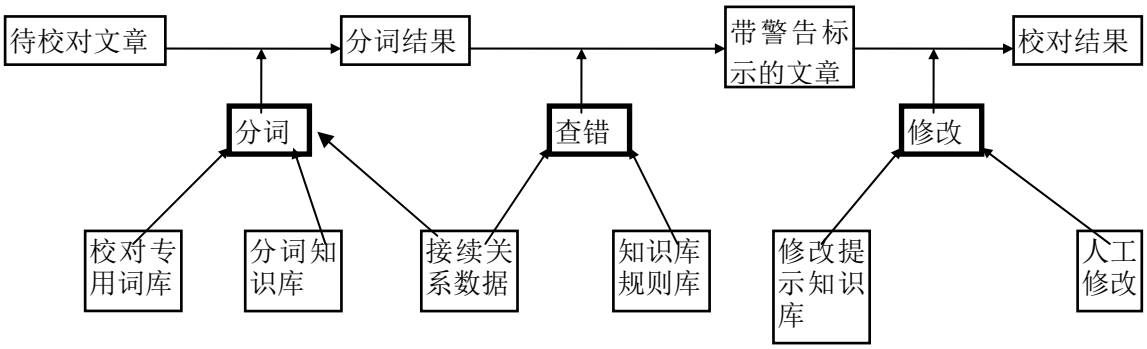


图 2. 计算机辅助汉语校对系统的工作原理

3.汉语校对系统的功能定位

（1）计算机校对同人工校对功能互补

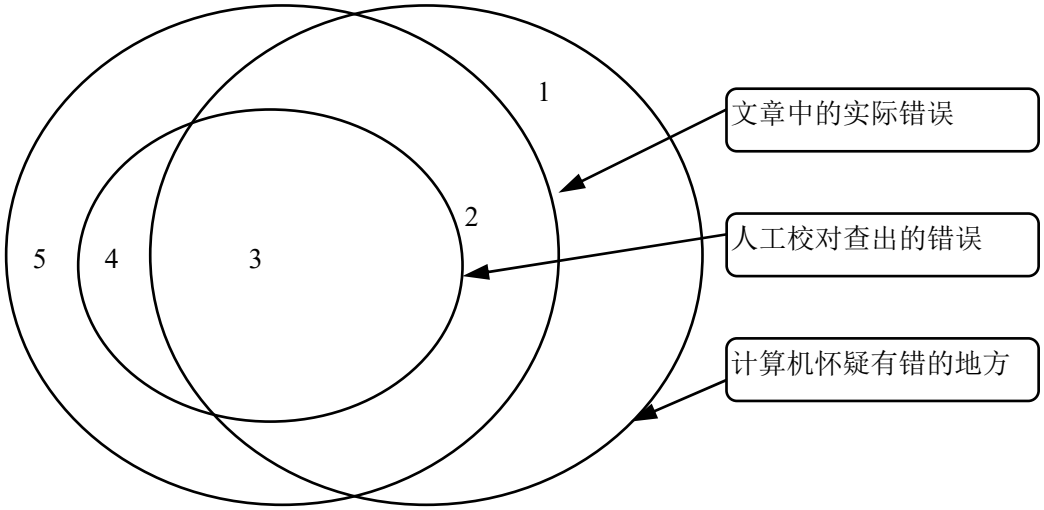
计算机校对所采用的语言模型不可能同汉语语言完全吻合，所以，计算机怀疑有错的地方不一定真有错，也就是说计算机有误检的情况；有些错误计算机查不出来，也就是说计算机有漏检的情况。一般来说，漏检和误检是避免不了的，所以，计算机不能代替人工校对，“自动校对”的说法是不切实际的，甚至在可预见到的将来也不可能做到。计算机校对不能代替人，只能起到辅助作用。所以，把校对系统称为“计算机辅助校对系统”比较适当。

那么，计算机校对的实用意义表现在哪里呢？表现在有相当一部分错误人常常漏检，但计算机却能快速地检出来，从而在人工校对之外再用计算机校对把一道关，可以明显提高文字质量。

