

王 岚 / 主编

解放军外国语学院英语博士文库

MT-oriented Sentence Category and Sentence Format Transfer from Chinese to English

张克亮 著

面向机器翻译的汉英句类及
句式转换

河南大学出版社

*Monographs by Ph.D Holders,
English Department, PLA UFL*



Off-Platform Resource Category and
Sentiment Percent Transfers from Chinese to English



Source: [Redacted]
[Redacted]



王 岚 / 主编

解放军外国语学院英语博士文库

MT-oriented Sentence Category and Sentence Format Transfer from Chinese to English

张克亮 著

面向机器翻译的汉英句类及
句式转换

河南大学出版社

*Monographs by Ph.D Holders,
English Department, PLA UFL*

图书在版编目(CIP)数据

面向机器翻译的汉英句类及句式转换/张克亮著. —开封:河南大学出版社, 2007. 4
(解放军外国语学院英语博士文库/王岚主编)
ISBN 978-7-81091-586-1

I. 面… II. 张… III. 英语-句法-机器翻译-研究
IV. H315.9

中国版本图书馆 CIP 数据核字(2007)第 033455 号

责任编辑 薛巧玲

装帧设计 张 胜·生生书房

出 版 河南大学出版社

地址:河南省开封市明伦街 85 号 邮编:475001

电话:0378-2864669(行管部) 网址:www.hupress.com

印 刷 河南省瑞光印务股份有限公司

版 次 2007 年 4 月第 1 版 印 次 2007 年 4 月第 1 次印刷

开 本 650mm×960mm 1/16 印 张 17.5

字 数 235 千字 印 数 1—1500 册

定 价 26.00 元

(本书如有印装质量问题请与河南大学出版社营销部联系调换)

序 言

——谋事在人，成事在天

克亮告诉我，他的博士论文已纳入他们学校的一项出版计划，要我写一篇序言。此前 苗传江博士和晋耀红博士已经分别把他们的博士论文扩展成为 HNC 探索的两部专著，并正式出版了。这是我第三次接受到这样的要求，在非常高兴之余，立即联想到“谋事在人，成事在天”的古语，因此就把它作为这篇序言的副标题了。与克亮博士论文的原稿相比，出版稿做了系统的扩充，已构成一部专著的规模，按照前两篇序言的惯例，下文将以专著称之。

专著首先概述了机器翻译的三次热潮，接着在“机器翻译的冷思考”中提出了理论误区、技术崇拜和市场化迷失三大论题，这里，既体现了克亮的学术勇气，又展现了他的学术冷静。字面上，专著没有对三大论题的未来之路给出直接的回应，但实际上已经蕴涵在专著的正文里了，专著的这种春秋笔法也许是最值得读者去体会的。

机器翻译能够超越专著所指出的雪线现象吗？对这个关键问题作者也同样采取了春秋笔法。但细心的读者不难觉察到，答案就在以扩展句类分析为核心的语言理解处理里，在以句类转换和句式转换为核心的 6 项过渡处理里，在以有指导机器学习为核心

的语言生成处理里。这三项处理构成一项巨大的科学工程。不同于通常意义下的软硬件工程,决定科学工程成败的主要是它的理论和知识侧面,而不是技术侧面,机器翻译科学工程尤其如此。专著清晰地论述了这一重要的科学工程观,并在结束语中对这项科学工程的艰巨性给出了十分清醒的描述。

专著所探讨的句类转换和句式转换问题,请允许我借用围棋的术语,乃是汉英机器翻译这一科学工程的大场和急所。作者是探索这一重大科学问题的第一人吗?作者取得了堪称突破性的进展吗?如果读者对此能做出自己的正确判断,那就没有虚读此书了。

作者与我共事期间,我正在集中精力从事扩展句类分析的理论探索,失去了与作者深入研讨机器翻译的难逢机遇,几年来常感愧悔。下面的话也许不该写在序言里,请把它当做一份补偿之情吧。

机器翻译不可能脱离源语言的理解处理,但机器翻译的研究则可以而且必须脱离源语言的理解处理而独立进行。其研究资源不能是那泛指平衡语料,也不能只是那经过机器分析处理以后的语料,而必须主要是那经过适度人工标注的语料。就专著所确定的研究思路来说,就必须是适度标注的 HNC 句群语料。机器翻译的理解(分析)、转换、生成三环节必须走“先分后合”之路,说白了就是必须走“转换先行、生成公关、理解逆推”之路。这确实是一条“曲线救国”之路,然而却是唯一可行之路。按照这一思路,那专著所展现的万里征途就可以起步于一个精干的研究小团队了。

