

HNC 理论全书

第三卷 基本概念和逻辑概念 — 第六册

论图灵脑技术实现之路 图灵脑理论基础之六

黄曾阳 / 著

 科学出版社

HNC 理论全书

第三卷 基本概念和逻辑概念 — (第六册)

论图灵脑技术实现之路

图灵脑理论基础之六

黄曾阳 / 著

科学出版社

北京

内 容 简 介

本书是《HNC理论全书》第六册,直接关系到“理论→技术→产品→产业”接力的第一棒。人工智能理论探索的核心是大脑之谜,而语言脑是其关键。HNC试图开辟一条探索语言脑的新路,提出语言超人(简称语超)的设想。

本书分四编,主要采用对话体。第五编重点阐述基于语言理解的机器翻译,第六编和第七编则阐释语超如何实现,第八编基于对后工业时代曙光的响应进行了科技和文明展望。语超是语言脑的计算机模拟,也是图灵检验的终极目标,其技术基础是第五册阐释的句类分析、语境分析和记忆生成这三大环节。机器翻译的实现应基于句类分析与语境分析,而语超的实现还应加上记忆生成。语超是对语言信息的深层处理,对语言脑奥秘的揭示也具有重大科学价值。

本书适合对自然语言理解、人工智能、认知科学等感兴趣的所有读者,特别适合语言信息处理方面的研究者及学生参阅。

图书在版编目(CIP)数据

论图灵脑技术实现之路 / 黄曾阳著. —北京: 科学出版社, 2016.9

(HNC 理论全书)

ISBN 978-7-03-049659-1

I. ①论… II. ①黄… III. ①系统科学—研究 IV. ①N94

中国版本图书馆 CIP 数据核字 (2016) 第 200060 号

责任编辑: 付 艳 高丽丽 / 责任校对: 李 影

责任印制: 肖 兴 / 封面设计: 黄华斌

联系电话: 010-6403 3934

电子邮箱: fuyan@mail.sciencep.com

科学出版社 出版

北京东黄城根北街 16 号

邮政编码: 100717

<http://www.sciencep.com>

中国科学院印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2016 年 8 月第 一 版 开本: 787×1092 1/16

2016 年 8 月第一次印刷 印张: 26 3/4

字数: 631 000

定价: 128.00 元

(如有印装质量问题, 我社负责调换)

本书得到下述项目资助

中国科学院“八五”重点项目“汉语人机对话项目”

国家“九五”科技攻关计划项目“汉语理解系统的核心技术专题”(98-779-02-04)

国家 973 项目课题“基于概念层次网络(HNC)的自然语言理解与处理”
(G1998030506)

国家 863 “十二五”计划项目课题“基于云计算的海量文本语义计算框架与开放域自动问答验证系统”(2012AA011102)

中国科学院战略性先导科技专项“面向感知中国的新一代信息技术研究”
(XDA06030100)

作 者 的 话

本书是《HNC 理论全书》的第六册。《全书》共三卷六册，第一卷三册，第二卷一册，第三卷二册。第六册也就是第三卷的第二册。

HNC 理论以自然语言理解为其核心探索目标，试图为语言理解的探索开启一条新的途径。HNC 认为：语言理解的奥秘，是大脑之谜的核心，也是意识之谜的核心。对这个谜团的探索不是当前的生命科学可以独立完成的，需要哲学和神学的参与。故《全书》之“全”是一个“三学（科学、哲学与神学）协力”的同义词，非 HNC 理论自身之“全”也。HNC 理论充其量是一名语言理解新探索的侦察兵，从这个意义上说，《全书》之“全”应看作是一种期待，一声呼唤。

《全书》的初稿是半成品，是 HNC 团队的内部读物。原定以十年（2006~2015 年）为期，完成初稿。不意十年未竟，推动者和出版者联袂而至。他们深谋远虑，要把一个半成品升级为成品，把一个内部读物正式出版，其间所展现出来的非凡胆识、灼见与谋划，居功至伟四字，不足以表达笔者心中感受之万一。

《全书》结构庞大，体例繁杂，带大量注释。结构方面，分上层与下层，上层分“卷、编、章”3 级，以汉字表示顺序，汉字“零”表示共相概念林或共相概念树。在某些编与章之间，还插入篇。下层分“节、小节、子节”3 级，子节之后，可延伸出次节，每级之内，可派生出分节。体例方面，主体文字之外，安置了大量预说和呼应。注释方面，分两类编号：数字与字母。前者是对正文本身的注释，后者是对正文背景的注释。数字和字母都放在方括号内，如[*01]和[*a]。其中的“*”可以多个，如同宾馆等级的标记。两星[**]以上的注释比较重要，表示读者应即时阅读。

