词类划分的数学理论*

白 硕

(北京大学数学系,北京 100871)

摘要 本文探讨在分布标准下词类划分的数学理论,证明了用分布分析方法进行词类划分,是一个求最大不动点的迭代过程,本质上不存在逻辑循环.理想的词类划分——最大不动点是在极限意义下可计算的.从而澄清了语言学界多年来对分布分析方法的误解,论证了词类划分的分布标准的科学性.

关键词 自然语言处理,词类划分,分布分析.

在汉语研究中,争论数十年不衰的最热门的话题,莫过于词类问题了.的确,形态的缺乏、意义的灵活和语法功能的多样化,使得汉语的词的划类问题很难有一个可以广为接受的定论,而这又影响了汉语语法的进一步探讨.关于汉语的词类标准问题,大致有如下3种观点:

1. 形态标准

在有词尾变化的语言中,词的内部结构和词法性质是用来划分词类的明显标记.严格地说,汉语的词是缺乏这种标记的.因此,即使是主张形态标准的人也不得不把"形态"的本来意义加以引伸,认为象"了""着""过"这一类时态助词都是动词的"词尾".很明显,在汉语这样形态很不发达的语言中用形态标准划分词类,是很难行得通的.

2. 意义标准

主张意义标准的人想按照词语所指称的概念的分类体系对词进行分类,例如,实体、事件、关系、性质、状态等.这种观点孤立起来看不能说没有意义,但放到语法研究的大背景下就成了问题.朱德熙先生曾经指出:"战争"和"打仗"从概念上说指称的并无不同,但两个词的语法性质绝然不同[1].因此,脱离开词的语法性质谈意义标准,所划的类很可能缺乏语言学价值.说到底,类是要为描述语法现象、刻划语言规律服务的.

3. 分布标准(功能标准)

这种观点认为,词的划类标准应该是词的语法功能分布的总和[2].就是说,两个词同属一类,当且仅当它们能进入同样一些环境(即上下文),即语法功能相同.这样,词类才有语法上的代表性,才有可能作为描述复杂语言现象的基本单位.从这一点上说,分布标准是客观的、科学的.

* 本文1991-11-21收到,1992-02-15定稿

作者白硕,38岁,副教授,主要研究领域为人工智能及计算语言等.
本文通讯联系人:白硕,北京 100080,国家智能计算机研究开发中心

分布标准的另一大优点是它的可操作性.要想判断两个词是不是同一类,拿一些环境来考验它们就是了.这种可操作性给计算机辅助词类划分的实现提供了极大的便利.比起容易流于"想当然"的意义标准和无从下手的形态标准来,显然具体得多、现实得多了.

分布分析方法是一种应用分布标准进行词类划分的形式化方法.以往的分布分析方法由于理论上的不完善,在语言学界受到一些批评和误解.本文提出的词类划分的"不动点"理论,就是针对这些批评和误解,对分布分析方法所作的完善和发展.

1 词类划分的不动点理论

1.1 严格同分布

令 V 为词的有穷集合,称为词汇,$L \subseteq V^*$ 称为 V 上的一个语言.

设 $\alpha,\beta \in V^*$.二元组 $\langle \alpha,\beta \rangle$ 称为一个具体环境,简称环境,称词 X 在 L 中满足环境 $\langle \alpha,\beta \rangle$,当且仅当 $\alpha x \beta \in L$.x 在 L 中所满足的环境的集合记为 $E_L(x)$,若 $E_L(x)=E_L(y)$,则称 x 与 y 严格同分布(Strictly Identical in distribution).在 Harris[3]的严格同分布意义下的词类划分原则可以表述如下:

命题 1. 两个词 x,y 属于同一词类,当且仅当它们是严格同分布的.

这种基于严格同分布关系的划类原则只是在一些非常理想的情况下才有意义.一般来说,在任何一种活的自然语言中难得找到两个严格同分布的词,因而,按照命题的划类标准,其结果几乎是一词一类,细到了根本无法实用的程度!