HNC 标注语料的适度性首先是指语境信息的适度性,其次是指标注自身的适度不确定性(这需要精心设计)。在 HNC 的专业术语里,语境信息叫做领域句类的先验知识,其适度性集中体现在领域句类框架知识的完备性里,后者是自然语言理解处理的关键性知识。HNC 的训诂学渊源就在于它以语境统摄了语法、语义

和语用的三维度说，并据此构成了相应的数学物理表示式，如此而已。

谋事在人，成事在天。莫疑无路，总有明村。

仅以此与克亮共勉。

黄曾阳

2006 年 8 月

前 言

20 世纪 90 年代中期以来,伴随着 Internet 的飞速发展、经济全球化的到来和电子商务的广泛开展,不同语言群体之间的交流越来越普遍,语言的障碍也显得越来越严重。毫无疑问,传统的人工翻译难以应对以网络化、实时化和海量信息为特征的信息时代的语言翻译要求,如何有效利用现代技术突破横亘人们之间的语言障碍,成了全人类面临的共同挑战。

机器翻译是信息时代克服语言障碍的不可缺少的手段,它在现代信息社会中的巨大作用将会越来越明显。现代机器翻译就是用计算机来进行不同自然语言之间的翻译,它是自然语言计算机处理的一个历史悠久的部门,是横跨语言学、数学、计算机科学的综合性学科,也是信息时代语言应用的一个重要领域。

近十年来,世界范围内掀起了又一次机器翻译研究和开发的热潮。前两次热潮主要是由国家安全需求和科学探索推动而掀起的,这一次热潮则主要是强大市场需求驱动的结果。在这种强大市场需求的推动下,国内外许多研究机构和公司加入到了研究和开发机器翻译产品的队伍中,实用化的双语或多语机器翻译系统不断涌现。机器翻译系统的实用化引起了机器翻译系统的商品化,

并带动了语言信息处理和知识工程产业的兴起和发展。

然而, 尽管研究人员尝试了形形色色的语言学理论, 使用了各种各样的翻译技术, 但开发出来的机译系统却始终无法突破70%的正确率, 形成了机器翻译领域的雪线现象。对于汉英机器翻译来说尤其如此, 由于理论上缺少一种对路的分析和理解汉语的语言学理论的指导, 开发的众多汉英机器翻译系统始终不能取得突破性的进展, 距达到用户满意的水平还相差甚远。

概念层次网络(HNC)理论是关于人类语言认知机制的学说, 也是面向计算机的自然语言理解的学说。该理论从语言概念空间入手, 采用基元化、层次化、网络化、形式化的方法, 通过句类精妙地把自然语言的表层结构和深层语义联系起来, 从而在自然语言理解和处理领域走出了一条新路子, 被业内权威刊物和专家称为国内自然语言理解和处理的三大流派之一。近年来, 基于HNC理论的应用研究首先在信息检索、信息过滤方面取得了重大的突破, 在信息抽取、文本分类等领域也取得了显著的进展。

机器翻译是HNC理论涉足的又一重要应用领域。运用HNC理论向机器翻译这一自然语言理解和处理的制高点发起冲击, 其本身具有深远的意义。首先, 以英语为试点和突破口, 检验HNC理论的普适性, 从而进一步发展这一创新的自然语言理解理论, 加速HNC立足汉语、走向世界的进程; 其次, 通过对比研究汉英两种语言在句类及句式方面的异同, 建立符合英语特点的句类及句式系统, 为基于英语的自然语言理解研究及应用打下基础, 为各种知识工程的开展提供支撑; 最重要的是, 以机器翻译为目标, 探索源语-译语转换的规律和机制, 从而推动基于HNC理论的机器翻译引擎的研究, 为研制HNC机器翻译系统创造必需的条件。

本书介绍基于HNC理论的面向机器翻译的汉英句类及句式转换研究。基于HNC理论的汉英句类及句式转换研究是HNC汉

英机器翻译引擎研究的重要组成部分，旨在探讨汉英句类及句式转换的一般性规律。全书共分十一章：

第一章是本书的引言部分，综述半个多世纪以来机器翻译的进展，分析中国机器翻译的现状和存在的几个值得重视的问题，并简单说明运用 HNC 理论进行机器翻译新思路探索的意义。