《全书》常相互引用，为标记之便，采取了[k₁k₂k₃-m]简化表示，其中的“k₁”表示卷，“k₂”表示编，“k₃”表示篇，无篇取“0”。“m”也是一个数字序列，依次表示章节序号。例如[210-0.2.1]和[210-1.2.1]分别表示第二卷第一编第零章和第一章的第 2 节第 1 小节。

《全书》使用了大量概念关联式。概念关联式是语言理解基因的重要组成部分，也是隐记忆的重要组成部分。每一个概念关联式总是联系于特定的概念基元、句类或语境单元。概念关联式分为无编号与有编号两类，无编号的表示尚待探索，《全书》只是给出了若干示范，为上下文引用方便给出的临时数字编号也属于这一类；有编号的统一使用

“——(编号)”或“——[编号]”，前者表示内使，后者表示外使。概念关联式编号区分普通与重要两级，后者加“-0”区别，若“-0”后缀于编号，表示不同文明对此有共识，而插入编号中间，则表示特定的文明视野。有编号的概念关联式都有牵头符号，代表着该概念关联式的重要性级别，目前主要用[HNC1]符号牵头。

在撰写初稿期间，池毓焕博士一直是我的学术助手。在本书出版期间，池博士一直是我个人的全权代表。科学出版社以付艳、王昌凤编辑为主的有关同志，为初稿的升级付出了巨大的辛勤与智慧，其审校之精细，无与伦比；池博士的配合，力求尽善。笔者的钦佩与感激之情，难以言表。

老子曰：“天地万物生于有，有生于无。”伟哉斯言。

黄曾阳

2015年9月22日于北京

引文出处缩略语对照表

《理论》	黄曾阳. HNC (概念层次网络) 理论[M]. 北京: 清华大学出版社, 1998
《定理》	黄曾阳. 语言概念空间的基本定理和数学物理表示式[M]. 北京: 海洋出版社, 2004
《全书》	即本丛书——HNC 理论全书, 共有三卷六册, 各册书名如下: 第一卷 第一册 论语言概念空间的主体概念基元及其基本呈现 第二册 论语言概念空间的主体语境基元 第三册 论语言概念空间的基础语境基元 第二卷 第四册 论语言概念空间的基础概念基元 第三卷 第五册 论语言概念空间的总体结构 第六册 论图灵脑技术实现之路
《导论》/《苗著》/《HNC 理论导论》	苗传江. HNC (概念层次网络) 理论导论[M]. 北京: 清华大学出版社, 2005
《转换》	张克亮. 面向机器翻译的汉英句类及句式转换[M]. 郑州: 河南大学出版社, 2007
《变换》	李颖, 王侃, 池毓焕. 面向汉英机器翻译的语义块构成变换[M]. 北京: 科学出版社, 2009
《现汉》	中国社会科学院语言研究所词典编辑室. 现代汉语词典 (第 3 版) [M]. 北京: 商务印书馆, 1996
《现范》	李行健. 现代汉语规范词典[M]. 北京: 外语教学与研究出版社/语文出版社, 2004

目录 | contents

作者的话

引文出处缩略语对照表

第五编 论机器翻译	1
第一章 句式转换	5
第 1 节 格式与样式	7
第 2 节 小句与大句	17
第 3 节 块扩与句蜕	31
第二章 句类转换	41
第 1 节 关于汉语“专利”句类	43
第 2 节 关于汉语对广义效应的多模式表达	47
第三章 语块构成变换	61
第 1 节 关于多元逻辑组合的对话	64
第 2 节 关于逻辑组合的内外之别	72
第 3 节 关于 RB01 宏观结构的汉英根本差异	73
第 4 节 关于 RB01 局部结构的对话	74
第四章 主辅变换与两调整	87
第 1 节 关于主辅变换	89
第 2 节 关于两调整	90
第五章 过渡处理的基本前提与机器翻译的自知之明	93
第 1 节 关于过渡处理的基本前提条件	95
第 2 节 关于机器翻译的自知之明	100