下图显示了校对系统的作用。图中 1 是机器误检区；2 是人漏检而机器检出的区域；3 是机器和人都检出的区域；4 和 5 是机器漏检区，其中 4 是人检出而机器漏检的，5 是人和机器都漏检的。由于人和机器都有自己能检出而对方漏检的区域，所以合理的工作方式是人机互补，尽量缩小人和机器都漏检的 5 区，扩大人机检出区域的总和，即 2、3、4 区的面积总和。撇去人的因素，对于机器来讲，最主要的事情之一就是扩大 2 区。把人漏检而被机器检出的错误数称为机器的补检数，补检数比上全文字数称为补检率。扩大 2 区就是要提高补检率。

目前的状况是，一般的书籍在水平不太高的校对人员校对以后，使用计算机校对系统补检率往往达到近千分之一，在高水平的校对人员两三遍校对之后，补检率也有万分之一以上。我国实行的标准是：出版物中文字错误超过万分之一便为不合格,更不能评优。由此可见，校对系统的补检对于提高书刊文字质量有重要作用。

前面分析过，4 区的存在是不可避免的，因此人的校对不能省去。既然如此，花大力气扩展 3 区减小 4 区，即让机器去查人能检出的错误，就没有太大的意义，况且即使花费巨大代价也难以推进一小步。所以，校对系统不应把力量过多用在人比较擅长而机器能力较薄弱的高层次错误的检查上。



- 区域 1 为计算机校对系统误检之处
- 区域 2 为计算机校对系统查出而人工校对漏检的错误
- 区域 3 为计算机校对系统和人工校对均查出的错误
- 区域 4 为人工校对查出而计算机校对系统漏检的错误
- 区域 5 为人工校对和计算机校对系统均漏检的错误

校对系统功能图示

提高补检率的同时还要降低误检率，即减小图中 1 区的面积。否则，校对系统的用户要从大量的误检中筛出真错，其代价会超过功效。但是，检出和误检往往是伴生的，检出的东西多，往往误检就多。解决这一矛盾是提高校对系统性能的关键。

（2）汉语校对系统的查错功能以词语检查为主

据调查统计，语文水平较好的作者的文章，很少有语法错误和标点错误，语义性、知识性、事实性错误可能由于写作、录入等环节的马虎而出现，但数量也很少。相反，由于作者记不准字的写法和录入中的种种失误，词语错误往往较多。在刚刚录入未曾校对过的文章中，词语错误占全部错误的 80%以上；在人校对过多遍的文章中，尚未校出的词语错误占全部尚未校出错误的 95%以上。另一方面，词语错误多数只看两三个词的上下文便能确定，而确定的方法主要是检索和比较，这对于机器来说正是发挥长处的机会。人识别高层次的错误有一种机器无法企及的感悟能力，但对词语错误却容易遗漏。（人校对过多遍的文章中残存的错误绝大多数是词语错误，正好印证了这一结论。）因此，在机器不可能彻底解决校对问题的现实面前，校对系统应采用“取大舍小”和“扬长避短”的策略，把主要力量放在词语检查上，这是提高补检率的最主要的途径。

除了词语检查外，校对系统还可以检查有确切错误模式的语法错误，可以检查括号、引号是否配对，可以检查某些知识性错误，如领导人单位职务姓名的匹配、行政地名的隶属关系匹配等等。

注意：外国人写的汉语文章中语法语义错误较多，针对外国人汉语作文的校对系统不能仅用上面的方法去解决。例如：

他洗干净了衣服→他把衣服洗干净了

他洗干净了一件衣服→？他把一件衣服洗干净了

他洗不干净衣服→*他把衣服洗不干净

他给朋友买了一件衣服→*他给朋友把一件衣服买了

他买了一件衣服给朋友→*他把一件衣服买了给朋友

其中标“？”表示不长这样说，标“*”表示不能这样说。

汉族人凭借语感不会发生这类错误，但学汉语的外国人很难具备汉语语感，会出现这种错误。

三、语言现象研究

1. 背景

语言教学和语言学的研究需要语言事实作为背景和依据。过去，主要通过研究者用卡片大量摘录语言材料，这种方法效率太低，研究成果受到数据规模的限制。

近十几年出现了大量的电子文档，可以组成大规模的电子语料库，并出现了相当成熟的文本信息检索技术。这一技术可用于电子文档中的语言事实的检索。比起人工收集语言事实，这种方法的好处是：