第二章介绍 HNC 理论有关机器翻译的思想、HNC 机器翻译系统的组成和原理，最后阐释基于 HNC 理论的机器翻译系统研发应遵循的策略。

第三章是对句类转换的理论研究。首先介绍 HNC 理论有关句类的思想，然后对句类转换进行分类，最后定义一种形式化的句类转换描述框架。

第四章是对句式转换的研究。首先介绍 HNC 理论有关句式的思想，然后定义英语的几种特有句式，分析汉英两种语言在句式方面的异同，最后研究汉英句式转换的一般性规律。

第五章到第八章是对几个重点句类的深入研究，分别研究块扩作用句（第五章）、承受句（第六章）、简明状态句（第七章）、是否判断句（第八章）的汉英句类及句式转换规律。

第九章是对其他代表性句类的汉英转换研究，包括效应句、存在判断句、比较判断句等。

第十章介绍有关汉英/英汉双语句级对齐语料库的建设，以及该语料库在汉英句类及句式转换研究中的作用。

第十一章是本书的结束部分，总结本课题的研究工作，分析存在的问题，提出下一步的研究方向和研究内容。

附录一～五收录了从事 HNC 理论研究和应用开发所必需的一些重要基础资源，这部分内容也是阅读和理解本书的必备知识，构成本书的有机组成部分。

本书初稿（即作者的博士论文）完成于 2004 年 5 月，后在博士后研究工作之余对其进行了数次补充，并订正了其中的一些文

字及格式错误。鉴于作者对 HNC 这一博大精深的理论的理解尚处于入门水平，书中或有不当之处，敬请读者批评指正。

张克亮

2006 年 5 月

目 录

序言	(i)
前言	(v)
第一章 绪论	(1)
1.1 机器翻译历史概述	(2)
1.2 机器翻译的发展和进步	(12)
1.3 机器翻译热的冷思考	(16)
1.4 基于 HNC 理论的汉英机器翻译引擎研究	(21)
第二章 基于 HNC 理论的机器翻译	(23)
2.1 HNC 理论简介	(23)
2.2 HNC 理论的机器翻译观	(33)
2.3 基于 HNC 理论的机器翻译策略	(42)
2.4 本章小结	(48)
第三章 句类转换	(49)
3.1 引言——关于语句类型	(49)
3.2 HNC 的句类理论	(50)
3.3 句类转换	(53)
3.4 句类转换的形式化表示	(62)
3.5 本章小结	(68)
第四章 句式转换	(71)
4.1 引言——关于语序	(71)

4.2 HNC 理论的句式	(72)
4.3 句式的语种个性	(75)
4.4 汉英句式转换	(84)
4.5 本章小结	(90)
第五章 块扩作用句的句类及句式转换	(92)
5.1 引言——块扩及块扩句	(92)
5.2 块扩作用句	(93)
5.3 块扩作用句的汉英句类及句式转换	(100)
5.4 本章小结	(109)
第六章 承受句的句类及句式转换	(111)
6.1 引言——作用及作用的承受	(111)
6.2 一般承受句	(112)
6.3 主动承受句	(116)
6.4 被动承受句	(123)
6.5 特殊承受句	(126)
6.6 本章小结	(127)
第七章 简明状态句的句类及句式转换	(128)
7.1 引言——无动词句	(128)
7.2 简明状态句	(130)
7.3 简明状态句的汉英句类及句式转换	(135)
7.4 本章小结	(138)
第八章 是否判断句的句类及句式转换	(140)
8.1 引言——判断和判断句	(140)
8.2 是否判断句的分类	(141)
8.3 是否判断句的汉英句类及句式转换	(143)
8.4 本章小结	(156)
第九章 其他形式的汉英句类及句式转换	(157)
9.1 效应句	(157)
9.2 存在判断句	(164)

9.3 比较判断句	(168)
9.4 本章小结	(171)
第十章 基于语料库的句类及句式转换研究	(172)
10.1 引言——语料库简述	(172)
10.2 汉英/英汉句级对齐语料库	(175)
10.3 HNC 双语语料检索系统	(179)
10.4 基于语料库的汉英句类及句式转换研究	(183)
10.5 本章小结	(197)
第十一章 结束语	(200)
11.1 本课题研究工作总结	(200)
11.2 下一步的研究计划	(202)
附录一 HNC 基本句类代码及表示式	(204)
附录二 HNC 语句格式代码及表示式	(210)
附录三 HNC 标注符号集	(213)
附录四 HNC 概念和语句表示符号	(216)
附录五 HNC 概念树表	(219)
参考文献	(253)
后记	(263)

第一章 绪 论

据《旧约全书》的记载，早期人类使用的是同一种语言，后因野心膨胀欲建造直达天庭的巴别塔（the Babel Tower）而冒犯上帝，上帝遂让人类说不同的语言，人们之间因语言不通而渐生间隙，建塔工程只得废止。

早期的人类语言应有成千上万种之多，经过时代变迁，延续至今的还有 6000 多种。语言的障碍时时刻刻在困扰着人们，语言鸿沟成了人类交流的巨大屏障。为了克服语言屏障，人类社会诞生了最古老的行业之一——翻译。翻译工作者成了跨越语言鸿沟、沟通不同文化群体的桥梁。

