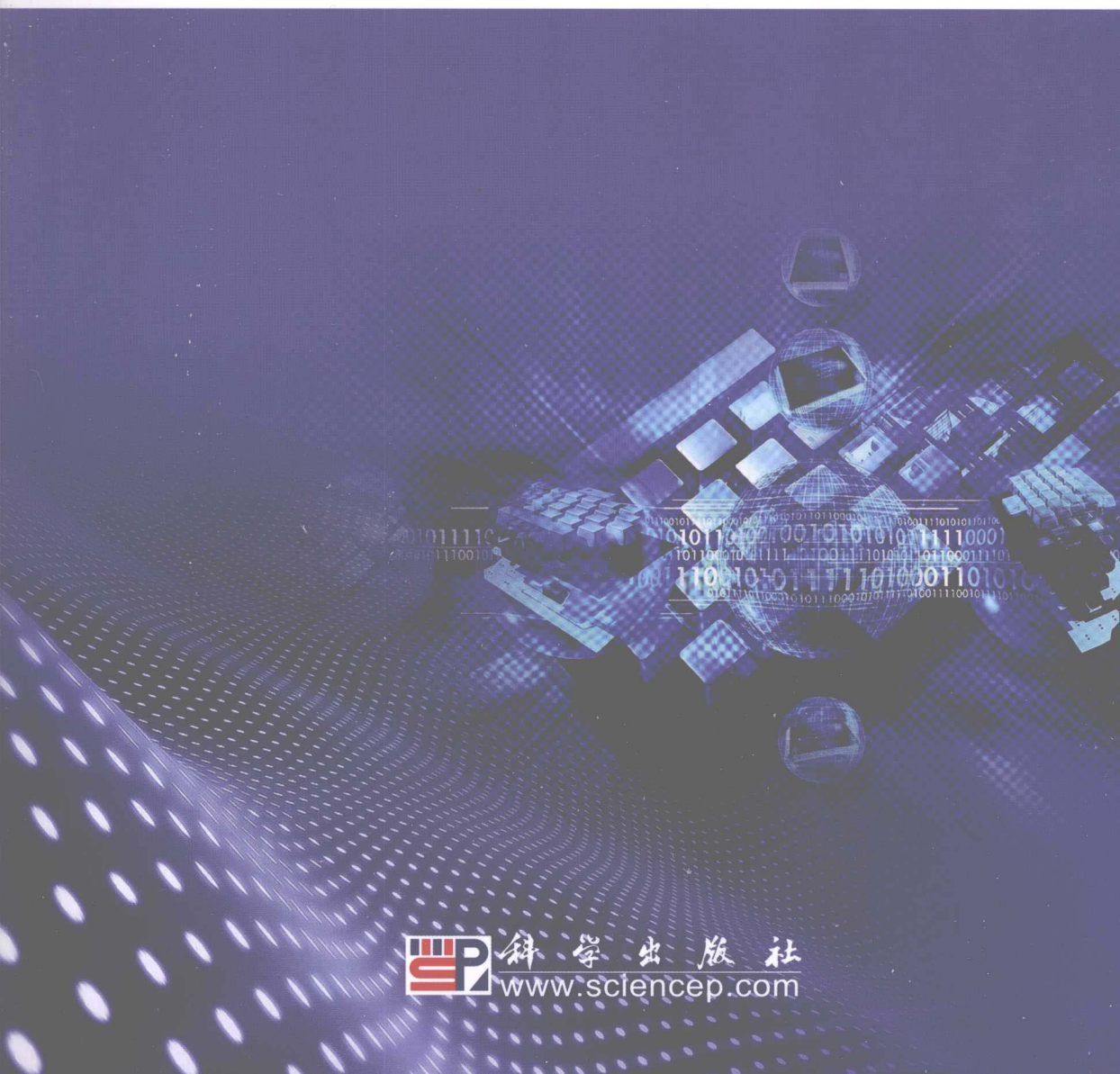
智能科学技术著作丛书

面向汉英机器翻译的语义块构成变换

李颖 王侃 池毓焕 著



科学出版社
www.sciencep.com

智能科学技术著作丛书

面向汉英机器翻译的语义块构成变换

李 颖 王 侃 池毓焕 著

科学出版社

北京

内 容 简 介

本书针对当前机器翻译准确率存在的两大难点(自然语言理解处理和过渡处理),在 HNC 理论框架下阐释了机器翻译引擎原理,第一次对其中的关键之处——语义块构成变换进行了全面、系统、深入的阐述,给出了具体的解决方案,制定了统摄具体规则的一系列原则。本书内容可应用于语言信息处理、机器翻译及语言分析。