当然,上述对严格同分布划类原则的批评不意味着对分布分析方法的全盘否定.以词所满足的环境总体(即"分布")来划类的思想是分布分析方法的精华所在.问题只是在于:用具体的词串二元组充当环境,区分了很多不必要区分的东西.为此,Harris[4]提出了一个"对型不对例"的改进方案,用本文的数学语言来描述,就是下一小节的——

1.2 π-同分布

设 $\pi=\{V_1,V_2,\cdots,V_n\}$ 是 V 上的一个划分(即 $\bigcup V_i=V, V_i \cap V_j=\varnothing$ 当 $i \neq j$),x 为一词.记 $\pi(x)=\{V_i \mid x \in V_i\}$,定义

$$\pi(x_1x_2\cdots x_m)=\pi(x_1)\pi(x_2)\cdots\pi(x_m)$$
$$=\{x'_1x'_2\cdots x'_m \mid x'_i \in \pi(x_i), i=1,\cdots,m\}.$$

设 $\alpha,\beta \in L^*$.二元组 $\langle \pi(\alpha),\pi(\beta) \rangle$ 称为划分 π 下的一个抽象环境,简称 π-环境.我们说 x 在 L 中满足环境 π,当且仅当存在 $\alpha' \in \pi(\alpha)$,$\beta' \in \pi(\beta)$,使得 $\alpha' x \beta' \in L$.x 在 L 中所满足的所有环境的集合记为 $E^\pi_L(x)$.在不需指明 L 时简记为 $E_\pi(x)$.若 $E^\pi_L(x)=E^\pi_L(y)$,则称 x 与 y 是 π-同分布的(π-identical in distribution).

在 π-同分布意义下的词类划分原则可以表述如下:

命题 2. 两个词 x,y 相对于划分 π 属于同一个词类,当且仅当它们是 π-同分布的.

这个"对型不对例"的改进方案,确实可以弥补严格同分布方案的"划类过细"的缺点,但语言学界对它也有很中肯的批评:既然划类是相对于 π 的,那么在具体操作时,π 到底应该怎样取?如果 π 本身就是我们要求的词类划分,那么整个改进方案就是"从未知求未知",这岂不是逻辑循环吗?

1.3 不动点

作者从数学角度考察了这个问题.事实上,所有的 π- 同分布类构成了对 V 的一个新的划分(我们可以证明这一点) $T(\pi)$.满足 $T(\pi)=\pi$ 的 π 就是算子 T 的不动点.直观地说,π 是表达抽象环境的工具,$T(\pi)$ 是在用 π 表达环境的条件下形成的 π- 同分布词类划分.$T(\pi)$ 与 π 相重合意味着 π 作为一个划类方案,在分布上是自恰的.于是,作者据此提出了"不动点"意义下的词类划分原则:

命题 3. 两个词 x,y 属于同一词类,当且仅当它们是 π- 同分布的,这里,π 是 T 的不动点,即 $T(\pi)=\pi$.

但是,问题到此并来结束.我们从数学上还可以继续追问:T 的不动点确实存在吗?如果存在,唯一吗?如果不唯一,那么哪一个不动点是语言学家所关心的?它是可计算的吗?在以下各节,我们就要来尽可能完满地解决上述一系列问题.

1.4 最大不动点

记 V 上的全部词类划分的集合为 Λ,因 V 有限,故 Λ 有限.

定义. 令 $\pi_1,\pi_2\in\Lambda$.称 π_2 是 π_1 的一个加细,当且仅当对任何 $x\in V$,

$$\pi_2(x)\subseteq\pi_1(x)$$

我们把这一事实记为 $\pi_2\prec\pi_1$.

容易证明,"$\prec$"定义了 Λ 上的一个偏序关系,$\langle\Lambda,\prec\rangle$ 是一个偏序集,其最小元为 $\pi_{\min}$:$\pi_{\min}(x)=\{x\}$(每词各成一类),其最大元为 $\pi_{\max}$:$\pi_{\max}(x)=V$(所有的词构成一类).

显然有

引理 1. 设 $\pi_1\prec\pi_2\prec\cdots\prec\pi_n\prec\cdots$,

则一定有 N 存在,当 $n>N$ 时,恒有 $\pi_{n+1}=\pi_n$.