第六编 微超论	103
第一章 微超就是微超	105
第 1 节 微超的前世今生	107
第 2 节 微超的知识与技术依托	113
第 3 节 微超的冷板凳还要坐多久?	158
第二章 微超的科学价值	161
第三章 微超的技术价值	175
第六编附 自然语言理解处理的 20 项难点及其对策	183
第七编 语超论	315
第一章 语超的科技价值	317
第二章 语超的文明价值	323
第八编 展望未来	345
第一章 科技展望	347
第 1 节 科技迷信与经济公理	349
第 2 节 探索无限而科技有限?	356
第二章 文明展望	361
第 1 节 六个世界与三类国家	363
第 2 节 王道能替代社会丛林法则吗?	367
附录	373
附录 1 《HNC (概念层次网络) 理论》弁言	374
附录 2 《HNC (概念层次网络) 理论》编者的话	377
附录 3 《HNC (概念层次网络) 理论》后记	383
附录 4 《语言概念空间的基本定理和数学物理表示式》序言	387
附录 5 《语言概念空间的基本定理和数学物理表示式》编者的话	389
附录 6 《HNC (概念层次网络) 理论导论》	

序言	391
附录 7 《HNC (概念层次网络) 语言理解技术 及其应用》序二	393
附录 8 《面向机器翻译的汉英句类及句式转换》 序言	395
附录 9 《面向汉英机器翻译的语义块构成变换》 序言	397
附录 10 《汉语理解处理中的动态词研究》序言	399
术语索引	401
人名索引	406
《HNC 理论全书》总目	411

第五编

论机器翻译

编 首 语

从此编开始的相继 4 编（第五编到第八编）是否纳入《全书》，笔者一直在犹豫。到 2011 年，才最终决定献丑。这最后 4 编将构成《全书》的第六册，即最后一册。其汉语名称如下：

- 第五编 论机器翻译
- 第六编 微超论
- 第七编 语超论
- 第八编 展望未来

这 4 编的撰写方式需要有所改变，但如何改变，笔者曾茫然良久，最终的决定也不过是“摸着石头过河”吧。

语言翻译和机器翻译的专著和专文浩如烟海，方兴未艾的大数据思路在统治着信息处理领域的所有分支，本编面对的第一个问题似乎是：如何与文献历史和大数据技术现实接轨呢？这个问题一直困扰着笔者，到本编动笔的时候才明白过来。原来，东坡红烧肉的制作不必参考沙拉和麦当劳；本编所论，不过是一盘东坡肉而已，不必理会沙拉和炸鸡腿。这个比喻，是以烹调技艺为参照的，也就是说，仅涉及理论与技术的第一趟接力，不涉及产品与产业的后两趟接力。

HNC 认为，语言翻译过程必须包括下列 6 个环节：句类转换与句式转换、语块构成变换与主辅变换、块序调整和句序调整。可简称两转换、两变换和两调整，统称语言翻译的 6 项过渡处理。

这 6 个环节是语言翻译过程的共相，而针对词语或短语的翻译属于语言翻译的殊相。能不能说以往的机器翻译研究仅关注殊相问题，对共相问题还不曾进行过比较系统的思考呢？能不能说一流的翻译家也没有意识到上述共相问题的明确存在呢？进一步说，上述共相问题的重要性如何？机器翻译不可以回避这些问题吗？这些问题的呈现对不同语种的翻译是否存在巨大差异呢？

本编将回答上述问题，但不属于本编的主体。

上文说，“语言翻译过程必须包括下列 6 个环节”，但是，这 6 个环节并不是语言翻译的特殊需求或“专利”，而是语言生成的共相。语言生成的过程，可以说就是这 6 个环节不断运转的过程。因此，两转换、两变换和两调整不仅是语言翻译的 6 项过渡处理，也是语言生成的 6 项过渡处理。这就是说，语言生成过程实质上是一个语句的句类与句式如何选择的过程，是一个语块如何构成、辅块位置如何安顿、辅块与主块如何分工的过程，也是一个在大句里如何安排各小句之间的分工与顺序之过程。对于不同的语言，这些过程的差异性可

能很小或比较小，也可能存在天壤之别，汉语与英语就属于后者。这里不能不说一声，传统语言学恰恰对这一过程的“天壤之别”缺乏必要的敏感，这种敏感是需要源泉的，该源泉就是指：语言生成与语言理解不可分割，生成是理解的结果，理解是生成的前提，翻译不过是一种特殊条件下的语言生成。所谓 SVO 语言和非 SVO 语言的说法，看似对上述“天壤之别”给出了一个简洁的描述，实际上起了一种堵塞上述敏感源泉的消极影响^[*00]。