本书适合于机器翻译、自然语言理解与处理、人工智能等智能信息处理专业领域的研究者、开发者和学习者参考阅读。

图书在版编目(CIP)数据

面向汉英机器翻译的语义块构成变换/李颖,王侃,池毓焕著. —北京:科学出版社,2009

(智能科学技术著作丛书)

ISBN 978-7-03-022744-7

Ⅰ. 面… Ⅱ. ①李…②王…③池… Ⅲ. 英语-机器翻译-研究 Ⅳ. H315.9

中国版本图书馆 CIP 数据核字(2008)第 121499 号

责任编辑:孙 芳 王志欣 / 责任校对:张 琪

责任印制:赵 博 / 封面设计:陈 敬

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

漷海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2009 年 1 月第 一 版 开本:B5(720×1000)

2009 年 1 月第一次印刷 印张:14 1/2

印数:1—3 000 字数:273 000

定价:42.00 元

(如有印装质量问题,我社负责调换〈环伟〉)

《智能科学技术著作丛书》编委会

《智能科学技术著作丛书》序

“智能”是“信息”的精彩结晶，“智能科学技术”是“信息科学技术”的辉煌篇章，“智能化”是“信息化”发展的新动向、新阶段。

“智能科学技术”(intelligence science&technology，IST)是关于“广义智能”的理论方法和应用技术的综合性科学技术领域，其研究对象包括：

·“自然智能”(natural intelligence，NI)，包括：“人的智能”(human intelligence，HI)及其他“生物智能”(biological intelligence，BI)。

·“人工智能”(artificial intelligence，AI)，包括：“机器智能”(machine intelligence，MI)与“智能机器”(intelligent machine，IM)。

·“集成智能”(integrated intelligence，II)，即：“人的智能”与“机器智能”人机互补的集成智能。

·“协同智能”(cooperative intelligence，CI)，指：“个体智能”相互协调共生的群体协同智能。

·“分布智能”(distributed intelligence，DI)，如：广域信息网，分散大系统的分布式智能。

1956 年，“人工智能”学科诞生，50 年来，在起伏、曲折的科学征途上不断前进、发展，从狭义人工智能走向广义人工智能，从个体人工智能到群体人工智能，从集中式人工智能到分布式人工智能，在理论方法研究和应用技术开发方面都取得了重大进展。如果说，当年“人工智能”学科的诞生是生物科学技术与信息科学技术、系统科学技术的一次成功的结合，那么，可以认为，现在“智能科学技术”领域的兴起是在信息化、网络化时代又一次新的多学科交融。

1981 年，“中国人工智能学会”(Chinese Association for Artificial Intelligence，CAAI)正式成立，25 年来，从艰苦创业到成长壮大，从学习跟踪到自主研发，团结我国广大学者，在“人工智能”的研究开发及应用方面取得了显著的进展，促进了“智能科学技术”的发展。在华夏文化与东方哲学影响下，我国智能科学技术的研究、开发及应用，在学术思想与科学方法上，具有综合性、整体性、协调性的特色，在理论方法研究与应用技术开发方面，取得了具有创新性、开拓性的成果。“智能化”已成为当前新技术、新产品的发展方向和显著标志。

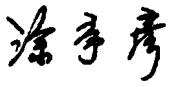
为了适时总结、交流、宣传我国学者在“智能科学技术”领域的研究开发及应用成果，中国人工智能学会与科学出版社合作编辑出版《智能科学技术著作丛书》。需要强调的是，这套丛书将优先出版那些有助于将科学技术转化为生产力以及对社会和国民经济建设有重大作用和应用前景的著作。

