Ontology

面向信息处理的汉语同类词短语歧义研究

杨　泉◎著

PATR

中国社会科学出版社

Ontology

面向信息处理的汉语同类词短语歧义研究

杨　泉◎著

PATR

中国社会科学出版社

图书在版编目(CIP)数据

面向信息处理的汉语同类词短语歧义研究 / 杨泉著. —北京：
中国社会科学出版社，2019.5
ISBN 978-7-5203-4757-0

Ⅰ.①面… Ⅱ.①杨… Ⅲ.①汉字信息处理-短语-语义分析
Ⅳ.①TP391.12②H146.2

中国版本图书馆 CIP 数据核字(2019)第 149158 号

出 版 人 赵剑英
责任编辑 任 明
责任校对 韩天炜
责任印制 郝美娜

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http：//www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印刷装订 北京君升印刷有限公司
版 次 2019 年 5 月第 1 版
印 次 2019 年 5 月第 1 次印刷

开 本 710×1000 1/16
印 张 12.25
插 页 2
字 数 205 千字
定 价 85.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683

本书受国家语委“十三五”科研规划2018年度项目
《面向深度学习的汉语歧义结构句法语义知识库建设研究》资助

序　言

“潜在歧义论”（Potential Ambiguity Theory，简称 PA 论）是我在 1986 年在研究科技术语自动剖析时提出的一种理论。杨泉博士在她的专著《面向信息处理的汉语同类词短语歧义研究》中运用这种理论来分析汉语同类词短语的歧义，进一步发展了这种理论，我感到很高兴。现在，这本著作就要出版了，杨泉博士要我写个序言，我欣然同意了。

有人说，“潜在歧义论”只是冯志伟对于歧义的一种看法，大家都在研究歧义，觉得歧义是实实在在存在的，可是冯志伟却主张歧义是“潜在”的。我觉得，这样的说法有片面性。其实，“潜在歧义论”绝不仅仅是我拍拍脑袋想出来的对于歧义的一种“看法”（point of view）或“主张”（proposal），而是我在研究汉语词组型术语时，通过计算机分析大量的“歧义格式”之后而得到的一个重要“发现”（discovery）。这个发现，不是拍拍脑袋主观地想出来的，而是基于大量的语言数据运算之后揭示出来的自然语言中的规律。

在这里，我想把自己发现“潜在歧义论”的过程稍微详细地说一说。

1985 年，中国科学院软件研究所所长许孔时教授邀请马希文（北京大学计算机系教授）和我担任该所的兼职研究员，以加强该所语言信息处理的研究。1986 年，中国科学院与德国签订协定，要派一个学者到德国合作进行科技术语的自动处理研究，由于这项研究涉及到多种语言，要求中方派出的学者不仅懂得英语，还要懂得德语和法语。我是法国留学归来的，在法国留学期间，我曾经研制过汉-法/英/日/俄/德多语言机器翻译系统，懂得德语和法语，并且使用计算机做过德语和法语的自动生成，我的条件正好符合协定的要求。于是，许孔时所长征得中国科学院的同意，决定派我到德国去进行这项合作研究。

当时我是中国社会科学院语言文字应用研究所的正式研究人员，在软件所只是兼职，因此，许孔时所长又亲自到语言文字应用研究所给领导做

工作，作为一个文科的研究人员，却要代表中国科学院去德国执行计算机方面的合作任务，语言文字应用研究所领导最初感到难以理解，觉得太古怪了。经过许孔时所长多方解释，终于说服了语言文字应用研究所的领导，取得了他们的应允。这样，我就可以理直气壮地到德国进行合作研究了。

于是，我在 1985 年 9 月至 1986 年 9 月到德国夫琅禾费研究院（Fraunhofer Gesellschaft，简称 FhG）新信息技术与通讯系统研究所担任客座研究员（Gastwissenschaftler），代表中国科学院与德方合作，专门研究科技术语数据库和科技术语的自动分析问题。

我在夫琅禾费研究院的 VAX 11/750 计算机上，使用 INGRES 关系数据库，建立了汉语术语数据库 GLOT-C，并在 GLOT-C 与夫琅禾费研究院的多语言术语数据库 GLOT 之间，确立了多语言术语之间的映射关系。

在我建立的汉语术语数据库 GLOT-C 中，汉语科技术语只有一小部分是单词型术语（word term），如“程序，算法，流程”等，而大部分都是词组型术语（phrase term）。词组型术语可以由两个词构成，如“程序/设计”，或者由三个词构成，如“数字/字符/子集”，或者由四个词构成，如“条件/控制/转移/指令”，或者由五个词构成，如“平均/无/故障/工作/时间”，或者由六个词构成，如“四/分/之/一/平方/乘法器”，为了解释这些词组型术语的内在结构规律，我决定使用计算机对这些词组型术语进行结构自动分析（parsing），从而为汉语科技术语的规范化和新术语的命名，在语言学上提供理论根据，使汉语科技术语的研究工作与汉语语法和语义的研究工作更加紧密地结合起来。