由于各种因缘的凑巧，上列 6 项过渡处理被误会成是机器翻译的一种特殊需求或“专利”，这是一个莫大的误解，笔者磨破了嘴皮似乎也不管用。所以，这里干脆来一次“矫枉过正”，把本编各章的名称约定如下：

第一章 句式转换

第二章 句类转换

第三章 语块构成变换

第四章 主辅变换与两调整

第五章 过渡处理的基本前提与机器翻译的自知之明

请注意，前四章的章名里没有翻译，第五章的章名里有翻译但没有机器。在随后各章的论述里，当然需要对这一“过正”举措给予适当弥补。上述莫大误解的缘由，下文也会有所呼应或交代。

第一章 句式转换

引言

句式这个词语已被最新版本的《现汉》所收录^[*01], 但《新时代汉英大词典》却没有。这似乎表明, 西方语言学界还没有感受到引入这一概念或术语的必要性。那么, 这是否意味着汉语学界在这方面的认识领先了一步? 这个问题值得思考。

HNC 最初引入 6 项过渡处理的时候^[*02], 并未使用“句式转换”, 而是使用“格式转换”, 排序在“句类转换”之后, 位列第二。从“格式”到“句式”的演变过程很值得回味, 下文会对此有所呼应。

本章分 3 节, 相应的标题是: 格式与样式、小句与大句、块扩与句蜕。这里的标题都略去“转换”, 以呼应前述“矫枉过正”的举措。前两节的设置, 实质上就是对上述演变过程的关键性呼应。

第 1 节

格式与样式

如果把句式分成基础结构和上层建筑两个梯级，那么就可以说，句式的基础结构或第一梯级是格式与样式，句式的上层建筑或第二梯级是小句与大句。格式与样式是 HNC 引入的术语，属于创新，前文已经系统讨论过，本节不过结合机器翻译的需求，来一次“回锅肉”处理。小句与大句是语言学的原有术语，HNC 也略有玩新，放在下节讨论。

格式是作用型语句的句式，样式是效应型语句的句式，这个论断概括了格式和样式的本质特征^[*03]，下面，对这一“本质特征”进行具体说明。

用传统语言学的话语来说就是：当主语和宾语两者直接连接时，作用型语句要求在两者之间加上语言标记，否则就是违例；而效应型语句面对这种情况，一定不附加语言标记。

如果把主语和宾语换成广义对象语块 GBK，把语言标记换成语块标记，把作用型语句叫作广义作用句，把效应型语句叫作广义效应句，那就可以用 HNC 话语改写上面的描述。描述话语如下：格式与样式可名之第一梯级句式，是关于一个语句主块排列顺序的描述，如果是两主块句，其可能排列顺序为 2，如果是 3 主块句，其可能排列顺序为 6，即可呈现出 6 种句式形态；如果是 4 主块句，其可能排列顺序为 24，即可呈现出 24 种句式形态。这样的描述在现代语言学论述里也可以见到，当然描述者使用的是主语和宾语之类的传统语言学学术语，而不会使用“主块”这个术语。但下面的描述在传统语言学描述里是见不到的，对于该描述将另起一段文字。

广义作用句的主块数量为 3~4，广义效应句的主块数量为 2~3。这就是说，广义作用句的主块数量最少 3 个，最多 4 个；广义效应句的主块数量最少 2 个，最多 3 个。这是任何自然语言的共相，是自然语言的第一铁律。但广义作用句与广义效应句的巨大差异，不仅表现在主块数量方面，更表现在对主块排序的处置方面。当两广义对象语块 GBK 在语句中相继出现时，广义作用句一定要求对 GBK 加上必要的语块标记；而广义效应句在面对同样情况时，却一定不附加任何语块标记。由此看来，对广义作用句和广义效应句的句式赋予不同命名，就显得非常有必要了，这就是格式与样式这两个术语的起源，格式专用于广义作用句，样式专用于广义效应句。不难想见，格式的形态变化一定远比样式复杂^[***04]。

一言以蔽之，上面的文字描述了格式与样式的本质特征或两者的本质差异，也就是描述了广义作用句与广义效应句的本质差异。这些话不过是对“论句类”（本卷第二编）的一个小小呼应，但依然是站在语句理解处理立场的话语，并不是本节的要点。

本节需要换一个立场说话，那就是语言生成的立场。这样的立场转换有悖于 HNC 探索