我们相信，有广大智能科学技术工作者的积极参与和大力支持，以及编委们的

共同努力，《智能科学技术著作丛书》将为繁荣我国智能科学技术事业、增强自主创新能力、建设创新型国家做出应有的贡献。

祝《智能科学技术著作丛书》出版，特赋贺诗一首：

智能科技领域广
人机集成智能强
群体智能协同好
智能创新更辉煌



中国人工智能学会荣誉理事长
2005 年 12 月 18 日

序

这篇序言，我想说的就是这样一句话：祝福你，又一雪线雄鹰。

这句话里的“祝福”、“又一”、“雪线”和“雄鹰”都有其特定的义境，这需要有所解释。然而，义境这东西的个性色彩一般大于共性，也许读者阅读此书的不同领会都是最好的解释，我何必多此一举呢？但是，所谓序言本来就是做多此一举的事，我也只好随众了。

请允许我从“雪线”说起，这里的雪线是指当前机器翻译准确率的上限，这个上限大约是70%。机器翻译学界完全有理由为这个雪线感到自豪，攀登到这个雪线是自然语言处理的一项奇迹，用行话来说，那既是RBMT（基于规则的机器翻译）的奇迹，也是SBMT（基于统计的机器翻译）的奇迹。但是，广大机器翻译的用户并不买这个奇迹的账，他们要求机器翻译继续向雪线之上攀登，直到那高山之巅。这是全球化时代对信息产业的一项呼唤，然而目前只是一项隐性的微弱呼唤。信息产业在全球化时代中扮演着举足轻重的角色，它面对着日新月异的巨大技术挑战，已经忙得不可开交，以至于相关的许多重大科学挑战难免遭受到不同程度的冷遇，这一点西方与东方并没有差别，美国不仅在机器翻译领域没有显示出任何高明，实际上处于十分落后的状态。上述呼唤表面上是技术需求的呼唤，实质上是一项伟大科学探索的呼唤，它直接关系到自然语言理解之谜和机器翻译之谜的探索。而这两项科学之谜的答案又直接关系到机器翻译的雪线之上的攀登，虽然这段攀登不过是全程的30%，但是登山者都知道这30%意味着什么！

下面该说“雄鹰”了，但实际上只说一个“雄”字，前面提到的RBMT和SBMT都是鹰，但他们不够“雄”，在雪线之上遇到了难以逾越的困难。这一遭遇的必然性大家都心知肚明，因为他们既回避了自然语言理解之谜，也回避了机器翻译之谜。该书之“雄”表现在两个方面：一是它对上述两大科学之谜给出了一个系统而简明的论述，二是它对汉英机器翻译之谜的急所给出了一个击中要害的全面解决方案，后者是第一作者博士论文的基本成果。上面使用了一个围棋术语——急所，我实在找不到更好的术语来描述该书的学术价值与贡献了。这价值与贡献不仅关系到机器翻译雪线之上的攀登（可简称雪线攀登），也关系到汉语和英语这两种自然语言之间本质差异的比较研究（可简称汉英质异）。从这个意义上来说，该书的书名有点过于学究气，但它是严谨而谦虚的，这值得称赞。

为什么“语义块构成变换”就是汉英机器翻译的急所呢？要回答这个问题很不容易。首先是语言术语的障碍。“语义块”、“语义块构成”、“语义块构成变换”是三个具有特定意义的术语，这三个术语在中外语言学著作里是看不到的，可是，这三个术语正是本书的立足点，也是其创新立说的出发点。简单地说，“语义块”这个术