（1） 效率高。可以在极短的时间内，用关键词语匹配的办法，在巨大规模电子语料库中找到相关文本的文章、段落、句子。

（2） 检索结果可以编辑、复制、打印。

（3） 查询表达式可以是关键词语同逻辑符号组合成的复杂的关系式。

但这种方法的直接应用目标是信息检索。信息检索关心的是信息的意义，不关心信息的语言表述形式。比如，当用户要检索有关“计算机”的文章时，应当把有关“电脑”的文章也检索出来；用户要检索“中国的大学”时，系统检索的是“中国”和“大学”，“的”字则被滤掉。相反，面向语言学研究的检索特别重视语言形式，它要求：

（1） 查询结果按照语言学研究的要求对语境排序，以便整理、提炼规律。

（2） 检索表达式有强大的表达能力，能表示各种各样的语言形式约束条件，能对关键词语的语境进行限定。

（3） 使用语言学词类的概念，要求能按词类检索。

（4） 其他用于语言学研究的辅助功能，比如字频、词频统计，字词的二元关系、三元关系统计等等。

这些功能是各种文本信息检索工具所不具备的。

近十年来，语料库技术迅速发展，可以对大规模语料库进行加工，包括分词、标注词类、甚至归并短语，使生语料库变成熟语料库。在这种加工的基础上利用文本信息检索技术进行面向语言学研究的检索，比基于生语料库的检索效果要好，特别是能进行词类关系的检索。目前，国内外由许多学者在做这类语料库标注的工作。但是，这里存在两个问题：

（1） 语料库自动加工技术尚不够成熟。大规模真实文本分词技术的精度可达 98%以上，但仍有一定量的错误。词类标注的精度只是 90%左右，短语归并的精度则还要低得多。这种精度水平不能满足语言学研究的需要，于是需要手工的后加工。而这一后加工所需的时间和人力是极其巨大的，使得熟语料库价格极为昂贵，且规模不可能太大，难以满足大规模语言学研究的需要。

（2） 语言学界对于汉语词类标记集尚无统一意见，加工语料库所用的词类标记集不可能适合于所有语言知识检索者的需要。即使检索者希望使用的词类标记集同语料库加工所用的词类标记集完全相同，对于一段具体的上下文中的一个具体的词到底应归

入哪一类，仍会有不同意见。

针对语言学研究的需求和目前各类相关系统存在的问题，我们提出了如下的解决方案：

（1） 由于熟语料库加工代价太高，而且对于熟语料库的标注体系、标注原则尚无统一意见，我们直接采用未经人加工的生语料库。

（2） 对于生语料库，我们提供字串检索功能。当需要检索词语时，我们把词语当作字串来检索。这样做会带来某些误报，但一般来说不会造成漏报。由人在检索结果中去掉无关的，保留有用的，负担并不十分重。

（3） 为了减少字串检索造成的误报，我们为检索表达式设计了比较强的表达功能，包括多项联合、多项选择、排斥、定常间隔、不定长间隔等。从而能有效地表达对检索结果的限制条件。

（4） 提供分词后的词串检索功能，从而能大大减少误报。

（5） 分词采用词库匹配、歧义消解、未登录词语归并的策略。分词用的词库由检索用户提供，以便检索用户按照自己的需要确定哪些词语需要被切出来。哪些词语不需要被切出来。检索用户也可以不提供自己的词库而使用系统自带的词库。