同理有

引理 2. 设 $\pi_1\succ\pi_2\succ\cdots\succ\pi_n\succ\cdots$,

则一定有 N 存在,当 $n>N$ 时,恒有 $\pi_{n+1}=\pi_n$.

下面我们来证一个重要的引理,即 T 泛函的单调性引理

引理 3. 若 $\pi_1\succ\pi_2$,则 $T(\pi_1)\succ T(\pi_2)$.

证明:设 $y\in T(\pi_2)(x)$,则 y 满足 $E_{\pi_2}(x)$ 中的所有 π_2-环境.事实上,对任何 $e\in E_{\pi_1}(x)$,e 可写成

$$e=\langle\pi_1(a_1a_2\cdots a_m),\pi_1(b_1b_2\cdots b_n)\rangle$$

的形式,其中 $a_1a_2\cdots a_mxb_1b_2\cdots b_n\in L$.由于 $\pi_1\succ\pi_2$,故

$$\begin{aligned}&\pi_2(a_1)\subseteq\pi_1(a_1),\\&\qquad\cdots\cdots,\\&\pi_2(a_m)\subseteq\pi_1(a_m);\\&\pi_2(b_1)\subseteq\pi_1(b_1);\\&\qquad\cdots\cdots,\\&\pi_2(b_n)\subseteq\pi_1(b_n).\end{aligned}$$

令 $e'=\langle\pi_2(a_1a_2\cdots a_m),\pi_2(b_1b_2\cdots b_n)\rangle$.由于 $e'\in E_{\pi_2}(x)$,由假设 y 满足 e',因此,y 满足 e.由 e 的任意性,知有

$$T(\pi_1)(x) \supseteq T(\pi_2)(x).$$

再由 x 的任意性，得 $T(\pi_1) \succ T(\pi_2)$，证毕.

定理 1. 对于有穷的 V 上语言 L，必存在词类划分 π，使得 $T(\pi) = \pi$.

证明：显然有 $\pi_{\min} \prec T(\pi_{\min})$，$\pi_{\max} \prec T(\pi_{\max})$. 由引理 3，我们可构造无穷升链

$$\pi_{\min} \prec T(\pi_{\min}) \prec \cdots \prec T^n(\pi_{\min}) \prec \cdots$$

及无穷降链

$$\pi_{\max} \succ T(\pi_{\max}) \succ \cdots \succ T^n(\pi_{\max}) \succ \cdots$$

由引理 1 及 2，一定存在充分大的 N_1, N_2 使得

当 $n > N_1$ 时，$T^{n+1}(\pi_{\min}) = T^n(\pi_{\min})$；

当 $n > N_2$ 时，$T^{n+1}(\pi_{\max}) = T^n(\pi_{\max})$；

分别将这两个极限值记为 π^- 和 π^+，它们都是 T 的不动点，证毕.

实际上，定理 1 的上述证明是构造性的，其中给出的 π^+ 和 π^- 两个不动点具有如下性质：

定理 2. 若 $\pi = T(\pi)$，则 $\pi^- \prec \pi \prec \pi^+$.

证明：(略).

根据定理 2 所断言的 π^-, π^+ 的性质，我们称 π^- 为 T 的最小不动点，π^+ 为 T 的最大不动点，一般来说，不一定有 $\pi^- = \pi^+$. 因此，T 的不动点不一定是唯一的.

定理 3. $\pi^-(x) = \pi^-(y)$ 当且仅当 x 与 y 严格同分布.

证明：(略).

既然 π^- 等价于严格同分布标准，那么对于严格同分布标准的全部批评也就是对 T 的最小不动点 π^- 的批评. 也就是说，语言学家不会采纳 π^- 这样的不动点作为划类标准. 那么，π^+ 又如何呢？

我们考察一下序关系"$\prec$". 按定义，"$\prec$"是两套词类划分方案谁"粗"谁"细"的一个刻划，越"大"则方案越"粗". 最大不动点 π^+ 恰恰是保持 $T(\pi) = \pi$ 性质的最"粗"的划类方案，也就是说，在保证"自恰"的前提下划出的类最少、最精简. 我们看到，语言学家所一直寻找的，其实就是这样的划类方案！

于是，我们可以给出作者提出的最关键的划类标准.