语或概念是“短语”的扩充，但这一扩充对于雪线攀登和汉英质异的研究不是可有可无，而是绝对必要的。英语拥有构造从句和非限定形态动词短语的完备语法手段，而汉语完全不具备这些手段。那么，汉语采用什么语法手段以达到同样的语言表达功能呢？我们是否需要引入一种超越于不同自然语言个性之上的术语或概念以统摄语言分析或语言表达的描述方式呢？该书提供了答案，那就是该书名称的关键词——“语义块构成变换”。对“语义块构成变换”的进一步描述需要引入“句蜕”与“逻辑组合”这样两个统摄性的术语，与“句蜕”对应的是全局谓语和局部谓语的概念，与“逻辑组合”对应的是并列、偏正、定中、主谓、动宾、动补、介词短语、动词名词化等诸多传统语法概念。“语义块构成变换”的核心科学问题就是对“句蜕”和“逻辑组合”的区分、判定及其交织性的处理，并在这一基础上制定相应的变换原则。那么，该书对这一核心科学问题作出了什么样的贡献和取得了什么程度的成果呢？我愿意把这个事关重大的答案留给读者去思考。

最后，该说到“又一”和“祝福”了。“又一”是因为在该书之前，张克亮教授的《面向机器翻译的汉英句类与句式转换》已经出版，这是汉英机器翻译之谜的大场。大场也是一个围棋术语，同样，我也找不到更好的术语来描述此书的学术价值与贡献了。

要赢得围棋的胜利，光有大场与急所的一流功力是不够的。机器翻译的收官之战是一项巨大的语言工程，如果把机器翻译的雪线攀登比作一场 400 米接力赛，那么，可以明确地说，《面向机器翻译的汉英句类与句式转换》与《面向汉英机器翻译的语义块构成变换》这两部著作都只是第一棒的主体内容。不过，这第一棒的意义不同于 400 米接力赛，它不仅具有奠基意义，而且具有对后续三棒的指导意义。至于这一指导作用能否实现，已不是一个纯粹的科学技术问题，更是一个科技指导方针和科技团队组织的问题了。因此，“祝福”就是这篇短序的最佳告别语了。

黄曾阳

2008 年 9 月 17 日

前　言

本书的研究内容属于机器翻译(machine translation,MT)的范畴,是在 HNC (hierarchical networks of concepts,概念层次网络)理论框架下对 HNC 机器翻译引擎原理中的一部分——语义块构成变换进行的研究,其目的是为机器翻译提供理论与技术支持。

6 项过渡处理是 HNC 机器翻译引擎原理的重要内容,它们是两转换(句类转换、句式转换)、两变换(语义块构成变换、主辅变换)、两调整(块序调整、句序调整),其中的语义块构成变换是 6 项过渡处理中使用频率最高的一种处理。句蜕现象的出现是形成语义块复杂性的主要因素,因此,句蜕构成及汉英变换处理是语义块构成变换的核心内容;而逻辑组合变换的使用频度超过任何其他的转换、变换或换位,虽然复杂性和重要性不及句蜕,但不失为语义块构成变换两翼之一。

本书对句蜕进行了较为深入的研究,提出了句蜕处理的规则及算法,具体包括:① 要素句蜕构成及其汉英变换的处理规则和算法;② 原型句蜕构成及其汉英变换的处理规则和算法;③ 包装句蜕构成及其汉英变换的处理规则和算法;④ 句蜕中的一种常见歧义结构的消歧算法。

这些算法和规则的提出标志着句蜕从语言描述阶段进入了可计算阶段,推进了 HNC 机器翻译引擎的研制。同时,本书从理论上对句蜕及其相关内容也进行了比较深入地探讨,主要包括:① 句蜕的理论阐释;② 句蜕的汉英对比研究;③ 句类空间的再抽象;④ 交互引擎的形式化描述。

上述所有内容都从不同侧面丰富了 HNC 关于句蜕及其相关的理论。

本书对逻辑组合也进行了具体研究,提出了逻辑组合汉英变换的规则,具体包括:① 逻辑组合的汉英异构、同构及异态研究;② 逻辑组合单元的概念关系分析;③ 逻辑组合汉英变换策略;④ 逻辑组合汉英变换规则。