人类对不同文化之间的交流有着巨大的需求，尤其是近一二十年来，随着计算机和通信技术的迅速发展，随着通信终端产品及互联网的普及和大规模应用，不同语言群体之间的交流越来越普遍，语言的障碍也显得越来越严重。如何有效利用现代技术突破横亘人们之间的语言障碍，成了信息化时代全人类面临的共同挑战。

机器翻译是信息时代克服语言障碍的不可缺少的手段，它在现代信息社会中的巨大作用将会越来越明显。现代机器翻译就是用计算机来进行不同自然语言之间的翻译，它是自然语言计算机

处理的一个历史悠久的部门，是横跨语言学、数学、计算机科学的综合性学科，也是信息时代语言应用的一个重要领域。

1.1 机器翻译历史概述

关于机器翻译的历史和发展，不少学者都做过深入的研究和论述，本书参考了许多文献，恕不在此一一列出。

1.1.1 机器翻译思想的萌芽

与机器翻译有关的一些朴素思想最早可以追溯到古希腊时代。当时，为方便不同民族之间的思想交流，人们尝试创造一种通用语言来代替种类繁多形式各异的自然语言，并考虑到了如何用机械手段来分析语言的问题。

到了 17 世纪的时候，一些学者曾经尝试采用机器词典来克服语言障碍，如笛卡儿（Descartes）和莱布尼兹（Leibniz）都试图在统一的数字代码的基础上来编写词典，而贝克（Cave Beck）、基尔施（Athanasius Kircher）和贝希尔（Johann Joachim Becher）等人都出版过此类词典。另外，有关通用语言和哲学语言的研究也很活跃，一些学者试图在逻辑原则和图形符号的基础上创造出一种无歧义的语言，代表性人物如维尔金斯（John Wilkins），他在一篇论文中提出了中介语（Interlingua）的思想，尝试通过精心设计的中介语来对概念和实体进行分类、编码和描述，并根据它们各自的特点和性质给予不同的标记和名称。

1.1.2 机器翻译的初创期（1933～1956）

据考证，首次使用“机器翻译”（einmechanisches Uebersetzen）这一术语的是德国学者里格（W. Rieger），他在 20 世纪 20 年代初提出过一种数字语法（Zifferngrammatik），该语法和词典结合

起来可以利用机器将一种语言翻译成其他多种语言。不过,真正具有实用意义的机器翻译成果直到 1933 年才出现,当时法国人阿尔楚尼(Georges Artsrouni)和俄国人特洛扬斯基(Petr Trojanskij)先后取得了通用多语机器词典的发明专利。阿尔楚尼的专利称作“机械脑”(mechanical brain),实际上是一台兼有多语机器词典功能的通用机器,可用于多语词项的检索,尚没有超出早期机器翻译的朴素思想。相比而言,特洛扬斯基则对机器翻译的全过程进行了深入的分析,明确把机器翻译分成分析、转换和生成三个独立的过程,他的翻译机采用基于世界语的通用符号体系对语法功能进行编码和解释,虽然仅能把源语言的原形词和逻辑符号转换成目标语言的原形词和符号,但他相信,只要能够建造出一部专门处理逻辑分析过程的机器,上述的整个翻译过程都能够用机器来实现。显然,特洛扬斯基这种认识已经超越了“机器词典”的简单想法,更加接近真正意义上的机器翻译。

现代意义上的机器翻译,即基于计算机的翻译,发轫于美国工程师韦弗(Warren Weaver)于 1949 年发表的以《翻译》为题的备忘录。在此之前的 1946 年,世界上第一台电子计算机 ENIAC 的诞生激发了韦弗和英国工程师布斯(Andrew Booth)利用计算机进行语言自动翻译的想法。在备忘录中,韦弗根据自己所掌握的有关密码学、统计学、逻辑学、信息理论以及语言共性等知识,针对语言中的歧义问题提出了具体的应对建议。不过,韦弗更多的是把机器翻译看成一种机械的解码过程,没有认识到在词法分析、句法分析、语义分析等方面的复杂性。

1951 年,巴希莱尔(Yehoshua Bar-Hillel)在充分调研的基础上写出了一篇有关机器翻译发展现状的报告,勾画了解决机器翻译问题的一些基本方法。次年,巴希莱尔在 MIT 主持召开了首次机器翻译大会,活跃在本领域的知名学者几乎悉数与会。此次大会提出了不少至今仍在对机器翻译研究发挥影响的观点和方法,如译前编辑、译后编辑等提倡人机互动的思想、利用微型术语表