命题 4. 两个词 x, y 同属一类，当且仅当它们是 π^+ - 同分布的，这里 π^+ 是 T 的最大不动点.

至此，我们可以回过头来讨论所谓"逻辑循环"的问题了，我们通过严格的数学论证表明：分布分析，说到底是要求解 T 的最大不动点 π^+，而 π^+ 是可以明确无误地定义出来的，这里面并不存在任何"逻辑循环"或者"怪圈". 在我们的构造性证明中，实际上也暗示了达到 π^+ 的途径——从 $\pi_{\max}$ 出发逐次迭代. 这样，我们就以严格的方式，澄清了语言学界对分布分析方法和分布标准的误解，同时也充实和完善了分布分析理论.

2 π^+ 的可计算性

上节的结果表明：只要分别从"每个词各成一类"($\pi_{\min}$)和"所有词构成一类"($\pi_{\max}$)这两

个极端的情况入手，反复用 T 泛函迭代，一定能在有限次迭代步骤之内达到 π^- 或 π^+. 前者是一个逐步合并(merge)的过程，后者是一个逐步分裂(split)的过程. 尤其是后者，由于其出发点和最终结果的简明性质，实际已成为分布分析的理想目标.

遗憾的是，这种迭代法还不是能行的(effective)算法. 因为一般有语言学意义的语言 L 都是无限的，所以相应的各 $E_\pi(x)$ 也都有可能是无限的. 于是，在一次迭代步骤之内，我们实际上不可能用逐个测试的办法检查每个词满足 $E_\pi(x)$ 中环境的情况. 为了实现用计算机辅助计算 T 的不动点，我们需要探讨 π^+ 的可计算性. 让我们先从 π- 环境的等价性谈起.

定义. 两个 π- 环境 e_1, e_2 称为等价的，若对任一词 x，要么 x 同时满足 e_1 与 e_2，要么 x 同时不满足 e_1 与 e_2，e_1 与 e_2 等价性记为 $e_1 \sim e_2$.

可以证明，π- 环境之间的等价关系满足自反，对称传递三公理.

所有可能的 π- 环境按其等价关系可导致一个等价类划分：

$$\varepsilon = \{E_1, E_2, \cdots\}.$$

引理 4. 对有限词汇 V 上的任何语言 L，ε 中一定只有的限个等价类.

证明：(略).

设从每个 π- 环境等价类 E_i 中选出代表元 e_i，可组成一个 π- 环境代表系 $E_0 = \{e_i | e_i \in E_i, E_i \in \varepsilon\}$，显然，$E_0$ 是有穷集.

引理 5. 对任何词 $x, y, E_\pi(x) = E_\pi(y)$ 当且仅当

$$E_\pi(x) \cap E_0 = E_\pi(y) \cap E_0$$

证明：(略).

定理 4. 如果已知 π 和相应的 π- 环境代表系 E_0，则 $T(\pi)$ 是能行可计算的.

证明：因 E_0 是有穷集，故 $E_\pi(x) \cap E_0, E_\pi(y) \cap E_0$ 亦然. 因此我们可以通过逐个检查 x，y 满足 E_0 中每个 π- 环境的情况这种穷举式的办法在有穷步内作出 x 与 y 是否 π- 同分布的判断. 又由于 V 是有穷的，故 $T(\pi)$ 可在有穷步骤内完成计算，即 $T(\pi)$ 是能行可计算的. 证毕.

定义一个 π- 环境 $e = \langle \pi(\alpha), \pi(\beta) \rangle$ 的长度

$$length(e) = |\alpha| + |\beta| + 1$$

定义一个 π-环境集合 E 的长度

$$length(E) = \max_{e \in E} length(e).$$

定理 5. 设 $\pi_0 = \pi_{max}, \pi_i = T(\pi_{i-1}), i = 1, \ldots, r, \pi_r = \pi^+$. 记 $E_0(i)$ 为一个 π_i- 环境代表系，若已知给定正整数 N 满足

$$N \geqslant length(\bigcup_{i=0}^{r} E_0(i))$$

则 π^+ 是能行可计算的.