论述次序上,本书首先对机器翻译原理进行了概述。概述是依照"机器翻译是一种映射"的观点,按映射的类型来展开的,在简要介绍了 HNC 的机器翻译观之后,重点介绍 HNC 机器翻译引擎原理,而后指出本书研究内容及意义。接着对与本书相关的 HNC 基本概念及基本观点进行了简述,这是本书研究的前提条件。之后就是本书的重点内容,即句蜕构成及汉英变换处理和逻辑组合汉英变换处理。最后是机器翻译实例分析与本书取得的结论以及今后进一步研究的方向。

HNC 理论博大精深,作者的理解也许肤浅而有限,难免存在不妥之处,敬请读者批评指正。

目　录

第1章 引 论

科技的发展，从某种意义上说是对人类自身功能的延伸。各种先进的交通工具可以看作是人类脚的延伸物，声呐和雷达是人类耳目的延伸物。但是，要让计算机真正成为名副其实的“电脑”，来充当人脑的延伸物，还任重而道远。究其原因，瓶颈之一就在于计算机目前还不能理解自然语言，具体表现在对下列4个方面的无所作为(这4个方面是衡量计算机是否理解自然语言的图灵(Turing)标准，也是目前人工智能或计算语言学界普遍认同的标准)：

(1) 问答(question-answering)。机器能正确地回答输入文本中的有关问题。

(2) 文摘生成(summarizing)。机器有能力产生输入文本的摘要。

(3) 释义(paraphrase)。机器能用不同的词语和句型来复述其输入文本。

(4) 翻译(translation)。机器具有把一种语言(源语言)翻译成为另一种语言(目标语言)的能力。

事实上，这4个方面是相互联系、互相支持、互相推动的。假如计算机能进行翻译，即实现了上述第4个方面的要求，那么，必将为上述前3个问题的解决提供强有力的支持。不仅如此，机器翻译如果获得成功，将至少在下列几个方面得到广泛应用：① 互联网语言障碍的消除；② 人机对话的实现；③ 技术文档的翻译。由此可见，机器翻译无论对计算机理解自然语言还是应对市场的广泛需求都具有重要的意义。然而，目前机器翻译的现状是翻译技术水平低下，翻译理论滞后，与市场日益扩大的需求形成强烈反差。机器翻译期待新理论、新方法的出现。

基于HNC理论的机器翻译引擎就是这种背景催生的产物，是新理论、新方法的具体体现。

HNC理论是在对传统理论和方法进行深刻反思中成长起来的面向整个自然语言理解处理的新理论。“在语义表达上有自己的特色，在语义处理上走了一条新路……对突破汉语理解问题尤其有实际意义。”[1)]

那么，HNC机器翻译引擎的原理是什么呢？概括起来，主要由下列3部分组成：① 源语言的理解(扩展句类分析)；② 源语言与目标语言在语言概念空间上进行对接(6项过渡处理)；③ 目标语言的生成(反映射过程)。这3部分是一个整体，只有每步都处理得好，整个系统才能保证质量。其中，第3部分和传统机器翻译原理比较接近，第1部分和第2部分则有本质差异。第1部分的核心任务是实现语言空间到语言概念空间的转换，它完全不同于传统的句法-语义-语用分析。第2部分并不对应于所谓的中间语言，而是在语言概念空间中进行6项过渡处理：① 句类

1) 陈力为院士为专著《HNC(概念层次网络)理论》所作的题词。

转换；② 句式转换；③ 语义块构成变换；④ 语义块主辅变换；⑤ 主块分离和辅块位置的调整；⑥ 句群小句顺序的调整。

目前的机器翻译系统大都在对第 3 部分的处理上很有功底，但第 1 部分比较弱，遇到歧义严重、构成复杂、指代不明、省略不清的句子就非常弱，几乎无能为力，第 2 部分虽然在形式上有所涉及，但思路上与 HNC 有本质区别，实质上处于空白状态。这是目前机器翻译质量差的根本原因。改变这一局面的出路在于加强弱项，填补空白，可简称“补弱填空”。“补弱”就是以扩展句类分析替换句法-语义-语用分析，“填空”就是实现上列 6 项过渡处理。因此，能否有效地处理好 6 项过渡处理，事关机器翻译能否克服雪线的关键。6 项过渡处理应该是目前机器翻译最值得研究的课题。