（6） 提供词类串检索、词类与词混合检索的功能。词类标记集和每个词的词类标注由检索用户提供，提供方式是将词条和它的词类登记在 access 库中。系统使用检索用户提供的带有词类标记的词库对生语料自动分词，并生成索引，供给用户检索。系统对于词类标记集没有任何限制，这个集合可以同现行的任何词类标记集没有任何关系。如此，检索用户可以完全自由地研究词类的语法语义特征，并按自己的需要改进词类标记。检索用户也可以不提供自己的词类标记而使用系统自带的词类标记。

（7） 词库中可以有兼类词。系统不负责对词类进行动态排歧。系统的检索原则是宁滥勿缺。比如，某个词具有 a、b、c 三种词类标记，则用 a、b、c 三个词类都可以检索到这个词的所有出现。这种做法虽然会在检索结果中带来垃圾，但避免了排歧中的争论，避免了自动排歧中的错误造成的漏检或误检，决不会遗漏任何可能的检索结果，并节省了人工标注带来的巨大的人力开支和时间消耗。检索用户可以在任何时候对任何规模、任何类型的生语料立即进行带有词类分析的检索。这种做法的基本思想是充分发挥机器检索的效率和人类专家甄别的智能，避开机器在智能方面的弱点和人类专家在机械搜索方面的短处。实践证明，这种方法是合理可行的。

此外，我们的系统还有如下特点：

（8） 系统在对生语料分词的时候，能以相当高的准确率自动归并一些未登录词语，包括各种数字串、中西人名、中西地名、机构名、后缀短语等。系统支持用户检索这些未登录词语。这一工作既提高了检索结果的召回率（比如将人名作为一个类来检索时，不仅能检索出词库中登录的“李白”、“鲁迅”、“毛泽东”等人名，也能检索出各种未登录的人名），又提高了检索结果的准确率（比如“马克坚”作为一个人名被归并，它就不会被关键词“马克”检索出来）。

（9） 系统对检索结果进行双向拼音排序。用户的检索表达式可以有多个词语，其中一个是关键词语。对于多个检索结果，用户可以要求它们按关键词语下文的拼音排序，也可以要求它们按关键词语上文的拼音排序。于是，从排序的检索结果中，检索用户可以总结关键词语的语境特点，从而可以总结出语言学规律。这种排序的结果也有利于机器通过统计来归纳出语言规律。

（10） 系统提供种种辅助功能，包括：词频、字频、词串频、词类串频、混合串频的统计，检索结果的文章出处等，便于人进行研究，便于机器进行自动分析。

（11） 对于巨大规模的生语料，系统生成索引的速度特别快。

（12） 对于巨大规模的生语料和复杂的检索表达式，系统进行检索的速度特别快。

2. 使用实例

（2.1）成对出现的连词

利用 CCRL 我们调查了现代汉语小说中典型的成对连词的出现状况。调查对象是老舍的 438 万字的小说、王蒙的 100 万字的小说、王朔的 165 万字的小说和巴尔扎克的 198 万字的翻译小说。调查内容涉及 3 对连词：“因为”和“所以”，“不但”和“而且”，“虽然”和“但是”，对于每一队，调查它们间隔 30 个字以内有序共现的次数（不考虑颠倒次序的情况以及分别出现的次数。调查结果如下：

4 位作家作品中 3 对连词出现次数统计

作者	老舍	王蒙	王朔	巴尔扎克	合计
字数（万字）	438	100	165	198	901
因为+<30+所以	304	5	16	23	348
因为	3730	400	1072	1480	6682
所以	2510	140	784	606	4040
不但+<30+而且	239	77	4	26	346
不但	555	118	128	86	887
而且	1981	543	416	585	3525
虽然+<30+但是	44	25	2	32	103
虽然	1713	247	159	392	2511
但是	1669	432	116	570	2787

4 位作家作品中 3 对连词使用频率统计

作者	老舍	王蒙	王朔	巴尔扎克	平均
因为+<30+所以	69	5	10	12	24
因为	852	400	650	747	662
所以	573	140	475	306	373
不但+<30+而且	55	77	2	13	36
不但	127	118	78	43	91
而且	452	543	252	295	385
虽然+<30+但是	10	25	1	16	13
虽然	391	247	96	198	233
但是	381	432	70	288	292
6 个连词频率合计	2776	1880	1621	1877	2038