我于 20 世纪 60 年代曾经在北京大学求学，我的老师朱德熙教授生前在讨论汉语的特点的时候曾经指出：“如果我们把各类词组的结构都足够详细地描述清楚了，那末句子的结构实际上也就描述清楚了。因为句子不过是独立的词组而已。”可见，要解决汉语句子的自动句法分析这个大问题，可以首先从汉语词组的自动分析入手，而要研究汉语词组的自动分析问题，可以首先从汉语词组型术语自动分析入手。这是我研究汉语词组型术语自动分析的初衷。

汉语的科技术语绝大部分是词组型术语，这些词组型的科技术语，其结构一般比较严谨，其含义一般比较单纯，它们在一定程度上反映了汉语词组结构的规律。根据朱德熙教授的理论，我相信，如果我把汉语词组型

科技术语的结构描述清楚了，也就有可能把汉语的词组结构描述清楚了，并有可能进一步把汉语句子的结构也描述清楚了。正是基于这样的信念，我试图从汉语词组型科技术语的自动分析研究中，找到解决汉语句子自动分析问题的钥匙。

这样，我便在夫琅禾费研究院的 VAX 11/750 计算机上，开始了汉语词组型科技术语的自动分析研究。出人意外的是，我在这项研究中，有了不少的新发现。这些发现驱使着我的好奇心不断地、一步一步地、深入地探索术语结构的奥秘。

科技术语一般都符合单一性的要求，因此，研究开始时，我幼稚地以为，在对汉语词组型科技术语进行自动分析时，不会存在歧义（ambiguity）方面的困难。哪知，实际情况与我的估计大相径庭。在自动分析术语结构时，我才发现，汉语词组型科技术语中，也存在着大量的歧义问题。

例如，“分割/字符”这个词组型术语，其中的“字符”这个名词，在词汇意义上，可以被分割，也可以具有“分割”这种特性，因此，“分割/字符”可以解释为“分割”某一个“字符”，在句法功能上是“述语+宾语”的“述宾式”，也可以解释为某一个被“分割”的“字符”，在句法功能上是“定语+中心语”的“定中式”。这样，从计算机处理自然语言的角度看来，“分割/字符”就是一个兼具“述宾式”和“定中式”两种结构的歧义术语了。把这两个词组型术语进一步抽象为 V+N 的结构，我们就可以说，V+N 这个结构是一个歧义结构。

进一步分析 V+N 这个结构，我还发现更多有趣的现象。在词组型术语“响应/时间”中，“响应”是动词，标注为 V，“时间”是名词，标注为 N，这是一个形式为 V+N 的词组类型结构；在词组型术语“取/比例尺”中，“取”是动词，标注为 V，“比例尺”是名词，标注为 N，这也是一个形式为 V+N 的词组类型结构，这两个词组型术语的词组类型结构相同；可是，在句法功能上，“响应/时间”是“定语+中心语”，属于“定中式”，“取/比例尺”是“述语+宾语”，属于“述宾式”；“响应/时间”和“取/比例尺”的词组类型结构相同，而它们句法功能结构却大不相同。就是 V+N 的句法功能结构被判断为“述宾式”之后，这个句法功能结构的逻辑语义结构还可能不同。述宾式的 V+N 可以解释为“谓词 + 受事者”（“取/比例尺”），又可以解释为“谓词 + 施事者”（“跑/

带”），又可以解释为“谓词 + 结果”（“印/字”），又可以解释为“谓词 + 目的”（“归/零”），又可以解释为“谓词 + 方向”（“面向/问题”）。可见，术语的句法功能结构与逻辑语义结构之间也不存在一一对应的关系。

这些现象说明，词组型术语并不像我原来估计的那样单纯，词组型术语仍然有着日常的自然语言中那样复杂的歧义问题。因此，我决定从词组类型结构、句法功能结构和逻辑语义结构三个方面来研究词组型术语的歧义问题。