本书中的语义块构成变换包含语义块构成与语义块变换两部分内容，前者事关语言理解，后者事关语言生成。上述 6 项过渡处理中的语义块构成变换主要定位于变换。

就语义块构成而言，句蜕与逻辑组合是其主要内容，但从机器翻译的 6 项过渡处理看，语义块构成变换还存在与其他部分形成交织处理的情况。因此，本书首先处理句蜕，然后是逻辑组合，最后是交织问题。

句蜕是 HNC 为描述汉语短语的一项独特（相对于印欧语系）表示形态而引人的术语，其字面意义是由一个句子蜕化而来的短语，这个短语可能对应于一个语义块，也可能只是语义块的一部分。从句法学的视角来看，汉语由于不存在关系代词和动词的非限定形态（包括所谓动词的名词化）这两项词法手段，因而也就不可能存在英语常见的从句和动词非限定形态短语这两种句法形态。那么，汉语以什么句法形态取代相应的英语形态呢？那就是句蜕和小句。“咬死了猎人的狗”这一著名的歧义语段实质上就是小句和句蜕的歧义。另一方面，句蜕作为语句的一种蜕化形式，是造成语句复杂的主要原因之一，也是复杂语义块的轴心。句蜕的汉英变换是语义块变换最核心的一部分，句蜕汉英变换是本书重点研究的内容。

英语拥有丰富的短语逻辑组合词（即 *l*4 和 *l*5），而汉语比较贫乏，这使得汉语不得不强行以“意合”的方式构造短语，而英语则在短语的词语之间给出明确的介词标记。汉语对逻辑组合的“意合”表达方式当然也是汉语的伟大创造，但排在句蜕和小句共享之后比较恰当。因此，逻辑组合的汉英变换作为语义块变换的重要内容，也是本书重点研究的内容。

句蜕与逻辑组合存在交织现象，两者又与 6 项过渡处理的其他环节交织在一起，交织意味着异常复杂，需要超常的洞悉能力和卓越的处理谋略，本书对这部分内容点到为止。

1.1 机器翻译原理概述

机器翻译是指利用计算机来实现两种自然语言[1]之间的对译。尽管在计算机出现之前就有过使用一些特殊的用于不同语言翻译的机械装置，但由于计算机具有存储容量大、运行速度快等特点，所以到目前为止，计算机被公认为是最适合用于机器翻译的机器。

从数学的观点看，机器翻译就是源语言到目标语言的一种映射，探索开发机器翻译系统的过程就是寻找这种映射的过程。最开始，人们把这种映射定位于一部庞大的双语词典和简单的词频统计上。这种朴素的想法尽管也是机器翻译所必需的，但终不能实现设计者的夙愿。相反，这种翻译模式下的机器翻译系统在当时被当作笑话频频引用，其中最为典型的例子如下：

The spirit is willing, but the flesh is weak. (心有余而力不足。)

译成俄语后再回译过来如下：

The wine is good, but the meat is spoiled. (酒是好的，但肉却变质了。)

也有些系统回译为

The liquor is holding out all right, but the meat has spoiled. (酒好而肉臭。)

机器翻译的这种状况降低了人们对机器翻译的热望，也使机器翻译过早地进入了低潮期[2]。这是忽视对语言本体研究而导致失败的必然结果。这一事实从反面告诫人们，忽视对语言进行深层次的分析而在源语言与目标语言之间建立直接映射是不可能的。于是，人们开始考虑将这个映射分解成多个映射，其中无一例外地加入了语言本体的研究内容，这就是理性主义(rationalist)的方法。

理性主义的方法目前的主流是基于 Chomsky(1957, 1959a, 1959b, 1965)的语言原则(principles)。Chomsky 认为，人为什么能理解和掌握语言，是因为人大脑里存在一种遗传过来的机制，他称之为普适语法(universal grammar)。所以，理性主义主要研究人的语言结构(language competence，语言能力)。实际的语言数