4 位作家作品中 3 对连词出现次数的比例如图 1 所示。

图中，中轴表示全文字数。“因为”和“所以”的出现次数表示成两个横条，横条搭接部分表示“因为”右边 30 个字之内出现“所以”的次数。这三部分的长度比例表示三者出现次数之比例。横条的高度与中轴长度之比例表示“因为”右边 30 个字之内出现“所以”的次数同全文字数之比例，只是这个比例扩大了 1000 倍。“不但”和“而且”、“虽然”和“但是”的图示也有类似的的意义。

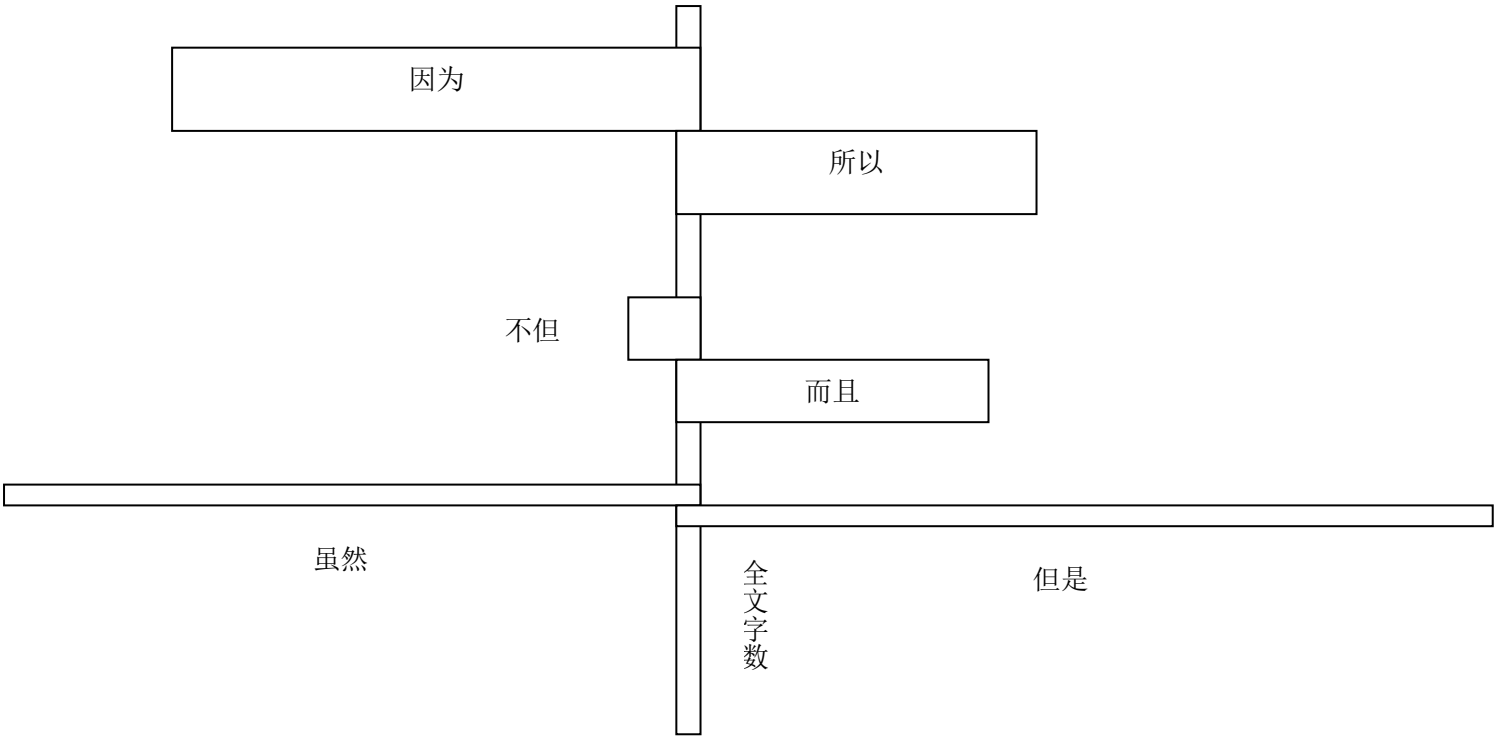


图 4 位作家作品中 3 对连词出现次数的比例

从表中和图中可以看出以下规律：

- 这些连词近距共现的次数远远低于分别出现的次数。
- 3 对连词之间比较，“虽然”“但是”的近距共现次数很少，分别使用却不少；“不但”“而且”近距共现次数与分别使用次数之差别较小。
- 6 个连词之间比较，“因为”出现得最多，“不但”出现得最少。
- 4 个作者之间比较，王朔最少用连词，最不喜欢成对使用连词，尤其不喜欢成对使用“不但”和“而且”、“虽然”和“但是”。老舍用连词最多，特别喜欢用“因为”和“所以”。

这里需要说明，CCRL 只能从外部形式上进行匹配，并无句法分析能力，所以两个连词近距共现并不意味着一定配对。我们对 “因为”“所以”30 个字以内近距共现的全部例子又手工进行了仔细分析，结果是：

作者	老舍	王蒙	王朔	巴尔扎克	合计
因为+<30+所以	304	5	16	23	348
二者配对	256	3	5	15	279
二者不配对	48	2	11	8	69

上面已知连词近距共现的次数远远低于分别出现的次数，这里又看出连词近距共现中有相当一部分并不是配对的，可见汉语中成对连词配对出现的比例是极低的。

（2.2） 把字句

使用词串和词类串检索，要求找出单字词“把”前面不是数词的所有句子，等于就找出了所有的把字句。使用这一方法，我们在2000年12月人民日报（共193万字）中，找出了全部把字句共1498句。下面是其中一部分：