根据 Chomsky（乔姆斯基）的上下文无关语法（Context - Free Grammar，简称 CFG）中的 Chomsky 范式（Chomsky Normal Form，简称 CNF），我采用二叉单标记树形图来表示汉语词组型术语的结构。这种由多个层次的二叉树枝构成的树形图，以二叉树枝作为其结构的基本单元。二叉树枝上的两个相邻结点的词类或词组类型组成的结构，叫做术语的词组类型结构（Phrase Type Structure，简称 PT-结构）。树形图中某一层级的两个相邻树枝结点上的句法功能信息，叫做术语的句法功能结构（Syntactical Functional Structure，简称 SF-结构）。树形图中某一层级的子树中两个相邻树枝结点的逻辑语义信息，叫做术语的逻辑语义结构（Logic-Semantic Structure，简称 LS-结构）。

任何术语都包括 PT-结构、SF-结构和 LS-结构这三种层次各异的结构，它们之间的相互作用，决定了术语的字面含义的基本内容。我们常常可以对术语的含义做出“望文生义”或者“顾名思义”的解释，正是由于这三种结构在我们头脑中相互作用的结果。因此，我们用严格的科学方法来分析这三种不同的结构，就有可能揭示这种“望文生义”或“顾名思义”现象的某些实质，从而对术语的字面含义做出科学的解释。

如果我们能够根据术语的 PT-结构，通过有穷步骤，自动地推算出术语的 SF-结构，并进而推算出术语的 LS-结构，那么，就可以做到词组型术语的自动分析（automatic parsing）。

然而，对于汉语来说，这是一个颇为复杂和相当困难的研究课题。

汉语术语的特点是，这三个结构之间，在绝大多数情况下，不存在一一对应关系。同样的 PT-结构，可以解释为不同的若干个 SF-结构；同样的 SF-结构，又可以解释为不同的若干个 LS-结构。

朱德熙教授在《汉语句法中的歧义现象》一文中，提出了“歧义格

式”这个概念来从理论上概括汉语中歧义结构的类型。他认为，句子的歧义“是代表了这些句子的抽象的‘句式’所固有的”，因此，他主张用“歧义格式”来概括汉语中的同形歧义结构。

朱德熙教授的这种见解是很有价值的，因为语言中的任何一个有结构歧义的形式，都不是孤零零地存在的，它往往代表具有某种格式的许许多多形式。抓住歧义格式这个关键，显然是研究歧义的必要途径。

但是，朱德熙教授的关于“歧义格式”的见解，还有不完全之处。我们在词组型术语的自动分析中发现，“歧义格式”所反映的类别的歧义，在具体的术语中有时存在，有时并不存在。当我们把具体的单词代真到歧义格式的范畴符号（也就是类别符号）中，而使歧义格式变为具体的词组型术语的时候，有的词组型术语中仍然可以保持歧义格式原有的歧义，而有的词组型术语中，歧义格式原有的歧义却消失得无影无踪了。

例如，根据我们前面的分析，在汉语词组型术语中，V+N 这个结构是一个“歧义格式”，当我们把其中的范畴符号代真为“分割/字符”的时候，其句法功能可以解释为“述宾式”，也可以解释为“定中式”，具有“述宾-定中”歧义，确实是有歧义的。可是，当我们把其中的范畴符号代真为“取/比例尺”的时候，其句法功能只能解释为“述宾式”，不能解释为“定中式”；当我们把其中的范畴符号代真为“响应/时间”的时候，其句法功能只能解释为“定中式”，不能解释为“述宾式”。在后两种情况下，“歧义格式”V+N 中原有的歧义完全消失了。

“歧义格式”竟然不再存在“歧义”，岂非咄咄怪事！

这是我在汉语词组型术语自动分析中一个重要的发现。这个发现意味着，朱德熙教授的关于“歧义格式”的见解并不是无可非议的“不刊之论”，这种见解难以解释我在汉语词组型术语自动分析中发现的“歧义格式”不再存在“歧义”的这种普遍存在的现象。事实上，朱德熙教授的“歧义格式”只具有歧义的可能性，并不一定具有歧义的现实性，因此，“歧义格式”这个名称也是不尽恰当的。

根据我在汉语词组型术语自动分析中的这种发现，我又进一步来研究英语中最常见的、公认的“歧义格式”VP + NP1 + Prep + NP2。

在这个“歧义格式”中，当我们把 VP 代真为 saw，把 NP2 代真为 a boy，把 Prep 代真为 with，把 NP1 代真为 a telescope 时，得到的“saw a

boy with a telescope" 是有歧义的，其意思可以是 "看见一个戴着望远镜的男孩"，也可以是 "用望远镜看一个男孩"。

可是，如果我们把 VP，NP1，Prep，NP2 等范畴符号代真为别的单词或词组的时候，这个 "歧义格式" 中的歧义却消失了。请看如下的例子：

She sent the ticket to New York (1)

(她把票寄到纽约)

She lost the ticket to New York (2)