证明：只需在每一步迭代中以长度不超过 N 的所有 π_i- 环境的集合 $E_{\pi_i}(N)$ 来代替 π_i- 环境代表系 $E_0(i)$，易证每个 π_i 都是能行可计算的，于是 $\pi_N = \pi^+$ 是能行可计算的. 证毕.

定理 5 告诉我们，π^+ 是在极限意义下可计算的[5]. 就是说，如果我们不断放大对 N 的估值，总会在某个时刻有

$$N \geqslant length(\bigcup_{i=0}^{r} E_0(i))$$

这时相应的 π^+ 实际已经可以被计算出来，然而究竟何时上述不等式能满足则是无法判定的. 在实用意义上，这一结果已经足够了. 人们经常使用的自然语言句子的长度总是有上限的，相应的环境的长度自然也有上限. 用超出这个上限的环境去区分词类显然已经没有实用价值了.

3 π^+ 计算实例

下面，我们通过一个简单的例子来说明计算最大不动点的迭代过程.

设 $V=\{$放，缩，大，小$\}$， $L=\{$放大，缩小$\}$.

首先从 $\pi_{max}=\{V\}$ 出发，得到：

$E_{\{V\}}$(放)$=E_{\{V\}}$(缩)$=\{\varnothing,V\}$， $E_{\{V\}}$(大)$=E_{\{V\}}$(小)$=\{V,\varnothing\}$.

于是，第一次迭代的结果为： $\pi_1=\{V_1,V_2\}$，

其中 $V_1=\{$放，缩$\}$，$V_2=\{$大，小$\}$.

由 π_1，我们进一步得到：

E_{π_1}(放)$=E_{\pi_1}$(缩)$=\{\varnothing,V_2\}$， E_{π_1}(大)$=E_{\pi_1}$(小)$=\{V_1,\varnothing\}$.

故 $\pi^+=\pi_1=\{\{$放，缩$\},\{$大，小$\}\}$.

注意到

$\pi^-=\pi_{min}=\{\{$放$\},\{$缩$\},\{$大$\},\{$小$\}\}$，

可见 π^+ 较简洁而 π^- 较繁琐.

4 结束语

本文讨论最基本的语言学范畴——词类及其划分的数学理论，指出了分布分析的任务是求解最大不动点 π^+，澄清了语言学界有关分布分析中含有“逻辑循环”的误解. 我们还证明了 π^+ 在极限意义下的可计算性. 作者基于这一套理论，并结合某些机器学习技术，设计了一个面向汉语的计算机辅助词类划分系统 CASD-1，取得了具有语言学意义的分类结果[6].

参考文献

1 朱德熙. 语法答问. 北京：商务印书馆，1985.

2 朱德熙. 句法结构. 中国语文，1962：8—9.

3 Harris Z. Methods in structural linguistics. University of Chicago Press, 1951.

4 Harris Z. Co-occurrence and transformation in linguistic structure. Language, 1957, **33**: 283—340.

5 Gold E. Language identification in the limit. Information and Control, 1967(**10**): 447—474.

6 白硕. 语言学知识的计算机辅助发现. 博士学位论文，北京大学，1990.

A MATHEMATICAL THEORY FOR WORD CLASSIFICATION

Bai Shuo

(*Department of Mathematics, Peking University, Beijing* 100871)

Abstract　Word classification is still an open issue for the linguistic research of Chinese. A fixed-point theory of word classification is proposed in this paper. According to our result, the Distributed Analysis (DA) method for word classification can be used in a process of iteration, so that the "logical circle" critique on DA method does not hold. The ideal classification, i. e., the maximal fixed-point, is computable in the limit. This paper therefore cleared out the misunderstandings on DA method among linguists, and proved the distribution criterion to be scientific.

Key words　Natural language processing, word classification, distributed analysis.