把鉴定交给农民，让农民对与他们朝夕相处的乡干部进行实打实的打分评价，

江苏省委明确要求省委常委、党员副省长把联系点办成学习教育活动的示范点，

邢台市委把领导干部深入基层、深入群众作为搞好“三讲”教育、实践“三个代表”的重要内容，
把领导干部在职自学理论纳入评先创优活动，
把领导干部在职自学理论的成绩作为评选优秀个人、先进党支部和精神文明单位的必备条件
中航第二集团公司在落实这项工作时始终把领导干部的学习放在重要位置，
坚持把培养党和人民的忠诚卫士作为“三项经常性”工作的核心。
切实把培养干部、“拴心留人”作为事关档案事业发展全局的大事抓紧抓好，
必须把强化考评作为主要手段，
把强化少年国防意识视为己任，
把伤害致死改成故意杀人，
我们把生活设计得高效和富有创造生机，
会把失败原因归罪于自己的家庭环境不好呀，学校不好呀；
再把示范点推广到面。
把说明中“一天服用 2 — 3 次”的“一天”，理解为白天，
想要把所谓的Ａ３８０补贴事件闹到世界贸易组织去。
把所谓“大东亚战争”描写为“解放亚洲的战争”。
联合国粮农组织和国际水稻研究所也把推广杂交稻技术作为全球大米增产的一项战略计划。
把维护和促进香港的繁荣稳定当作义不容辞的责任。
把维护和实现职工群众的具体利益与维护和实现全国人民的总体利益统一起来，
三是把维护困难职工群体合法权益作为当前重点工作来抓，
把维护困难职工群体合法权益作为重点工作来抓，
二是把维护困难职工群体合法权益作为重点工作来抓，
有了这些素材，就可以对把字句进行深入研究了。

当然，从理论上说，这样找把字句是有缺陷的，因为“把”还可能作动词用，这时“把”后面往往接“了”、“着”、“过”、“得”、“住”等词。但是在本例中未发现这类情况。

四、罕用汉字输入

见另文（留学生错字处理技术.ppt）

五、印刷体汉字识别

机器分析字形特征，与特征库中的字比对，找出最可能的字。
人进行横检和竖检。

六、信息检索

机器快速找出可能的结果并按可能性大小排序，人工选择