(她把到纽约的票丢失了)

He cooks dinner for the children (3)

(他为孩子们做饭)

The company sells toys for children (4)

(这家公司出售儿童玩具)

在 (1) 中，动词 sent 表示传送，具有趋向性，介词词组 to New York 作它的状语，不作名词词组 the ticket 的定语，"歧义格式" 中的歧义消失了；在 (2) 中，动词 lost 表示丧失，不具有趋向性，介词词组 to New York 作名词词组 the ticket 的定语，不作动词 lost 的状语，"歧义格式" 中的歧义也消失了；同样地，在 (3) 中，介词词组 for the children 作动词 cooks 的状语，表示目的，而不作名词 dinner 的定语，"歧义格式" 中的歧义也消失了；在 (4) 中，介词词组 for the children 作名词 toys 的定语，而不作动词 sells 的状语，"歧义格式" 中的歧义也消失了。

这说明，在研究歧义问题时，我们归纳概括出来的 "歧义格式" 中所反映的歧义，并不是 "现实的歧义"，而是一种 "潜在的歧义" (potential ambiguity)；这种潜在歧义只有歧义的可能性，没有歧义的现实性。当用具体的单词去代真 "歧义格式" 中的范畴符号时，在所形成的具体的句子或词组中，这种潜在歧义有可能继续保持，也有可能不再继续保持而消失的无影无踪了。在歧义格式的研究中，"潜在歧义" 是一个值得特别注意的、带有普遍性的语言现象。

在汉语的 "歧义格式" 中，也同样存在着潜在歧义的问题。例如，"VP + 的 + 是 + NP" 是汉语中的一个歧义格式，其中的 VP 是一个双向动词，"VP + 的" 做主语，"是 + NP" 作谓语，整个格式是一个主谓结构，由于主语部分的 "VP + 的" 可以是施事，又可以是受事，因而产生了歧义。

例如，如果我们把 VP 代真为“反对”，把 NP 代真为“少数人”，得到“反对的是少数人”这一句子，可以理解为“提反对意见的是少数人”，这时，主语“反对的”是施事，表示反对者，也可以理解为“所反对的是少数人”，这时，主语“反对的”是受事，表示被反对者。“反对的是少数人”有歧义。

当歧义格式“VP + 的 + 是 + NP”代真为如下的句子时，这种歧义都一直保持着：

“看的是病人”可以理解为“正在观看某种情况的是病人”（“看的”是施事），也可以理解为“被看的是病人”（“看的”是受事）；

“关心的是她母亲”可以理解为“她母亲关心某人某事”（“关心”是施事），也可以理解为“被关心的人是她母亲”（“关心”是受事）；

“扮演的是一个演员”可以理解为“一个演员扮演了剧中某个非演员的角色”（“扮演的”是施事），也可以理解为“被扮演成一个演员”（“扮演的”是受事）；

“援助的是中国”可以理解为“中国援助了别国”（“援助的”是施事），也可以理解为“别国援助了中国”（“援助的”是受事）；

“相信的是傻瓜”可以理解为“相信某种情况的人是傻瓜”（“相信的”是施事），也可以理解为“所相信的人是傻瓜”（“相信的”是受事）。

但是，如果我们把歧义格式“VP + 的 + 是 + NP”代真为“关心的是分数”时，只可以理解为“所关心的事是分数”，“关心的”只能是受事，而不可能是施事，因为“分数”不可能去关心什么东西，这样，歧义格式中的潜在歧义也消失了。

如果把歧义格式“VP + 的 + 是 + NP”代真为“反对的是战争”时，只可以理解为“被反对的东西是战争”，“反对的”只能是受事，而不可能是施事，因为“战争”作为无生命的事物，不会去反对什么东西，这样，歧义格式中的潜在歧义也消失了。

上述语言现象说明，在自然语言的歧义研究中，当我们把具体的歧义词组或歧义句子概括为某种抽象的歧义格式的时候，这种抽象的歧义格式中所包含的歧义只是一种潜在的歧义。这种潜在的歧义在该歧义格式被代真为其他的词组或句子时，有可能继续保持，也有可能消失。这是自然语言歧义格式研究区别于自然语言的一般句法研究的一个重要特点，我们在

自然语言的歧义格式的研究中，不可不注意这一个重要特点。在这些研究的基础上，我提出了“潜在歧义论”（potential ambiguity theory，简称 PA 论）。

“潜在歧义论”从实质上改进了朱德熙教授关于“歧义格式”的理论，把“歧义格式”的理论更加深化了（deepening），也更加泛化了（generalization）。我们可以把“歧义格式”看成是“潜在歧义”转化为“现实歧义”的一种特殊情况。