1) 自然语言是相对于计算机的各种程序设计语言、世界语等人造语言而言的，它是人类语言集团的本族语，如汉语、英语、法语等。

2) 1964 年 4 月，美国国家科学院(NAS)成立了自动语言处理咨询委员会(Automatic Language Processing Advisory Committee, ALPAC)。这一 7 人小组包括语言学家、心理学家、机器翻译专家和人工智能研究者，对美国政府机构资助的机器翻译研究进行调查。ALPAC 对当时机器翻译的速度、质量、花费以及人们对机器翻译的需求等方面进行了分析，并对一些机器翻译系统进行了测试，1966 年发表了著名的 ALPAC 报告。报告指出，机器翻译的译文质量明显低于人工翻译，认为尽管未经编辑的机器译文大部分可读，但造成了"缓慢而痛苦的阅读"，从而建议不要再对机器翻译进行更多投资了。在 ALPAC 报告的影响下，各类机器翻译项目剧减，机器翻译陷入了低谷。

据(language performance,语言行为)只提供了这种内在的间接证据,它通过语言所必须遵守的一系列原则来描述语言。基于这种理论的机器翻译方法实际上是一种基于规则(rule-based)的方法[1)]。

Chomsky 的转换生成理论就是这一思想的典型代表。Chomsky 的转换生成文法理论所描述的词、短语之间的结构关系实际上就是建立从词性串向句法结构映射的关系,是句法分析的理论模型。在其体系之中,Chomsky 将自然语言的这种映射(句法关系)用规则的形式进行描述。而句法分析的过程,就是一种有序规则的派生过程。

这样,基于 Chomsky 理论的机器翻译(也就是传统的基于规则的机器翻译)就是把源语言向目标语言的映射一分为三。从整体上来看,包括 3 个步骤:源语言分析、转换和目标语言生成。源语言分析是通过形态分析、词性标注、句法分析、语义分析和语用分析等将源语言表达转换成事先定义的中间表达。转换是将源语言的中间表达转换成目标语言的中间表达,不同语言之间的中间表达也不同,只有中间语言方法所采用的中间表达为各种语言所共有的中间表达。目标语言生成是将目标语言的中间表达变成目标语言表达,主要完成文本规划和表层实现两项任务。文本规划确定欲实现的目标语言文本的有关内容、修辞方式的信息,包括内容界定、文本构造、词汇选择、句法选择、共指现象处理和成分调序等子任务;表层实现根据目标语言语法,将由词汇组成的句法表达式映射为表层字符串。上述过程的每个步骤都不可避免地包括许多歧义现象,每种歧义都需要大量的知识来帮助消歧。

但是,需要用来帮助消歧的大量世界知识如何获取呢? 世界知识是无限的,在普适语法的框架下,Chomsky 没有找到完整的答案,也不可能找到完整的答案,有以下几个方面的原因:

(1) 用 Chomsky 的思想描述语言的规律,总是事先有一个假定的模型(模型的内核按照 Chomsky 的观点是来源于人的大脑中遗传过来的一种机制,称之为普适语法),然后再对其进行描述。实际上,是通过语言所必须遵守的一系列原则来描述语言的。尽管它对计算机程序设计语言作形式化描述相当成功,但自然语言远比程序设计语言复杂得多,自然语言绝不会在普适语法的框架内安分守己。

(2) 如前所述,实现 3 个步骤中的第 1 个就是完成从词性串向句法结构映射。虽然在自然语言中,这种映射关系存在,但是要完整地找到所有关系并通过规则方法描述出来非常困难,因为在某些语言中,词类与句法成分之间不存在简单的一一对应关系。汉语就是一个实证。不仅如此,汉语词类本身缺乏形态标志,进入句子后又缺乏种种形态变化,多数词类没有固定的语法功能,这就造成了词类与句法成

1) 基于规则的方法除了转换生成法外,还有中间语言法。基于中间语言的方法是对源语言进行分析以后产生一种称为中间语言的表示形式,然后直接由这种中间语言的表示形式生成目标语言。所谓中间语言,就是自然语言的计算机表示形式的系统化,它试图创造一种独立于各种自然语言同时又能表示各种自然语言的人工语言。