由上所述可以看出，“潜在歧义论”并不是我灵机一动、脱口而出的一种“看法”（point of view）或“主张”（proposal），而是我在研究汉语词组型术语时，通过计算机分析“歧义格式”而得到的一个重要“发现”（discovery）。这个发现，对于揭示自然语言中的歧义，是行之有效的。

解放军外国语学院的博士生张禄彭根据我提出的“潜在歧义论”来研究俄语，发现在俄语中也存在潜在歧义。他写了《计算语言学视野下的俄语潜在歧义研究》一书，来揭示俄语中的潜在歧义。由此可见，我发现的“潜在歧义”，不仅存在于汉语和英语中，也存在于其他语言中。“潜在歧义论”是可以泛化的，这种理论具有普适性。

杨泉博士在《面向信息处理的汉语同类词短语歧义研究》这本专著中，根据我的“潜在歧义论”来研究汉语中的同类词短语，进一步提出“格式真歧义短语”和“实例真歧义短语”，这实际上是把潜在歧义结构词汇化后得出的结果，也是她在理论上的重要创新。

她还设计了一部机器可读词典和规则库的模式，并以此为核心，把我在日汉机器翻译研究中提出的知识本体 ONTOL-MT 作为语义特征来源，把意合手段作为知识表示内容，把复杂特征作为知识表达形式，把 PATR 作为实现手段，最终提出了“基于知识本体的语义驱动的句法功能消歧方法”。然后根据每种结构的不同特点编写消歧规则，提出消歧策略，建立规则库，来对汉语中的同类词短语进行歧义消减（disambiguation），取得了良好的效果。

杨泉的研究说明，“潜在歧义论”不仅是一种关于自然语言歧义的理论，而且还是一种行之有效的歧义消解方法。杨泉在写作这本专著的过程中，在理论上进行了深入的思辨，在方法上进行了具体的实践，把理论和方法紧密地结合起来，这是难能可贵的。

杨泉是我的博士生，她原来是学中文的，在研究生学习期间，进行了知识更新，成为了能够进行跨学科研究的新一代语言学家。在她的专著出版之际，我写了这篇序言，以此来表达我对她衷心的祝贺。

冯志伟

2019 年 5 月于德国海德堡

序　言

歧义问题是语言学中的一个难点，也一直是自然语言理解和机器翻译中难以解决的问题。1960 年巴尔-希列尔（Bar-Hillel）就曾说过歧义是自然语言处理中的主要绊脚石。

事实上，虽然歧义在自然语言中普遍存在，但真正影响到人们交流的情况却很少。这可能是因为在交际过程中，人们总是可以根据一定的知识背景，文化常识，结合特定的上下文和某些语法、语义特点找出几个意义当中正确的一个。人类这种“与生俱来”的本领恰恰是机器欠缺的，如果我们能将人类排除歧义的主要依据找到，通过一定的算法教会计算机，似乎就应该可以帮助计算机解决自然语言处理过程中的歧义问题。

本书正是从这种思想出发，以中文信息处理中高频语言现象——同类词短语为切入点，采用基于 Ontology 语义驱动的句法功能消歧方法，重点分析每种结构中两个词的语义特征及语义关系，再结合语法和上下文语境提出了一套具体的消解同类词短语句法功能歧义的规则，并最终用 PATR 语言在计算机上实现。

本书在编写消歧规则的同时，还建立了一个可供计算机使用的汉语知识本体雏形，提出了一个面向中文信息处理的现代汉语短语结构歧义研究的词典和规则库的模式，这些理性主义的研究成果对于其他歧义结构研究甚至其他中文信息处理研究都具有一定的启发性。

本书在潜在歧义理论基础上，进一步提出了格式真歧义短语和实例真歧义短语，这两种短语实质上是用潜在歧义理论词汇化的结果来验证大词库、小规则在语言工程实践当中的合理性。这两种短语本身的区分又一次提醒人们，在语言信息处理的时代要以崭新的眼光去重新审视古老的语言事实。

同类词短语中歧义问题的解决不但可以提高机器翻译的正确率，而且对计算机消解其他类型的结构歧义具有方法论上的指导意义。对于结构格

式的消歧研究实质上是一种面向中文信息处理的现代汉语句法结构分析方法，其研究成果可以用于计算机自动进行句法关系分析、句法成分及句型的标注和提取，本书得出的结果不仅是一些具体的规则，希望其探讨的方法对计算机消解其他结构的歧义也能产生积极的作用，从而对汉语本体、应用语言学、机器学习及其他相关领域的研究提供支持。

2018 年岁尾

目　录